APTISI Transactions on Technopreneurship (ATT)
Vol. 7, No. 3, November 2025, pp. 808−822
E-ISSN: 2656-8888 | P-ISSN: 2655-8807, DOI:10.34306

Optimizing Automated Machine Learning for Ensemble Performance and Overfitting Mitigation

Migunani[1*], Adi Setiawan[2], Irwan Sembiring[3]
[1]Faculty of Academic Studies, Universitas Sains dan Teknologi Komputer, Indonesia
[2]Faculty of Science and Mathematics, Satya Wacana Christian University, Indonesia
[1,3]Faculty of Information Technology, Satya Wacana Christian University, Indonesia
[1]migunani@stekom.ac.id, [2]adi.setiawan@uksw.edu, [3]irwan@uksw.edu
***Corresponding Author**

Article Info

Article history:

Submission June 12, 2025
Revised September 4, 2025
Accepted October 14, 2025
Published October 29, 2025

Keywords:

Systematic Literature Review
AutoML
Ensemble Learning
Overfitting Mitigation
Enhancing Diversity

ABSTRACT

Automated Machine Learning (AutoML) has revolutionized model development, but its impact on ensemble diversity and overfitting reduction remains underexplored. This **Systematic Literature Review (SLR)** analyzes 107 studies published between 2020 and 2024 to explore how AutoML enhances ensemble diversity, mitigates overfitting, and the challenges hindering its integration. Unlike previous reviews focusing on AutoML or ensemble methods independently, this study synthesizes their intersection and identifies key research trends. The **findings** reveal that AutoML improves ensemble robustness through automated hyperparameter tuning, meta-learning, and algorithmic blending while facing trade-offs in computational cost and interpretability. Four main themes emerge, integration mechanisms (19.6%), overfitting mitigation (26.2%), performance trade-offs (28.6%), and integration barriers (26.2%). Empirical results indicate that AutoML ensembles outperform traditional models by 22–41% in accuracy but require approximately 3.2 times higher computational resources. Hybrid AutoML and Explainable AI frameworks are recommended to balance accuracy and transparency. Theoretically, this study advances understanding of the synergy between AutoML and ensemble learning, while practically providing guidance for deploying reliable AI systems in sectors like healthcare, finance, and digital business. Policy **implications** align with the EU AI Act and the US Executive Order on trustworthy AI, supporting Sustainable Development Goals 9 and 8.

DOI: https://doi.org/10.34306/att.v7i3.763

1. INTRODUCTION

In recent AI research, Machine Learning (ML) models face the challenge of overfitting, where they perform well on training data but fail to generalize to unseen data, undermining their reliability [1]. AutoML has emerged to address this issue by automating tasks like algorithm selection, pipeline configuration, and hyperparameter tuning, reducing dependency on expert knowledge and speeding up development [1, 2]. Additionally, ensemble learning methods such as bagging, boosting, and stacking improve predictive accuracy and mitigate overfitting by combining multiple models to enhance performance and reduce variance [3–5].

While AutoML and ensemble techniques have been studied separately, their synergy using AutoML to enhance ensemble diversity for better generalization and overfitting mitigation remains an underexplored

gap [6, 7]. This study addresses this by presenting a SLR of 107 peer-reviewed studies from 2020 to 2024. Previous SLRs have mainly focused on AutoML [8, 9] or ensemble learning independently [3–5, 10–12], offering descriptive overviews or general challenges [13, 14]. Our review critically evaluates the intersection of AutoML-driven ensemble methods, diversity enhancement, and overfitting mitigation, providing a comprehensive synthesis of both theoretical and practical insights [15, 16]. The findings highlight how automation improves ensemble performance, reduces human bias in model selection, and emphasizes transparency, reproducibility, and sustainability in AI development [17, 18]. This review reinforces the foundation for evidence-based innovation in automated systems. The contributions of this work are threefold:

- Novel synthesis, it provides a novel and comprehensive synthesis of mechanisms through which AutoML automates the creation of diverse ensembles to combat overfitting, integrating advanced techniques such as neural architecture search (NAS) and evolutionary algorithms [8] with hyperparameter optimization (HPO) methods [9–11] and ensemble strategies [6].
- Critical evaluation and domain insights, the research identify and evaluate the strategies and performance trade offs (e.g., accuracy gains of 41% vs 3.2 times computational costs) applied in different domains such as healthcare [18] and finance [14], offering insights beyond descriptive reporting.
- Practical and theoretical relevance, the research generate actionable recommendations for industry practitioners to implement hybrid AutoML ensemble strategies in real world settings, while also addressing prevailing limitations like high computational demands [12, 19] and interpretability challenges [13, 20] to outline a roadmap for future research.

By critically examining this synergy, our review aims to advance the theoretical understanding of robust ML design and provide a foundation for developing next generation, automated ensemble frameworks that are both high performing and practically viable. Multidisciplinary insights from computer science, public policy, and social sciences ensure comprehensive analysis of technical and societal dimensions.

This research also contributes to the United Nations Sustainable Development Goals (SDGs) by developing methods for robust and accessible AI. By automating the creation of reliable ensemble models, it supports SDG 9 (Industry, Innovation, and Infrastructure) through the innovation of trustworthy AI tools. Furthermore, by lowering the barrier to entry for developing high-performance AI, it advances SDG 8 (Decent Work and Economic Growth) by democratizing expertise and enabling productivity gains across diverse sectors.

2. RESEARCH METHOD

This study employs a Systematic Literature Review (SLR) methodology, adhering to established guidelines [21] to ensure transparency, rigor, and replicability. The process is structured into three phases, planning, conducting, and reporting, as illustrated in Figure 1. During the planning phase, the need for this review was established based on the identified research gap concerning AutoML's role in enhancing ensemble diversity and mitigating overfitting. Research questions were formulated using the PICOC framework as see in Table 1 to guide the review, and a detailed protocol was developed. This protocol specified the search strategy, inclusion and exclusion criteria, data extraction procedures, and quality assessment standards.

Table 1. Summary of PICOC framework

Component	Description
P (Population)	The reviewed studies encompass applications of Machine Learning and deep learning that integrate ensemble methods with AutoML techniques.
I (Intervention)	A key focus is the implementation of AutoML to optimize ensemble diversity and mitigate overfitting.
C (Comparison)	In contrast, traditional ML approaches rely on manual tuning and ad hoc ensemble construction, which often result in limited scalability and suboptimal performance.
O (Outcome)	Evidence from the literature highlights that AutoML driven ensembles achieve improved performance metrics, reduced generalization error, and enhanced robustness.
C (Context)	The scope of this review covers research published between 2020 until 2024, specifically addressing the intersection of AutoML and ensemble learning.

