International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 21, Number 3 (2026) pp. 391-399
Research India Publications
https://www.ripublication.com

Artificial Intelligence-Based Non-Intrusive Load Monitoring for Residential Energy Disaggregation Using Smart Meter Data

A.N. Hanafi [1]*, M.N. Saiful Jefri [1] and A.F. Abdul Kadir[1]

[1] *Faculty of Electrical Technology and Engineering, Universiti Teknikal Malaysia Melaka, Malaysia*

Correspondence: ainain@utem.edu.my

Abstract

Non-Intrusive Load Monitoring (NILM) enables appliance-level energy consumption estimation from aggregate power measurements, reducing the need for individual appliance monitoring devices. However, accurate energy disaggregation remains challenging when using low-frequency smart-meter data, particularly for appliances with overlapping and variable operating characteristics. This study investigates the performance of two supervised deep learning models, a Deep Neural Network (DNN) and a Temporal Convolutional Network with Sequence-to-Point prediction (TCN-S2P), for residential energy disaggregation. Experimental power measurements were collected using an Omnimeter smart meter for five household appliances: a lamp, inverter air conditioner, electric kettle, electric iron, and blower fan. The acquired data were pre-processed, normalized, and segmented into sliding windows for model training and evaluation. Both models were independently evaluated using mean absolute error (MAE), root mean square error (RMSE), precision, recall, and F1-score under different window sizes and training epochs. The selected DNN configuration, using a 121-sample window and 50 epochs, achieved an average MAE of 2.44 W and an F1-score of 0.932. In comparison, TCN-S2P achieved a lower RMSE of 16.37 W and an F1-score of 0.910 using a 61-sample window and 50 epochs. Appliance-level analysis revealed that DNN provided better overall identification performance, while TCN-S2P demonstrated advantages in estimating the variable power consumption of the inverter air conditioner. These findings highlight the trade-offs between prediction accuracy and appliance identification performance, providing insights into the selection of deep learning architectures for residential NILM applications.

Keywords: Deep neural network, energy disaggregation, non-intrusive load monitoring, temporal convolutional network

Introduction

Residential electricity consumption has increased with the growing adoption of household appliances and smart-home technologies. Conventional electricity billing systems provide information on total energy consumption but do not identify the contribution of individual appliances, limiting consumers' ability to monitor electricity usage and implement effective energy-saving strategies. Intrusive Load Monitoring (ILM) addresses this limitation by installing individual sensors or sub-meters for each appliance; however, its implementation involves additional hardware costs, complex wiring, and maintenance requirements. Recent developments in smart metering infrastructure have enabled more accessible and centralized electricity monitoring, supporting improved residential energy management [9,10]. Non-Intrusive Load Monitoring (NILM) offers an alternative approach by estimating individual appliance consumption from aggregate electrical measurements obtained at a single monitoring point [1].

The fundamental principle of NILM involves identifying distinctive electrical signatures associated with individual appliances through variations in power consumption and operating patterns. However, accurate load disaggregation remains challenging when multiple appliances operate simultaneously, particularly when low-frequency smart-meter measurements provide limited information about transient electrical events. Early NILM approaches relied on event detection, threshold-based methods, and manually engineered features, which are often sensitive to electrical noise and overlapping appliance signatures [2–4]. These limitations have encouraged the application of artificial intelligence and machine learning techniques to improve appliance identification and power estimation.

Artificial neural networks have been investigated for electrical load management, demonstrating their potential for learning nonlinear relationships in energy consumption data [11]. More recently, deep learning approaches have gained attention in NILM research because of their ability to extract representative features directly from time-series measurements without extensive manual feature engineering [5,6]. Deep Neural Networks (DNNs), particularly sequence-to-point architectures, can learn nonlinear relationships between aggregate power measurements and appliance-level consumption. Temporal Convolutional Networks (TCNs), on the other hand, employ dilated convolutional operations to capture temporal dependencies across different time scales, making them suitable for analysing appliances with dynamic operating characteristics [7,8]. Nevertheless, the relative performance of these architectures depends on appliance characteristics, input-window selection, and model training configurations.

