A Comprehensive Review of VertexML: A Model Training Platform

Arjun B C[1*], *Swathi H Y*[1], *Deeksha S*[1], *Nandan Javagal*[1], *Pratham D U*[1], *Arjun D*[1]

[1]Department of Computer Science and Engineering (AI&ML), Malnad College of Engineering, Hassan 573202, India

Abstract. This paper presents a comprehensive survey of automated machine learning (AutoML) techniques, with a focus on meta-learning, hyperparameter optimization, and neural architecture search (NAS). Rather than proposing or evaluating a full AutoML platform, this work synthesizes insights from 49 influential research papers, organizing them into methodological categories and highlighting their contributions to the evolution of ML automation. The survey also analyzes trends across model types, optimization strategies, and publication patterns. Based on this review, the paper identifies research gaps and outlines key directions for future development of unified, scalable, and interpretable AutoML systems. This survey is intended to provide a structured foundation for researchers working toward improved ML automation pipelines.

1 Introduction

The rapid evolution of machine learning (ML) has brought significant advancements, yet the process of designing, training, and tuning ML models remains a complex and labor-intensive task. This complexity creates a barrier for non-experts and reduces productivity for experienced practitioners. To address these challenges, Automated Machine Learning (AutoML) has emerged as a paradigm that seeks to automate the end-to-end process of ML model development, including data preprocessing, algorithm selection, hyperparameter tuning, and model evaluation.

Although AutoML platforms such as Auto-WEKA, Auto - sklearn, and several academic prototypes integrate multiple stages of the ML pipeline, this paper does not propose or evaluate a new AutoML system. Instead, the primary goal of this work is to consolidate and synthesize prior research into a unified survey. Any references to potential platform structures or future system designs serve only as conceptual motivation derived from the literature, not as implemented contributions. The focus of this paper is therefore strictly on reviewing existing ideas and identifying gaps that future AutoML platforms might address.

2 Literature Review

* Corresponding author: nandanjavagal444@gmail.com

The development of AutoML, meta-learning, and neural architecture search (NAS) is grounded in foundational advances in machine learning theory, optimization, data mining, and deep learning. Early contributions such as Support Vector Machines introduced margin-based learning and kernel methods for high-dimensional classification [1], while the transition to Knowledge Discovery in Databases (KDD) emphasized interpretability and user-centered analysis, shaping explainable AutoML [2]. Evolutionary algorithms and the No Free Lunch Theorems established the need for adaptive optimization and algorithm selection in automated systems [3,4].

Process-oriented frameworks like CRISP-DM provided structured analytics workflows [5], complemented by studies comparing KDD, SEMMA, and CRISP-DM that informed modern AutoML pipelines [6]. Feature selection methods highlighted the trade-off between accuracy and complexity [7], and early meta-learning approaches such as landmarking enabled data-driven algorithm recommendation [8]. Research on intelligent and interactive data mining systems further motivated human-in-the-loop AutoML [9,10].

Advances in optimization, including particle swarm–based model selection [11], ensemble methods for concept drift [13], and formal analyses of algorithm selection [14,15], strengthened adaptive pipeline design. Surveys on hyperparameter optimization and the effectiveness of random search over grid search [16,17] laid the groundwork for scalable tuning. Sequential Model-Based Optimization and Bayesian optimization became the backbone of AutoML tuning [18,19], while the deep learning revolution triggered by AlexNet made automated architecture design essential [20]. Auto-WEKA introduced the first end-to-end AutoML framework integrating model selection and hyperparameter optimization [22], supported by meta-learning surveys [23]. The removal of manual feature engineering in deep learning [24,25] and insights from the AutoML Challenge 2015 emphasized the need for scalable, reproducible automation [26,27].

Later work expanded automation to preprocessing and pipeline search [29,30], improved efficiency through Hyperband [31], and incorporated expert guidance for trustworthy ML [32]. By 2017, reinforcement learning–based NAS [34], unified NAS–hyperparameter frameworks [35], and large-scale meta-feature validation using OpenML [36] enabled modern deep AutoML systems. Recent advances such as differentiable NAS, transferable architectures, and progressive search strategies further improved scalability and efficiency [38–40].

