Diversity of Generated ML Pipelines in SageMaker Autopilot and Cross-Domain Generalization on OpenML

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does the diversity of generated ML pipelines in Amazon SageMaker Autopilot impact cross-domain generalization performance on the OpenML-2019 dataset, as measured by accuracy and latency. 6 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Amazon SageMaker Autopilot: a white box AutoML solution at scale. Research question: How does the diversity of generated ML pipelines in Amazon SageMaker Autopilot impact cross-domain generalization performance on the OpenML-2019 dataset, as measured by accuracy and latency trade-offs in the top-ranked models?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.5/10.

3 Results

12 papers retrieved. 6 claims extracted; 0 independently verified. Quality review score: 2.5/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The simple ϵ-greedy algorithm works better than other more complicated bandit algorithms, such as EXP3 and Rotting bandi	×	0.07
The system requires 5 finished HP evaluations before using Bayesian Optimization (BO) to ensure the model can be trained	×	0.04
With only 5 random HP evaluations before moving to ϵ-greedy, the algorithm can successfully identify the best pipeline o	×	0.02
For around 80% of datasets, the algorithm's choice is among the top 3 pipelines.	×	0.06
Using learned zero-shot HP configurations instead of random HPs increases the probability of committing to the best pipe	×	0.02
The meta-model in Amazon SageMaker Autopilot is trained with a collection of datasets and evaluated on a separate test c	×	0.10

References

- http://arxiv.org/abs/2012.08483v2
- http://arxiv.org/abs/2109.03285v1
- http://arxiv.org/abs/2111.13657v3