iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** XI **Month of publication:** November 2023

DOI: https://doi.org/10.22214/ijraset.2023.56634

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue XI Nov 2023- Available at www.ijraset.com

Automated Machine Learning: A Comprehensive Survey and Framework for Advancements

Pal Sonam Vinod[1], Wagh Shraddha Balasaheb[2], Yadav Mandar Pandurang[3], Yadgude Samrudhi Ravindra[4], Dr. S F Sayyad[5]
[1, 2, 3, 4, 5] *Computer Department, AISSMS College of Engineering, Savitribai Phule Pune University*

Abstract: ***Machine learning has become an indispensable tool in numerous domains, and the selection of an appropriate machine learning algorithm can significantly impact the success of a project. To address the challenge of algorithm selection, we present an Automated Machine Learning (AutoML) platform designed to empower users with the ability to effortlessly compare the performance of two selected machine learning algorithms on their datasets. This platform offers a user-friendly interface, guiding users through the entire process, from uploading their datasets to selecting algorithms and making informed decisions. The platform provides a streamlined workflow that allows users to easily upload datasets of their choice, select two machine learning algorithms from a pre-defined set, and compare their performance based on a variety of evaluation metrics. Users can visualize and analyze the results to gain insights into how different algorithms impact their data. Additionally, the platform offers assistance in identifying the most suitable algorithm for achieving optimal results. With the rapid evolution of machine learning techniques, this AutoML platform addresses the need for a user-friendly, efficient, and accessible tool to assist practitioners and researchers in algorithm selection. By simplifying the comparative analysis of machine learning models, this platform empowers users to make informed decisions that enhance the effectiveness of their data-driven projects.***
Keywords: Automated Machine Learning, Machine Learning Algorithms, Algorithm Comparison, Data-Driven Decision-Making, Model Performance

I. INTRODUCTION

In the realm of machine learning, the selection of the most suitable algorithm for a given dataset stands as a pivotal determinant of success. The rapidly evolving landscape of machine learning algorithms, each designed to address specific problem domains and data characteristics, has made this decision-making process both intricate and consequential. To this end, we introduce an innovative Automated Machine Learning (AutoML) platform meticulously engineered to offer users the power to effortlessly and rigorously compare the performance of two selected machine learning algorithms, ultimately guiding them towards data-driven decisions that yield optimal results. As machine learning applications permeate diverse industries and sectors, the demand for accessible, user-friendly tools to streamline the algorithm selection process has grown exponentially. Our AutoML platform addresses this need by providing a comprehensive and intuitive solution that empowers users, regardless of their expertise in machine learning, to upload datasets of their choosing and embark on a journey of algorithmic exploration.

The challenges that our AutoML platform seeks to overcome are multifaceted. First and foremost is the complexity of modern machine learning algorithms, each offering unique advantages and limitations. Choosing the right algorithm often demands in-depth understanding and prior experience, a barrier that our platform aims to dismantle. Secondly, data scientists and practitioners must consider the idiosyncrasies of their datasets – size, structure, distribution, and specific requirements – which further complicates the decision-making process. Our AutoML platform leverages the synergy of automation, accessibility, and in-depth analysis to provide a tailored solution for algorithm selection. The platform ensures that users can navigate seamlessly through the entire process, commencing with the effortless uploading of their datasets. A user-friendly interface will then guide them to select two machine learning algorithms for performance comparison.

The heart of our AutoML platform lies in its ability to harness the intrinsic characteristics of the data and the algorithms, providing users with a holistic understanding of the interaction between these variables. By quantifying and visualizing key performance metrics, such as accuracy, precision, recall, and F1-score, users gain insights into how different algorithms impact their dataset.

Additionally, the AutoML platform incorporates explain ability mechanisms to demystify the outcomes, offering insights into why specific algorithms excel or underperform. This functionality provides users with a deeper understanding of the underlying dynamics, which is invaluable in aiding decision-making and refining model selections.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue XI Nov 2023- Available at www.ijraset.com

II. LITERATURE REVIEW

A. Automated Machine Learning: The New Wave of Machine Learning [1]

The paper provides an in-depth survey of Automated Machine Learning (AutoML) and its current state, highlighting significant contributions and recent trends in the field. The primary contributions of the paper are threefold: it segments the AutoML pipeline, reviews contributions in each segment, evaluates state-of-the-art AutoML tools, and explores advancements in machine learning often overshadowed by deep learning. The paper includes a case study on AutoML's application in the insurance industry and summarizes various AutoML frameworks and tools available in the market. The taxonomy of AutoML is introduced as a framework for understanding its components and operations, helping to identify areas where automation can enhance the efficiency and accuracy of machine learning models. Key components covered include Data Preprocessing (data cleaning, transformation, and feature engineering), Model Selection (choosing appropriate algorithms based on problem type and dataset size), Hyperparameter Optimization (fine-tuning algorithm parameters), and Model Ensembling (combining multiple models for improved performance). The paper provides a comprehensive table summarizing various AutoML tools, data sources, preprocessing techniques, feature engineering methods, machine learning tasks, model selection and hyperparameter optimization techniques, neural architecture search methods, meta-learning approaches, user interfaces, and authorization methods. The conclusion anticipates ongoing advancements in AutoML, making it more accessible and efficient for data scientists in the future.

Fig. 1. Taxonomy of AutoML

B. Automated Machine Learning in Practice: State of the Art and Recent Results [2]

The paper presents a survey of Automated Machine Learning (AutoML) and its components, focusing on automating the machine learning model building process. It emphasizes the importance of data analysis and machine learning in enhancing business outcomes and highlights the complexities involved in constructing effective machine learning models. AutoML is introduced as a solution to automate tasks such as algorithm selection, hyperparameter tuning, and ensemble construction, which typically require substantial human expertise. The paper concentrates on the "Combined Algorithm Selection and Hyperparameter Optimization" (CASH) problem, which involves choosing algorithms and fine-tuning their hyperparameters to maximize model performance during validation. The paper delves into various aspects of AutoML, including Feature Engineering, Meta-Learning, Architecture Search, CASH problem, and Pipeline Optimization. These components aim to streamline the machine learning pipeline, making it more efficient and accessible. The paper evaluates different approaches, such as Data Science Machine (DSM), Auto-sklearn, TPOT, and Portfolio Hyperband, on various datasets for classification and regression tasks. Results suggest that advanced approaches outperform DSM's baseline but exhibit similar accuracy among themselves. Portfolio Hyperband stands out for its computational efficiency.

