Scalability of SageMaker Autopilot's Preprocessing Pipeline and Fairness Metrics on Large-Scale Tabular Datasets

Assignee Research

June 12, 2026

Abstract

Modern approach to artificial intelligence (AI) aims to design algorithms that learn directly from data. This approach has achieved impressive results and has contributed significantly to the progress of AI, particularly in the sphere of supervised deep learning. It has also simplified the design of machine learning systems as the learning process is highly automated. However, not all data processing tasks in conventional deep learning pipelines have been automated. In most cases data has to be manually collected, preprocessed and further extended through data augmentation before they can be e

1 Introduction

This paper examines: Automated data processing and feature engineering for deep learning and big data applications: A survey. Research question: How does the scalability of SageMaker Autopilot's preprocessing pipeline affect fairness metrics (e.g., group fairness) when applied to large-scale tabular datasets (e.g., Criteo, Kaggle datasets) compared to distributed fairness-aware preprocessing frameworks like Turi Create?.

2 Methodology

Systematic literature search across multiple databases yielded 8 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.1/10.

3 Results

8 papers retrieved. 11 claims extracted; 11 independently verified. Quality review score: 8.1/10.

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
Modern approach to artificial intelligence (AI) aims to design algorithms that learn directly from data.	✓	0.28
This approach has achieved impressive results and has contributed significantly to the progress of AI, particularly in t	✓	0.32
It has also simplified the design of machine learning systems as the learning process is highly automated.	✓	0.26
Not all data processing tasks in conventional deep learning pipelines have been automated.	✓	0.35
In most cases data has to be manually collected, preprocessed and further extended through data augmentation before they	✓	0.28
Recently, special techniques for automating these tasks have emerged.	✓	0.22
The automation of data processing tasks is driven by the need to utilize large volumes of complex, heterogeneous data fo	✓	0.43
Today, end-to-end automated data processing systems based on automated machine learning (AutoML) techniques are capable	✓	0.49
Automated data preprocessing tasks include data cleaning, labeling, missing data imputation, and categorical data encodi	✓	0.28
Data augmentation techniques include synthetic data generation using generative AI methods.	✓	0.24
Feature engineering is specifically automated in these systems.	✓	0.19

References

- https://doi.org/10.1007/s10462-024-10726-1
- https://doi.org/10.48550/arxiv.2012.12600
- https://doi.org/10.1016/j.jiixd.2024.01.002