This study investigates the performance of DNN and Temporal Convolutional Network Sequence-to-Point (TCN-S2P) models for residential energy disaggregation using experimental smart-meter measurements from five household appliances. Both models are independently developed and evaluated using the same dataset and performance metrics, including mean absolute error (MAE), root mean square error (RMSE), precision, recall, and F1-score. The effects of input-window size and training epochs are also examined to identify suitable model configurations. The main contribution of this study is an experimental comparison of the two architectures at

both overall and appliance levels, highlighting their respective strengths and limitations in appliance identification and power estimation. The findings provide insights into selecting appropriate deep learning models for smart-meter-based residential NILM applications.

Methodology

3.1 Data Acquisition and Preprocessing

The experimental dataset was obtained from a laboratory electrical monitoring setup employing an Omnimeter smart meter. Five household appliances were considered: a lamp, an inverter air conditioner, an electric kettle, an electric iron, and a blower fan. These appliances exhibit different operating characteristics, including steady-state consumption, thermostatically controlled operation, intermittent high-power demand, and variable-power operation.

The experimental dataset comprised aggregate active-power measurements and corresponding appliance-level power records for the five selected loads. The data were processed in Python using pandas and NumPy. Rows containing missing values were removed, and min–max normalization was applied separately to the aggregate input and appliance-level target variables. The normalized power data were subsequently transformed into fixed-length sliding windows. Each input window contained a sequence of aggregate power measurements, with the corresponding target defined as the appliance-level power at the central prediction point. Window sizes of 61, 121, and 241 samples were investigated to evaluate the influence of temporal context on disaggregation performance.

The overall NILM workflow is shown in Figure 1. The smart-meter data were first transformed into a structured dataset. The dataset was then divided into training and testing sets. During training, each model learned the relationship between the aggregate input window and the target appliance power. During testing, the trained model predicted appliance-level output from unseen aggregate data.

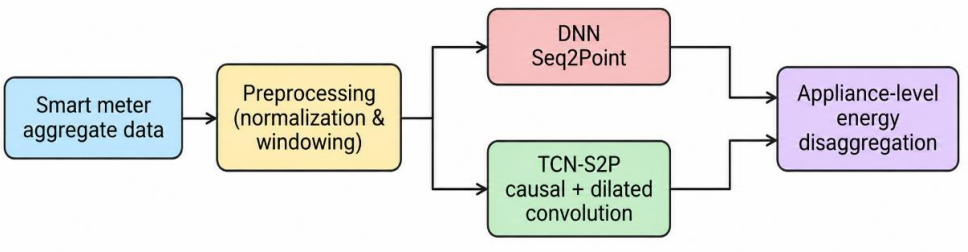


Figure 1. AI-based NILM workflow using smart-meter aggregate data.

3.2 DNN and TCN-S2P Model Development

Two supervised deep learning architectures were implemented independently to evaluate appliance-level power disaggregation from aggregate measurements. The DNN architecture employed two one-dimensional convolutional layers containing 32 and 64 filters, respectively, followed by a flattening layer and three fully connected hidden layers with 256, 128, and 64 neurons. Rectified Linear Unit (ReLU) activation

functions were used, and dropout regularization with a rate of 0.2 was included. The output layer generated five continuous appliance-power predictions.

The TCN-S2P architecture employed four temporal convolutional residual blocks with dilation rates of 1, 2, 4, and 8. The filter configurations were 32, 32, 64, and 64, with a convolution kernel size of five. The temporal feature extraction stage was followed by fully connected layers containing 256, 128, and 64 neurons and a linear output layer for five appliance-power predictions.