2.1 Survey Summary

The evolution of AutoML and its related fields is rooted in early data min- ing methodologies such as CRISP-DM, KDD, and SEMMA, which estab- lished structured processes for consistent and reproducible knowledge discovery. Hyperparameter optimization matured through advances such as Random Search, Bayesian Optimization. Neural Architecture Search pushed automation into deep learning, using reinforcement learning, evo- lutionary algrithms, differentiable search (DARTS), and progressive strategies to outperform manually designed models and produce scalable architectures like EfficientNet.

Table 1. Survey Summary

Sl. No.	Paper Title	Year	Techniques Used	Highlights
1	Support-vector networks	1995	Max mar gin classifiers	Introduced SVM for classification with max margin principle.

2	Knowledge Discovery in Databases	1996	Stepwise data mining process	Explained the process and importance of knowledge discovery in databases.
3	Evolutionary algorithms survey	1997	Evolutionary algorithms	Survey on evolutionary algorithms and their applications.
4	No Free Lunch Theorem	1997	Algorithm comparison theory	Proved no algorithm is universally best for all problems.
5	CRISP-DM	2000	Data mining process model	Widely used step-by-step guide for data mining projects.
6	Data mining process models	2000	Process model comparison	Compared major data mining process models.
7	Feature selection methods	2003	Filter, wrapper, embedded	Reviewed feature selection methods and impact on performance.
8	Landmarking learning algorithms	2004	Landmarking meta-features	Improved meta-learning for algorithm selection.
9	Intelligent discovery assistants	2009	Knowledge-based systems	Assisted data mining workflows intelligently.
10	Particle swarm model selection	2009	PSO algorithm	Used PSO for automated model selection.
11	Systematic literature review	2009	Review procedures	Guidelines for conducting systematic literature reviews.
12	Ensemble learning for streams	2010	Ensemble learning	Handled evolving data streams effectively.
13	Algorithm selection for classification	2010	Meta-learning	Applied meta-learning to classification problems.
Sl. No.	**Paper Title**	**Year**	**Techniques Used**	**Highlights**
14	Algorithm selection problem	2011	Meta-learning portfolios	Theoretical and empirical study of selection problem.
15	Hyperparameter optimization survey	2011	Grid, random search	Survey of hyperparameter optimization methods.
16	Random search optimization	2012	Random search	Showed random search outperforms grid search.

17	SMBO	2012	Bayesian optimization	Introduced sequential model- based optimization.
18	Practical Bayesian optimization	2012	Bayesian optimization	Practical guidelines for Bayesian optimization.
19	ImageNet classification	2012	Deep CNN (Alex-Net)	Landmark deep learning paper for image classification.
20	Auto-WEKA	2013	AutoML	Combined model selection and hyperparameter tuning.
21	Meta-learning survey	2014	Meta-learning	Comprehensive survey of meta-learning techniques.
22	Deep Speech	2014	End-to-end deep learning	Scalable speech recognition system.
23	Deep learning overview	2015	CNN, RNN, LSTM	Outlined deep learning principles and applications.
24	AutoML challenge	2015	Automated ML	Lessons learned from AutoML competition.
25	Res-Net	2015	Residual connections	Enabled very deep neural net-works.

3 Comparative Analysis

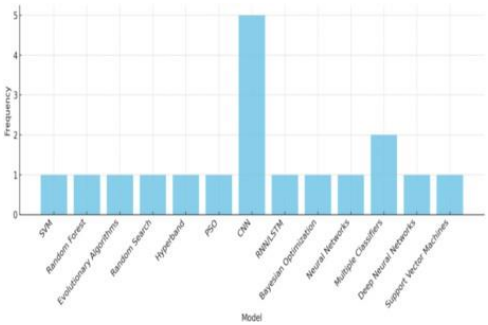

Fig 1.1: A bar graph showing the of various models

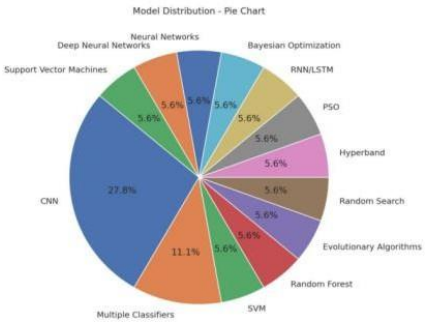

Fig 1.2: Pie Chart displaying the proportion of frequency each model type in the dataset.