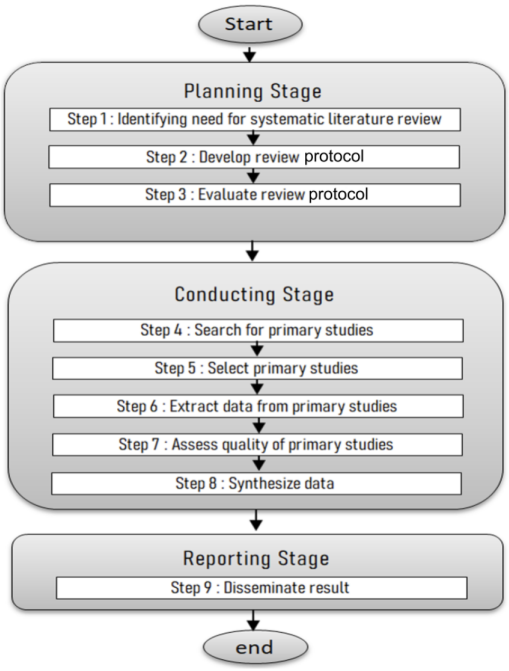

Figure 1. Systematic Literature Review Steps

During the conducting phase, a systematic search was carried out across three major databases, Scopus, IEEE Xplore, and the ACM Digital Library. The search strategy applied Boolean logic with keywords derived from the PICOC framework, resulting in the final query: `(("Automated Machine Learning" OR AutoML) AND ("Ensemble Learning" OR "Ensemble Models") OR ("Diversity" OR "Model Diversity") OR ("Overfitting" OR "Reducing Overfitting"))`. The query was adapted to the syntax of each database to maximize precision and relevance. The initial search produced a large set of records, which were screened by title and abstract, followed by full text reviews using predefined eligibility criteria as see in Table 2. Ultimately, 107 primary studies published between January 2020 and December 2024 were included. The overall process including identification, screening, and selection stages is illustrated in Figure 2.

Table 2. Inclusion and Exclusion Criteria

Inclusion Criteria	**Exclusion Criteria**
Studies from academic or industrial settings applying AutoML in ensemble diversity or overfitting contexts	Non English publications
Research evaluating the effectiveness of AutoML in ensemble learning	Studies lacking empirical validation or irrelevant to AutoML–ensemble integration
Most recent or comprehensive version selected in cases of duplicates	Conference versions when corresponding journal publications are available

The detailed search process and the number of studies identified at each stage are presented in the PRISMA flow diagram in Figure 2. The study selection Step 5 was performed in two stages namely exclusion based on title and abstract screening and exclusion based on full text review. The initial screening yielded 107 primary studies which were then subjected to full text assessment. In addition to the predefined inclusion and exclusion criteria further considerations included study quality relevance to the research questions and thematic alignment. Duplicate or highly similar publications by the same authors across different venues were removed. Following this process 107 primary studies were retained for analysis.

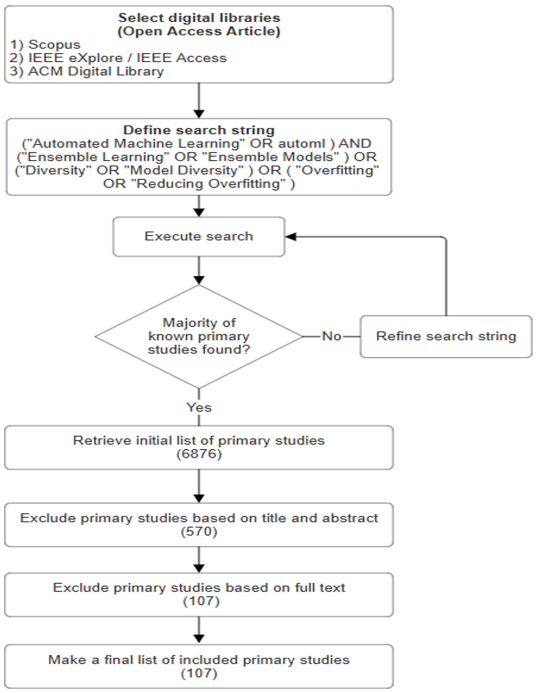

Figure 2. Presents the PRISMA Flow Diagram Illustrating the Study Selection Process

During the reporting phase the selected studies were synthesized to identify recurring themes and patterns aligned with the research objectives. A narrative synthesis approach was adopted to integrate qualitative insights with limited quantitative trends. The methodology was refined iteratively throughout the process to ensure comprehensiveness and coherence of the findings.

2.1. Research Questions (RQ) and Objectives

This systematic review applies the PICOC framework to ensure focus and clarity. The Population covers studies on machine and deep learning, the Intervention explores AutoML techniques that improve ensemble diversity and reduce overfitting, and the Comparison evaluates them against traditional methods. Outcomes are measured through predictive performance and generalization within studies published from 2020 to 2024, forming the basis for the research questions in Table 3.

Table 3. Research Questions and Motivations

RQ ID	Research Question	Motivation
RQ1	What is the role of AutoML in generating and selecting diverse base models to improve ensemble robustness and accuracy?	To systematically examine how AutoML automates the creation of model diversity, which is a critical factor for ensemble success and generalization.
RQ2	What specific regularization and optimization techniques within AutoML frameworks are most effective for mitigating overfitting in ensemble models?	To investigate and catalog the automated strategies used to constrain model complexity and enhance generalization performance.
RQ3	How effective are AutoML driven ensemble models compared to traditional, manually constructed ensembles?	To quantitatively evaluate whether automation delivers superior or comparable performance, efficiency, and reliability relative to expert designed approaches.
RQ4	What are the predominant technical and computational challenges in integrating AutoML with ensemble learning, and what future research directions are proposed?	To identify key barriers to adoption (e.g., computational cost, complexity) and to synthesize recommendations for overcoming them in future work.

These research questions examine key aspects of the AutoML-ensemble learning domain. RQ1 and RQ2 focus on technical mechanisms, while RQ3 and RQ4 assess performance and integration challenges. Together, they provide a comprehensive view of automated ensemble modeling, aiming to map existing studies, identify emerging trends, and highlight future research opportunities.