Both networks were trained using the Adam optimizer with a learning rate of 0.001 and a batch size of 32. An appliance-weighted mean squared error function was applied to emphasize selected appliance operating events. Training budgets of 50, 100, and 150 epochs were investigated to examine their influence on model performance. The final model configurations were selected based primarily on the macro-averaged F1-score, with early stopping applied during training. The model configurations were evaluated using different input-window sizes to determine the relative performance of the two architectures.

3.3 Performance Evaluation

The prediction performance was assessed using mean absolute error (MAE), root mean square error (RMSE), precision, recall, and F1-score. MAE and RMSE were computed from the predicted and measured appliance powers in watts. For appliance operating-state evaluation, the actual and predicted power signals were converted into binary ON/OFF states using appliance-specific power thresholds. Thresholds of 5 W, 50 W, 200 W, 100 W, and 10 W were applied to the lamp, air conditioner, kettle, electric iron, and fan, respectively. Precision, recall, and F1-score were calculated independently for each appliance. The overall reported performance was obtained by calculating the arithmetic mean of the five appliance-level metric values, providing a macro-averaged comparison between the models.

Table 1. Experimental and model configuration.

Item	Configuration
Measured appliances	Lamp, inverter air conditioner, electric kettle, electric iron, blower fan
Measurement source	Smart meter aggregate power data
Software platform	Python with NumPy, pandas, matplotlib, TensorFlow/Keras
Preprocessing	Data cleaning, min-max normalization, sliding-window segmentation
Tested window sizes	61, 121, and 241 samples
Tested epochs	50, 100, and 150 epochs
DNN settings	Hidden layers 256, 128, and 64; ReLU activation; Adam optimizer; MSE loss
TCN settings	Four convolutional layers; kernel size 5; dilation rates 1, 2, 4, and 8; causal convolution
Evaluation metrics	MAE, RMSE, precision, recall, and F1-score

RESULTS AND DISCUSSION

Appliance-level disaggregation performance

Two models, DNN and TCN-S2P, were evaluated independently to estimate the active power of five appliances from the aggregate input signal. The actual-versus-predicted profiles are presented in Figures 2 and 3. For both models, predicted operating intervals generally followed the reference power traces of the lamp, fan, and air conditioner. The estimates for the electric kettle and iron captured major high-power intervals, although departures from the measured signals, including additional or imperfectly reconstructed peaks, were also apparent. These observations are qualitative; quantitative comparisons are provided in Table 2.