Conclusion: -In conclusion, the authors highlight the increasing demand for efficient and generalizable AutoML techniques as digital data continues to grow. They emphasize the potential of incorporating meta-information and learning to optimize concepts in AutoML. The learning to optimize approach, which leverages reinforcement learning for non-smooth objective functions, holds promise for future advancements in AutoML and deep learning architecture search. The paper anticipates that these more general concepts will drive the future of AutoML, overcoming engineering constraints and improving efficiency.

C. Evolving Fully Automated Machine Learning via Life-Long Knowledge Anchors [3]

This paper introduces a fully automated AutoML pipeline that encompasses data preprocessing, feature engineering, model selection and training, and ensemble creation for any dataset and evaluation metric. The innovation lies in the extensive scope of this learning pipeline, which incorporates "life-long" knowledge anchors to accelerate the search process across the entire search space. These knowledge anchors store detailed information about pipelines and integrate them with an evolutionary algorithm for joint optimization across components. Experimental results demonstrate that the resulting pipeline achieves state-of-the-art performance on various datasets and modalities. The central concept of this research is the use of life-long knowledge anchors, serving as repositories of historical learning experiences, which allow the AutoML system to leverage past model performance and choices, reducing redundant exploration and speeding up model evolution. The proposed algorithmic approach combines reinforcement learning and neural architecture search, facilitating decision-making by rewarding successful choices and exploring diverse model architectures iteratively. This integration enables the AutoML process to continuously learn from its experiences and past knowledge, leading to more efficient and effective model evolution. The core innovations in this paper revolve around life-long knowledge anchors and the fusion of reinforcement learning and neural architecture search, paving the way for more informed and automated machine learning.

Conclusion: -In conclusion, while the Semi-AutoML paradigm discussed in the paper may save computational costs by using a smaller search space, it has two fundamental drawbacks. First, searching isolated components independently over the entire learning pipeline may lead to sub-optimal results. Second, the fine-grained search space needs to be meticulously composed, creating a new learning space for the model. These challenges highlight the need for a more comprehensive and integrated approach to AutoML.

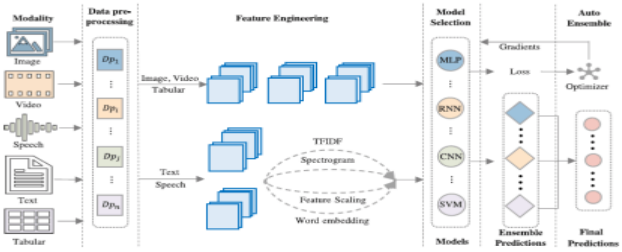


Fig. 2. The search space on a machine learning pipeline consisting of different components for different modalities in the proposed Fully-AutoML paradigm.

D. A Review on Automated Machine Learning (AutoML) Systems [4]

The paper addresses the increasing attention garnered by Automated Machine Learning (AutoML) but highlights the absence of a well-documented overview of existing approaches and systems in the literature. The authors stress the importance of categorizing and analyzing existing AutoML work to facilitate further research in the field. The paper primarily focuses on three domains: AutoML, hyperparameter tuning, and meta-learning. AutoML is defined as the extensive automation of various machine learning processes, encompassing data preprocessing, meta-learning, feature learning, model searching, hyperparameter optimization, classification/regression, workflow generation, data acquisition, and reporting. The authors analyze six AutoML systems, classifying them as fully automated or semi-automated approaches and comparing the machine learning algorithms they support for regression and classification tasks. The paper outlines four essential components required for developing an AutoML system: the Preprocessing Engine, Feature Engine, Predictor Engine, and Model Selection and Ensemble Engine. Additional components introduced in commercial AutoML systems, such as the Human Engine, Knowledge Base, Visualization Engine, and GUI Engine, are also mentioned. The paper provides a comparison of the algorithms supported by the reviewed AutoML systems for regression and classification learners. The authors identify several gaps and challenges in the existing AutoML research. They emphasize the need for fully functional, user-friendly AutoML products that can be applied across various domains, collaborative knowledge hubs for statistical models to support meta-learning, and exploring AutoML beyond the Python-centric landscape by considering languages like R. Additionally, they note the relatively unexplored potential of advanced neural network technologies in AutoML.

Conclusion: -In conclusion, the paper summarizes the state of AutoML research and calls for advancements and improvements in the field. It underscores the significance of creating efficient and fully automated AutoML systems, establishing collaborative knowledge repositories, and exploring emerging technologies such as deep neural networks. The paper offers a comprehensive review of AutoML, its current status, and future directions, providing insights for researchers and practitioners in the field of automated machine learning.

E. Efficient and Robust Automated Machine Learning [5]

This paper introduces a novel Automated Machine Learning (AutoML) framework that automates data preprocessing, feature preprocessing, algorithm selection, and hyperparameter tuning through Bayesian optimization without human intervention. The authors make two key contributions to existing AutoML methods: a metalearning component that leverages past datasets to enhance the Bayesian optimizer's performance, and an ensemble construction component that combines the top-performing methods identified by the Bayesian optimizer for enhanced robustness. The algorithms employed in this framework include Bayesian Optimization for efficient algorithm and hyperparameter selection, Ensemble Learning to combine multiple models, 15 different Machine Learning Classifiers to address various dataset types, 14 Feature Preprocessing Methods to optimize data, 4 Data Preprocessing Methods for data preparation, and optimization of 110 hyperparameters to fine-tune component performance.

Conclusion: -The conclusion emphasizes that AUTO-SKLEARN is a significant contribution to the AutoML field, offering accessibility to non-experts while achieving state-of-the-art performance. Its innovative use of Bayesian optimization, ensemble techniques, and dataset-specific considerations positions it as a powerful and practical AutoML solution with substantial potential. In summary, AUTO-SKLEARN represents a substantial advancement in AutoML, combining automation with advanced optimization and ensemble strategies to make machine learning more accessible and effective.