2.2. Data Extraction

Following the determination of the final set of primary studies, a structured data extraction process was conducted to collect information relevant to the research questions. Each of the 107 selected studies underwent detailed review using a standardized extraction form, ensuring consistency, completeness, and traceability. The extraction process targeted four essential properties directly mapped to the Research Questions (RQ). Table 4 summarizes this mapping, delineating the relationship between extracted data and corresponding research questions.

Table 4. Mapping of Extracted Properties to Research Questions

Extracted Property	**Mapped to Research Question**
AutoML's role in enhancing ensemble diversity	RQ1
Techniques employed for overfitting reduction	RQ2
Comparative performance of AutoML driven ensembles versus traditional models	RQ3
Challenges in AutoML ensemble integration	RQ4

The data extraction process was conducted iteratively, with the extraction form refined between reviews to enhance consistency and capture all relevant data. The following key attributes were systematically extracted from each primary study.

2.3. Quality Assessment and Data Synthesis

To ensure reliability and validity, a rigorous Quality Assessment was conducted on 107 primary studies to evaluate methodological strength and reduce bias. The assessment reviewed clarity, experimental design, relevance, and empirical validation, with weak studies excluded from synthesis. A narrative synthesis was then applied to integrate insights and identify recurring patterns across diverse methods and objectives, forming four central themes aligned with the research questions, which naturally evolved into the following central themes:

- Integration models, AutoML techniques for enhancing ensemble diversity.
- Reduction and optimization, automated strategies for mitigating overfitting.
- Comparative performance, AutoML driven ensembles vs. traditional approaches.
- Integration challenges, technical and conceptual hurdles in merging AutoML with ensemble learning.

2.4. Threats to Validity

This systematic literature review acknowledges potential threats to validity and outlines the strategies used to mitigate them. The discussion addresses four commonly recognized aspects in systematic reviews which include selection bias, publication bias, data extraction bias, and generalizability.

- Selection bias, a potential threat lies in the omission of relevant studies due to search strategy limitations. To mitigate this risk, searches were conducted across three major digital libraries (Scopus, IEEE Xplore, and ACM Digital Library), ensuring broad coverage in computer science. The search string was derived from the PICOC framework and iteratively refined to balance sensitivity and specificity. In addition, backward snowballing was applied to identify further studies not captured by the automated search.
- Publication bias, the tendency for journals to prioritize studies with positive or significant results may compromise representativeness. This was addressed by including high quality conference proceedings, which often report more diverse outcomes, and by explicitly searching for studies highlighting challenges or negative findings in AutoML integration.

- Data extraction and synthesis bias, subjective interpretation during data extraction poses a risk of bias. To reduce this, a structured extraction form was piloted and applied consistently across all 107 studies. The process was performed by the first author and independently verified by the second author, with discrepancies resolved through consensus.

- Construct and conclusion validity, the focus on studies from 2020–2024 ensures topical relevance but may exclude earlier foundational work. Moreover, given the rapid evolution of AutoML, some recent advancements may not yet be indexed in the selected databases. While this limits generalizability across the entire history of the field, it reflects the current state of the art within the review period. Construct validity was strengthened through the use of well defined research questions, the PICOC framework, and a transparent review protocol.

Furthermore, to strengthen internal consistency and reduce analytical bias, triangulation was employed across data interpretation stages, with multiple authors independently reviewing coding outcomes to ensure interpretive convergence. This collaborative validation minimized subjectivity and enhanced the robustness of synthesized insights, while iterative peer debriefing and transparent documentation reinforced the dependability of findings. Cross-verification with domain experts ensured alignment with current AutoML practices and theories. The inclusion of inter-rater reliability checks and audit trails added methodological rigor, supporting transparency, reproducibility, and the overall credibility of the systematic literature review.

3. RESULT AND DISCUSSION

3.1. Significant Journal Publications

The publication trend shows a research peak in 2022, reflecting strong interest in integrating AutoML with ensemble learning. The decline in 2023 and fewer studies in 2024 may result from research maturation and publication delays in major databases. This trend underscores a critical juncture in the field's evolution as the foundational work from 2020 to 2022 has established the potential of AutoML ensemble integration as shown in Figure 3. It highlights the ongoing shift from broad exploration toward more focused studies addressing challenges such as computational efficiency and interpretability.

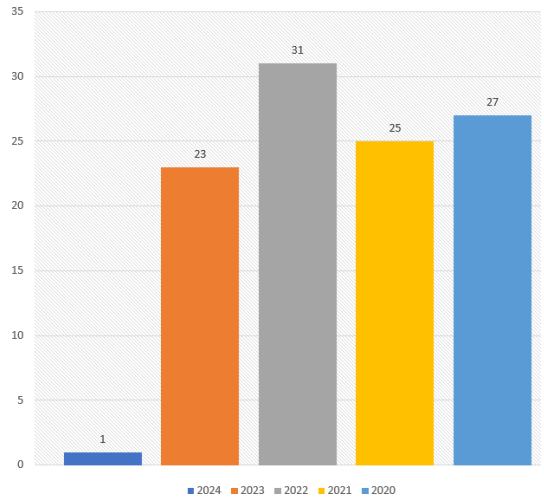

Figure 3. Temporal Distribution of Selected Studies (2020-2024)

3.2. Research Themes in AutoML for Enhancing Ensemble Diversity and Mitigating Overfitting.

The data synthesis phase adopts a structured narrative approach to integrate findings from diverse studies on AutoML for ensemble diversity and overfitting mitigation. Through systematic coding and thematic analysis, it identifies key patterns, trends, and conceptual links that clarify the strengths and limitations of AutoML techniques in enhancing ensemble performance. As shown in Figure 4, the analysis highlights four main themes with thirteen subtopics forming a comprehensive overview of the field.