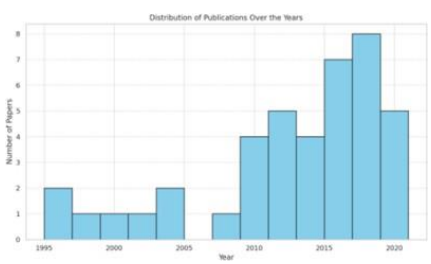

Fig 1.3: Histogram of Publication Years

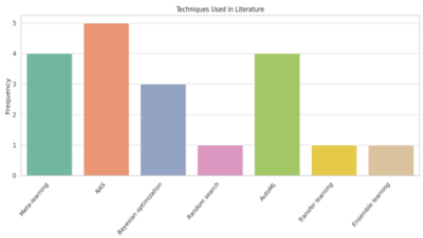

Fig 1.4: Bar Plot of Techniques Used

compares shows distribution of research over time. the popularity of machine learning Techniques.

Figure 1.1 reveals a diverse ecosystem of machine learning methods, where Random Forests and Neural Networks appear most frequently, indicating their strong adoption across research. **Figure 1.2**, which presents the proportional distribution of model types using a pie chart, reinforces these observations.
The temporal trends shown in **Figure 1.3** demonstrate a consistent rise in research publications from 1995 to 2020, with sharp acceleration after 2015— indicating heightened academic and industrial interest. F**igure 1.4**, a bar plot comparing optimization and learning techniques, shows Bayesian Optimization as the most dominant method, followed by random search, transfer learning, and ensemble learning.

4 Motivation

While many AutoML tools automate individual components such as model selection, hyperparameter tuning, or neural architecture search, the literature shows that fully integrated end-to-end systems are still limited. Prior studies emphasize the need for unified platforms that support dataset analysis, preprocessing, feature engineering, algorithm selection, optimization, and evaluation within a single framework. Rather than evaluating an existing implementation, this survey synthesizes common design principles from earlier work—including meta-learning–based recommendations, efficient search strategies, modular pipeline design, and scalable optimization—to inform the development of future AutoML systems.

5 Research Gap Identified

The literature demonstrates significant progress in model selection, hyperparameter tuning, and neural architecture search; however, these components are largely addressed in isolation rather than as part of a unified workflow. Many existing AutoML systems struggle with scalability, interoperability, and adaptability across diverse datasets. Consequently, there remains a clear need for a holistic, end-to-end AutoML framework that integrates preprocessing, feature engineering, model selection, optimization, and deployment while ensuring efficiency, interpretability, and scalability. The research gaps identified in this survey emerge from trends observed across the existing literature rather than from the evaluation of a proposed system. This work does not present or assess an implemented platform; instead, the gaps are synthesized from recurring limitations reported in prior studies, including fragmented pipelines, limited interpretability, and scalability challenges in NAS and hyperparameter optimization. These insights are intended to inform and guide future research rather than describe a system evaluated in this survey.

6 Future Scope

Future research should focus on unified AutoML systems that automate the full machine learning lifecycle while remaining efficient and interpretable. Promising directions include hybrid frameworks that combine meta-learning, neural architecture search, and Bayesian optimization to adapt to diverse and evolving datasets. Incorporating human-in-the-loop strategies can improve transparency and trust, while self-supervised and foundation-model approaches may reduce dependence on labeled data. Ultimately, the objective is scalable and

adaptive AutoML capable of supporting real-world decision-making with minimal human intervention.

