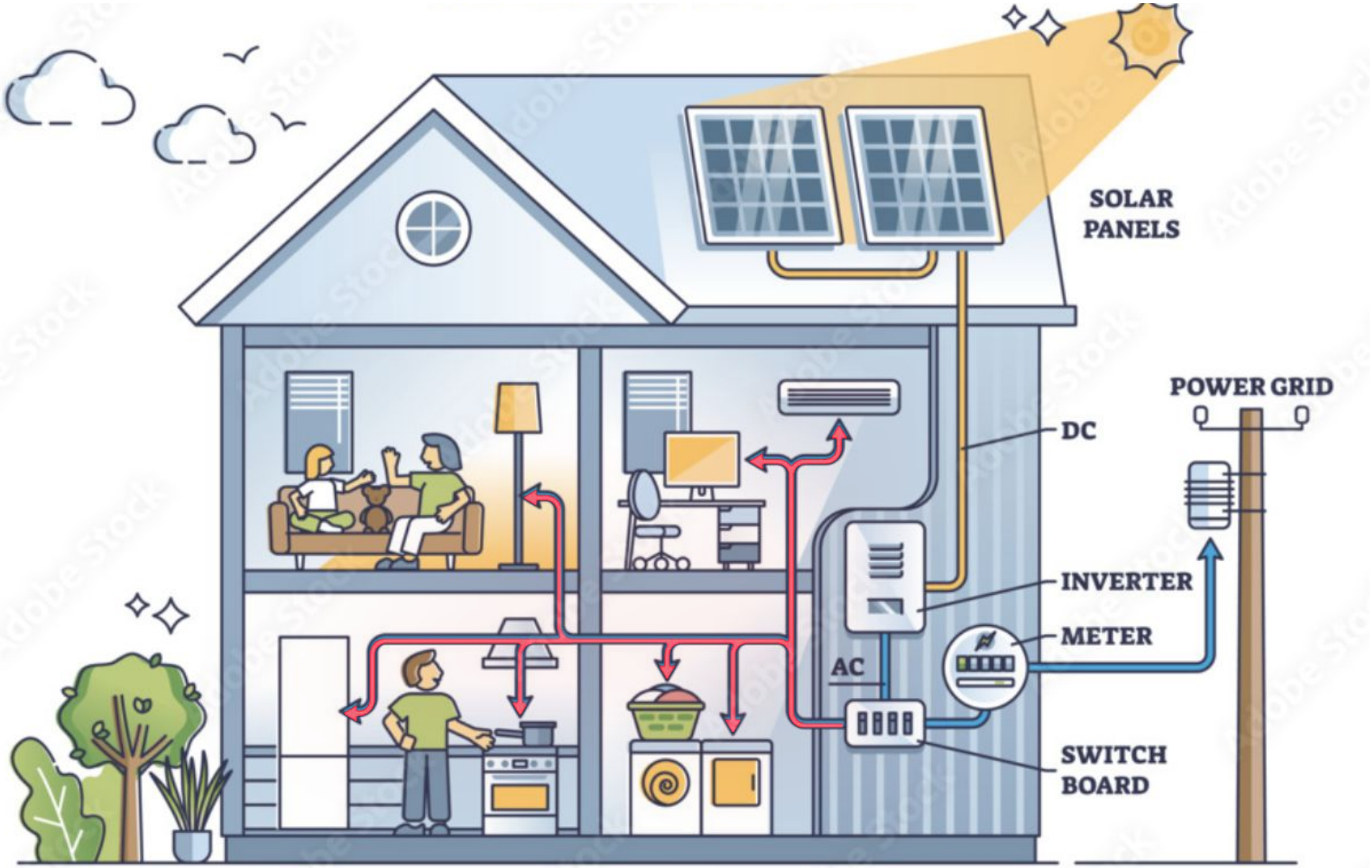
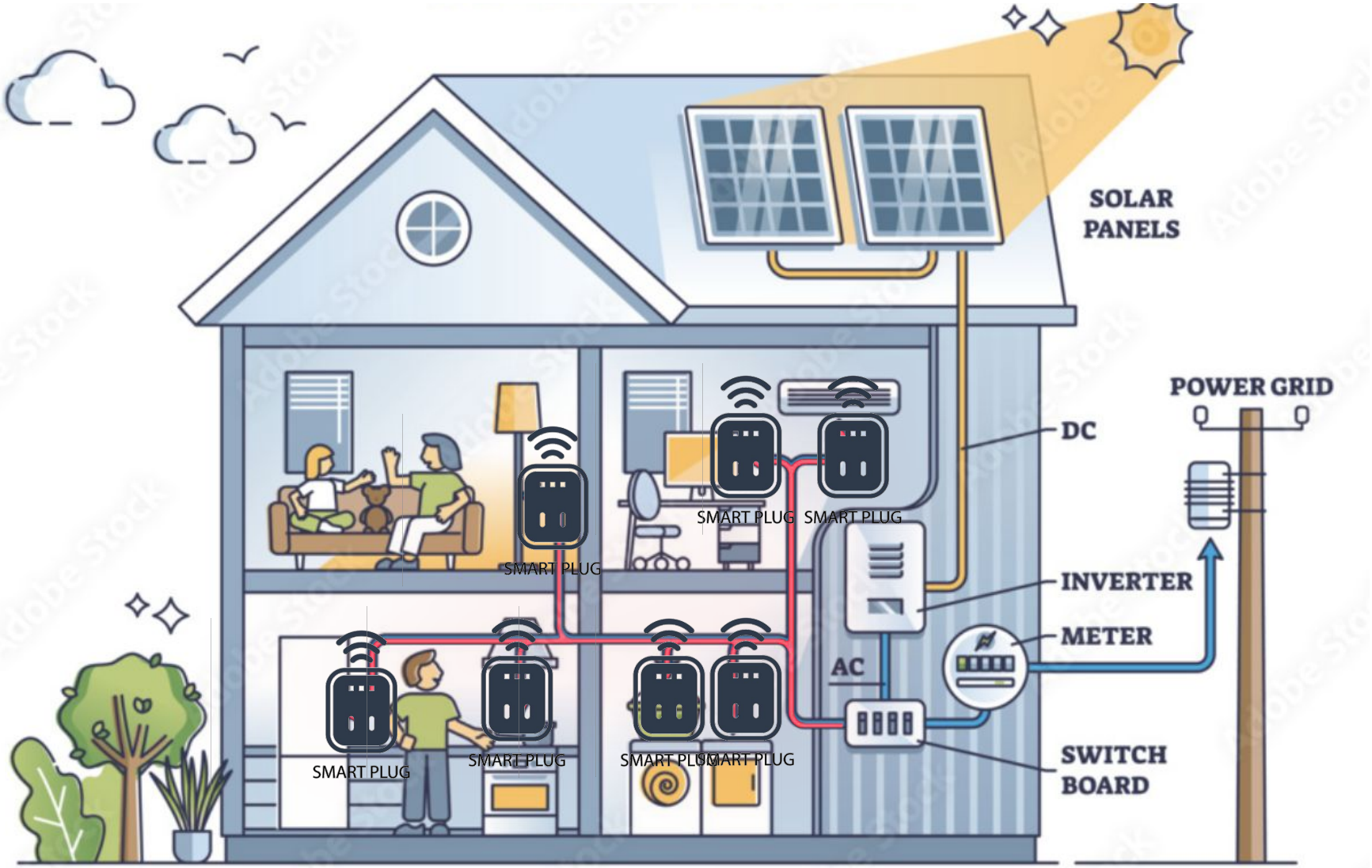
Deep Learning Based Non-intrusive Load Monitoring for Electric Power Systems

Load monitoring introduction



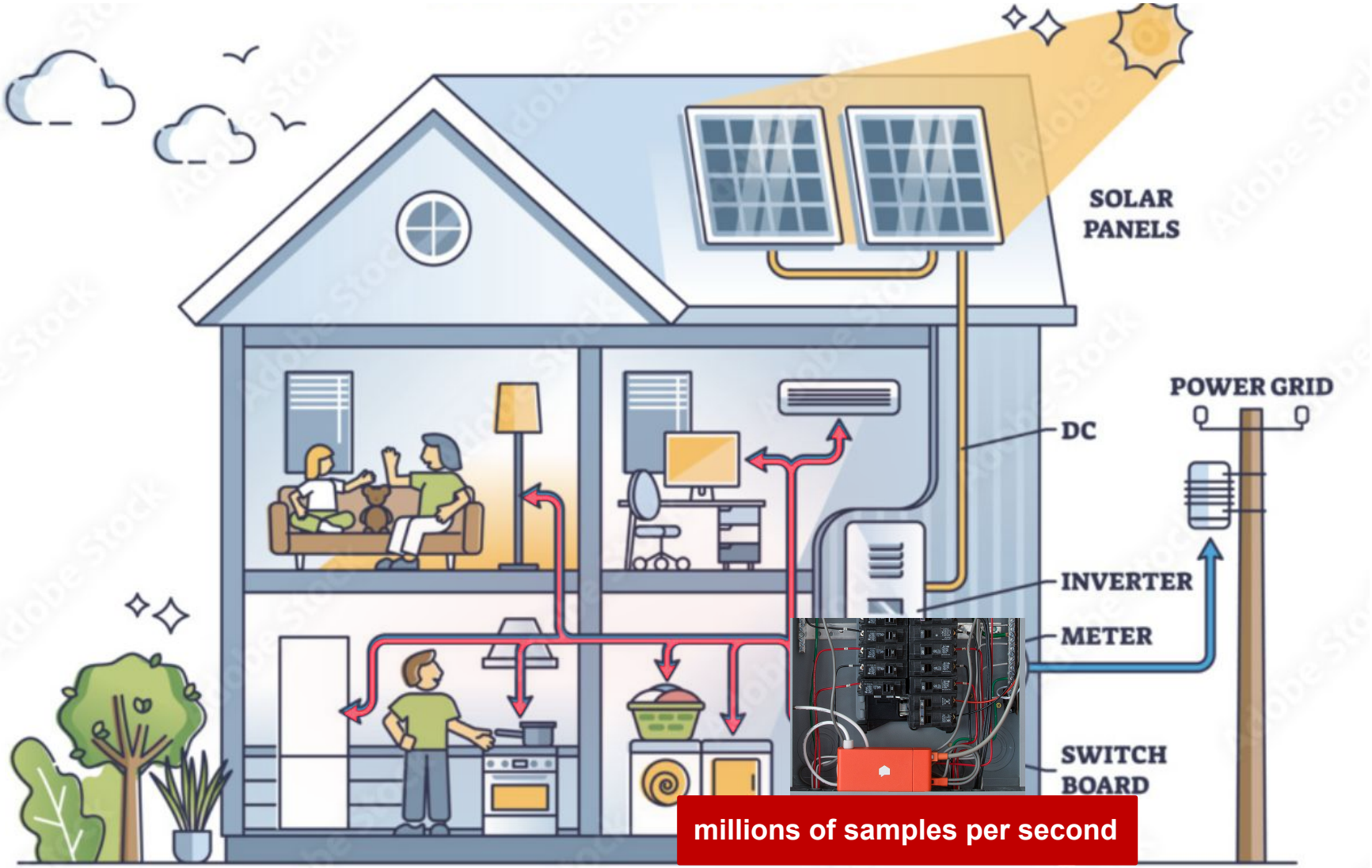
https://as2.ftcdn.net/v2/jpg/05/32/74/25/1000_F_532742547_bAvCSKsjAOswFg97twlOKNZ3ATbxtg4q.jpg