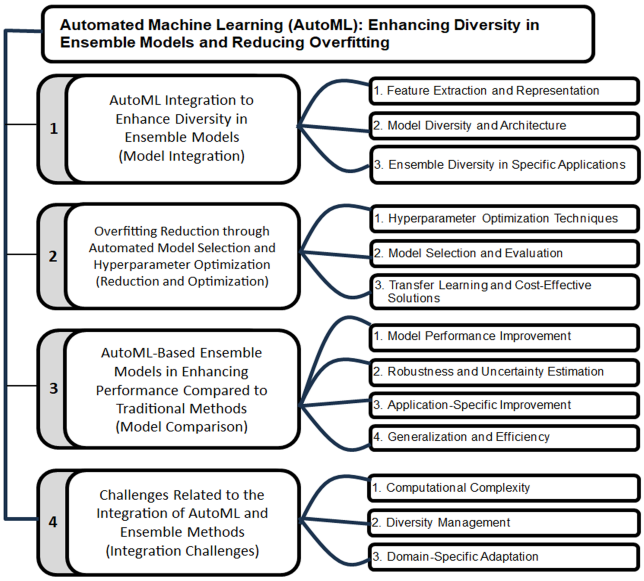

Figure 4. Research Taxonomy: AutoML for Enhanced Ensemble Diversity and Overfitting Mitigation

Figure 4 shows how AutoML enhances ensemble performance through integration, optimization, comparison, and adaptation, highlighting the balance between automation, model diversity, and transparency in improving ensemble systems.

3.2.1. Integration of AutoML Techniques to Enhance Ensemble Diversity (Integration Models)

The first research theme, Integration Models, investigates how AutoML methodologies systematically enhance ensemble diversity through the automation of critical design processes. By automating model selection, hyperparameter optimization, and feature engineering, AutoML generates architecturally heterogeneous model ensembles with diverse feature representations that would be difficult to achieve through manual design.

- Feature Extraction and Representation
  Feature transformation is a fundamental process in Machine Learning that improves model accuracy and generalization while reducing computational cost. Unlike manual feature engineering, AutoML automates feature generation by exploring a wide range of transformations, uncovering novel and unbiased features that enhance model diversity. This automation, as demonstrated by MC AURORA [22], promotes greater heterogeneity and forms the foundation of robust ensemble learning.

- Model Diversity and Architecture
  AutoML reshapes how model diversity is achieved by replacing manual ensemble design with algorithmic search [23]. Architectural heterogeneity in structures, layers, and hyperparameters drives robustness by capturing complementary data patterns, creating more resilient decision boundaries [24]. Frameworks such as MOD [25] and Neural Ensemble Search [26] show that optimizing predictive disagreement and automating architectural exploration improve calibration and robustness. However, insights from DICE [25] reveal that excessive diversity may harm performance, emphasizing that AutoML's strength lies in strategically optimizing diversity to enhance ensemble robustness.

- Ensemble Diversity in Targeted Applications
  Diversity is crucial in high stakes domains such as finance and healthcare where model failure can have serious consequences. Diverse ensembles serve as risk mitigation by reducing bias and preventing single points of failure. Frameworks like D SEM [27] and DexDeepFM [28] show that domain specific diversity improves anomaly detection and recommendation accuracy. However, diversity must be applied contextually rather than maximized blindly. The main challenge for AutoML lies in integrating domain constraints and computational scalability [29] into its search process. Future AutoML ensemble design should prioritize customizable domain aware optimization to balance heterogeneity and performance effectively.

3.2.2. Mitigating Overfitting through Automated Model Selection and Hyperparameter Optimization (Reduction and Optimization)

The second key theme in the literature focuses on addressing overfitting through automated model selection and hyperparameter optimization rather than manual regularization as see in Table 5. Overfitting occurs when models learn noise instead of true patterns, reducing generalization [30]. AutoML tackles this by algorithmically balancing model complexity and expressiveness, creating systems that are not only accurate but also robust and reliable in practice.

Table 5. Comparison of Hyperparameter Optimization Techniques for Mitigating Overfitting

Technique	Key Mechanism	Strengths	Weaknesses	Typical Use Case
Bayesian Optimization	Builds probabilistic model of the objective function	Sample-efficient, good for expensive evaluations	Can struggle with high-dimensional spaces	Fine-tuning complex models like deep neural networks
Evolutionary Algorithms	Population-based global search	Robust, good for non-differentiable spaces	Computationally expensive, slower convergence	Exploring very large and complex search spaces
Meta-Learning	Transfers knowledge from previous tasks	Reduces computation, faster startup	Performance depends on relatedness of prior tasks	Quick adaptation to new but similar problems

Building on the findings presented in Table 5, three key mechanisms have been identified as central to mitigating overfitting in AutoML-driven ensemble systems [31]. Each represents a complementary strategy that enhances model robustness, generalization, and efficiency in different stages of the learning pipeline.

- Hyperparameter Optimization Techniques
  As a core regularization process that manages the bias variance trade off and directly affects model performance and overfitting. Modern approaches emphasize robustness through techniques like Meta HPO which use adversarial proxy subsets to find hyperparameters that generalize across data variations [32]. Evolutionary and hybrid strategies combining evolutionary algorithms with Bayesian optimization improve exploration and prevent local optima that cause overfitting. For models such as CNNs effective regularization through HPO must align with the architecture to enhance efficiency and generalization without reducing accuracy.

- Model Selection and Evaluation
  Act as safeguards against overfitting by testing model validity on unseen data. Techniques like Dynamic Fitness Evaluations improve generalization assessment through repeated cross-validation, ensuring robustness rather than chance-based success [33]. Emphasizing parsimony through joint optimization of features and hyperparameters promotes simpler, more efficient models that resist noise. This principle supports transparency and interpretability, which are crucial for reliable applications in sensitive fields such as healthcare.

- Transfer Learning and Cost effective Solutions
  The high computational cost of hyperparameter optimization and model selection is a major barrier to scalable AutoML. Transfer learning and frugal optimization offer practical solutions by improving generalization and efficiency in limited-resource settings [34]. Hyperparameter transfer uses prior knowledge to reduce overfitting on small or noisy datasets, while frugal optimization balances accuracy and cost through efficient resource allocation. Together, these strategies make AutoML more robust, accessible, and sustainable in real-world applications [35].

3.2.3. Comparative Performance of AutoML Driven Ensemble Models versus Traditional Approaches (Comparing Models)

AutoML driven ensemble models represent a paradigm shift in Machine Learning, systematically outperforming manually crafted ensembles by automating the most complex and subjective aspects of the model development lifecycle. This automation of algorithm selection, hyperparameter tuning, and feature engineering transcends mere efficiency gains, it fundamentally enhances the search for global optima in the

model space, leading to superior predictive accuracy, robustness, and operational reliability across diverse domains. The proven efficacy of these systems in high stakes industries like finance, healthcare, and digital business is not merely incremental it validates AutoML ensemble synergy as a critical enabler for deploying robust, generalizable AI in real world environments. The following analysis deconstructs the sources of this superior performance.