F. A Unified Framework for Automatic Distributed Active Learning [6]

This paper presents a general framework for an automated distributed active learning algorithm that optimizes multiple hyperparameters in the classification and query selection stages. It achieves this through a two-step optimization process using automated machine learning and integer linear programming iteratively. The paper introduces two novel loss functions, Cluster-Specific Maximum Entropy (CME) and Shrinkage Optimized KL-Divergence Within Local Clusters-based Active Learning (SOAR), designed to address the challenges posed by unbalanced datasets. To efficiently harness distributed computing resources, the unlabelled data is randomly partitioned, and the labelled data is replicated in the classification stage, enabling scalable and superior performance for big data classification. In the query selection stage, the most informative and representative samples are centrally collected. The proposed AutoDAL algorithm leverages automated machine learning techniques for hyperparameter optimization, leading to significant classification performance improvements. The algorithms and methods involved include Shrinkage Regularization to handle imbalanced datasets, the CME loss function for active learning, and a combination of Genetic Algorithm and Grid Search Strategy for automatic hyperparameter selection. The Distributed Algorithm is introduced to execute the semi-supervised learning algorithm in a distributed computing environment, making use of partitioning and replication techniques to reduce running time and memory requirements for big data applications.

Conclusion: -In conclusion, while the paper primarily focuses on the AutoDAL algorithm and its effectiveness in addressing active learning challenges, it acknowledges its relevance to the field of automated machine learning (AutoML). By leveraging AutoML techniques for hyperparameter optimization, the work contributes to the development of AutoML platforms, offering a novel approach to tackle active learning issues in a distributed computing environment.

G. D-SmartML: A Distributed Automated Machine Learning Framework [7]

The paper discusses the challenges in developing high-quality machine learning models and introduces the D-SmartML framework as a solution. This distributed automated machine learning framework is built on Apache Spark, addressing the limitations of centralized AutoML frameworks.

It supports large datasets and incorporates a meta-learning mechanism for algorithm selection, distributed grid search, random search, and hyperband optimization for hyperparameter tuning. The D-SmartML framework's architecture comprises various components, including classifiers like Random Forest, Logistic Regression, Decision Tree, and more. It also features a Model Selector, Classifier Manager, Metadata Manager, and KB Manager. Hyperparameter optimization is based on Hyperband, which involves successive halving sessions with different budgets and configurations to choose the best settings. The framework also employs a meta-model for predicting the best-performing classifiers based on dataset meta-features.

Conclusion: -D-SmartML leverages meta-learning to automate algorithm selection, making predictions based on past experience with various learning algorithms. Its distributed grid search, random search, and hyperband optimization techniques facilitate effective hyperparameter tuning. Overall, D-SmartML offers a scalable and efficient solution for automated machine learning, outperforming existing frameworks in terms of optimization time and model accuracy.

H. Adaptation Strategies for Automated Machine Learning on Evolving Data [8]

The study delves into the adaptability of Automated Machine Learning (AutoML) systems in the context of concept drift, where data evolves over time. It explores the impact of concept drift on AutoML performance and introduces six adaptation strategies to enhance their robustness to changing data. The adaptation strategies proposed in the study are designed to modify AutoML methods to handle evolving data effectively. These strategies include Detect & Increment (D&I), Detect & Retrain (D&RT), Detect & Warm-start (D&WS), Detect & Restart (D&RS), Periodic Restart (PRS), and Train Once (T1). The study evaluates the performance of various AutoML libraries and adaptation strategies under different types and magnitudes of concept drift. It uses both synthetic data streams and real-world data streams to analyze the impact of drift on AutoML systems.The findings emphasize that the effectiveness of adaptation strategies varies based on different AutoML systems and data characteristics. The study highlights the interplay between adaptation strategies and drift detectors, influencing performance and recovery speed.

Conclusion: -Overall, the research provides valuable insights into the adaptability of AutoML methods to concept drift, contributing to the development of more resilient AutoML techniques that can effectively handle evolving data. It encourages further exploration of dynamic adaptation in changing data environments and advances the field of AutoML.

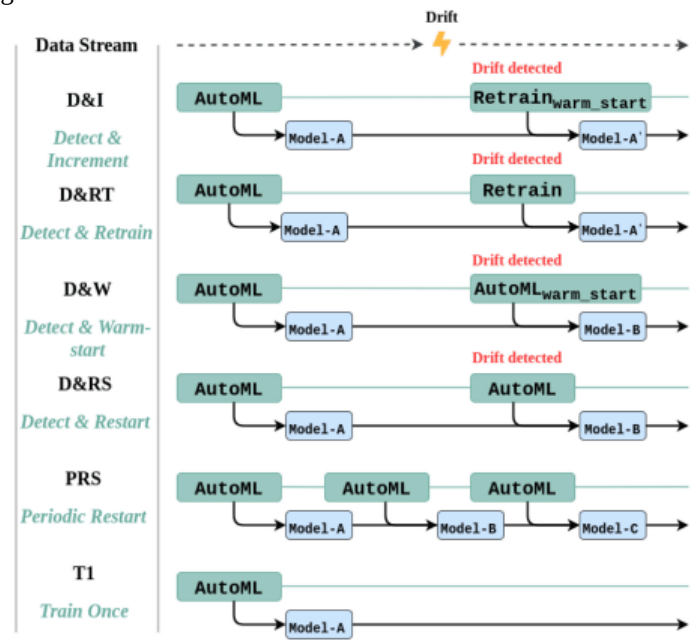


Fig. 3. Adaptation strategies.

I. An Empirical Study on the Usage of Automated Machine Learning Tools AutoML Tools Popularity [9]

The study investigates the utilization of AutoML tools in open-source projects on GitHub and addresses several key questions:

1) *Identification of the Most Used AutoML Tools:* The investigation identifies the top 10 most utilized AutoML tools, including Optuna, HyperOpt, Skopt, Featuretools, Tpot, Bayes opt, Autokeras, Auto-sklearn, AX, and Snorkel. It sheds light on the development history and popularity of these tools, providing valuable information for machine learning practitioners and AutoML tool developers.
2) *How ML Practitioners Use AutoML Tools:* The manual analysis reveals the common utilization of AutoML tools during various stages of the machine learning pipeline, such as Hyperparameter optimization, Model training, and Model evaluation. It identifies ten core purposes for their usage, empowering ML practitioners to streamline their tasks and improve workflow efficiency.
3) *Coexistence of Different AutoML Tools:* The study examines whether multiple AutoML tools are used concurrently within the same projects. It finds limited instances of co-utilization among the top 10 AutoML tools, with Hyperopt being combined with tools like Tpot, Skopt, and Optuna. These insights are valuable for ML practitioners seeking to integrate heterogeneous AutoML features and for tool developers aiming to enhance collaboration between their tools and others.
4) *Comparison of AutoML Tools:* While previous research has compared AutoML tools in terms of performance and coverage, this study addresses the gap by investigating the practical utilization of AutoML tools in real-world projects on GitHub. The findings bridge the knowledge gap and offer insights for both ML practitioners and tool developers.