Intrusive load monitoring



https://as2.ftcdn.net/v2/jpg/05/32/74/25/1000_F_532742547_bAvCSKsjAOswFg97twlOKNZ3ATbxtg4q.jpg

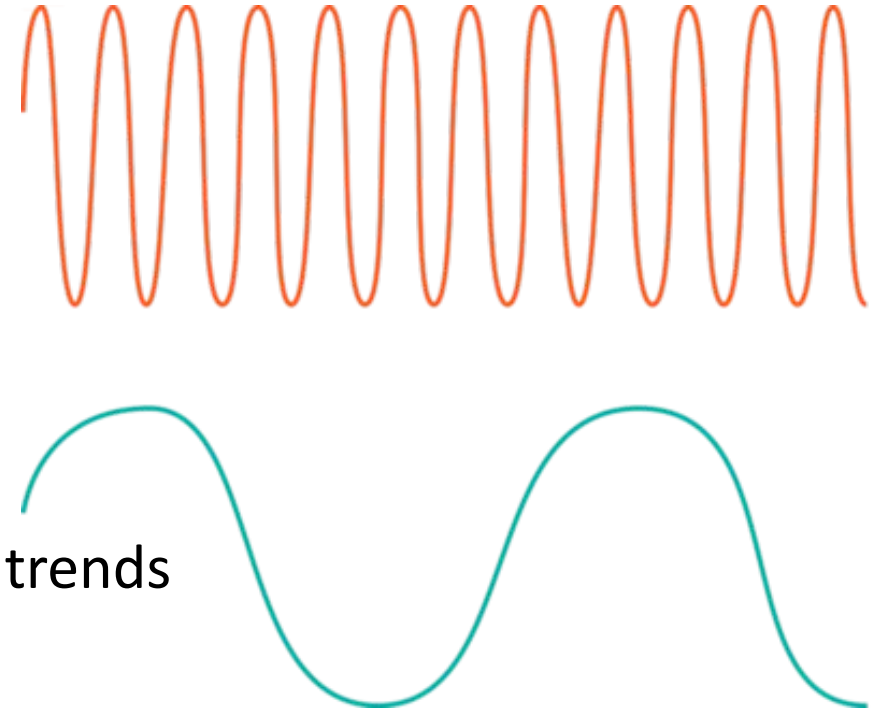
Home energy monitoring system



https://as2.ftcdn.net/v2/jpg/05/32/74/25/1000_F_532742547_bAvCSKsjAOswFg97twlOKNZ3ATbxtg4q.jpg

High / low sampling rate

- High sampling rate
  - Pros: Higher accuracy
  - Cons: Cost, data storage
- Low sampling rate
  - Pros: Cost-effective, storage-efficient, long-term trends
  - Cons: Reduced accuracy



Background

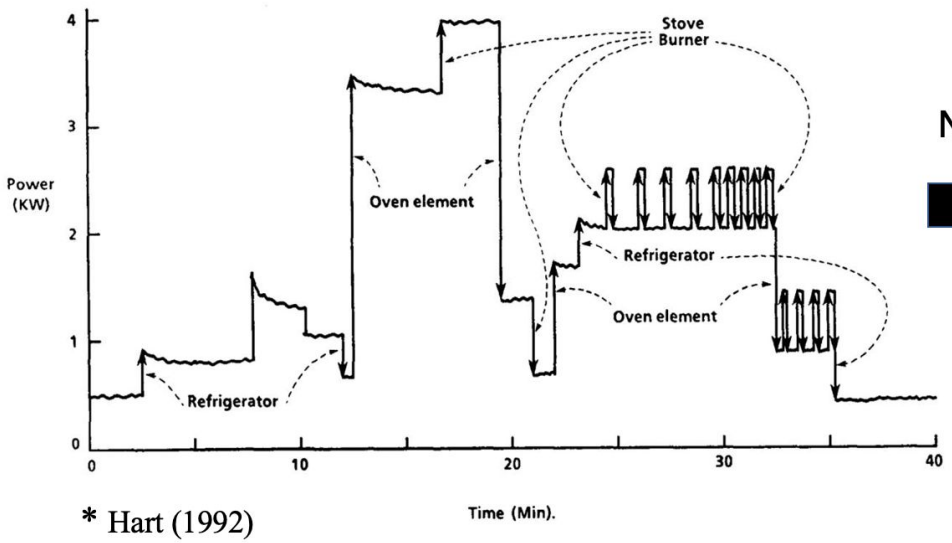
- System level
  - SCADA --- RTU
  - WAMS --- PMU
  - Tasks such as load event identification, fault detection[1]

- Household level
  - Smart meter, smart plug and home energy monitoring system
  - Tasks such as non-intrusive load monitoring (NILM)[2], anomaly appliance detection

1 Ma, Y., Maqsood, A., Oslebo, D., & Corzine, K. (2021). Wavelet Transform Data-driven Machine Learning based Real-time Fault Detection for Naval Dc Pulsating Loads. IEEE Transactions on Transportation Electrification.
2 Adabi, A., Manovi, P., & Mantey, P. (2016). Cost-effective instrumentation via NILM to support a residential energy management system. In 2016 IEEE International Conference on Consumer Electronics (ICCE).

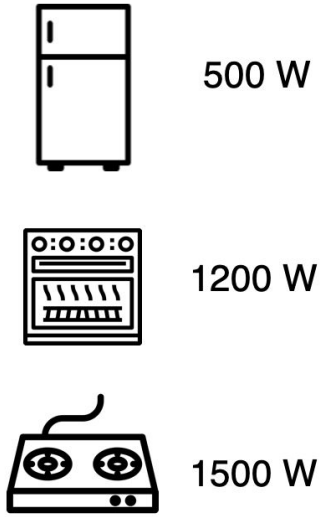
Non-intrusive load monitoring (NILM)

- Non-intrusive Load Monitoring (NILM) is the task of using an aggregate power signal to make inference about the different individual loads of the system.



* Hart (1992)

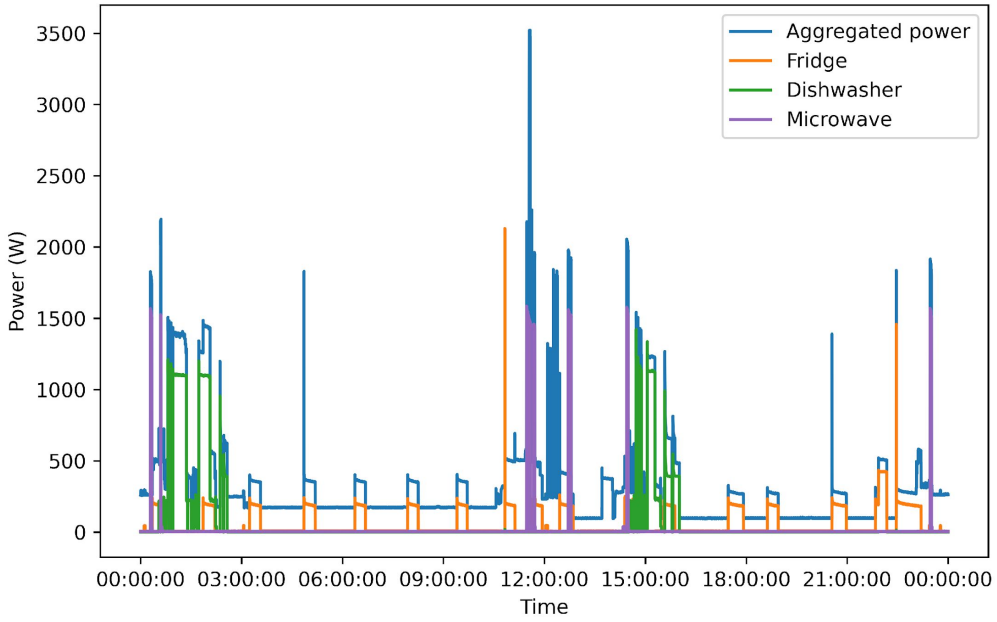
NILM



500 W

1200 W

1500 W

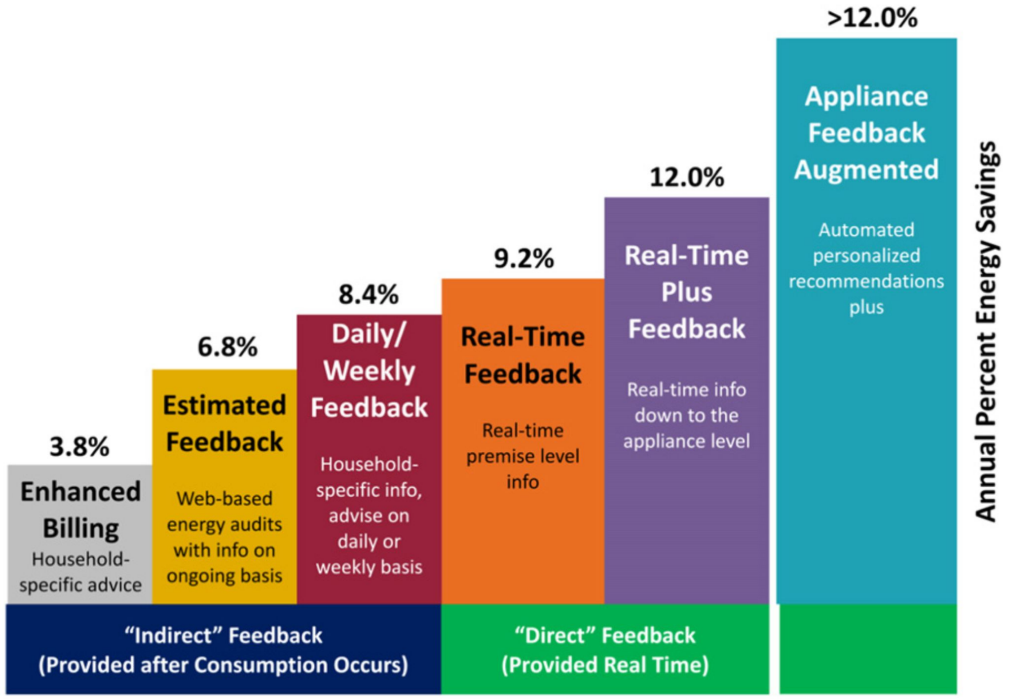


One day load profile from a household in REDD dataset

https://github.com/ch-shin/awesome-nilm

Benefit of NILM

- Appliance-level feedback helps reduce energy consumption.
- Benefits for utilities: Energy efficiency marketing, incentives and rate structures, program design and evaluation, policy recommendation.