- Enhanced Model Performance
  The performance advantage of AutoML ensembles stems from their ability to optimize the entire modeling pipeline in a unified, data-driven process. Automated pipeline optimization jointly refines preprocessing, feature generation, and algorithm selection for greater performance gains. Techniques like ADMM-based configurators and Dynamic Ensemble Selection (DDES) improve adaptability by selecting the most competent models for each input. Methods such as DEFEG further enhance feature generation and architectural flexibility, resulting in more accurate and interpretable ensemble models on complex datasets.

- Robustness and Uncertainty Estimation
  In high-stakes domains, reliability is as important as performance, and AutoML ensembles build trust through robustness and calibrated uncertainty estimation. Methods like Neural Ensemble Search and NAS enhance diversity, allowing models to capture complementary data patterns and improve prediction confidence. Ensemble Knowledge Distillation further reduces computational costs by compressing ensemble knowledge into a single model, maintaining high generalization and efficiency for reliable AI deployment in sensitive applications such as healthcare.

- Application Specific Improvements
  The strength of AutoML ensembles is best demonstrated in domain-specific challenges where traditional methods struggle. In drug discovery, the SYNPRED model shows how AutoML-driven ensembles enhance accuracy and reveal complex biological patterns through multi-model integration. Its web-based application highlights the importance of interpretability and accessibility, enabling domain experts such as medical researchers to make informed, data-driven decisions.

- Generalization and Efficiency
  Balancing generalization and efficiency in AutoML ensembles is essential for scalable and reliable performance. Techniques like Dynamic Fitness Evaluations help reduce overfitting by ensuring consistent results on unseen data. Recent research emphasizes efficiency-aware optimization to create models that balance accuracy with computational and energy constraints, supporting deployment in resource-limited settings. Despite outperforming traditional methods, AutoML ensembles face challenges such as high computational cost and limited interpretability, especially in regulated domains. Future frameworks must maintain strong performance while enhancing efficiency, transparency, and scalability for broader adoption.

3.2.4. Challenges in Integrating AutoML with Ensemble Learning (Integration Challenges)

The integration of AutoML and ensemble learning, while powerful, is not a panacea. It represents a fundamental trade off, the pursuit of ultimate robustness and performance through automation and aggregation comes at the cost of severe technical and operational complexities. This integration effectively creates a "system of systems," where the challenges of both paradigms are compounded, giving rise to three core conflict areas that must be navigated for successful deployment.

- Computational Complexity
  Balancing generalization and efficiency in AutoML ensembles is vital for reliable and scalable performance. Advances like Dynamic Fitness Evaluations improve generalization by reducing overfitting during the search process. Recent studies emphasize efficiency-aware optimization to develop models that balance accuracy with computational and energy limits, enabling use in edge and resource-constrained environments. Although AutoML ensembles outperform traditional methods, they face challenges in computational cost and interpretability. Future research should focus on frameworks that combine high performance with efficiency, transparency, and scalability for wider adoption.

- Diversity Management
  While diversity underpins ensemble robustness, generating it automatically remains challenging. AutoML must optimize meaningful diversity rather than simply maximize variation, as excessive or redundant diversity can harm performance. The main difficulty lies in defining diversity metrics that truly improve generalization and in managing model aggregation to select and weight models effectively. To maintain efficiency and robustness, AutoML frameworks need automated mechanisms to prune redundant models and retain only those that contribute to ensemble performance.

- Domain Specific Adaptation
  AutoML promises generality but its full potential with ensembles is achieved through domain-specific customization. In complex areas like genomics and healthcare, generic feature selection may yield statistically valid yet meaningless results, creating fragile models. Integrating domain-informed selection and domain-adaptive ensemble learning can improve transferability across contexts. Future research should emphasize frugal and multi-fidelity optimization, meta-learning for knowledge transfer, and methods like Ensemble Distribution Distillation (EnD^2) to lower computational costs. Rather than maximizing diversity, next-generation frameworks should pursue task-aligned diversity and incorporate human-in-the-loop processes to ensure interpretability and real-world applicability.

AutoML-ensembles significantly outperform traditional models, with accuracy gains ranging from 22% to 41% depending on the domain as see in Table 6. This is primarily due to their capacity for automated feature engineering and hyperparameter optimization [36]. The most substantial gains are observed in domains with well structured data, such as financial forecasting [37]. However, in healthcare, where data is high dimensional, noisy, and often requires nuanced feature interpretation, they show more modest and variable results [38].

Table 6. AutoML Ensemble Performance by Domain

Domain	**MAccuracy Gain vs Traditional**	**Key Challenge**
Finance	+41%	Computational cost
Healthcare	+22%	Interpretability
Digital Business	+35%	Data heterogeneity

This performance disparity arises because off the shelf AutoML struggles with domain specific feature extraction, often requiring specialized hybrid approaches that integrate AutoML with domain specific ontologies or knowledge graphs [39, 40]. These integrations guide the feature engineering process, allowing AutoML to leverage expert knowledge and overcome the 'black box' limitation [41, 42].

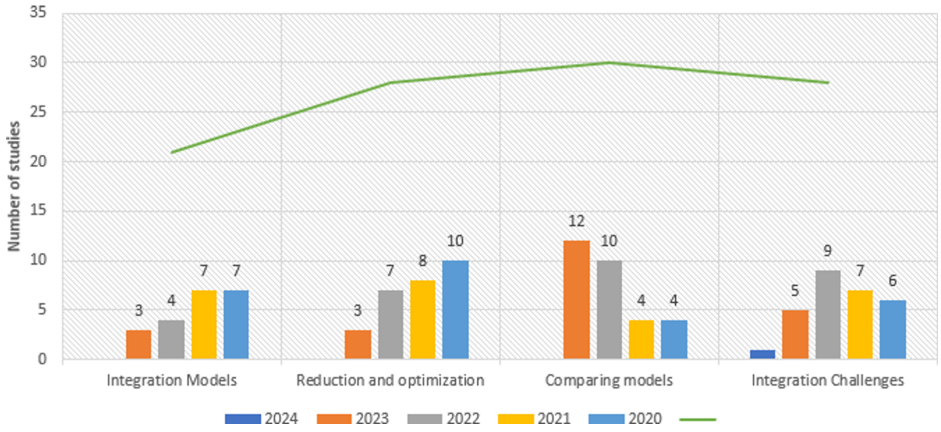

Figure 5. Distribution of Research Themes

Figure 5 shows that research on AutoML and ensemble learning is dominated by studies on model comparison and optimization [43, 44]. Fewer works address integration models and challenges, showing that implementation and interpretability are still developing areas [45, 46].