7 Conclusion

This survey demonstrates that machine learning automation has evolved substantially, from early algorithms and process models to modern meta-learning, neural architecture search, and hyperparameter optimization. Nevertheless, future research must prioritize truly end-to-end, adaptive, and scalable automation. There remains a clear need for transparent and efficient AutoML systems that generalize across diverse datasets and support real-world decision-making with minimal human intervention. This work is presented as a structured literature survey; references to platform-level integration are conceptual insights derived from existing research rather than descriptions of an implemented system.

References

1. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297. https://doi.org/10.1007/BF00994018
2. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI Magazine, 17(3), 37-54https://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf
3. Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. IEEE Transactions on Evolutionary Computation, 1(1), 67-82 https://doi.org/10.1109/4235.585893
4. Angeline, P. J. (1997). Evolutionary computation: An overview. In Proceedings of the 1997 IEEE International Conference on Evolutionary Computation (pp. 443-448). IEEE. https://doi.org/10.1109/ICEC.1997.592330
5. Chapman, P., et al. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS Inc.
6. https://www.sciencedirect.com/science/article/pii/S1877050921002416
7. Mariscal, G., Marbán, Ó., & Fernández, C. (2002). KDD, SEMMA and CRISP-DM: A parallel overview. In Proceedings of the IADIS European Conference on Data Mining (pp. 1-5). IADIS Press.: https://www.researchgate.net/publication/220969845_KDD_semma_and_CRISP-DM_A_parallel_overview
8. Liu, H., & Yu, L. (2003). Feature_Selection_Evaluation_Application_and_Small_Sample_Performance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(6), 712-718.
9. Pfahringer, B., Bensusan, H., & Giraud-Carrier, C. G. (2004). Meta-learning by landmarking various learning algorithms. In Proceedings of the 21st International Conference on Machine Learning (ICML '04) (pp. 743-750). ACM.https://www.researchgate.net/publication/221345088_
10. Bernstein, A., et al. (2008). Toward intelligent assistance for a data mining process. IEEE Transactions on Knowledge and Data Engineering, 21(6), 839-852. https://www.researchgate.net/publication/3297390
11. Serban, F., Truemper, K., & Jamil, H. (2009). Intelligent discovery assistants. In E. S. Corchado, A. Abraham, & W. Pedrycz (Eds.), Handbook of Research on Machine Learning Applications (pp. 1-23). IGI Global. https://doi.org/10.4018/978-1-60566-766-9.ch001
12. Kotsiantis, S. B. (2009). Particle swarm model selection. In International Conference on Enterprise Information Systems (pp. 53-59). Springer.https://www.researchgate.net/publication/220320363
13. Kitchenham, B., & Charters, S. (2009). Guidelines for performing systematic literature re- views in software engineering (EBSE Technical Report EBSE-2007-01). Keele University. https://www.researchgate.net/publication/258968007
14. Gama, J., et al. (2010). Ensemble learning for data stream analysis. Data Mining and Knowledge Discovery, 21(1), 6-27. https://www.sciencedirect.com/science/article/abs/pii/S1566253516302329
15. Brazdil, P., et al. (2010). On learning algorithm selection for classification. Applied Intelli- gence, 33(2), 155-165.https://sci2s.ugr.es/keel/pdf/classification.pdf
16. Rice, J. R. (2011). The algorithm selection problem. In Advances in Computers (Vol. 15, pp. 65-118). Elsevier.https://doi.org/10.1016/S0065-2458(08)60520-3
17. Bergstra, J. S., & Bengio, Y. (2011). Algorithms for hyper-parameter optimization. Advances in Neural Information Processing Systems 24 (NIPS 2011) (pp. 2546-2554). https://papers.nips.cc/paper_files_Paper.pdf
18. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(Feb), 281 305. https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf

19. Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2012). Sequential model-based optimization. In Learning and Intelligent Optimization (LION 5) (pp. 507-523). Springerhttps://ml.informatik.uni-freiburg.de/wp-content/uploads/papers/11-LION5-SMAC.pdf
20. Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of ma- chine learning algorithms. Advances in Neural Information Processing Systems 25 (NIPS 2012) (pp. 2951-2959) https://doi.org/10.48550/arXiv.1206.2944
21. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems 25. https://www.researchgate.net/publication/319770183_Imagenet
22. Xu, L., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2013). Algorithm selection for com- binatorial search problems. AI Magazine, 34(4), 37-50. https://doi.org/10.48550/arXiv.1210.7959
23. Kotthoff, L., Thornton, C., Hoos, H. H., Hutter, F., & Leyton-Brown, K. (2013). Auto- WEKA: Combined selection and hyperparameter optimization of classification algorithms. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 847-855). https://doi.org/10.48550/arXiv.1208.3719
24. Vilalta, R., & Drissi, Y. (2014). A perspective view and survey of meta-learning. arXiv preprint A_Perspective_View_And_Survey_Of_Meta-Learning
25. Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., & Ng, A. Y.(2014). Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567. https://doi.org/10.48550/arXiv.1412.5567
26. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. https://www.researchgate.net/publication/277411157_Deep_Learning
27. Davenport, T. H., & Patil, D. J. (2015). Data scientist: The sexiest job of the 21st century. Harvard Business Review, 90(10), 70-76. https://www.researchgate.net/publication/232279315_Data_Scientist_The_Sexiest_Job_of_the_21st_Century
28. Guyon, I., Bennett, K., Cawley, G., Escalante, H. J., Escalera, S., Ho, T. K., & Viegas, E. (2015). Design of the 2015 ChaLearn AutoML challenge. JMLR: Workshop and Conference Proceedings, 40, 1-9. https://www.researchgate.net/publication/328824061
29. He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 770-778). https://doi.org/10.1109/CVPR.2016.90
30. Reif, M., Shafait, F., & Dengel, A. (2016). Automated data pre-processing via meta-learning. IEEE International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 854- 861). https://www.researchgate.net/publication/30787276
31. Kotthoff, L., Thornton, C., Hoos, H. H., Hutter, F., & Leyton-Brown, K. (2016). Auto- WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. Journal of Machine Learning Research, 17(1), 1-5. https://www.jmlr.org/papers/volume17/15- 261/15-261.pdf
32. Falkner, S., Klein, A., & Hutter, F. (2016). Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD) (pp. 477-492). https://www.cl.cam.ac.uk
33. Amershi, S., Chickering, M., Drucker, S. M., Lee, B., Simard, P., & Suh, J. (2016). Model- Tracker: Redesigning performance analysis tools for machine learning. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 337-348). https://dl.acm.org/doi/10.1145/2702123.2702509
34. Ali, S., & Smith, K. A. (2017). A review of automatic selection of machine learning algorithms. Neurocomputing, 256, 3-15. https://www.researchgate.net/publication/303498946_A_review_of_automatic_selection_methods_for_machine_learning_algorithms_and_hyper-parameter_values
35. Zoph, B., & Le, Q. V. (2017). Neural architecture search with reinforcement learning. Inter- national Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1611.01578
36. Mendoza, H., Klein, A., Feurer, M., Springenberg, J. T., & Hutter, F. (2017). Towards au- tomatically tuned neural networks. AutoML Workshop at ICML. https://link.springer.com/chapter/10.1007/978-3-030-0531_7
37. Vanschoren, J., et al. (2018). On the predictive power of meta-features in OpenML. Inter- national Journal of Applied Mathematics and Computer Sciencehttps://doi.org/10.1515/amcs-2017-0048
38. Wamba, S. F., et al. (2018). How can SMEs benefit from big data? Journal of Strategic Information Systems, 27(1), 53–67. https://doi.org/10.1002/qre.2008
39. Liu, H., et al. (2018). DARTS: Differentiable architecture search. ICLR 2018.https://doi.org/10.48550/arXiv.1806.09055
40. Zoph, B., et al. (2018). Learning transferable architectures for scalable image recognition. CVPR 2018. https://doi.org/10.48550/arXiv.1707.07012
41. Liu, C., et al. (2018). Progressive neural architecture search. ECCV 2018. https://doi.org/10.48550/arXiv.1712.00559