Armel, K. Carrie, et al. "Is disaggregation the holy grail of energy efficiency? The case of electricity." Energy policy. 2013.

Motivation

- Data requirements for NILM models
  - Traditional belief: Need for extensive household data.
  - Limitations: Restricted monitoring and few households' data in datasets.

Motivation

- Data requirements for NILM models
  - Traditional belief: Need for extensive household data.
  - Limitations: Restricted monitoring and few households' data in datasets.
- Questions:
  - Is data recorded over months to years a necessity?
  - Is household-labeled data essential?

Motivation

- Data requirements for NILM models
  - Traditional belief: Need for extensive household data.
  - Limitations: Restricted monitoring and few households' data in datasets.
- Questions:
  - Is data recorded over months to years a necessity?
  - Is household-labeled data essential?
- Answers:
  - No and No.

Problem formulation

- The aggregate load consumption at each time step is given by:

$$x_t = \sum_{i}^{n} y_t^i + u_t + \epsilon_t$$

x_t	Aggregate power
y_t^i	i^{th} appliance power
u_t	Unknown appliances power
ϵ_t	Measurement noise

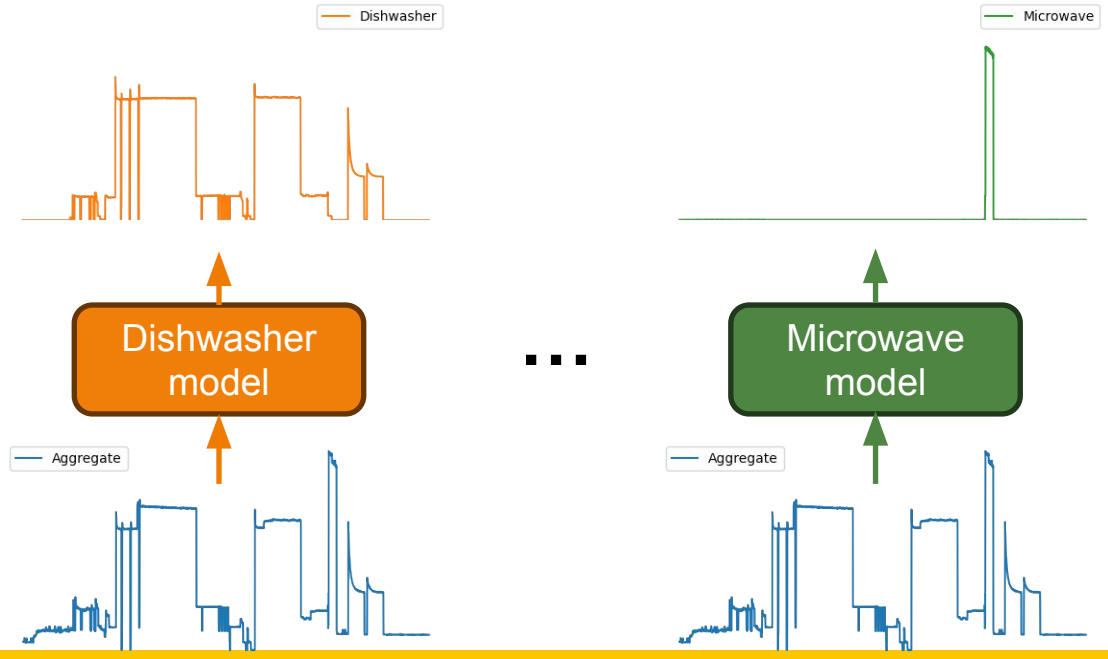
Problem formulation

➢ The aggregate load consumption at each time step is given by:

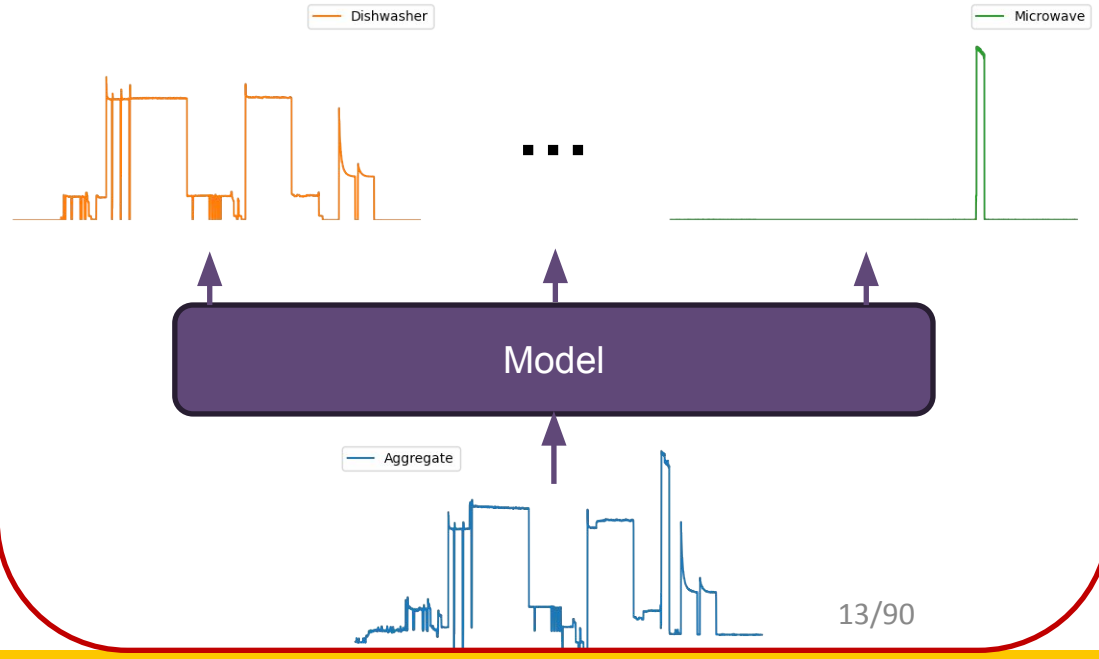
$$x_t = \sum_i^n y_t^i + u_t + \epsilon_t$$

x_t	Aggregate power
y_t^i	i[th] appliance power
u_t	Unknown appliances power
ϵ_t	Measurement noise