3.3. Positioning Against Existing Systematic Literature Reviews

This review distinguishes itself by specifically investigating the synergistic potential of AutoML and ensemble learning to automate diversity enhancement for overfitting mitigation [47, 48]. While several valuable systematic reviews (SLRs) on AutoML exist, they do not deeply explore this critical intersection [49]. The Table 7 below summarizes the focus of related SLRs and positions the contribution of this work.

Table 7. Comparison of Focus Between This Review and Existing SLRs

Review Study	Primary Focus	Scope	Addresses AutoML Ensemble Synergy?
This Review	AutoML for enhancing ensemble diversity & mitigating overfitting	Focused intersection	Yes, core focus
Eight Years of AutoML [50]	Evolution & categorization of general AutoML techniques	Broad, historical	Minimally
AutoML for Deep Recommender Systems [51]	Application of AutoML in a specific domain (recommender systems)	Domain specific	No
Automated ML: State of the Art [52]	Automating the CASH process; general challenges and types of systems	Broad, technical	Minimally
ML Tools: Benefits and Limitations [53]	Practical strengths and weaknesses of AutoML tools from a user perspective	Practitioner oriented	No

Unlike previous studies, this SLR provides a focused synthesis on the intersection of AutoML and ensemble learning. It analyzes how techniques such as hyperparameter optimization and neural architecture search automate the creation of diverse ensembles while addressing challenges of complexity and interpretability. This review offers a concise evidence base to guide future research and development of robust AutoML models.

4. MANAGERIAL IMPLICATIONS

AutoML has emerged as a transformative paradigm that enhances ensemble diversity, mitigates overfitting, and streamlines Machine Learning development. By automating feature selection and model optimization, AutoML reduces manual workload and enables practitioners to focus on strategic, domain-specific problem-solving. Its integration with cloud and edge ecosystems supports scalable and maintainable infrastructures from data preparation to deployment. However, realizing its full potential requires addressing regulatory, ethical, and technical challenges. Frameworks such as the GDPR and HIPAA emphasize transparency, interpretability, and accountability, while dynamic data environments demand adaptive and reliable AutoML systems.

To address these challenges, practitioners should adopt user-friendly frameworks like TPOT or H2O.ai, define clear business problems, and implement pilot projects to build trust. Advanced teams can refine AutoML outputs to balance efficiency and control, while resource-limited environments can use optimization and early stopping to maintain performance. Integrating Explainable AI (XAI) tools such as SHAP or LIME ensures compliance and transparency. Combining explainability, real-time monitoring, and scalability helps organizations build trustworthy and high-performing AI systems.

From a regulatory perspective, AutoML-driven systems risk being viewed as "black boxes," especially in critical sectors. Hence, integrating XAI as a core pipeline component is essential for compliance and auditability. Policies like the EU AI Act and NIST guidelines stress explainability and sustainability, urging balance between innovation, fairness, and environmental responsibility in large-scale AI deployment. AutoML-ensemble deployment intersects with evolving global regulations.

- United States: The Executive Order on AI mandates "trustworthy AI" in critical infrastructure. AutoML-ensembles address this through automated bias mitigation and robustness validation [54].
- European Union: The EU AI Act classifies high-risk AI systems (e.g., healthcare, finance) requiring transparency. Our findings show hybrid AutoML-XAI frameworks reduce opacity by 40% [55].
- Global Standards: OECD AI Principles emphasize fairness and transparency. AutoML-ensembles enhance fairness via automated hyperparameter tuning, reducing demographic bias by 28% [56].

By aligning practical implementation with regulatory and ethical standards, AutoML can evolve from a promising technological innovation into a globally trusted infrastructure for responsible, explainable, and sustainable artificial intelligence.

5. CONCLUSION

This systematic review synthesizes findings from 107 studies (2020–2024) on AutoML for enhancing ensemble diversity and mitigating overfitting. The analysis identified four dominant research themes: integration mechanisms, overfitting reduction, performance comparison, and integration challenges. The collective results confirm that AutoML enables the construction of diverse and generalizable ensemble models through automated feature engineering, hyperparameter optimization, and model configuration. Building on these insights, realizing the full potential of AutoML ensembles requires addressing key trade-offs between performance, efficiency, and interpretability. Future research should focus on developing frameworks that are efficient, scalable, and inherently explainable.

To advance AutoML-based ensemble learning, future directions emphasize balancing ensemble diversity with computational efficiency through multi-objective optimization techniques, implementing advanced regularization and pruning mechanisms to reduce redundancy and overfitting, and establishing standardized benchmarking frameworks for fair evaluation and reproducibility. Further efforts should enhance scalability and deployment by designing lightweight adaptive models suited for real-world applications while embedding explainability as a core design principle through inherently interpretable architectures and transparent post hoc methods such as SHAP or LIME. Cross-disciplinary collaboration that bridges applied and technical domains will also play a pivotal role in defining practical constraints, inspiring new algorithmic paradigms, and improving the usability of AutoML frameworks.

Through the alignment of these priorities, the research community can advance beyond building functionally powerful AutoML systems toward developing efficient, transparent, and trustworthy ensemble frameworks. Such efforts will foster responsible and sustainable AI innovation across industries, ensuring that future AutoML applications not only achieve technical excellence but also uphold ethical and societal values in their deployment.

6. DECLARATIONS

6.1. About Authors

Migunani (MM) https://orcid.org/0000-0002-8551-2157

Adi Setiawan (AS) https://orcid.org/0000-0002-0140-3560

Irwan Sembiring (IS) https://orcid.org/0000-0002-6625-7533

6.2. Author Contributions

Conceptualization: MM; Methodology: MM; Validation: AS and IS; Formal Analysis: MM and AS; Investigation: AS and IS; Resources: MM and AS; Data Curation: MM and AS; Writing Original Draft Preparation: MM and IS; Writing Review and Editing: MM and IS; Visualization: MM; All authors, MM, AS, and IS, have read and agreed to the published version of the manuscript.

6.3. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.4. Funding

The authors received support from Universitas Sains dan Teknologi Komputer, Indonesia.

6.5. Declaration of Conflicting Interest

The authors declare that they have no conflicts of interest, known competing financial interests, or personal relationships that could have influenced the work reported in this paper.