5) *Experimental Setup:* The study outlines its experimental setup, including the identification and filtering of AutoML tools and the collection of GitHub projects using these tools.

Conclusion: -In conclusion, this empirical study provides insights into the landscape of AutoML tool adoption in real-world machine learning projects on GitHub. It highlights the dominance of specific tools in less established projects, variations in tool usage across the machine learning pipeline, and the need for increased automation in data-related stages. Future research aims to explore unused tool features and factors impacting successful integration, further enhancing the understanding of AutoML tool adoption and its implications for real-world projects.

J. Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools [10]

The paper introduces the concept of AutoML as a solution to automate repetitive tasks in machine learning pipelines, encompassing data preprocessing, feature engineering, model selection, hyperparameter optimization, and result analysis. It sets clear research objectives, including evaluating the available AutoML functionalities, tool performance across different datasets, trade-offs between optimization speed and accuracy, and result reproducibility. Diverse datasets collected from OpenML are used, with variations in sample size, feature dimensions, categorical feature ratio, missing data proportion, and class imbalance. The paper does not delve into specific algorithms or models used by each AutoML tool. Instead, it focuses on evaluating the overall capabilities and performance of these tools in automating various aspects of the machine learning pipeline. The AutoML tools employ a variety of algorithms and techniques, including hyperparameter optimization methods such as grid search, random search, Bayesian optimization, and genetic algorithms. They also utilize model selection techniques that cover a range of machine learning algorithms, data preprocessing methods, model interpretation approaches, and ensembling techniques. The specific algorithms and models used by each AutoML tool may vary, but the paper emphasizes evaluating their effectiveness in automating machine learning tasks.

Conclusion: -In summary, this paper provides a valuable evaluation of AutoML tools in a real-world context, offering insights into their strengths and weaknesses across various machine-learning tasks and datasets. While it does not delve into algorithmic details, it provides a comprehensive understanding of the current state of AutoML and highlights areas where further research and development are needed.

Fig. 4. The common AutoML pipeline.

K. A Comparison of AutoML Tools for Machine Learning, Deep Learning and XGBoost[11]

This paper presents a comprehensive benchmark study of eight open-source AutoML tools, with a primary focus on supervised machine learning tasks. The study aims to evaluate the capabilities of these AutoML tools in automating the selection of machine learning algorithms and tuning hyperparameters. The authors analyze the tools across three main scenarios: General Machine Learning (GML) algorithm selection, Deep Learning (DL) algorithm selection, and XGBoost (XGB) hyperparameter tuning. Evaluation metrics such as Mean Absolute Error (MAE), Area Under the Curve (AUC), and Macro F1-score are used to assess predictive performance. The study provides a lexicographic approach for selecting the best tool based on both predictive performance and computational effort. The results highlight the strengths and preferences for each tool across different scenarios anddatasets.

Conclusion: -In conclusion, this benchmark study offers valuable insights into the performance of open-source AutoML tools for supervised machine learning tasks. It provides recommendations for tool selection based on specific scenarios and highlights the competitive nature of these tools compared to manually configured machine learning models. The study encourages further exploration with a broader range of AutoML tools, datasets, and consideration of big data and infrastructure settings, ensuring the continued development and effectiveness of automated machine learning solutions.

L. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data [12]

This paper introduces AutoGluon-Tabular, an open-source AutoML framework designed to simplify the process of training accurate machine learning models on raw tabular datasets, such as CSV files. AutoGluon-Tabular employs a unique approach that involves stacking multiple models in multiple layers for ensembling. The study evaluates this multi-layer model combination approach and its effectiveness in optimizing training time utilization.

AutoGluon is compared to various AutoML platforms in terms of speed, robustness, and accuracy using a suite of 50 classification and regression tasks from Kaggle and the OpenML AutoML Benchmark. AutoGluon often outperforms other platforms and surpasses the hindsight combination of its competitors, achieving remarkable results in Kaggle competitions with minimal preprocessing. The provided algorithmic details outline the framework's functionality and key strategies: The paper describes the training and testing process with AutoGluon, highlighting the use of the `fit()` function for data preprocessing, model selection, and ensembling. AutoGluon automatically infers the prediction problem type and identifies feature types for data preprocessing. Various types of models, including neural networks, LightGBM, CatBoost, and scikit-learn's models, are utilized for training. The multi-layer stacking approach is detailed, focusing on base models and stacker models in each layer. Repeated k-fold ensemble bagging is explained, enhancing stacking performance. The overall training strategy is provided in pseudocode, and AutoGluon's sequential training approach is emphasized. The evaluation section compares AutoGluon's performance against other AutoML platforms on the OpenML AutoML Benchmark and Kaggle competitions: AutoGluon consistently outperforms other frameworks in terms of predictive accuracy, average rank, and robustness. It achieves higher accuracy and ranks better on a wide range of datasets. An ablation study reveals the importance of different components in AutoGluon's exceptional performance.

Conclusion: -The paper concludes by highlighting AutoGluon's exceptional performance in terms of predictive accuracy, robustness, and adherence to time limits, making it a competitive choice for automated machine learning on tabular data.

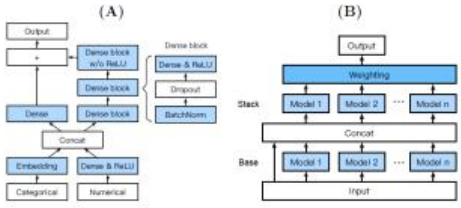


Fig. 5. (A) Architecture of AutoGluon's neural network for numerical and categorical features. Blue layers contain learnable parameters. (B) AutoGluon's multi-layer stacking strategy, shown here using two stacking layers and n types of base learners.

M. Machine Learning Operations (MLOps): Overview, Definition, and Architecture[13]

MLOps, or Machine Learning Operations, is a holistic approach and a set of best practices aimed at optimizing the entire lifecycle of machine learning projects. It focuses on efficiently managing the development, deployment, and scaling of machine learning products by bridging the gap between development and operations through automation and a product-oriented mindset. MLOps is underpinned by a set of fundamental principles, including Continuous Integration and Continuous Deployment (CI/CD) automation, workflow orchestration, reproducibility, version control for data, models, and code, collaboration, continuous machine learning training and evaluation, robust metadata tracking, continuous monitoring, and feedback loops.

MLOps encompasses several critical components, including CI/CD automation for seamless integration and deployment, workflow orchestration for coordinating various processes, data and model versioning to track changes, collaborative tools for effective teamwork, continuous training and evaluation of machine learning models, robust metadata tracking and logging, continuous monitoring for model performance assessment, and feedback mechanisms for iterative improvements.