Single-appliance

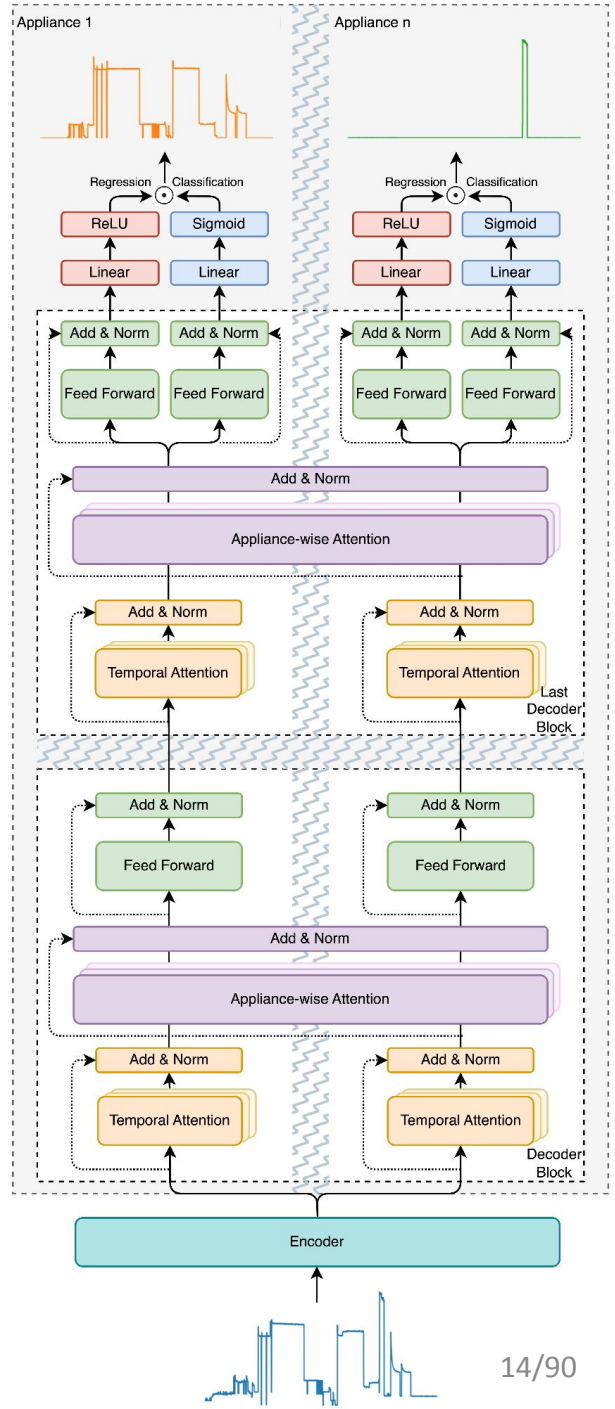


Multi-appliances



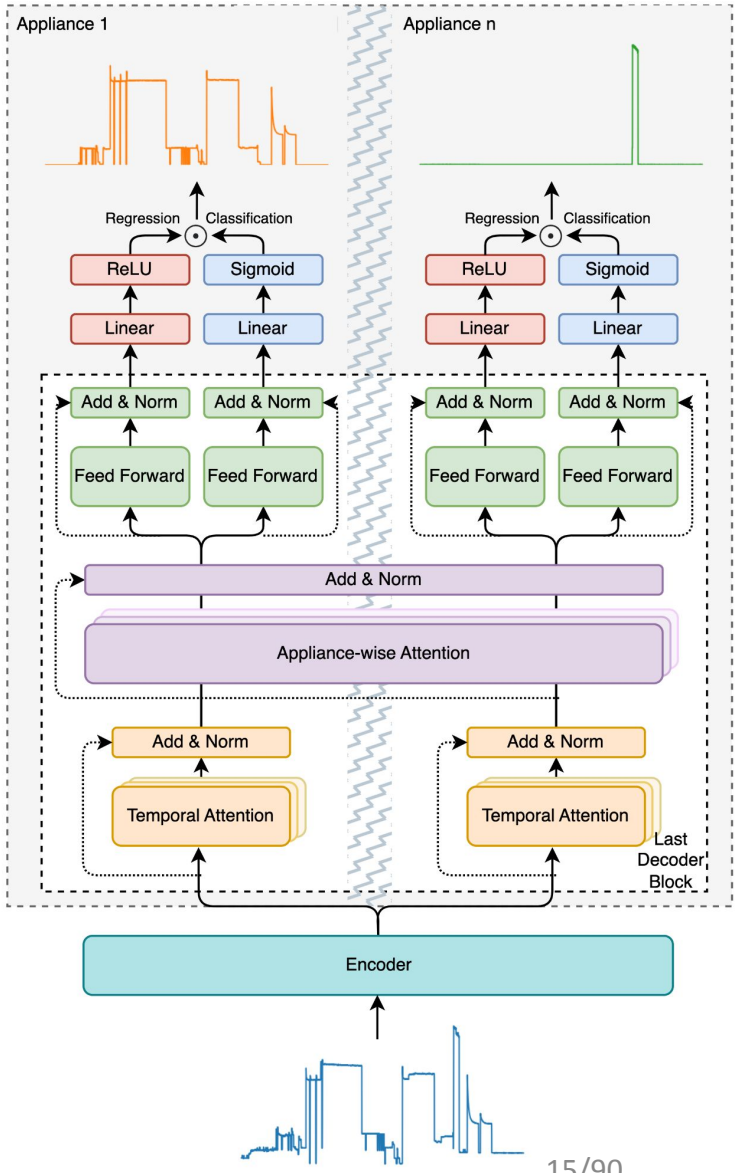
Model architecture

- Developed a shared-hierarchical split structure for its regression and classification tasks.
- Designed a two-dimensional attention mechanism to capture spatial-temporal correlations among all appliances.



Model architecture

- Developed a shared-hierarchical split structure for its regression and classification tasks.
- Designed a two-dimensional attention mechanism to capture spatial-temporal correlations among all appliances.



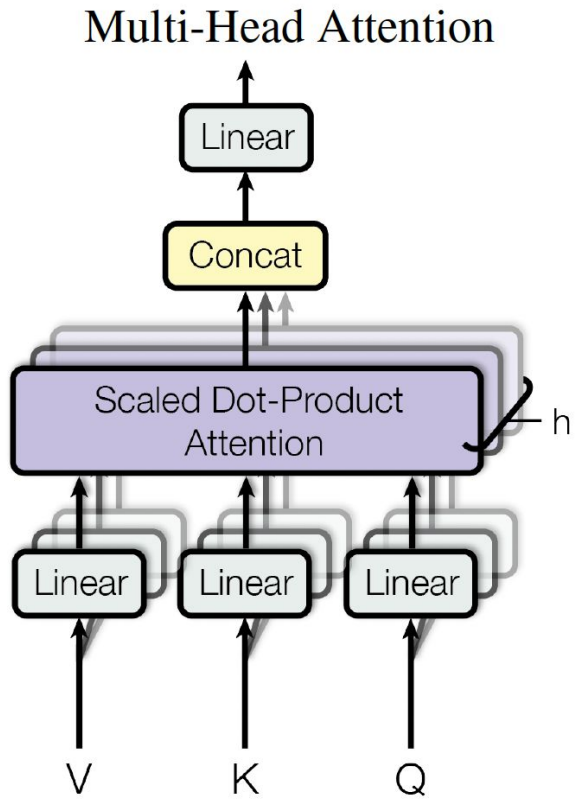
Multi-head attention

Allows the model to jointly attend to information from different representation subspaces at different positions.

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V},$$

$$\text{head}_\text{i} = \text{Attention}\left(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V\right),$$

$$\text{MH}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}\left(\text{head}_1, \ldots, \text{head}_\text{h}\right)\mathbf{W}^O$$



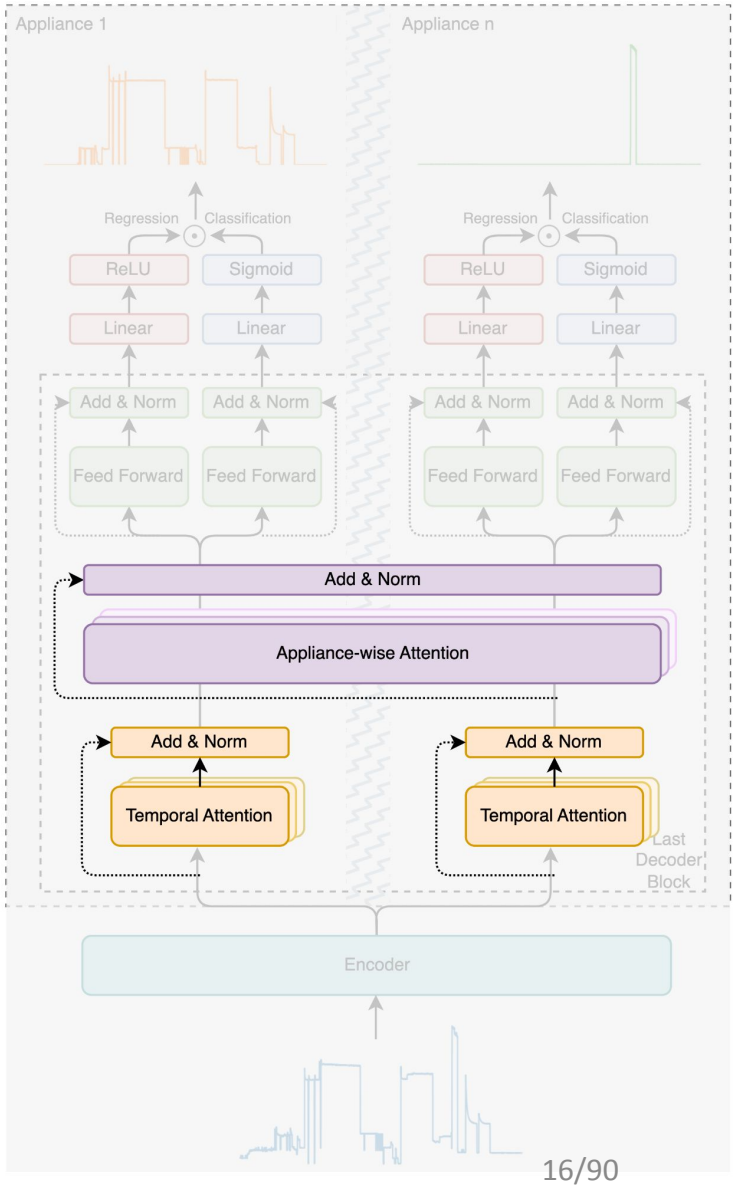


Figure: Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30. 2017.

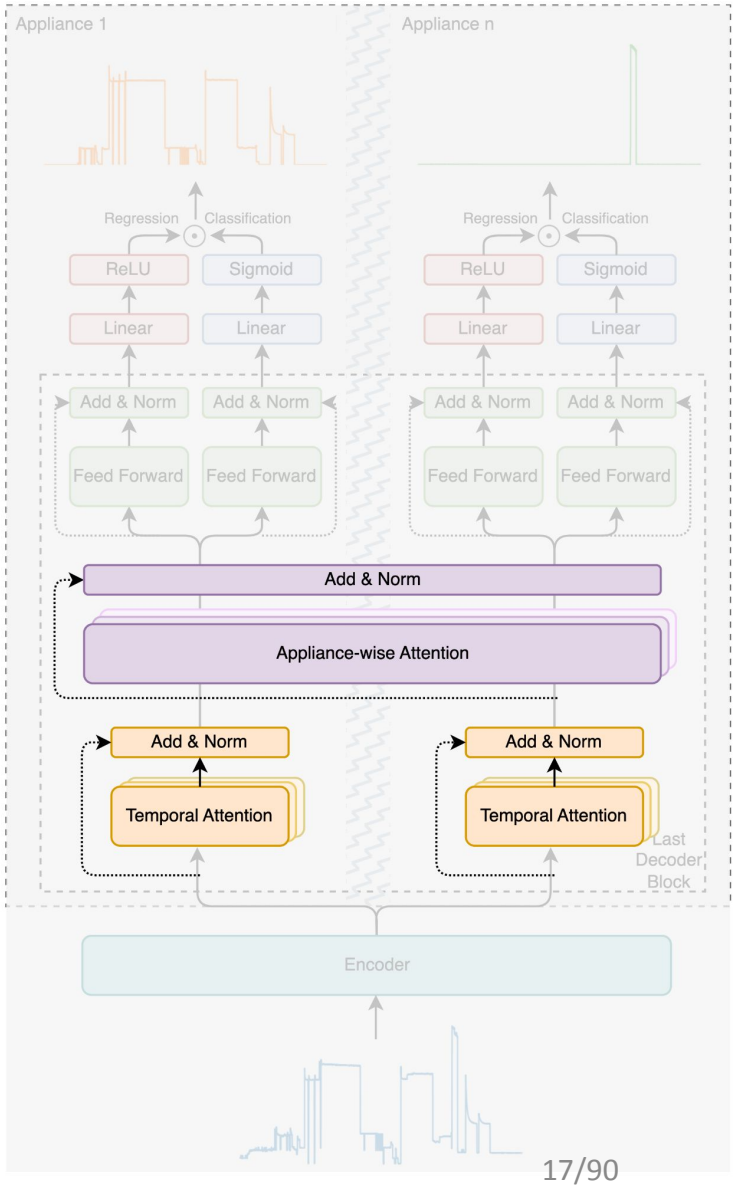
Two-dimensional multi-head self-attention 2DMA