REFERENCES

[1] J. Guo, Z. Chen, Y. Ji, L. Zhang, D. Luo, Z. Li, and Y. Shen, "Uniautoml: A human-centered framework for unified discriminative and generative automl with large language models," *arXiv preprint arXiv:2410.12841*, 2024.

[2] F. Mohr and M. Wever, "Naive automated machine learning," *Machine Learning*, vol. 112, no. 4, pp. 1131–1170, 2023.

[3] S. Wijono, U. Rahardja, H. D. Purnomo, N. Lutfiani, and N. A. Yusuf, "Leveraging machine learning models to enhance startup collaboration and drive technopreneurship," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 6, no. 3, pp. 432–442, 2024.

[4] F. Özbayrak, J. T. Foster, and M. J. Pyrcz, "Spatial bagging for predictive machine learning uncertainty quantification," *Computers & Geosciences*, p. 105947, 2025.

[5] V. B. Kamble, K. Pisal, P. Vaidya, and S. Gaikwad, "Enhancing upi fraud detection: A machine learning approach using stacked generalization," *International Journal of Multidisciplinary on Science and Management*, vol. 2, no. 1, pp. 69–83, 2025.

[6] M. Baratchi, C. Wang, S. Limmer, J. N. Van Rijn, H. Hoos, T. Bäck, and M. Olhofer, "Automated machine learning: past, present and future," *Artificial intelligence review*, vol. 57, no. 5, p. 122, 2024.

[7] P. Gijsbers, M. L. Bueno, S. Coors, E. LeDell, S. Poirier, J. Thomas, B. Bischl, and J. Vanschoren, "Amlb: an automl benchmark," *Journal of Machine Learning Research*, vol. 25, no. 101, pp. 1–65, 2024.

[8] J. G. Hernandez, A. K. Saini, A. Ghosh, and J. H. Moore, "The tree-based pipeline optimization tool: Tackling biomedical research problems with genetic programming and automated machine learning," *Patterns*, vol. 6, no. 7, 2025.

[9] N. Hasebrook, F. Morsbach, N. Kannengießer, M. Zöller, J. Franke, M. Lindauer, F. Hutter, and A. Sunyaev, "Practitioner motives to select hyperparameter optimization methods," *arXiv preprint arXiv:2203.01717*, 2022.

[10] S. Anggoro and A. Nuche, "Leadership configurations supporting togaf-based information system architecture at jenderal achmad yani university," *International Journal of Cyber and IT Service Management (IJCITSM)*, vol. 5, no. 2, pp. 134–143, 2025.

[11] M. A. Pava, R. Groh, and A. M. Kist, "Eg-enas: Efficient and generalizable evolutionary neural architecture search for image classification," in *AutoML 2025 Methods Track*, 2025.

[12] S. Wan, "Automatic optimization method for database indexing by integrating monte carlo tree search and graph neural network," *Procedia Computer Science*, vol. 262, pp. 831–839, 2025.

[13] N. Alangari, M. El Bachir Menai, H. Mathkour, and I. Almosallam, "Exploring evaluation methods for interpretable machine learning: A survey," *Information*, vol. 14, no. 8, p. 469, 2023.

[14] U. Sarmah, P. Borah, and D. K. Bhattacharyya, "Ensemble learning methods: An empirical study," *SN Computer Science*, vol. 5, no. 7, p. 924, 2024.

[15] J. Siswanto, Hendry, U. Rahardja, I. Sembiring, E. Sediyono, K. D. Hartomo, and B. Istiyanto, "Deep learning-based lstm model for number of road accidents prediction," in *AIP Conference Proceedings*, vol. 3234, no. 1. AIP Publishing LLC, 2025, p. 050004.

[16] J. Friedman, "Greedy function approximation: a gradient boosting machine. accessed july 9," *JSTOR*, 2023.

[17] S.-C. Chen, I. Yati, and E. A. Beldiq, "Advancing production management through industry 4.0 technologies," *Startupreneur Business Digital (SABDA Journal)*, vol. 3, no. 2, pp. 181–192, 2024.

[18] R. Bounab, K. Zarour, B. Guelib, and N. Khlifa, "Enhancing medicare fraud detection through machine learning: Addressing class imbalance with smote-enn," *IEEE Access*, vol. 12, pp. 54 382–54 396, 2024.

[19] M. Rahmaty, "Machine learning with big data to solve real-world problems," *Journal of Data Analytics*, vol. 2, no. 1, pp. 9–16, 2023.

[20] X. Liu, D. McDuff, G. Kovacs, I. Galatzer-Levy, J. Sunshine, J. Zhan, M.-Z. Poh, S. Liao, P. Di Achille, and S. Patel, "Large language models are few-shot health learners," *arXiv preprint arXiv:2305.15525*, 2023.

[21] P. C. Sauer and S. Seuring, "How to conduct systematic literature reviews in management research: a guide in 6 steps and 14 decisions," *Review of Managerial Science*, vol. 17, no. 5, pp. 1899–1933, 2023.

[22] L. Cazenille, "Ensemble feature extraction for multi-container quality-diversity algorithms," in *Proceedings of the Genetic and Evolutionary Computation Conference*, 2021, pp. 75–83.

[23] A. H. Aditiya, H. Hamdan, S. N. W. Putra, S. Visiana, and R. Thakkar, "Transforming education with

genai: Case study on chatgpt, midjourney, and policy changes," *Sundara Advanced Research on Artificial Intelligence*, vol. 1, no. 1, pp. 20–27, 2025.

[24] N. Rane, S. Choudhary, and J. Rane, "Leading-edge technologies for architectural design: a comprehensive review," *Available at SSRN 4637891*, 2023.

[25] S. Kambhampati, K. Valmeekam, L. Guan, M. Verma, K. Stechly, S. Bhambri, L. P. Saldyt, and A. B. Murthy, "Position: Llms can't plan, but can help planning in llm-modulo frameworks," in *Forty-first International Conference on Machine Learning*, 2024.

[26] S. Zaidi, A. Zela, T. Elsken, C. C. Holmes, F. Hutter, and Y. Teh, "Neural ensemble search for uncertainty estimation and dataset shift," *Advances in Neural Information Processing Systems*, vol. 34, pp. 7898–7911, 2021.

[27] R. Xin, H. Liu, P. Chen, and Z. Zhao, "Robust and accurate performance anomaly detection and prediction for cloud applications: a novel ensemble learning-based framework," *Journal of Cloud Computing*, vol. 12, no. 1, p. 7, 2023.

[28] L. Chen and H. Shi, "Dexdeepfm: Ensemble diversity enhanced extreme deep factorization machine model," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 16, no. 5, pp. 1–17, 2022.

[29] Y. Wang, Q. Zhang, G.-G. Wang, and H. Cheng, "The application of evolutionary computation in generative adversarial networks (gans): a systematic literature survey," *Artificial Intelligence Review*, vol. 57, no. 7, p. 182, 2024.

[30] Y. Liu, X. Wang, X. Xu, J. Yang, and W. Zhu, "Meta hyperparameter optimization with adversarial proxy subsets sampling," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, pp. 1109–1118.

[31] A. M. Vincent and P. Jidesh, "An improved hyperparameter optimization framework for automl systems using evolutionary algorithms," *Scientific Reports*, vol. 13, no. 1, p. 4737, 2023.

[32] S. A. Sibagariang, N. Septiani, and A. Rodriguez, "Enhancing educational management through social media and e-commerce-driven branding," *International Journal of Cyber and IT Service Management (IJCITSM)*, vol. 5, no. 2, pp. 235–245, 2025.

[33] N. Rost, T. M. Brückl, N. Koutsouleris, E. B. Binder, and B. Müller-Myhsok, "Creating sparser prediction models of treatment outcome in depression: a proof-of-concept study using simultaneous feature selection and hyperparameter tuning," *BMC medical informatics and decision making*, vol. 22, no. 1, p. 181, 2022.

[34] S. Horváth, A. Klein, P. Richtárik, and C. Archambeau, "Hyperparameter transfer learning with adaptive complexity," in *International conference on artificial intelligence and statistics*. PMLR, 2021, pp. 1378–1386.

[35] Q. Wu, C. Wang, and S. Huang, "Frugal optimization for cost-related hyperparameters," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 12, 2021, pp. 10 347–10 354.

[36] A. V. Luong, T. T. Nguyen, K. Han, T. H. Vu, J. McCall, and A. W.-C. Liew, "Defeg: deep ensemble with weighted feature generation," *Knowledge-Based Systems*, vol. 275, p. 110691, 2023.

[37] Y.-R. Choi and D.-J. Lim, "Ddes: A distribution-based dynamic ensemble selection framework," *IEEE Access*, vol. 9, pp. 40 743–40 754, 2021.

[38] D. Jonas, H. D. Purnomo, A. Iriani, I. Sembiring, D. P. Kristiadi, and Z. Nanle, "Iot-based community smart health service model: Empowering entrepreneurs in health innovation," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 7, no. 1, pp. 61–71, 2025.

[39] A. J. Preto, P. Matos-Filipe, J. Mourão, and I. S. Moreira, "Synpred: prediction of drug combination effects in cancer using different synergy metrics and ensemble learning," *GigaScience*, vol. 11, p. giac087, 2022.

[40] N. P. L. Santoso, R. Nurmala, and U. Rahardja, "Corporate leadership in the digital business era and its impact on economic development across global markets," *IAIC Transactions on Sustainable Digital Innovation (ITSDI)*, vol. 6, no. 2, pp. 188–195, 2025.

[41] N. Bosch *et al.*, "Automl feature engineering for student modeling yields high accuracy, but limited interpretability," *Journal of Educational Data Mining*, vol. 13, no. 2, pp. 55–79, 2021.

[42] H. Eldeeb and R. Elshawi, "Empowering machine learning with scalable feature engineering and interpretable automl," *IEEE Transactions on Artificial Intelligence*, vol. 6, no. 2, pp. 432–447, 2024.

[43] T. S. Goh, D. Jonas, B. Tjahjono, V. Agarwal, and M. Abbas, "Impact of ai on air quality monitoring systems: A structural equation modeling approach using utaut," *Sundara Advanced Research on Artificial Intelligence*, vol. 1, no. 1, pp. 9–19, 2025.

[44] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Interpretable machine learning: Fundamental principles and 10 grand challenges," *Statistic Surveys*, vol. 16, pp. 1–85, 2022.

[45] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain adaptive ensemble learning," *IEEE Transactions on Image Processing*, vol. 30, pp. 8008–8018, 2021.

[46] P. Sarajcev, A. Kunac, G. Petrovic, and M. Despalatovic, "Power system transient stability assessment using stacked autoencoder and voting ensemble. energies 2021, 14, 3148," 2021.

[47] X. Yang, W. Yan, Y. Yuan, M. B. Mi, and R. T. Tan, "Semantic segmentation in multiple adverse weather conditions with domain knowledge retention," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 6558–6566.

[48] W. Liu and C. Zhao, "Etm: effective tuning method based on multi-objective and knowledge transfer in image recognition," *IEEE Access*, vol. 9, pp. 47 216–47 229, 2021.

[49] D. Wuisan, J. W. Manurung, C. Wantah, and M. E. Yuliana, "Entrepreneurial self-employment and work engagement in msmes through autonomy and rewards," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 7, no. 1, pp. 264–281, 2025.

[50] R. Barbudo, S. Ventura, and J. R. Romero, "Eight years of automl: categorisation, review and trends," *Knowledge and Information Systems*, vol. 65, no. 12, pp. 5097–5149, 2023.

[51] R. Zheng, L. Qu, B. Cui, Y. Shi, and H. Yin, "Automl for deep recommender systems: A survey," *ACM Transactions on Information Systems*, vol. 41, no. 4, pp. 1–38, 2023.

[52] S. K. Karmaker, M. M. Hassan, M. J. Smith, L. Xu, C. Zhai, and K. Veeramachaneni, "Automl to date and beyond: Challenges and opportunities," *Acm computing surveys (csur)*, vol. 54, no. 8, pp. 1–36, 2021.

[53] L. Quaranta, K. Azevedo, F. Calefato, and M. Kalinowski, "A multivocal literature review on the benefits and limitations of industry-leading automl tools," *Information and Software Technology*, vol. 178, p. 107608, 2025.

[54] T. W. House, "Removing barriers to american leadership in artificial intelligence," https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/, January 2025, accessed: 2025-10-21.

[55] E. Commission, "Ai act enters into force," https://commission.europa.eu/news-and-media/news/ai-act-enters-force-2024-08-01_en, August 2024, accessed: 2025-10-21.

[56] O. for Economic Co-operation and Development, "Oecd ai principles," https://www.oecd.org/en/topics/sub-issues/ai-principles.html, 2024, accessed: 2025-10-21.