MLOps involves various roles that collectively contribute to its success, such as Business Stakeholders who define project goals, Solution Architects who design the system's architecture, Data Scientists who translate business problems into ML tasks, Data Engineers who manage data pipelines, Software Engineers who develop ML solutions, DevOps Engineers who manage CI/CD and deployments, and ML Engineers/MLOps Engineers who oversee the ML infrastructure, automate workflows, deploy models, and monitor performance. The MLOps architecture encompasses the end-to-end machine learning process, starting from project initiation, feature engineering, experimentation, and automated ML workflows, culminating in model deployment. It is flexible and adaptable, allowing the selection of specific technologies and tools based on unique project requirements, whether through open-source solutions or enterprise platforms.

The adoption of MLOps is accompanied by various challenges, spanning organizational, ML system, and operational aspects. These challenges range from fostering cultural shifts and addressing skill gaps to designing systems that can adapt to fluctuating demand, creating robust automation, ensuring governance, and effectively troubleshooting complex ML systems. Addressing these challenges is crucial for the successful implementation of MLOps in real-world settings.

N. Predicting Machine Learning Pipeline Runtimes in the Context of Automated Machine Learning [14]

The paper discusses a novel approach to predict timeouts in machine learning pipelines, particularly for classification tasks within AutoML. It addresses the challenge of timeouts that limit CPU usage and lead to inefficiencies in the AutoML process, where substantial CPU time is lost due to timed-out evaluations. The proposed approach utilizes regression models for various pipeline components, including atomic algorithms, pre-processors, and meta-learners, to achieve accurate runtime predictions. This innovation aims to enhance the efficiency of AutoML by predicting whether pipeline executions will time out. The paper outlines the problem definition, emphasizing the need to predict timeouts in AutoML pipelines and highlighting the advantages of using a regression-based approach over binary classification. It focuses on maximizing prediction quality regarding pipeline timeouts and integrating the predictor into the AutoML search process to assess pipeline success before execution. The compositional approach for runtime prediction in AutoML pipelines is described, avoiding the conversion of pipeline descriptions into vectors. It involves creating individual prediction models for different pipeline components and addresses challenges like predicting meta-learner runtimes and determining dataset features for algorithms. The empirical evaluations, involving a vast number of experiments, demonstrate the practicality of this approach. The paper delves into predicting runtime for atomic algorithms, exploring variable types such as algorithm parameters, dataset meta-features, and previous observations. It evaluates the impact of prediction errors on pruning decisions and assesses the performance of models when using dataset meta-features or adding a feature for default algorithm performance. The experimental setup and results of predicting atomic algorithm runtimes are presented, revealing variations in prediction accuracy among different algorithms and influences of input data complexity on prediction performance. The paper assesses the accuracy of predicting timeouts in AutoML for various atomic algorithms and timeout thresholds, highlighting false positives and the potential for improving prediction models. The enhancement of prediction performance by considering default parameterization of algorithms is discussed. Posterior models are introduced, which substantially improve prediction performance, especially for algorithms with poor performance in standard models. The paper also addresses the predictability of feature transformations realized by pre-processors, suggesting effective methods for predicting dataset transformations following preprocessing steps.

Conclusion: -In conclusion, the paper introduces a regression-based approach for predicting timeouts in AutoML pipelines, enhancing resource-intensive datasets' efficiency. It discusses potential future research directions to further advance AutoML runtime prediction.

O. Efficient AutoML via Combinational Sampling [15]

This study addresses the challenges of Automated Machine Learning (AutoML) by introducing a novel approach to optimize machine learning pipelines automatically, without substantial human intervention. Unlike traditional AutoML methods that often treat algorithm choices as categorical hyperparameters within a broader Hyperparameter Optimization (HPO) framework, this research treats AutoML as its distinct optimization problem. It introduces an innovative "Algorithm Choice" hyperparameter class for modeling algorithm selection within operators and groups similar operator algorithms for initial sampling budgets. Additionally, the study proposes a robust sampling technique to enhance Bayesian Optimization in AutoML.

Inference: -

1) AutoML Problem Definition: AutoML involves selecting machine learning algorithms and their hyperparameters to create efficient machine learning pipelines comprising a sequence of operators applied to input data. The goal is to maximize the pipeline's predictive performance.
2) Search Space: The search space in AutoML encompasses a wide range of operators, algorithms, and hyperparameter sets for each operator.
3) New Hyperparameter Class: The introduction of the "Algorithm Choice" hyperparameter class distinguishes between categorical hyperparameters and algorithm choices within operators, acknowledging their distinct roles in AutoML.
4) *Combination-Based Initial Sampling:* The proposed sampling technique reallocates the initial sampling budget to explore algorithm-hyperparameter combinations efficiently. This approach aims to improve coverage and robustness in the surrogate model.
5) *BO4AutoML and Robust AutoML:* The paper introduces BO4AutoML, a Bayesian optimization library for AutoML, and a framework called Robust AutoML, which incorporates the novel approaches.

Experimental Results

The research includes two experiments to evaluate the effectiveness of the proposed Bayesian optimization approach and AutoML framework:

a) *First Experiment Results:* The authors compared their approach with traditional Hyperopt, assessing performance with different initial sample sizes. Their approach significantly outperformed Hyperopt when using 50 initial samples.
b) *Second Experiment Results:* In this experiment, their RobustAutoML (TPE with their sampling approach) was compared to other AutoML frameworks across 73 datasets. Bayesian optimization approaches, including their RobustAutoML, outperformed Random Search in most cases. Their approach achieved the highest results in 28 out of 73 datasets and significantly outperformed other approaches in 23 cases.

Conclusion: The research concludes that the proposed Bayesian optimization approach, along with the novel sampling technique, offers significant improvements in AutoML optimization tasks, outperforming traditional approaches and other AutoML frameworks. It suggests further research avenues, such as applying the sampling approach to other AutoML frameworks and integrating pruning techniques to reduce evaluation time for non-promising configurations. Overall, the experimental results validate the effectiveness of this approach in addressing AutoML challenges, positioning it as a competitive solution for automating machine learning pipeline optimization.