- Temporal Attention

$$\boldsymbol{\theta}^{i}_{t+w:t+T-w-1} = \mathrm{LayerNorm}\left(\mathbf{h}^{i}_{t+w:t+T-w-1} + \mathrm{MH}\left(\mathbf{h}^{i}_{t+w:t+T-w-1}\right)\right)$$

- Appliance-wise Attention

$$\boldsymbol{\phi}^{i\in\mathcal{K}}_{t} = \mathrm{LayerNorm}\left(\boldsymbol{\theta}^{i\in\mathcal{K}}_{t} + \mathrm{MH}\left(\boldsymbol{\theta}^{i\in\mathcal{K}}_{t}\right)\right)$$



Proposed approach

- The encoder aims to learn a hidden shared representation from the aggregate signal: $\mathbf{h}_{t+w:t+T-w-1} = \text{Encoder}\left(\mathbf{x}_{t:t+T-1}\right)$
- Then, the decoder updates hidden representations of all appliances by using the 2DMA layers and a fully connected feed-forward network:

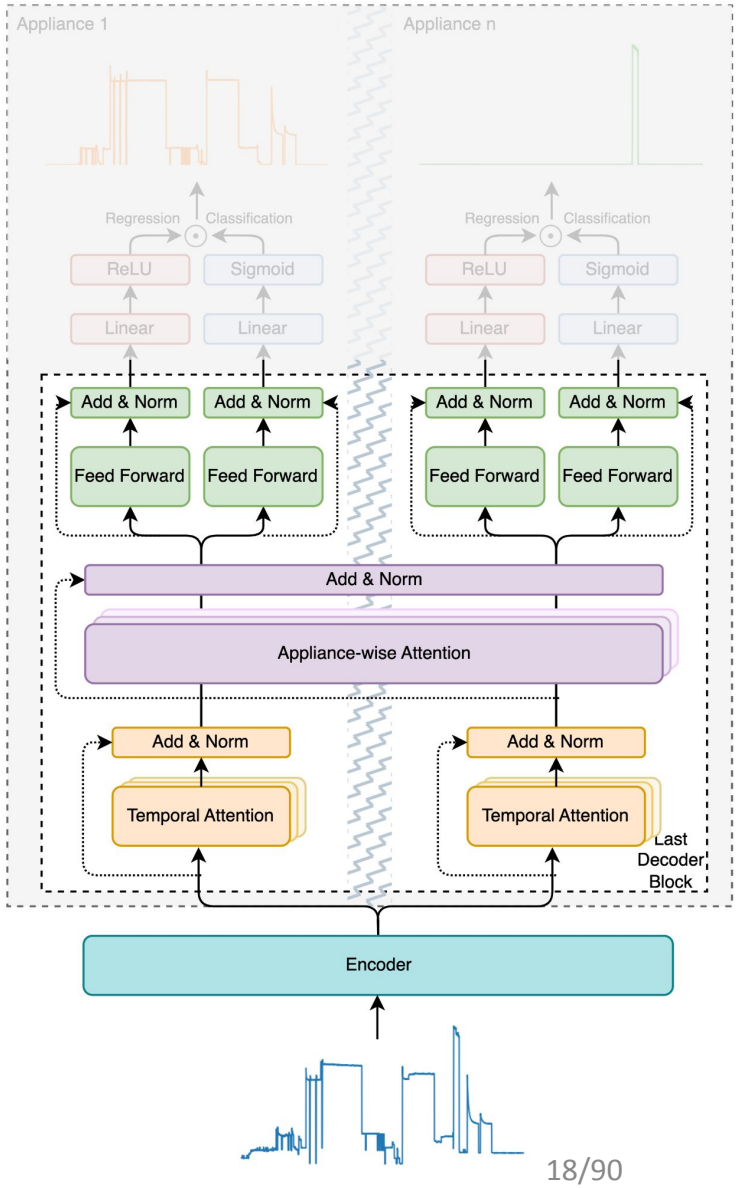
$$\boldsymbol{\phi}_t^i = \text{2DMA}\left(\mathbf{h}_{t+w:t+T-w-1}\right)$$

$$\mathbf{h}_t^i = \text{LayerNorm}\left(\boldsymbol{\phi}_t^i + \max\left(0, \boldsymbol{\phi}_t^i \mathbf{W}_1^i\right)\mathbf{W}_2^i\right)$$

- The last decoder block further splits the network to explore power consumption and on/off state

$$\mathbf{h}_{c,t}^i = \text{LayerNorm}\left(\boldsymbol{\phi}_t^i + \max\left(0, \boldsymbol{\phi}_t^i \mathbf{W}_{c,1}^i\right)\mathbf{W}_{c,2}^i\right)$$

$$\mathbf{h}_{r,t}^i = \text{LayerNorm}\left(\boldsymbol{\phi}_t^i + \max\left(0, \boldsymbol{\phi}_t^i \mathbf{W}_{r,1}^i\right)\mathbf{W}_{r,2}^i\right)$$



Proposed approach

- Regression subnetwork outputs the power consumption.

$$\hat{p}_t^i = \mathrm{ReLU}(\mathbf{h}_{r,t}^i \mathbf{W}_r^i)\mathbf{V}_r^i$$

- Classification subnetwork outputs on/off status.

$$\hat{o}_t^i = \mathrm{Sigmoid}(\mathbf{h}_{c,t}^i \mathbf{W}_c^i)\mathbf{V}_c^i$$

- The final estimated power consumption:

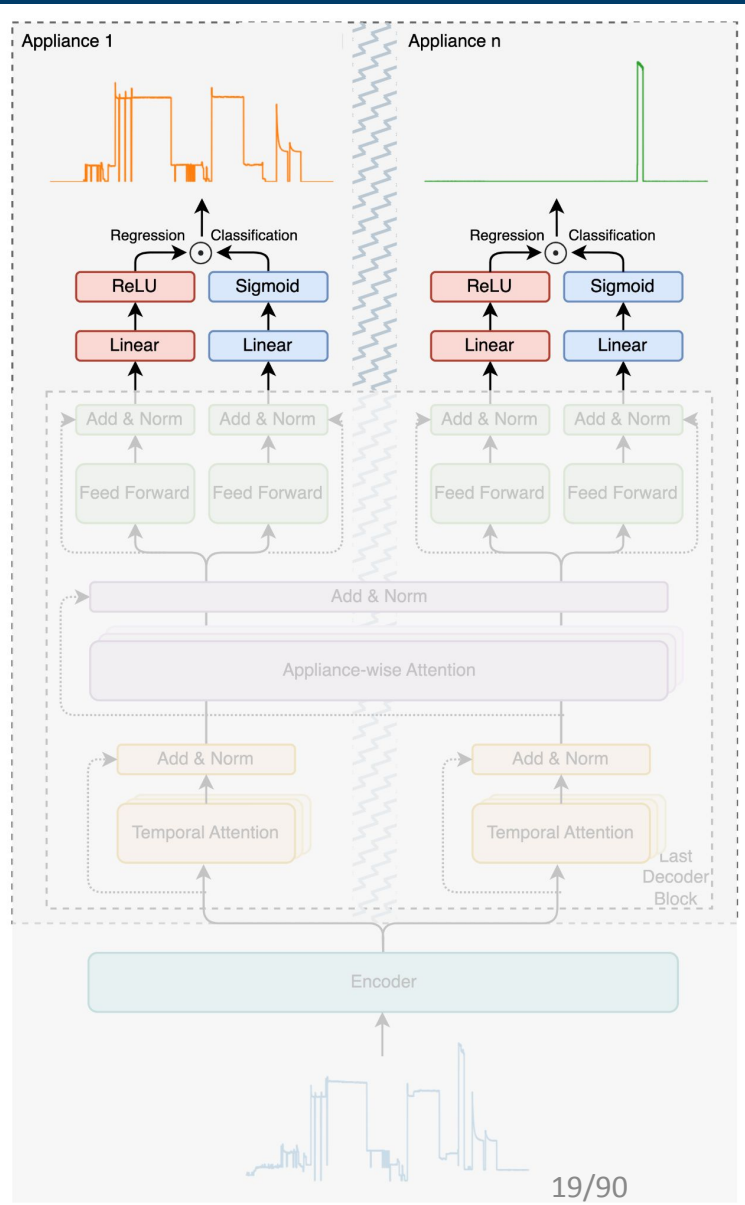
$$\hat{y}_t^i = \hat{p}_t^i \times \hat{o}_t^i$$

- Loss function:

$$\mathcal{L} = \sum_{i=1}^{|\mathcal{K}|} (\mathcal{L}_{\text{output}}^i + \mathcal{L}_{\text{on}}^i)$$

$$\mathcal{L}_{\text{output}}^i = \frac{1}{T} \sum_t^T \left(y_t^i - \hat{p}_t^i \hat{o}_t^i\right)^2$$

$$\mathcal{L}_{\text{on}}^i = -\frac{1}{T} \sum_{t=1}^T \left(o_t^i \log \hat{o}_t^i + \left(1 - o_t^i\right) \log\left(1 - \hat{o}_t^i\right)\right)$$



Sample augmentation

- The aggregate load consumption at each time step is given by:

$$x_t = \sum_{i}^{n} y_t^i + u_t + \epsilon_t$$

x_t	Aggregate power
y_t^i	i^{th} appliance power
u_t	Unknown appliances power
ϵ_t	Measurement noise

- Using appliance operation profile can mimic the aggregate load.
- Benefit of collecting appliance operation profile:
  - No privacy issue
  - "Easier to collect"

An appliance's operation profile is defined as a sequence of sampled power measurements over one complete operation cycle