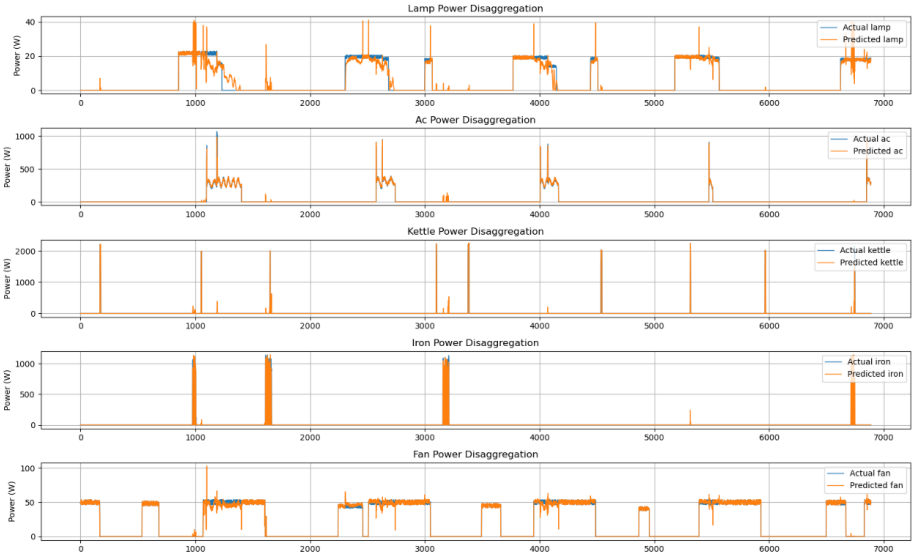


Figure 2. Actual and predicted appliance power using the DNN model

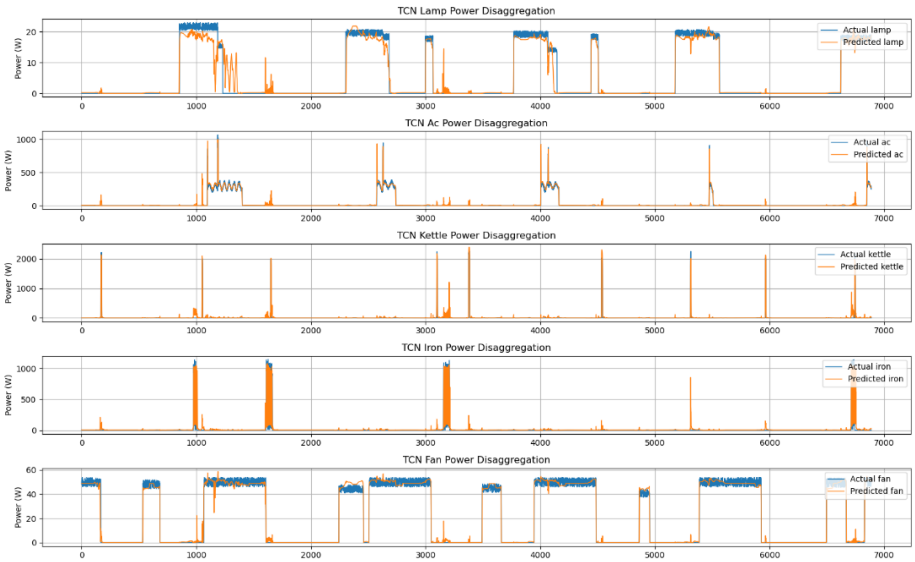


Figure 3. Actual and predicted appliance power using the TCN-S2P model

Table 2. Appliance-level disaggregation performance for the DNN and TCN-S2P models

Appliance	Model	MAE (W)	RMSE (W)	F1-score
Lamp	DNN	0.79	2.16	0.955
	TCN-S2P	1.02	2.68	0.944
Inverter air conditioner	DNN	4.07	19.85	0.980
	TCN-S2P	3.19	15.23	0.985
Electric kettle	DNN	3.38	41.75	0.841
	TCN-S2P	7.23	34.59	0.779
Electric iron	DNN	2.80	28.69	0.887
	TCN-S2P	4.26	26.88	0.846
Blower fan	DNN	1.16	2.62	0.999
	TCN-S2P	1.27	2.45	0.998
Average	**DNN**	**2.44**	**19.01**	**0.932**
	TCN-S2P	**3.39**	**16.37**	**0.910**

The blower fan recorded the highest F1-score for both models (0.999 for DNN and 0.998 for TCN-S2P), while the kettle recorded the lowest (0.841 and 0.779, respectively). The DNN also achieved a higher F1-score for the lamp (0.955 versus 0.944) and iron (0.887 versus 0.846). These results suggest that the model comparison depends on appliance operating characteristics rather than yielding a uniform advantage for one architecture.

For the inverter air conditioner, TCN-S2P achieved a lower MAE (3.19 versus 4.07 W), a lower RMSE (15.23 versus 19.85 W), and a slightly higher F1-score (0.985 versus 0.980) than DNN. Although DNN had better precision and F1-score for the kettle, TCN-S2P achieved a lower kettle RMSE (34.59 versus 41.75 W). The comparison therefore distinguishes appliance operating-state recognition from instantaneous power estimation and does not imply that a single model optimizes every performance metric.