P. An Efficient Contesting Procedure for AutoML Optimization [16]

This study explores the framework of an Automated Machine Learning (AutoML) Platform focused on selecting an optimal combination of hyperparameters and operators. It introduces a novel approach known as Divide And Conquer Optimization (DACOpt), which aims to enhance the robustness of AutoML. DACOpt partitions the AutoML search space into manageable sub-spaces based on algorithm similarity and budget constraints, providing a more effective alternative to traditional Bayesian Optimization (BO) methods. Unlike traditional BO, which integrates all operator search spaces into a single space, DACOpt dynamically allocates resources to each sub-space, prioritizing the most promising ones.

Inference: -

1) Bayesian Optimization (BO): BO is a widely used approach in AutoML for optimizing the selection of algorithms and hyperparameters. It employs a probabilistic surrogate model to approximate the objective function, iteratively selecting new points for evaluation based on predictions and an acquisition function. BO efficiently explores high-dimensional and noisy search spaces, adapting to observed data and focusing on promising regions.
2) Divide and Conquer Optimization (DACOpt): DACOpt is a contesting procedure proposed for AutoML optimization, with the goal of improving robustness. It partitions the AutoML search space into sub-spaces based on algorithm similarity and budget constraints. Each sub-space is independently optimized using BO, and resources are allocated based on performance. DACOpt offers advantages such as budget distribution, parallel efficiency, and improved performance compared to existing AutoML optimization methods.
3) DAC-HB and DAC-SB: DACOpt includes two approaches—Divide And Conquer Highest-Based (DAC-HB) and Divide And Conquer Statistical-Based (DAC-SB). DAC-HB selects the best-performing candidate based on the highest performance, while DAC-SB selects the best-performing candidate using a statistical procedure.

Conclusion: -The paper introduces DACOpt, a novel contesting procedure for AutoML optimization, which partitions the search space into sub-spaces, optimizes each sub-space independently, and efficiently allocates resources. DACOpt offers two variants, DAC-HB and DAC-SB, which demonstrate improved performance compared to traditional AutoML optimization approaches. This approach has the potential to enhance the robustness and efficiency of AutoML, making it a valuable contribution to the field.

Q. Automated Reinforcement Learning (AutoRL): A Survey and Open Problems [17]

The paper discusses the emergence of Automated Reinforcement Learning (AutoRL) as a promising approach to address the sensitivity of design choices in the training process of Reinforcement Learning (RL) agents. The effectiveness of RL agents largely depends on intricate design choices, requiring manual tuning. AutoRL aims to automate these design choices and has shown potential in various applications, from RNA design to complex games like Go. The survey provides a common taxonomy for AutoRL, delves into various areas in detail, and highlights open research problems. It unifies research efforts in different RL subfields, including meta-learning and evolution, to provide a comprehensive framework for automated RL. The primary objective is to stimulate further research and development in AutoRL, ultimately advancing the capabilities of RL agents and their potential for solving complex problems.

AutoRL bridges the gap between AutoML and RL, offering an approach to automate design and training processes for RL agents. By unifying research efforts in different RL subfields, it addresses the challenge of sensitive design choices. The paper aims to provide a common taxonomy for AutoRL and emphasizes open research challenges, stimulating further advancements in the field. AutoRL holds promise for making RL more accessible and efficient, with potential applications in a wide range of complex problems. The focus on these challenges will drive research and development, enhancing the capabilities of RL agents.

R. AMLBID: An auto-explained Automated Machine Learning tool for Big Industrial Data [18]

The ability to handle the entire machine learning pipeline, from data preprocessing to algorithm selection and hyperparameter tuning. The incorporation of a metalearning component, which leverages knowledge from past datasets, enhances the efficiency and effectiveness of the system. The ensemble construction component further bolsters the robustness of the results. By comparing AUTO-SKLEARN to an existing AutoML system, AutoWEKA, and analyzing the impact of its novel components, the authors demonstrate the superior performance of their approach. In particular, the meta learning component is identified as a significant contributor to the system's success. AUTO-SKLEARN's approach holds the potential to democratize machine learning by automating complex decisions and configurations, thereby reducing the barriers to entry for users without extensive machine learning expertise. It offers a valuable tool for researchers, practitioners, and data scientists looking to apply machine learning to a wide range of problems without the need for extensive manual intervention. The system's adaptability and efficiency in handling diverse datasets make it a promising asset in the field of AutoML.

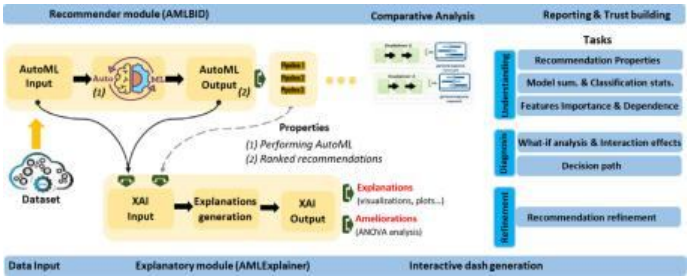


Fig. 6. Workflow of the white-box internal structures of AutoML.

S. AutoML for Multi-Label Classification: Overview and Empirical Evaluation [19]

This paper offers an overview and empirical evaluation of automated machine learning (AutoML) for multi-label classification. It explores the key components of AutoML approaches for multi-label classification, benchmarking setup, and optimizer interface. The paper also compares the complexity of the search space between single-label and multi-label classification and presents extensive empirical research to evaluate various AutoML methods on benchmark datasets. The results demonstrate that certain AutoML methods outperform others on specific datasets and performance metrics, supported by visualizations in scatter plots. This paper provides valuable insights into the effectiveness of AutoML for multi-label classification.

Inference: -

1) Major Components of AutoML Approaches: AutoML for multi-label classification considers the challenge of associating multiple labels with each instance, necessitating prediction for multiple labels. The optimizer employs techniques like random search, grid search, or evolutionary algorithms to navigate the expansive search space. The choice of optimizer significantly influences the pipeline's performance.
2) *Challenges in Comparing AutoML Methods:* Comparing AutoML methods presents challenges due to the complex search space and lack of standardization in evaluation procedures. The diverse exploration of the search space by different methods and variations in datasets, performance measures, and evaluation protocols hinder direct performance comparisons. Standardized benchmark datasets and evaluation procedures are essential for meaningful comparisons.
3) *Potential Benefits of AutoML:* AutoML has the potential to streamline and democratize machine learning by reducing the time and expertise required to develop effective pipelines. Traditionally, various complex steps in pipeline development necessitate expertise, making it inaccessible to non-experts. AutoML aims to make machine learning more accessible.
4) *Effectiveness of AutoML Methods:* Effective AutoML systems require careful consideration of factors such as the search space, optimization method, missing data handling, and performance evaluation. Balancing these factors enables the construction of highly effective machine learning pipelines adaptable to various applications.