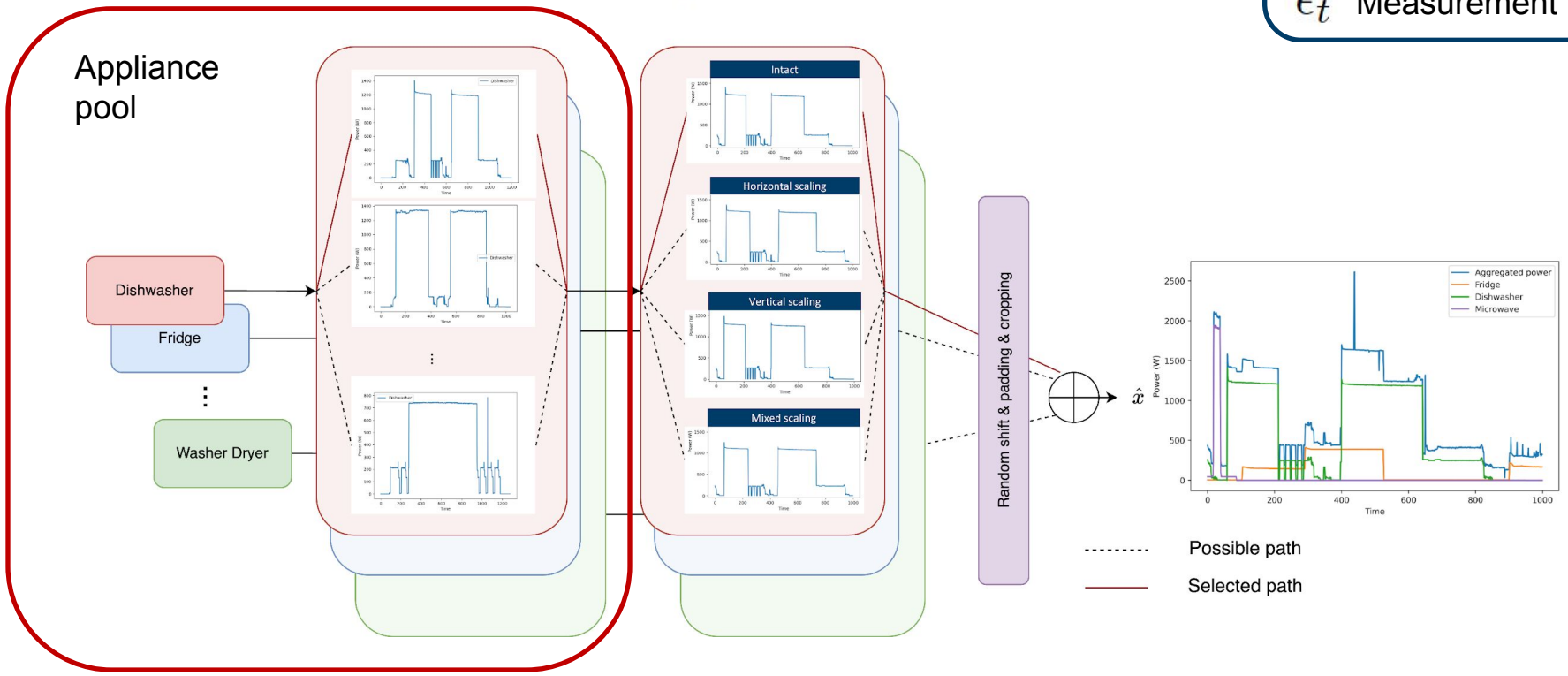
https://promova.com/english-vocabulary/home-equipment-and-appliances-list

Sample augmentation

➢ The aggregate load consumption at each time step is given by:

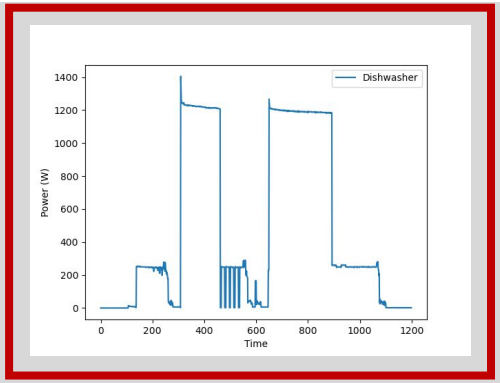
$$x_t = \sum_{i}^{n} y_t^i + u_t + \epsilon_t$$

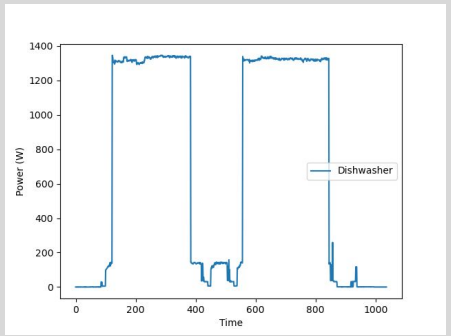
x_t Aggregate power
y_t^i i^{th} appliance power
u_t Unknown appliances power
ϵ_t Measurement noise



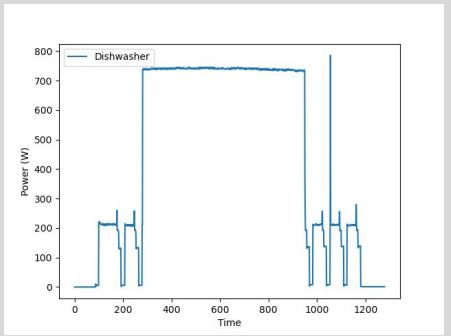
Appliance pool with appliance active profiles

Dishwasher

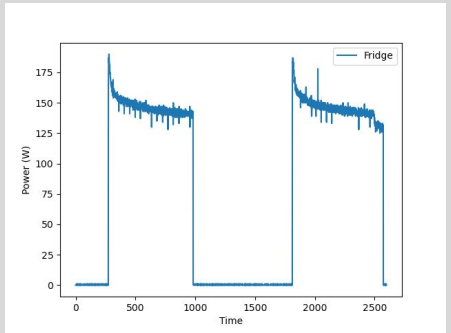


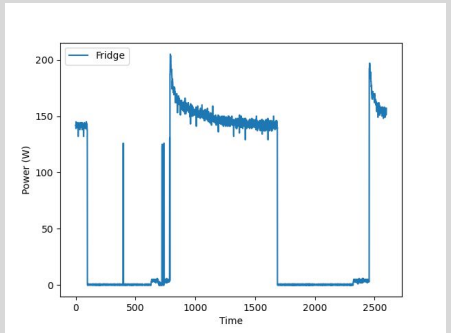


. . .

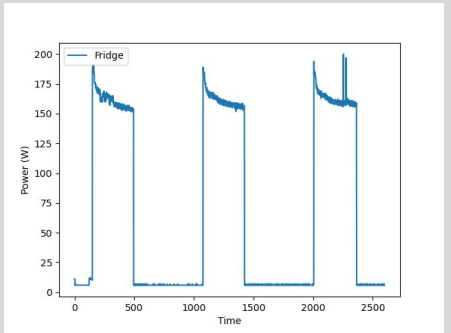


Fridge

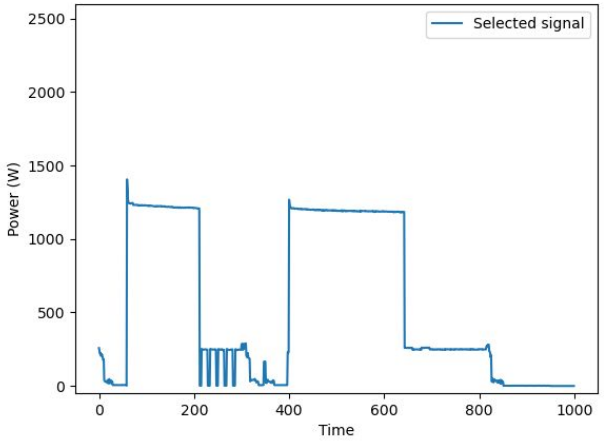




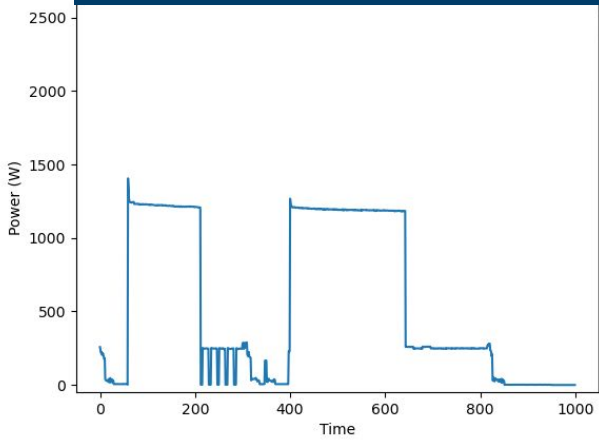
. . .



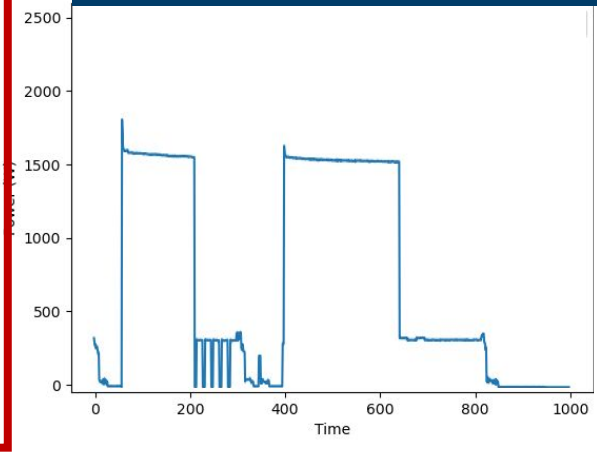
Sample augmentation steps



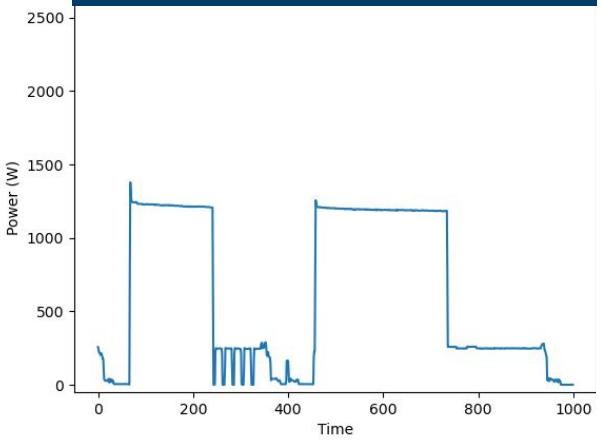
Intact



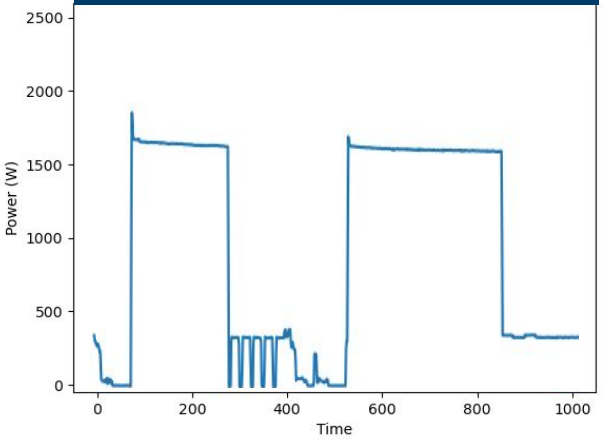
Vertical scaling



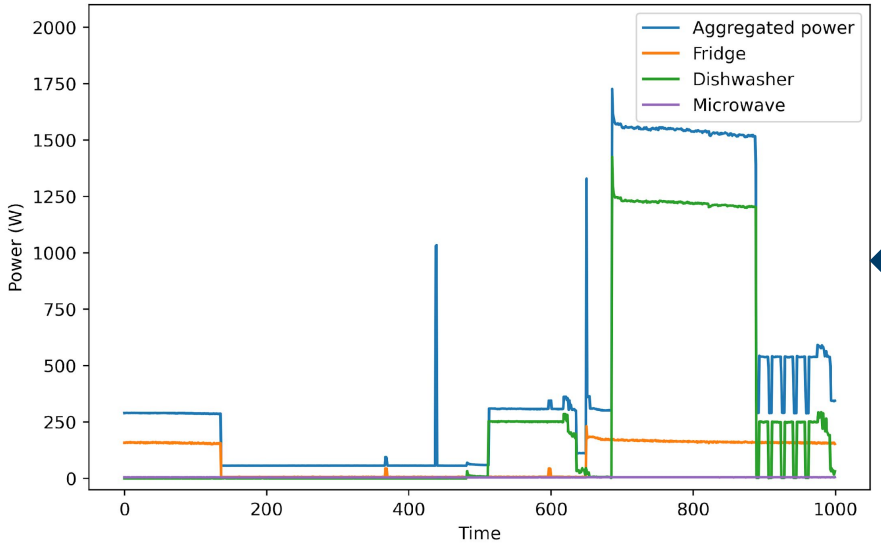
Horizontal scaling



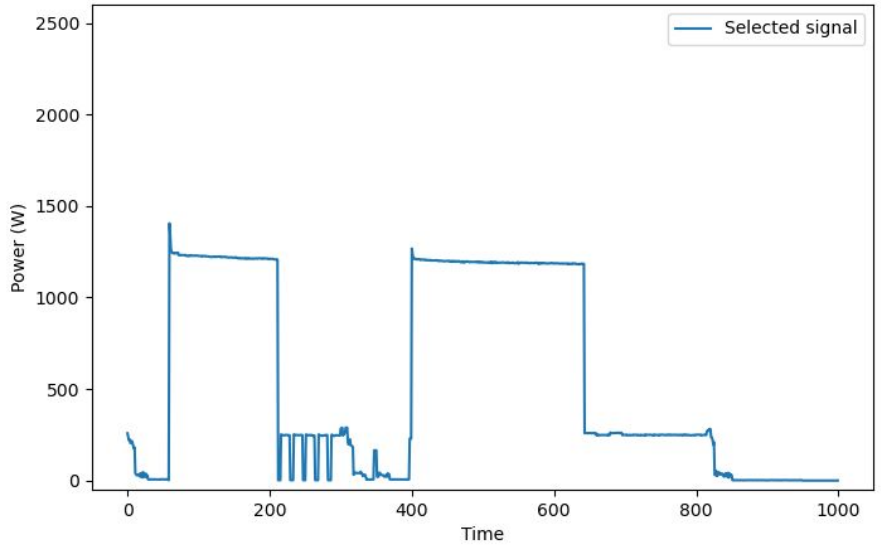
Mixed scaling



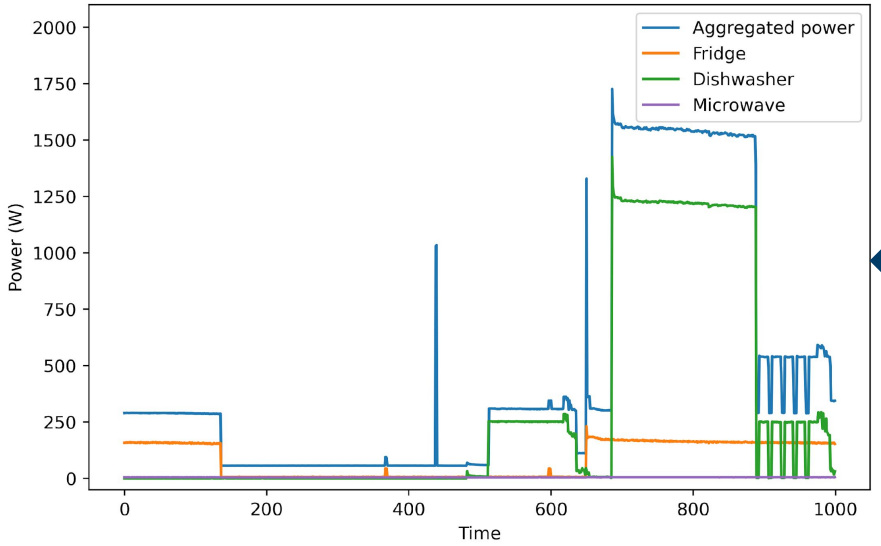
Sample augmentation steps



Original Sample

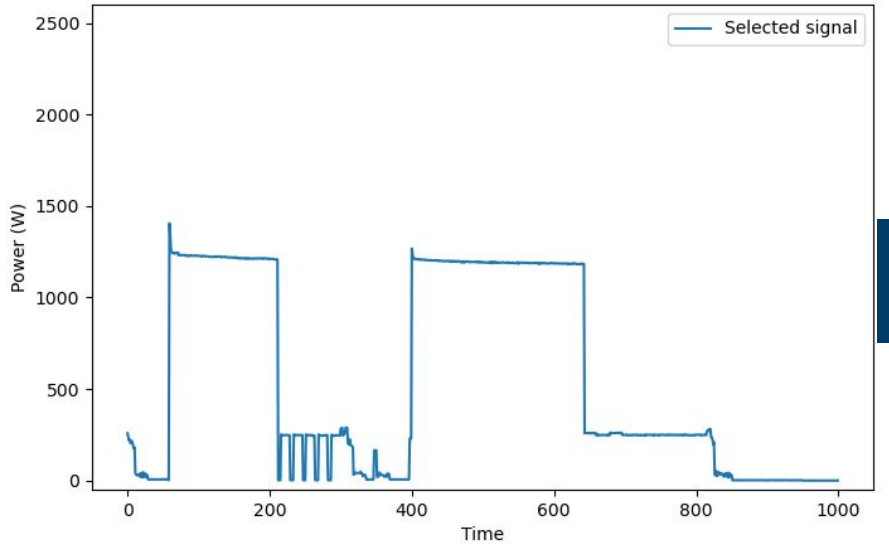


Sample augmentation steps

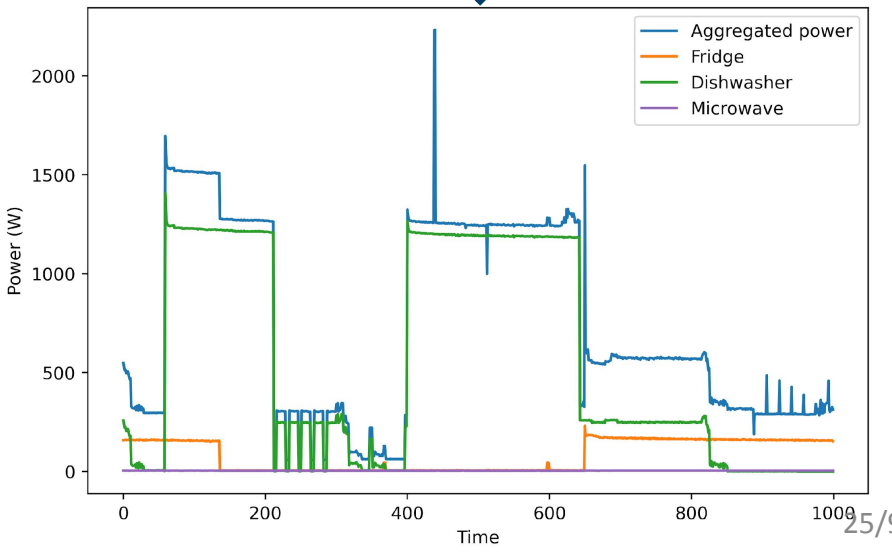


Original Sample

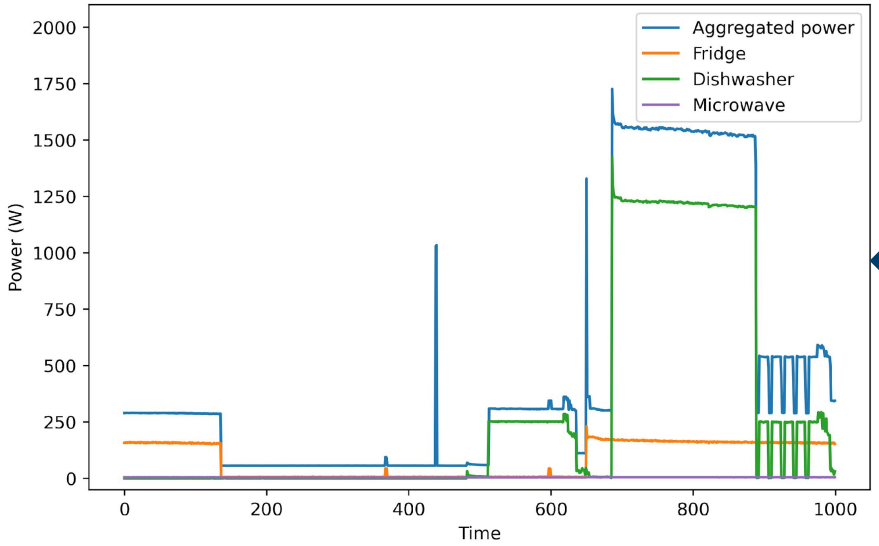
Repeat the steps for all appliances



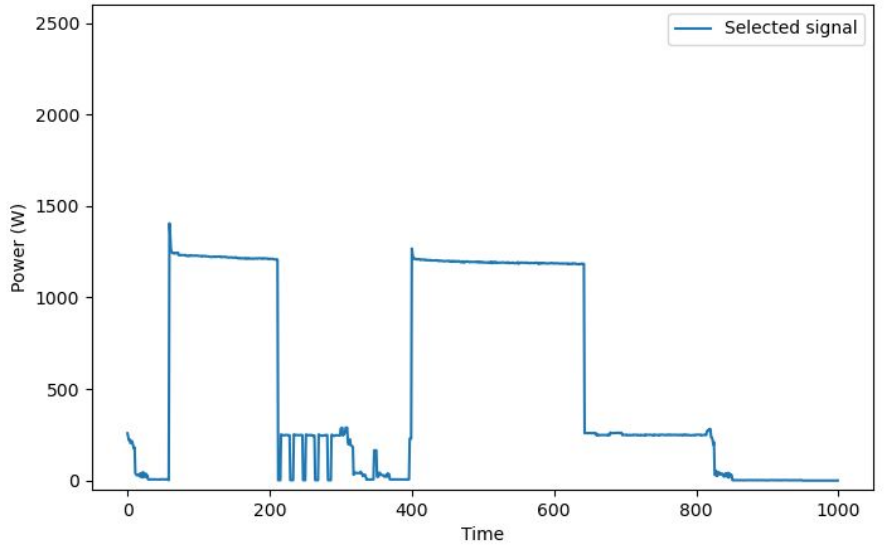
Selected



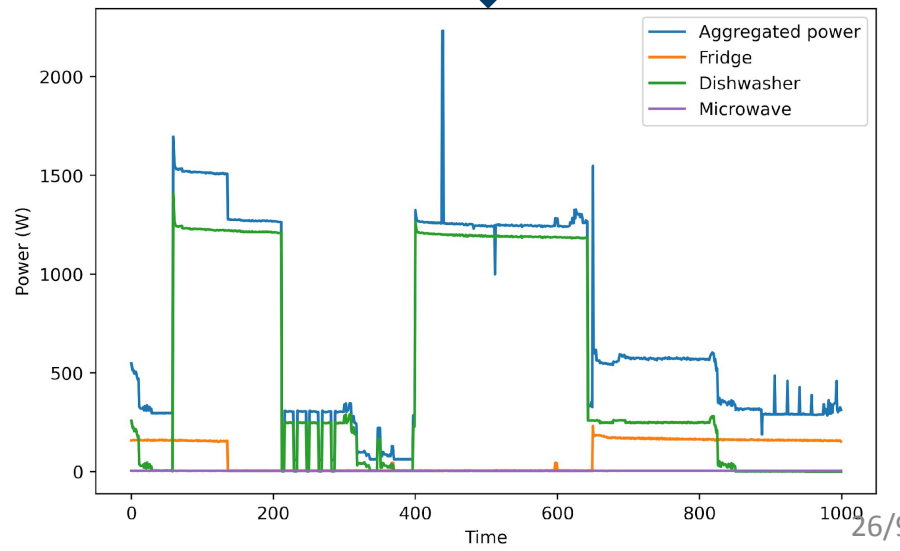
Sample augmentation steps



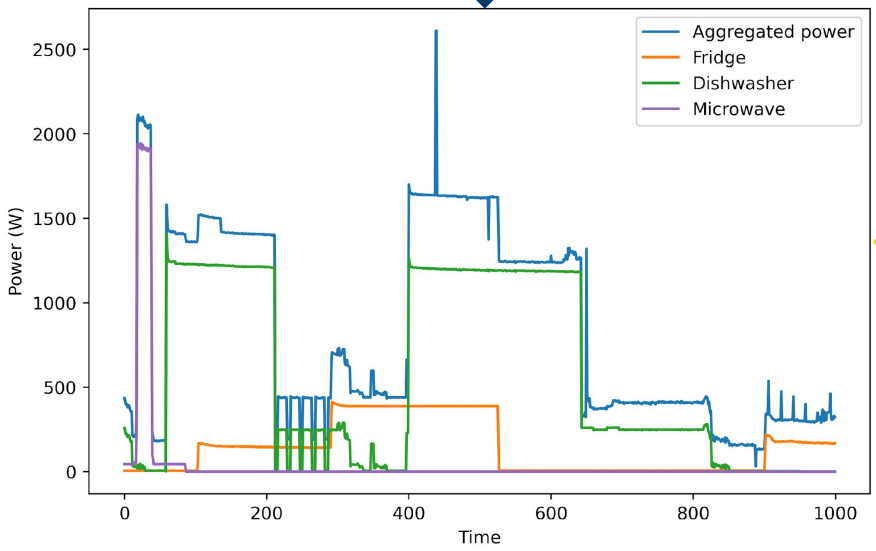
Original Sample



Repeat the steps for all appliances



New Sample



Datasets

- Reference Energy Disaggregation Dataset (REDD)
  - 6 US houses time ranges varying from 23 to 48 days.
  - Resolution: 1s for aggregate signal and 3s for appliance signal.

- UK Domestic Appliance-Level Electricity (UK-DALE)
  - 5 UK houses time ranges varying from 36 to 655 days.
  - Resolution: 6s for both aggregate and appliance signal.

Three scenarios

- S1 is a baseline using the **full dataset**, like the prior works assume.
  - REDD: House 2-6 for training, house 1 for testing
- S2 uses **limited labeled data**: one-day for training and one-day for validation.
- S3 = S2 + sample augmentation.

Table: Training, validation, and testing datasets S2 and S3

		House #	Time index
REDD	Training	3	2011-04-21 19:41:24 - 2011-04-22 19:41:21
	Validation	3	2011-05-23 10:31:24 - 2011-05-24 10:31:21
	Testing	1	2011-04-18 09:22:12 - 2011-05-23 09:21:51
UK-DALE	Training	1	2017-04-23
	Validation	1	2017-04-25
	Testing	2	2017-04-12 - 2017-04-25

How much data needed?

$$\mathrm{MAE} = \frac{1}{H}\sum_{t=1}^{H}|y_t - \hat{y}_t|$$

$$\mathrm{SAE} = \frac{1}{S}\sum_{\tau=0}^{S-1}\frac{1}{M}|y_\tau - \hat{y}_\tau|$$

$$F_1 = 2 \times \frac{\mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

REDD: COMPARISON OF THE SGN MODEL TRAINED ON THE FULL DATASET (S1) VERSUS ON THE LIMITED DATASET WITH SA (S3).

Metric	Scenario	DW	FD	MW	WD	Average
MAE	S1	19.51	**27.27**	23.61	39.22	27.40
	S3	**14.31**	28.68	**16.90**	**21.79**	**20.42**
SAE	S1	18.92	17.23	18.33	34.79	22.32