Effect of input window size

Three input window sizes (61, 121, and 241 samples) were examined for each architecture. Their reported performance is summarized in Table 3. For DNN, the 121-sample window produced the highest F1-score (0.932); however, the 61-sample window produced slightly lower MAE (2.40 versus 2.44 W) and RMSE (18.04 versus 19.01 W). Thus, selection of the 121-sample DNN configuration is justified by a classification-oriented selection criterion, not by minimum regression error.

Table 3. Effect of input window size on model performance.

Model	Window (samples)	MAE (W)	RMSE (W)	Precision	Recall	F1-score
DNN	61	2.40	18.04	0.883	0.968	0.920
DNN	121	2.44	19.01	0.907	0.967	0.932
DNN	241	5.20	48.52	0.764	0.765	0.764
TCN-S2P	61	3.39	16.37	0.869	0.969	0.910
TCN-S2P	121	5.47	25.87	0.800	0.971	0.862
TCN-S2P	241	5.24	24.55	0.818	0.963	0.867

For TCN-S2P, the 61-sample configuration gave the highest F1-score (0.910) and the lowest MAE (3.39 W) and RMSE (16.37 W) among the three tested windows. Increasing the DNN window from 121 to 241 samples reduced F1-score from 0.932 to 0.764. TCN-S2P F1-scores for 121 and 241 samples were 0.862 and 0.867, respectively. These results indicate that larger input windows were not consistently beneficial for the present dataset; the reasons for the performance changes cannot be established from these metrics alone.

Effect of training epoch budget

The experiment also evaluates training budgets of 50, 100, and 150 epochs (Table 4). Both models obtained higher reported F1-scores at 50 epochs than at the longer training budgets in this experiment. In the DNN results, F1-score decreased from 0.932 at 50 epochs to 0.590 and 0.789 at 100 and 150 epochs, respectively. Corresponding TCN-S2P F1-scores were 0.862, 0.832, and 0.830. The outcomes suggest that additional training did not improve the reported test performance in these runs. The deterioration observed at larger epoch budgets should not be attributed conclusively to overfitting without examination of comparable training and validation loss trajectories, stopping behavior, and repeated training runs. Moreover, the epoch comparison in Table 4 must not be confused with the final selected TCN-S2P result of 0.910 F1-score, which was reported at a 61-sample window in Table 3.

Table 4. Effect of epoch budget on model performance

Model	Epoch budget	MAE (W)	RMSE (W)	Precision	Recall	F1-score
DNN	50	2.44	19.01	0.907	0.967	0.932
DNN	100	7.69	66.22	0.586	0.594	0.590
DNN	150	4.58	31.94	0.782	0.795	0.789
TCN-S2P	50	5.47	25.87	0.800	0.971	0.862
TCN-S2P	100	8.62	31.58	0.754	0.968	0.832
TCN-S2P	150	6.74	29.36	0.760	0.960	0.830

Overall comparison and practical implications

Among the configurations reported in Table 3, the selected DNN model (121 samples, 50-epoch budget) achieved a mean MAE of 2.44 W, RMSE of 19.01 W, precision of 0.907, recall of 0.967, and F1-score of 0.932. The selected TCN-S2P configuration (61 samples, 50-epoch budget) achieved a mean MAE of 3.39 W, RMSE of 16.37 W, precision of 0.869, recall of 0.969, and F1-score of 0.910. DNN therefore provided higher macro-averaged precision and F1-score, whereas TCN-S2P provided lower macro-averaged RMSE and a marginally higher recall for the selected dataset.

Based on F1-score as the primary selection criterion, DNN was the preferred model for operating-state identification in the reported experiments. However, TCN-S2P remained competitive for power estimation, particularly for the inverter air conditioner. The model ranking is consequently metric-dependent and should not be interpreted as proof of universal DNN superiority.

