Adaptive Machine Learning for Resource-Constrained Environments

Sebastián Andrés Cajas, Jaydeep Samanta, Andrés L. Suárez-Cetrulo, and Ricardo Simón Carbajo

30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining August 25 - 29, 2024 - Barcelona, Spain

Ireland’s Centre for Artificial Intelligence (CeADAR), University College Dublin.
V2N9, Dublin, Ireland

ICOS

Funded by the European Union

Main contributions

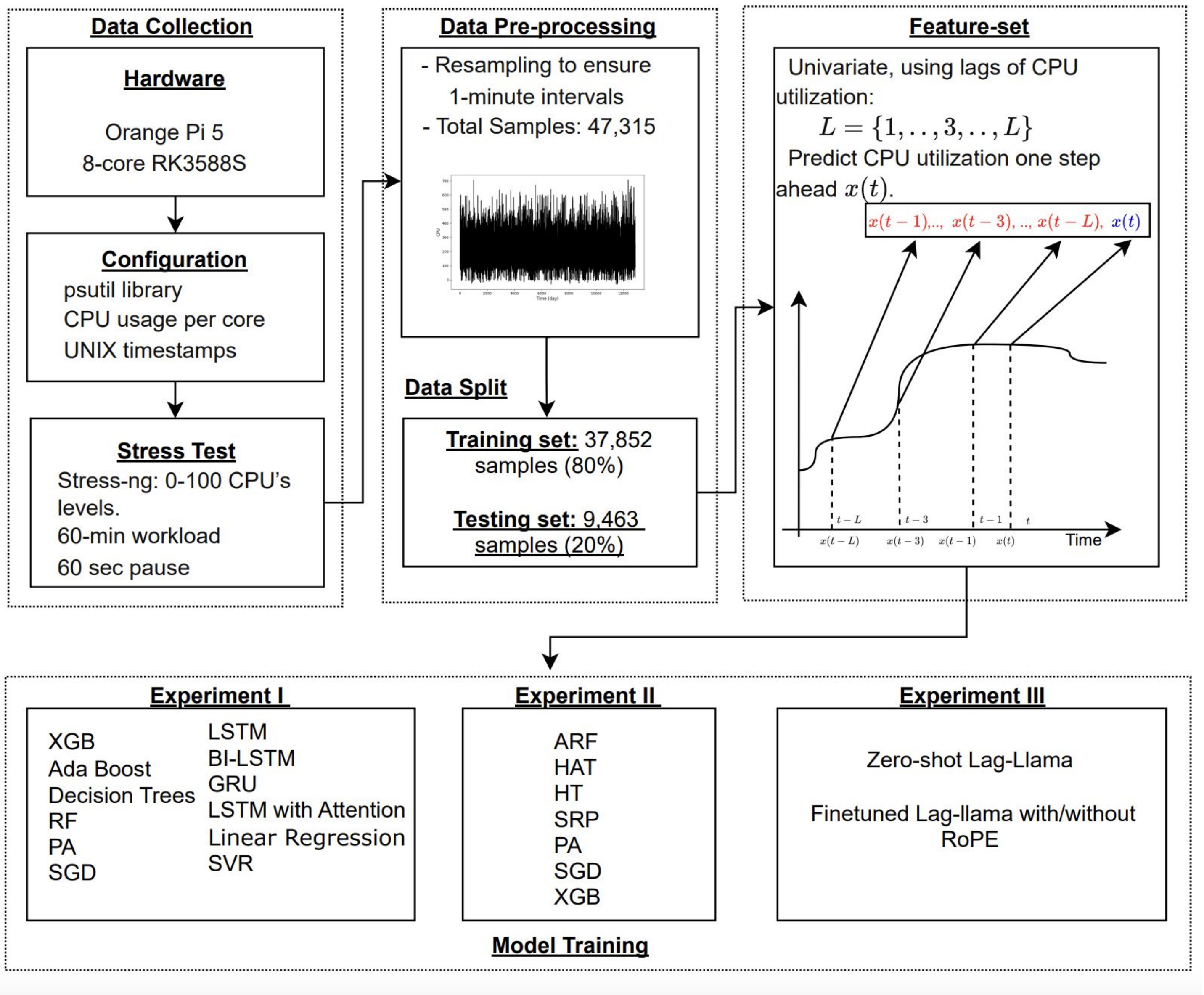
The main contributions of this paper are outlined below