5) *Potential Applications in Various Domains:* AutoML optimization methods include meta-learning, evolutionary algorithms, Bayesian optimization, gradient-based optimization, and reinforcement learning. These methods aim to efficiently explore the search space, predict pipeline performance, and guide search towards better solutions. The suitability of these methods depends on the problem type.

Conclusion: - AutoML platforms hold the potential to reduce the time and expertise needed for developing effective machine learning pipelines. However, empirical comparisons of AutoML tools are challenging due to the complexity of the search space and evaluation procedure variations. To address these challenges, benchmarking, and standardized evaluation procedures are crucial. Effective AutoML systems require a balanced consideration of factors for constructing highly adaptable machine learning pipelines.

T. How far are we from true AutoML: reection from winning solutions and results of AutoDL challenge [20]

The quest for true AutoML, an algorithm capable of consistently delivering high performance across diverse data, was explored through the AutoDL challenges in 2019. While groundbreaking theoretical insights remained elusive, the challenge showcased promising results with deep learning using pre-trained networks. The top two winners passed final tests on unseen data, and their open-sourced solutions marked progress toward automation. However, the absence of novel theoretical breakthroughs and the reliance on domain-specific workflows revealed the complexity of the AutoML problem. Future directions, including a meta-learning challenge, were proposed to advance AutoML's potential for generalization across new domains. The AutoDL challenge involved various data types, including images, videos, audio, text, and tabular data, presented in a standardized tensor format. A total of 100 datasets were prepared, and the challenge featured two phases: a feedback phase with practice datasets and a final phase with fresh datasets. The challenge employed the Area under the Learning Curve (ALC) as the scoring metric, encouraging any-time learning, and ranked participants based on ALC across individual datasets to determine the winners. The challenge saw participation from 54 teams, resulting in numerous submissions. The final rankings were based on performance across 10 unseen datasets, with a focus on reducing variance from various factors. Winning approaches utilized a combination of methods from different domains and were heavily inspired by a baseline approach. Although true AutoML is yet to be achieved, the winning approaches demonstrated good generalization ability and addressed the any-time learning problem effectively. The AutoDL series in 2019 underscored the dominance of deep learning in various domains and highlighted the readiness of automated deep learning methods across different domains. The emergence of meta-learning and ongoing challenges are promising avenues for future exploration, aiming to automate meta-learning and potentially develop universal workflows for AutoML.

III. CONCLUSIONS

In conclusion, our proposed AutoML platform represents a significant advancement in the field of automated machine learning, offering a user-friendly solution for the comparative analysis of machine learning models. The platform empowers users to make informed decisions about the selection of machine learning algorithms by providing a seamless and efficient process, from data upload to algorithm evaluation. Through the use of performance metrics and data-driven insights, users can gain a deeper understanding of how different algorithms impact their specific datasets.

This research contributes to the democratization of machine learning, making it more accessible to a broader audience by reducing the barriers associated with algorithm selection. By promoting transparency and explain ability, our platform assists users in choosing the most suitable algorithm for their unique data, thereby facilitating the development of more accurate and effective machine learning models.

REFERENCES

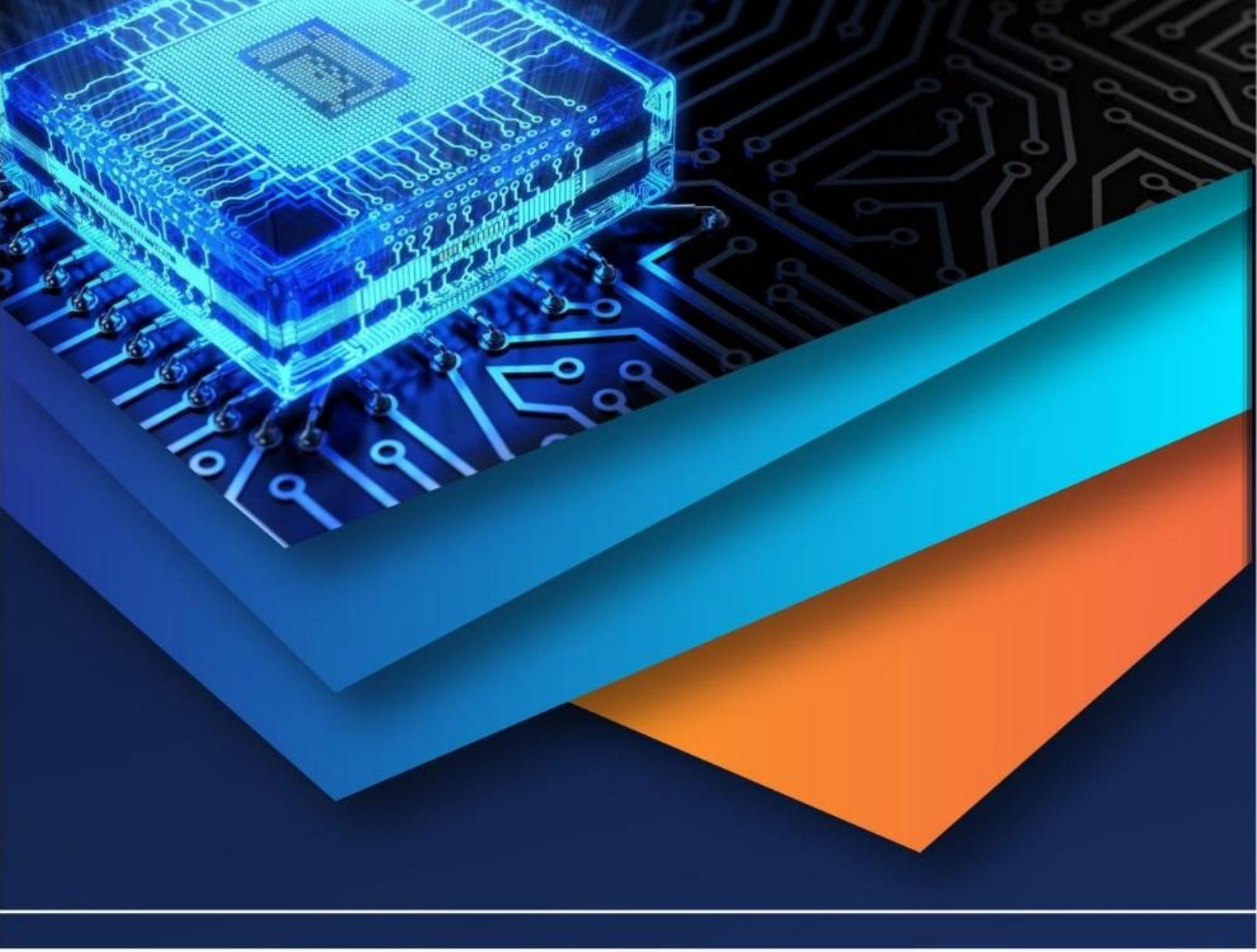
[1] Karansingh Chauhan1, Shreena Jani1, Dhrumin Thakkar1, Riddham Dave1,Jitendra Bhatia1, Sudeep Tanwar2, Mohammad S. Obaidat, Fellow of IEEE and Fellow of SCS3: Automated Machine Learning: The New Wave of Machine Learning in IEEE Xplore Part Number: CFP20K58-ART; ISBN: 978-1-7281-4167-1