The findings are limited to data collected from one experimental setup and five selected appliances under offline evaluation. The study does not establish independent-household generalization, deployment latency, or accurate performance during all possible combinations of simultaneous appliance operation. In addition, the available results report active-power errors in watts rather than an integrated energy-estimation error in Wh or kWh. Further investigations using independent operating sessions, rigorous separation of training, validation, and testing data, and energy-based performance indicators are necessary before wider residential deployment can be assessed.

Conclusion

This study demonstrated the potential of DNN and TCN-S2P for residential energy disaggregation using aggregate smart-meter data. The findings revealed distinct strengths between the two models. DNN demonstrated better overall appliance identification performance and lower average prediction error, particularly for appliances with stable and intermittent operating patterns. In contrast, TCN-S2P was more effective in reducing larger prediction errors, with a slight advantage in modelling the variable-power characteristics of the inverter air conditioner. These findings suggest that DNN is preferable when accurate appliance identification and lower average error are prioritized, whereas TCN-S2P may be more suitable when minimizing large prediction deviations is the primary objective. Nevertheless, model performance depends on appliance characteristics and operating conditions.

Future research should incorporate longer monitoring periods, more diverse appliance combinations, and independent validation datasets. Further investigation into computational efficiency and real-time smart-meter integration is also recommended to establish the practical suitability of these models for residential energy management applications.

Acknowledgement

The authors would like to acknowledge the Faculty of Electrical Technology and Engineering, Universiti Teknikal Malaysia Melaka, for providing academic guidance and financial support throughout the development of this project.

References

[1] Hart, G.W. Nonintrusive appliance load monitoring. Proceedings of the IEEE 1992, 80(12), 1870–1891.
[2] Devlin, M.; Hayes, B.P. Non-Intrusive Load Monitoring using Electricity Smart Meter Data: A Deep Learning Approach. In Proceedings of the IEEE Power & Energy Society General Meeting (PESGM); 2019.
[3] Kaselimi, M.; Protopapadakis, A.; Doulamis, N.; Doulamis, A.; Towards Trustworthy Energy Disaggregation: A Review of Challenges, Methods and Perspectives for Non-Intrusive Load Monitoring. Sensors 2022, 22(15), 5872.
[4] Cruz-Rangel, D.; Ocampo-Martinez, C.; Diaz-Rozo, J. Online Non-Intrusive Load Monitoring: A Review. Energy Nexus 2025, 17, 100348.
[5] Shabbir, N.; Vassiljeva, K.; Nourollahi Hokmabad, H.; Husev, O.; Petlenkov, E.; Belikov, J. Comparative Analysis of Machine Learning Techniques for Non-Intrusive Load Monitoring. *Electronics* 2024, *13*, 1420. https://doi.org/10.3390/electronics13081420.
[6] Faisal, M.R.; Hossain, M.J.; Mahmud, M.A. A Review of NILM Applications with Machine Learning. Computers, Materials & Continua 2023, 79(2).
[7] Zhang, C.; Zhong, M.; Wang, Z.; Goddard, N.; Sutton, C. Sequence-to-Point Learning with Neural Networks for Non-Intrusive Load Monitoring. In Proceedings of the AAAI Conference on Artificial Intelligence 2018, 32(1).
[8] Bai, S.; Kolter, J.Z.; Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modelling. arXiv 2018, arXiv:1803.01271.
[9] Sivarasu, S.R.; Chandira Sekaran, E.; Vasantharathna, S. Development of Smart Metering Infrastructure for Residential Distributed Generation/Micro Grid in Indian Power System. International Journal of Applied Engineering Research 2014, 9(24), 27915–27940.
[10] Merza, A.M.; Nasr, M.S. Electrical Energy Billing System Based on Smart Meter and GSM. International Journal of Applied Engineering Research 2015, 10(21), 42003–42012.
[11] Mounia, R.; Velmurugan, K. An Advanced Energy Efficient Load Management System Using Artificial Neural Network. International Journal of Applied Engineering Research 2015, 10(20), 16992–16996.