I. **CPU utilization prediction:** Predicting CPU load in IoT gateways using advanced ML algorithms: online ensemble methods balance accuracy and cost, while continual learning shows promise for edge devices

II. **Evaluation benchmark:** A benchmark is proposed to compare traditional versus online and foundation models

III. **Code and data sharing**:
   - GitHub repository: https://github.com/sebasmos/AML4CPU

Pipeline



CeADAR
Ireland's Centre for AI

Experiments

Experiment I

- A hold-out benchmarking process was conducted between state-of-the-art ML algorithms.

Experiment II

- Online incremental learners were evaluated using the training and test sets from Experiment I for pre-training and for a prequential evaluation respectively

Experiment III

- A zero-shot and fine-tuning setup of the time-series foundation model Lag-Llama was run as in the previous experiments to compare the generalization capabilities of foundation models against other state-of- the-art and online ML methods

Experiments

Experiment I

- A hold-out benchmarking process was conducted between state-of-the-art ML algorithms.

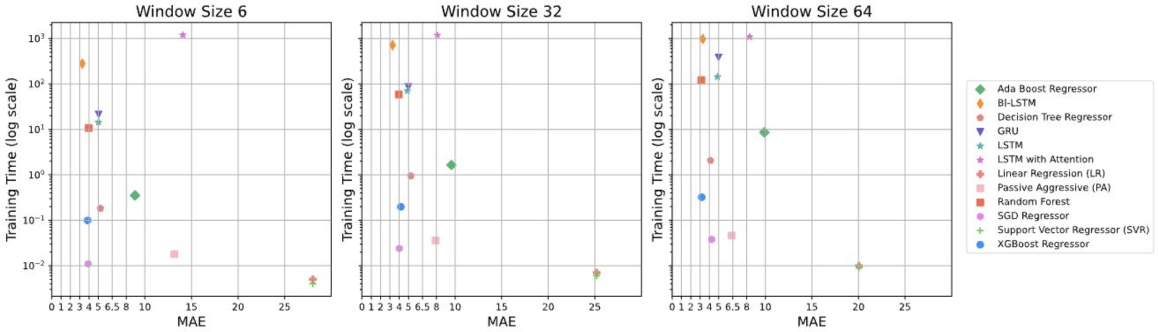


Fig. 2: Training time vs. MAE per model in Experiment I.

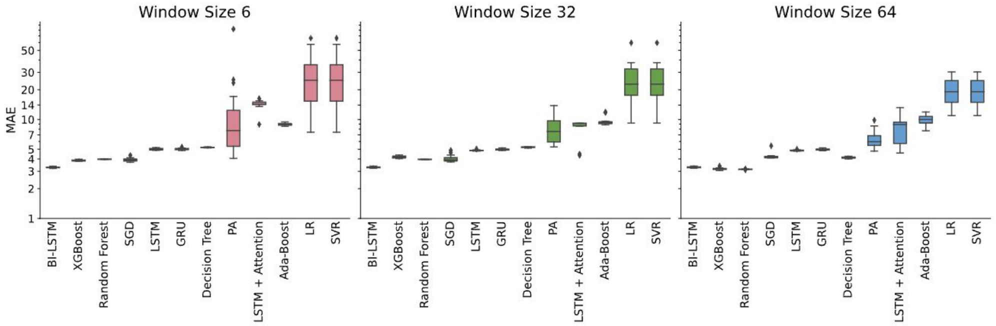


Fig. 1: MAE per model in Experiment I at different window sizes.

Experiments

Experiment I

Table 1: Experiment I results for WS with the lowest MAE across 20 runs.

Model	WS	MAE		RMSE		SMAPE		R^2		MASE		Training (s)	Inference (s)	Memory
		mean	std	mean	std	mean	std	mean	std	mean	std	mean	mean	mean
XGBoost Regressor	**64**	**3.185**	**0.086**	**7.246**	**0.326**	**21.988**	**0.418**	**0.942**	**0.005**	**0.821**	**0.022**	**0.322**	**0.005**	**0.004**
Ada Boost Regressor	9	8.901	0.238	12.381	0.3	32.669	0.635	0.831	0.008	2.3	0.061	0.449	0.003	0.013
Decision Tree Regressor	64	4.123	0.074	9.783	0.294	26.17	0.177	0.895	0.006	1.065	0.019	2.065	0.003	0.003
Random Forest Regressor	**64**	**3.141**	**0.02**	**7.525**	**0.087**	**20.195**	**0.094**	**0.938**	**0.001**	**0.811**	**0.005**	**121.804**	**0.164**	**0.085**
Passive Aggressive Regressor	64	6.38	1.379	9.729	1.228	34.621	5.144	0.894	0.028	1.647	0.356	0.046	0.002	0.005
SGD Regressor	20	3.886	0.203	9.817	0.02	22.288	0.977	0.894	0.001	1.003	0.053	0.019	0.001	0.004
Linear Regression	64	20.05	5.906	24.994	6.018	54.479	9.883	0.276	0.344	5.177	1.525	0.01	0.001	0.001
Support Vector Regression	64	20.05	5.906	24.994	6.018	54.479	9.883	0.276	0.344	5.177	1.525	0.009	0.001	0.001
LSTM	12	4.811	0.098	10.765	0.081	24.574	0.411	0.872	0.002	1.243	0.025	28.351	0.013	0.066
Gated Recurrent Units	20	4.961	0.097	10.497	0.09	25.648	0.219	0.878	0.002	1.281	0.025	76.746	0.037	0.049
BiLSTM	6	3.279	0.043	7.448	0.07	19.899	0.592	0.939	0.001	0.847	0.011	279.032	0.178	0.131
LSTM with Attention	20	7.258	3.805	12.622	3.893	30.334	7.686	0.809	0.137	1.874	0.982	1083.556	0.423	0.227

Experiments

Experiment II

- Online incremental learners were evaluated using the training and test sets from Experiment I for pre-training and for a prequential evaluation respectively

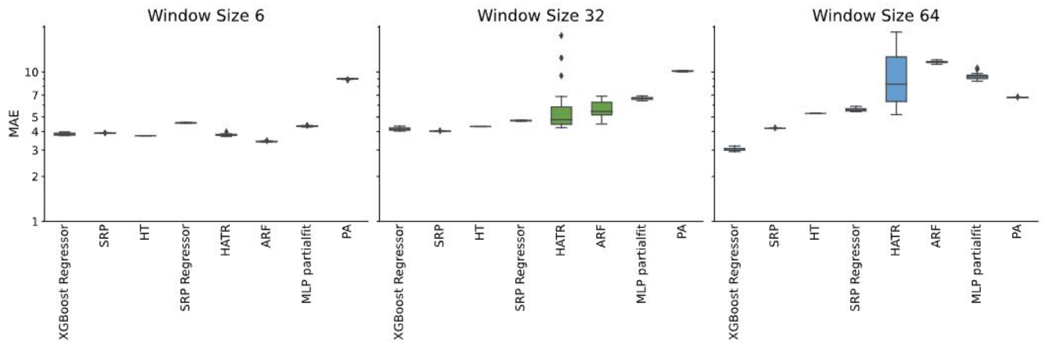


Fig. 3: MAE per model in Experiment II at different window sizes.

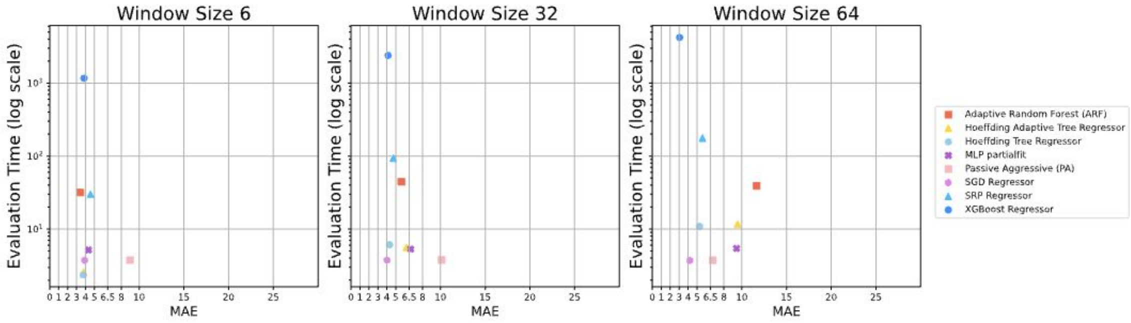


Fig. 4: Prequential evaluation time vs. MAE per model in Experiment II.