	S3	**12.73**	**16.77**	**11.94**	**16.12**	**14.39**
F1	S1	0.34	0.83	0.04	0.29	0.38
	S3	**0.70**	**0.85**	**0.65**	**0.59**	**0.70**

S1 is a baseline using the **full dataset**, like the prior works assume.

S2 uses **limited labeled data**: one-day for training and one-day for validation.

S3 = S2 + sample augmentation.

Acronyms

DW: Dishwasher

FG: Fridge

MW: Microwave

WD: Washer dryer

Limited data scenario

Performance comparisons of existing models and the proposed framework with limited data for REDD and UK-Dale dataset

Metric	Model	Scenario	REDD						UK-DALE						
			DW	FD	MW	WD	Ave	Imp	DW	FD	MW	KT	WD	Ave	Imp
MAE	SGN - Conv [18]	S2	22.14	39.01	19.40	40.13	30.17	-	19.81	27.05	8.03	15.09	20.92	18.18	-
	SGN - LSTM	S2	21.98	41.46	21.29	43.91	32.16	-	21.40	32.35	7.90	7.22	21.24	18.02	-
	MAT - Conv	S3	**8.44**	19.38	13.40	**16.44**	14.41	52.23%	10.88	17.06	**7.08**	5.95	6.52	9.50	47.75%
	MAT - LSTM	S3	9.12	**17.86**	**12.49**	17.08	**14.14**	56.04%	**6.51**	**15.86**	8.20	**5.36**	**5.37**	**8.26**	54.17%
SAE	SGN - Conv [18]	S2	21.83	25.78	17.87	39.11	26.15	-	15.00	13.07	8.06	13.09	20.79	14.00	-
	SGN - LSTM	S2	21.61	31.71	17.89	39.49	27.67	-	19.52	16.23	7.49	5.49	20.24	13.79	-
	MAT - Conv	S3	**6.94**	**12.30**	10.47	**13.35**	10.76	58.84%	9.44	6.93	**5.53**	**3.01**	4.54	5.89	57.92%
	MAT - LSTM	S3	8.20	12.67	**9.81**	14.29	**10.23**	63.05%	**5.39**	**6.79**	6.70	3.34	**3.65**	**5.18**	62.48%
F1	SGN - Conv [18]	S2	0.19	0.80	0.10	0.33	0.36	-	0.85	0.65	0.00	0.39	0.08	0.39	-
	SGN - LSTM	S2	0.23	0.71	0.08	0.61	0.41	-	0.76	0.60	0.18	0.85	0.09	0.50	-
	MAT - Conv	S3	0.80	0.88	0.66	**0.67**	0.75	110.98%	0.82	0.83	**0.67**	**0.91**	0.79	0.80	104.41%
	MAT - LSTM	S3	**0.82**	**0.91**	**0.74**	0.64	**0.78**	89.23%	**0.91**	**0.84**	0.65	0.90	**0.82**	**0.83**	66.76%

S1 is a baseline using the **full dataset**, like the prior works assume.

S2 uses **limited labeled data**: one-day for training and one-day for validation.

S3 = S2 + sample augmentation.

Acronyms

DW: Dishwasher

FG: Fridge

MW: Microwave

WD: Washer dryer

Contribution

- Designed a MATNilm with 2-D attention mechanism capturing correlations across branches and temporal steps.

- Proposed efficient sample augmentation algorithm that achieved full-dataset training performance with only 3% of the ground truth and limited appliance profile for sample augmentation.