[2] Lukas Tuggener1,2, Mohammadreza Amirian1,3, Katharina Rombach1, Stefan L¨orwald4, Anastasia Varlet4, Christian Westermann4, and Thilo Stadelmann1: Automated Machine Learning in Practice: State of the Art and Recent Results in 2019 6th Swiss Conference on Data Science (SDS)

[3] Xiawu Zheng , Yang Zhang , Sirui Hong, Huixia Li, Lang Tang, Youcheng Xiong, Jin Zhou, Yan Wang ,Xiaoshuai Sun, Member, IEEE, Pengfei Zhu , Chenglin Wu , and Rongrong Ji , Senior Member, IEEE: Evolving Fully Automated Machine Learning via Life-Long Knowledge Anchors in IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 43, NO. 9, SEPTEMBER 2021

[4] Thiloshon Nagarajah University of Westminster, Guhanathan Poravi Informatics Institute of Technology: A Review on Automated Machine Learning (AutoML) Systems in 2019 IEEE 5th International Conference for Convergence in Technology (I2CT)DOI:10.1109/I2CT45611.2019 29-31 March 2019

[5] Matthias Feure,r Aaron Klein, Katharina Eggensperger Department of Computer Science University of Freiburg, Germany: Efficient and Robust Automated Machine Learning in NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2December 2015Pages 2755–2763
[6] Ahmed Abd Elrahman, Mohamed El Helw Nile University Giza, Egypt Radwa Elshawi, Sherif Sakr University of Tartu Tartu, Estonia: D-SmartML: A Distributed Automated Machine Learning Framework in 2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)
[7] Xu Chen and Brett Wujek: A Unified Framework for Automatic Distributed Active Learning in IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 44, NO. 12, DECEMBER 2022
[8] Bilge Celik and Joaquin Vanschoren:Adaptation Strategies for Automated Machine Learning on Evolving Data in IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 43, NO. 9, SEPTEMBER 2021
[9] Forough Majidi†, Moses Openja†, Foutse Khomh†, Heng Li†: An Empirical Study on the Usage of Automated Machine Learning Tools in arXiv:2208.13116v1 [cs.SE] 28 Aug 2022
[10] Anh Truong∗, Austin Walters∗, Jeremy Goodsitt∗, Keegan Hines∗, C. Bayan Bruss∗, Reza Farivar∗: Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools in 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)
[11] Lu´ıs Ferreira, Andr´e Pilastri, Carlos Manuel Martins, Pedro Miguel Pires, Paulo Cortez: A Comparison of AutoML Tools for Machine Learning, Deep Learning and XGBoost in 2021 International Joint Conference on Neural Networks (IJCNN) DOI: 10.1109/IJCNN52387.2021 18-22 July 2021
[12] Nick Erickson, Jonas Mueller, Alexander Shirkov , Hang Zhang, Pedro Larroy , Mu Li, Alexander Smola: AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data in 7th ICML Workshop on Automated Machine Learning (2020)
[13] DOMINIK KREUZBERGER, NIKLAS KÜHL AND SEBASTIAN HIRSCHL: Machine Learning Operations (MLOps): Overview, Definition, and Architecture in IEEE Access (Volume: DOI: 10.1109/ACCESS.2023.3262138 Date of Publication: 27 March 2023
[14] Felix Mohr , Marcel Wever , Alexander Tornede , and Eyke H€ullermeier: Predicting Machine Learning Pipeline Runtimes in the Context of Automated Machine Learning in IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 43, NO. 9, SEPTEMBER 2021
[15] Duc Anh Nguyen, Anna V. Kononova, Stefan Menzelx, Bernhard Sendhoffx, and Thomas Bäck, Leiden Institute of Advanced Computer Science (LIACS), Leiden University, The Netherlands: Efficient AutoML via Combinational Sampling in 2021 IEEE Symposium Series on Computational Intelligence (SSCI) DOI: 10.1109/SSCI50451.2021 5-7 Dec. 2021
[16] DUC ANH NGUYEN 1, ANNA V. KONONOVA, STEFAN MENZEL, BERNHARD SENDHOFF AND THOMAS BÄCK: An Efficient Contesting Procedure for AutoML Optimization in IEEE Access (Volume: 10) DOI: 10.1109/ACCESS.2022.3192036 Date of Publication: 18 July 2022
[17] Jack Parker-Holder, Raghu Rajan rajanr, Xingyou Song, André Biedenkapp, Yingjie Miao, Theresa Eimer, Baohe Zhang, Vu Nguyen, Roberto Calandra, Aleksandra Faust, Frank Hutter, Marius Lindauer: Automated Reinforcement Learning (AutoRL): A Survey and Open Problems in Journal of Artificial Intelligence Research 74 (2022) 517-568 Submitted 01/2022; published 06/2022
[18] Moncef Garouani, Adeel Ahmad, Mourad Bouneffa, Mohamed Hamlich: AMLBID: An auto-explained Automated Machine Learning tool for Big Industrial Data in ELSEVIER SoftwareX 17 (2022) 100919
[19] MarcelWever , Alexander Tornede , Felix Mohr , and Eyke H€ullermeier , Senior Member, IEEE: AutoML for Multi-Label Classification: Overview and Empirical Evaluation in IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 43, NO. 9, SEPTEMBER 2021
[20] Zhengying Liu, Adrien Pavao, Zhen Xu, Sergio Escalera, Isabelle Guyon, Julio C. S. Jacques Junior, Meysam Madadi, Sebastien Treguer: How far are we from true AutoML: reection from winning solutions and results of AutoDL challenge in 7th ICML Workshop on Automated Machine Learning (2020)



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY