Te Ipu o te Mahara Artificial Intelligence Institute
THE UNIVERSITY OF WAIKATO

ASML: A Scalable and Efficient AutoML Solution for Data Streams

Nilesh Verma[1], Albert Bifet[1], Bernhard Pfahringer[1], Maroua Bahri[2]

AI Institute, University of Waikato[1]
Inria, Paris, France[2]

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Introduction

Machine learning (ML) is a sub-field of artificial intelligence that enables computers to learn from data without explicit programming [1].

Automated machine learning (AutoML) tackles the Combined Algorithm Selection and Hyperparameter (CASH) optimization problem [2], which involves finding the best combination of ML algorithms and their hyperparameters for a given task.

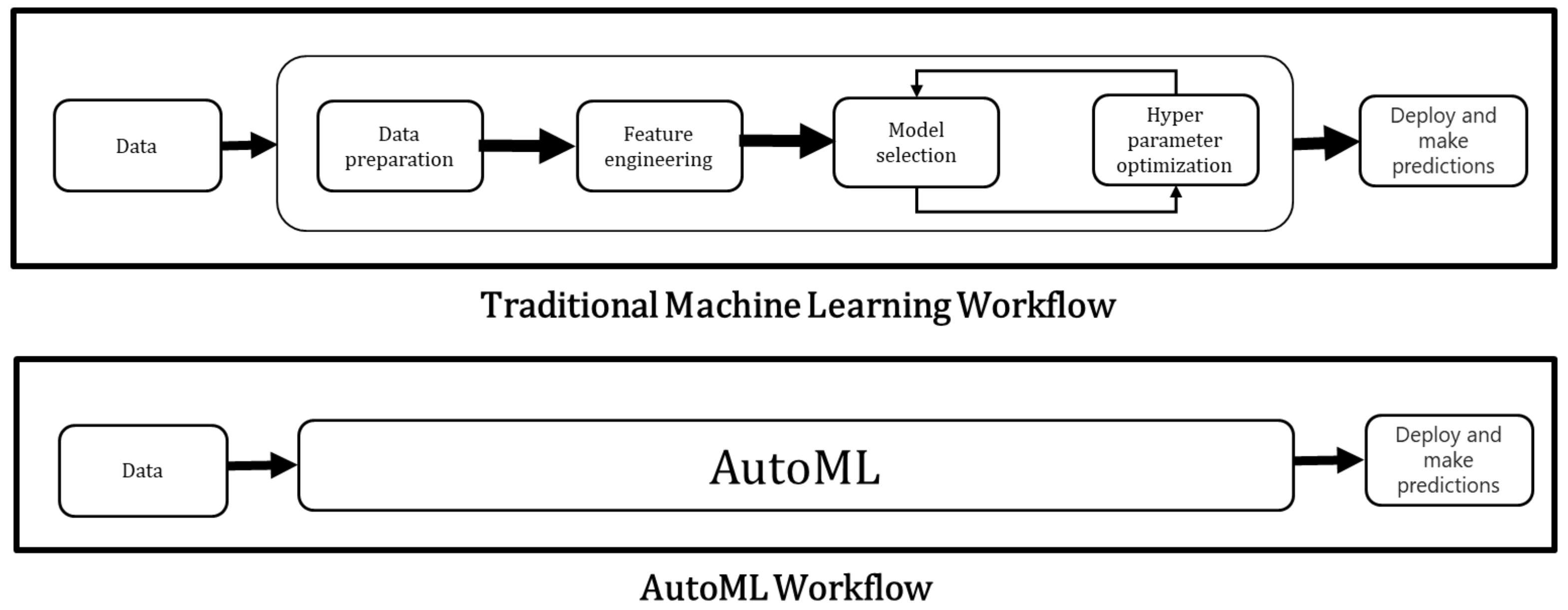
Data
Data preparation
Feature engineering
Model selection
Hyper parameter optimization
Deploy and make predictions
Traditional Machine Learning Workflow
Data
AutoML
Deploy and make predictions
AutoML Workflow

Figure 1. Traditional ML vs AutoML workflow.

Data streams are unbounded data sequences that arrive continuously and may change over time. They are generated by various devices such as IoT sensors, mobile phones, and websites and have applications in many domains such as transportation, finance, healthcare, and more [3].

Key Contributions

We propose Automated Streaming Machine Learning (ASML), an automated machine learning framework tailored for data streams. The main contributions of this work are:

- A novel AutoML formulation for online learning on evolving data streams.
- An adaptive random-directed nearby search strategy that continuously explores the pipeline configuration space to find optimal solutions over time.
- An experimental study that shows its effectiveness and resource efficiency over existing Online AutoML and state-of-the-art online learning algorithms.

Problem Formulation

Given an unbounded data stream $\mathcal{D} = (x_t, y_t)$, where x_t represents the input features at time step t and y_t denotes the target labels at time step t, the goal is to find the optimal combination S of data preprocessing p, input feature f, machine learning classifier c, and hyperparameter h that maximizes a defined objective function $\mathcal{O}$ over the evolving data stream $\mathcal{D}$. The search process is guided by an algorithm $\mathcal{A}$ that adapts model optimization based on detected concept drift in the stream. Formally, the Online AutoML optimization problem is:

$$S^t_{\text{best}} = \underset{(p,f,c,h)\in P\times F\times C\times H}{\operatorname{argmax}} \mathcal{O}_t(\mathcal{A}\{\mathcal{D}_t\}) \qquad (1)$$

Where P, F, C, and H represent the spaces of possible configurations for the preprocessing, features, models, and hyperparameters, respectively.

Automated Streaming Machine Learning (ASML)

We propose ASML, a novel online AutoML algorithm for classifying dynamic data streams, based on continuous exploration and adaptation strategy using the River online learning library [4].

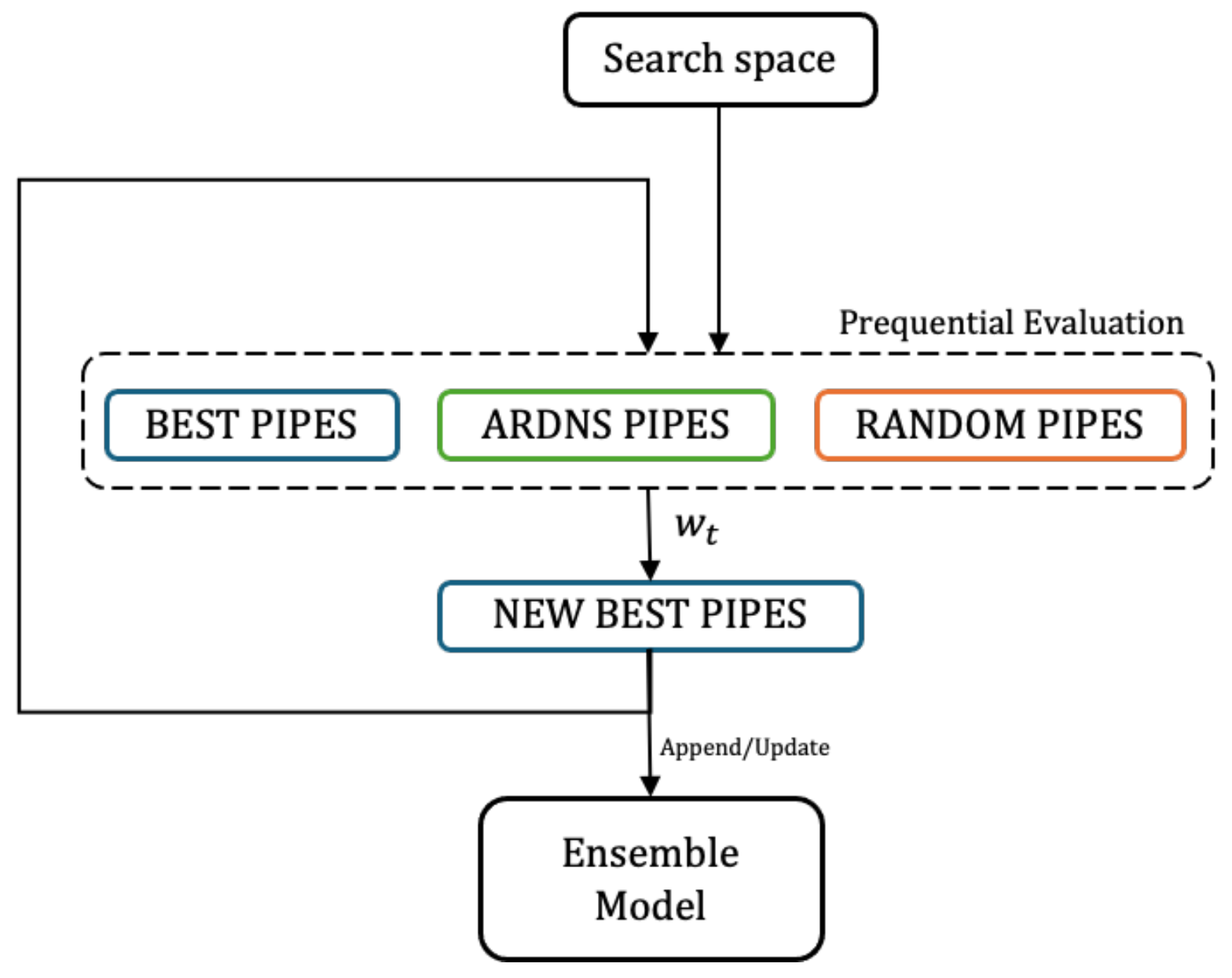

Figure 2. ASML Pipelines Search Process at Every Window W.

ASML uses fixed-size windows and initially explores all pipeline combinations with default hyperparameters. It sets a concurrent pipeline budget B and employs a search strategy: best pipeline, (B-1)/2 pipelines using ARDNS for hyperparameter optimization, and (B-1)/2 random pipelines. For prediction, ASML uses a confidence-weighted ensemble of all pipelines or the best pipeline.

Adaptive Random Directed Nearby Search (ARDNS)

ARDNS randomly changes the hyperparameters of the current best pipeline by choosing one of four directions: Same (*s*), Upper (*u*), Lower (*l*), or Random (*r*); This function explores the hyperparameter space around the current best pipeline and increases the chance of finding better local optimum solutions.

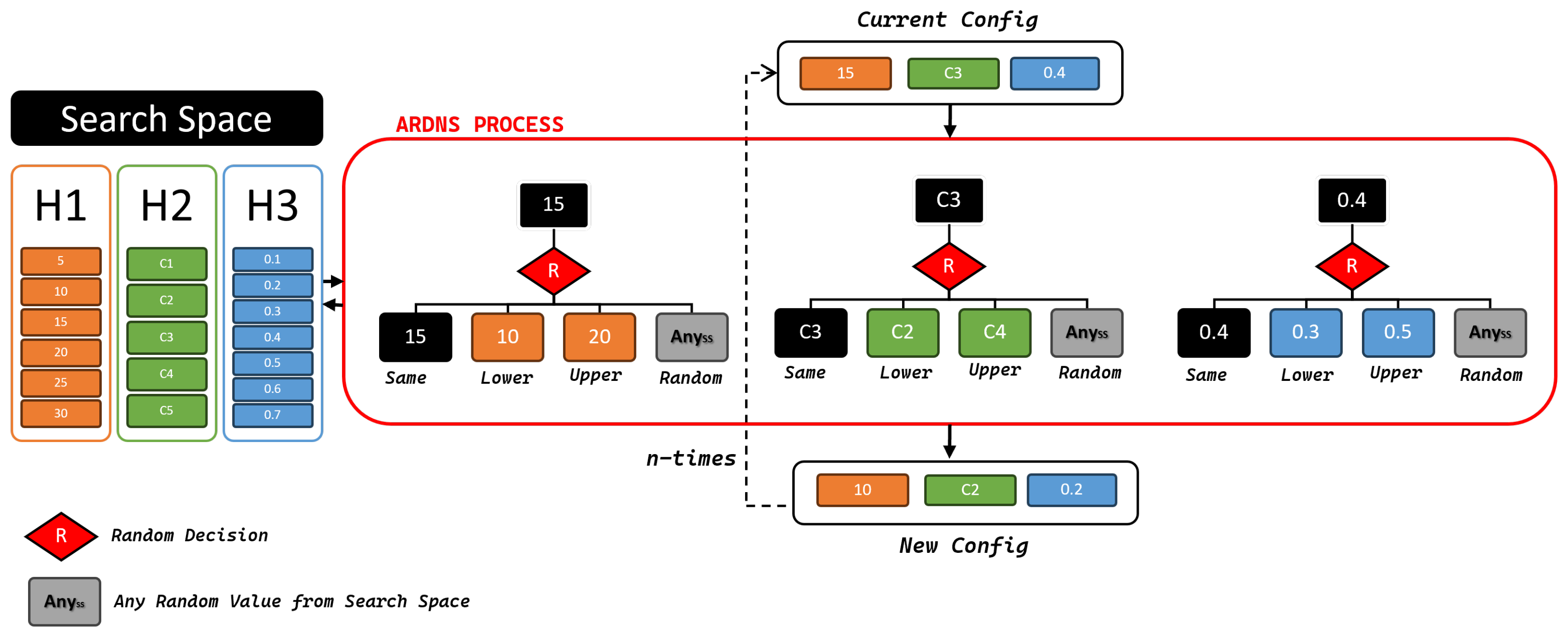

Figure 3. Adaptive Random Directed Nearby Search (ARDNS) Process.

Experiments

We experimented with two versions of the proposed ASML method: an ensemble version (ASML_E) and a best model selection version (ASML_B). We compared them with AutoClass, EvoAutoML (EAML), Online AutoML (OAML), and state-of-the-art online algorithms such as Hoeffding Adaptive Tree Classifier (HATC), Adaptive Random Forest classifier (ARFC) and Streaming Random Patches ensemble classifier (SRPC). To evaluate the performance of different methods in terms of prequential (test-then-train) accuracy, time (Sec.), and memory (M.B.) usage, we used 14 datasets that cover various streaming challenges. These datasets include both artificial (synthetic) and real-world data.

Results

Table 1. Accuracy comparison.

Datasets	ASML_E	ASML_B	AutoClass	OAML	EAML	ARFC	SRPC	HATC
Adult	80.01±0.61	80.36±0.27	76.86±1.28	72.07±0.44	80.56±1.29	81.35±0.33	80.11±0.30	**81.64±0.27**
Electricity	**91.50±0.12**	90.65±0.26	87.98±1.16	86.96±0.49	89.13±0.44	85.84±0.25	86.63±0.16	83.19±0.32
Forest Cover	**95.63±0.07**	95.39±0.18	95.32±0.02	83.16±0.52	94.07±0.06	88.85±0.40	92.97±0.09	70.99±1.48
Insects	70.95±0.46	**71.25±0.28**	64.24±0.25	63.69±0.27	70.05±1.66	68.61±0.48	68.60±0.51	60.25±1.50
New Airlines	66.58±0.09	65.46±0.05	63.03±0.48	67.03±0.49	**67.64±0.34**	65.31±0.09	64.53±0.17	65.27±0.10
Shuttle	99.34±0.07	98.58±0.11	**99.66±0.03**	97.31±0.18	98.65±0.25	99.54±0.07	99.48±0.08	94.57±0.69
Vehicle Sensit	**79.64±0.83**	75.70±0.67	73.73±0.16	73.11±0.25	79.11±1.74	75.44±0.50	78.10±0.39	75.38±0.27
Hyperplane High Gradual Drift	**91.85±0.03**	91.56±0.03	75.78±0.11	91.27±0.53	87.69±3.35	75.73±0.16	71.95±0.83	84.88±0.13
Moving RBF	**88.00±0.34**	86.82±0.17	85.11±0.70	67.23±0.31	83.20±2.35	51.27±0.32	49.62±0.67	39.36±0.40
Moving Squares	98.21±0.35	**98.61±0.20**	88.69±0.15	88.34±0.69	87.12±3.02	59.59±1.98	75.44±1.10	80.75±0.48
Sea High Abrupt Drift	88.52±0.03	88.04±0.02	84.33±1.31	88.78±0.23	88.60±0.23	**88.92±0.03**	84.39±3.45	88.80±0.02
Sea High Mixed Drift	87.87±0.02	87.47±0.03	84.13±1.42	**88.27±0.38**	87.49±0.34	88.03±0.11	81.73±3.20	88.13±0.03
Synth Random RBF Drift	**67.48±0.14**	65.34±0.20	60.78±2.86	65.59±0.20	60.46±3.54	62.82±1.25	56.22±0.77	55.03±0.28
Synth Agrawal	99.17±0.38	94.74±1.11	99.85±0.13	**99.97±0.59**	99.94±0.01	97.77±0.97	98.19±1.57	99.58±0.00
Avg. Accuracy	**86.05**	85.00	81.39	80.91	83.84	77.79	77.71	76.27
Avg. Rank	**2.64**	3.64	5.07	4.50	3.64	4.79	6.00	5.71

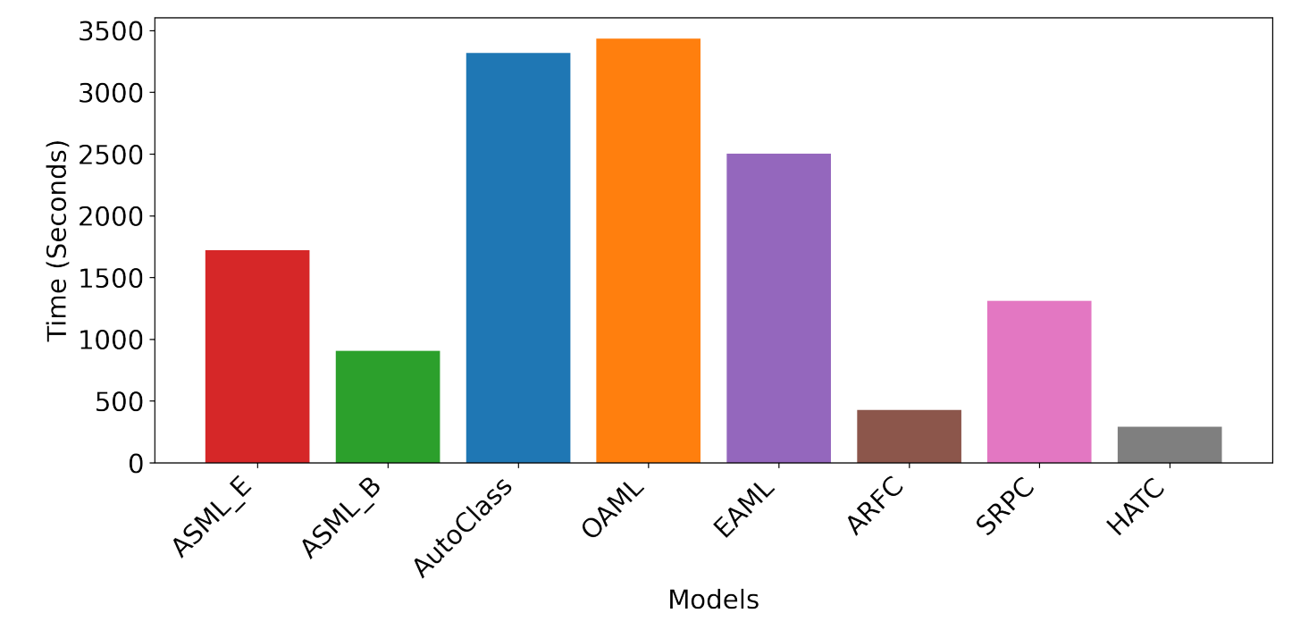

Figure 4. Overall Execution Time (Sec.)

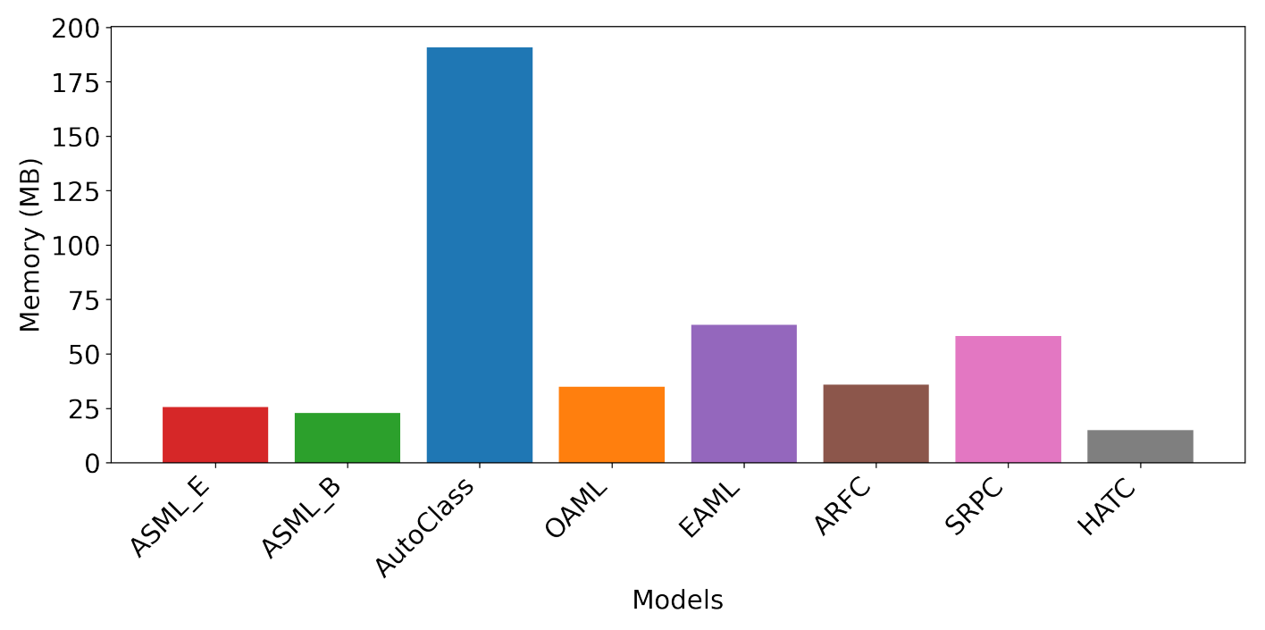

Figure 5. Overall Memory Usage (M.B.)

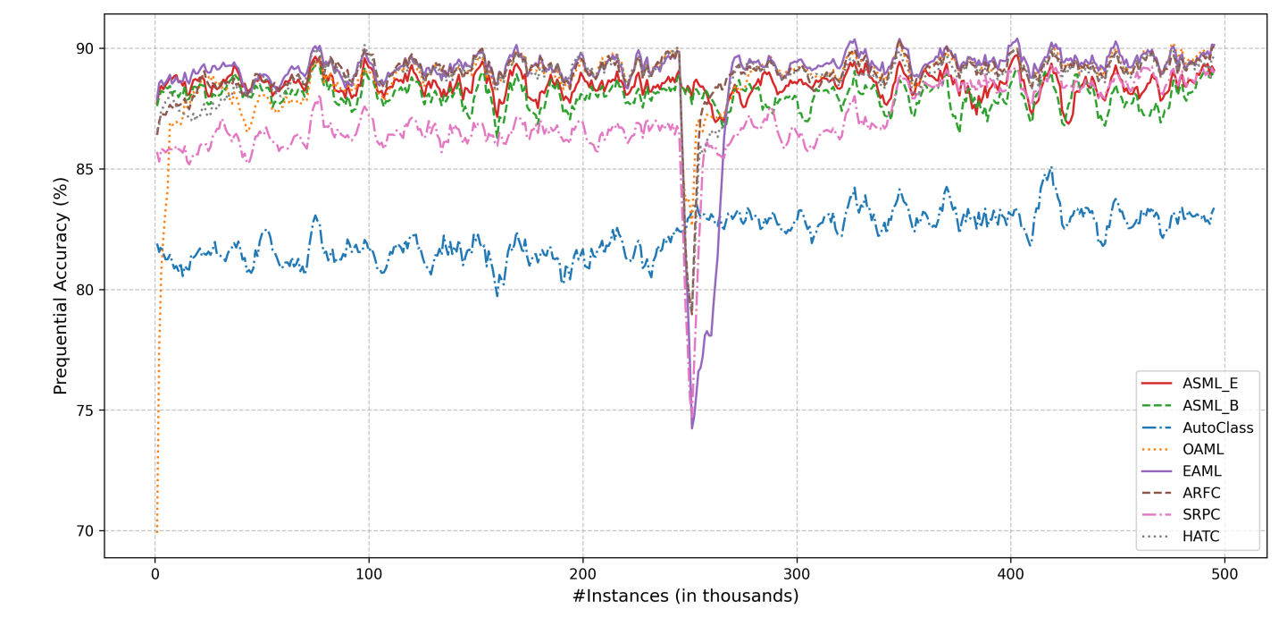

Figure 6. Sea High Abrupt Drift Dataset

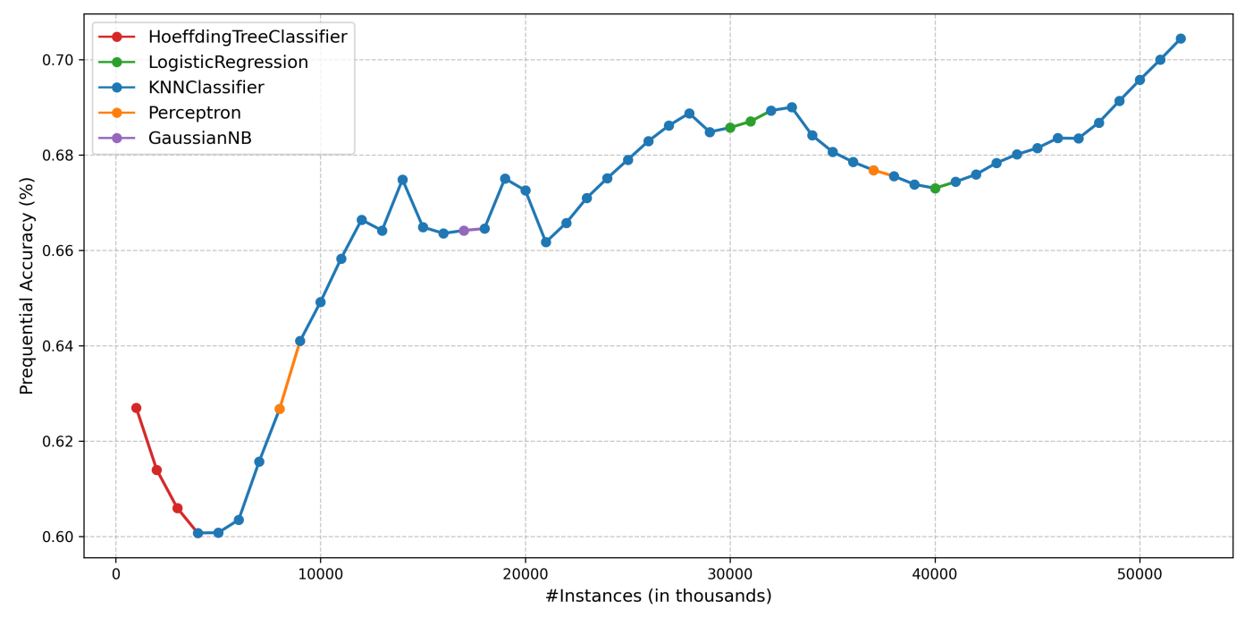

Figure 7. Algorithm Update (Instect Dataset)

Conclusion and Future Works

In this poster, we present Automated Streaming Machine Learning (ASML), a novel automated machine learning system for non-stationary data streams. We conducted comprehensive experiments on real-world and synthetic data streams with different types of drifts and compared ASML with state-of-the-art online learning algorithms. Our results show that ASML achieves better or comparable predictive performance while being more efficient regarding time and memory consumption.

In future work, we plan to extend ASML to support other machine learning tasks common in data stream applications. We believe that ASML is a versatile and powerful tool that can advance the state-of-the-art in AutoML for data streams.

References

[1] A. L. Samuel.
Some studies in machine learning using the game of checkers. ii—recent progress.
IBM Journal of Research and Development, 11(6):601–617, 1967.

[2] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter.
Efficient and robust automated machine learning.
Advances in neural information processing systems, 28, 2015.

[3] Albert Bifet, Ricard Gavaldà, Geoff Holmes, and Bernhard Pfahringer.
Machine learning for data streams.
The MIT Press, 2018.

[4] Jacob Montiel, Max Halford, Saulo Martiello Mastelini, Geoffrey Bolmier, Raphael Sourty, Robin Vaysse, Adil Zouitine, Heitor Murilo Gomes, Jesse Read, Talel Abdessalem, et al.
River: machine learning for streaming data in python.
2021.

AUTOML 2024, Paris, France

nv78@students.waikato.ac.nz