Experiments

Experiment II

Table 2: Experiment II results for WS with the lowest MAE across 20 runs.

Model	WS	MAE		RMSE		SMAPE		R^2		MASE		Pretraining (s)	Evaluation (s)	Memory
		mean	std	mean	std	mean	std	mean	std	mean	std	mean	mean	mean
ARF	6	3.427	0.018	9.078	0.023	20.170	0.131	0.909	0.000	0.885	0.005	71.431	31.641	146.031
HAT	6	3.795	0.063	9.340	0.085	21.583	0.420	0.904	0.002	0.981	0.016	4.567	2.575	2.819
HTR	6	3.750	0.000	9.233	0.000	20.943	0.000	0.906	0.000	0.969	0.000	3.208	2.348	2.012
SRP	6	4.574	0.016	10.772	0.028	24.048	0.106	0.872	0.001	1.182	0.004	117.284	30.116	0.360
PA	64	6.764	0.017	10.395	0.014	34.720	0.059	0.881	0.000	1.747	0.004	0.010	3.739	0.004
SGD	6	4.682	0.029	11.011	0.058	24.255	0.175	0.866	0.001	1.21	0.007	50.789	14.765	0.15
XGB	**64**	**3.057**	**0.066**	**6.766**	**0.180**	**21.826**	**0.422**	**0.950**	**0.003**	**0.789**	**0.017**	**0.430**	**4271.722**	**0.004**

CeADAR
Ireland's Centre for AI

Experiments

Experiment III

- A zero-shot and fine-tuning setup of the time-series foundation model Lag-Llama was run as in the previous experiments to compare the generalization capabilities of foundation models against other state-of- the-art and online ML method.

Table 3: Experiment III results for WS with the lowest MAE across 20 runs.

Model	CL	RoPE	MAE		RMSE		R^2		SMAPE		MASE	
			mean	std	mean	std	mean	std	mean	std	mean	std
Zero shot	256	Yes	5.500	0.021	11.579	0.034	0.857	0.001	32.021	0.169	1.169	0.004
Finetuned model on 32 lags	32	Yes	5.271	0.645	10.703	0.742	0.844	0.025	24.460	0.783	2.037	0.238
Finetuned model on 64 lags	**256**	**No**	**3.514**	**0.161**	**7.158**	**0.211**	**0.940**	**0.004**	**22.460**	**0.639**	**0.896**	**0.048**
Finetuned model on 128 lags	256	Yes	3.653	0.149	7.680	0.262	0.933	0.005	22.475	0.507	0.929	0.035
Finetuned model on 256 lags	256	Yes	3.683	0.176	7.444	0.261	0.935	0.004	22.872	0.462	0.971	0.030

CeADAR
Ireland's Centre for AI

Discussions

- **Performance vs. Computational Cost**.
  - Selecting the best model involves balancing prediction accuracy with computational efficiency, crucial for resource-constrained devices.
- **Best models**
  - XGBoost performs well but is costly in evaluation time. ARF offers good performance with higher memory usage, while ensemble models balance accuracy and cost effectively
  - Non-stationarity of CPU computational data

- **Online Learners and Ensembles**: Online learners are competitive but don't surpass ensemble models, which are recommended for edge devices. Further research could optimize model performance.
- Lag-llama is suitable for longer than the one-step ahead horizons, while still having a higher carbon footprint and inference time.

CeADAR
Ireland's Centre for AI

Thank you

ICOS project has received funding from the European Union's Horizon Europe Framework Programme under the Grant Agreement N° 101070177.
Views and opinions expressed in this presentation are however those of the ICOS Consortium only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them

Funded by the European Union

ICOS