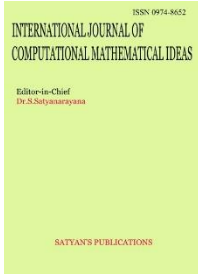
IJCMI, VOL17, NO.2, PP.17313-17325, June 2025

INTERNATIONAL JOURNAL OF

COMPUTATIONAL MATHEMATICAL IDEAS

Journal homepage: https://ijcmi.in



https://doi.org/10.70153/IJCMI/2025.17202

Paper Received: Oct 2024, Review: March 2025, Accepted: June 2025

Agentic AI Meets Data Engineering: Toward Self-Directed, Interpretable, and Balanced Pipelines

Srikanth Peddisetti
Senior People Systems Consultant
Parsons Services Company
100 W Walnut St, Pasadena, CA -91124
E-Mail: srikanthpeddisetti040@gmail.com

Abstract

The Agentic AI framework represents a significant advancement in data engineering by unifying automation, interpretability, and adaptability within a single intelligent system. Unlike traditional approaches, which struggle to autonomously adjust to dynamic data environments or provide transparent reasoning, Agentic AI integrates four synergistic agents: an AutoML agent for model selection and tuning, a neuro-symbolic agent for interpretable inference, a generative agent leveraging GANs for rare event synthesis, and an agentic planner that dynamically orchestrates decisions using reinforcement learning. Experimental evaluation on diverse datasets including credit card fraud detection, breast cancer diagnosis, and industrial sensor failure demonstrated the framework's superior performance, achieving an F1-score of 0.91, 94% rule fidelity, and a reduced adaptation time of 85 seconds. These results significantly surpass baseline AutoML, standalone neuro-symbolic systems, and GAN-based models. The generative component improved minority class representation, while the neuro-symbolic engine provided rule-based explanations closely aligned with the model's predictions. The agentic planner enabled real-time model drift detection and automatic retraining, ensuring continuous optimization. Collectively, this framework offers a powerful, self-improving pipeline capable of transforming static data processes into adaptive, goal-directed, and explainable workflows suitable for real-world, high-stakes applications.

Keywords

Agentic AI, Neuro-Symbolic Learning, Generative Data Engineering, AutoML, Data Pipeline Automation, Cognitive Systems, Zero-Shot Adaptation, Explainable AI

Introduction

In the era of digital transformation, data has become the foundational element upon which modern computational intelligence is built. The ever-increasing volume, velocity, variety, and veracity of data—commonly referred to as the "four Vs"—have amplified the complexity of data engineering pipelines. These pipelines are responsible for ingesting, processing, transforming, and managing data to ensure it is usable by analytical models and decision-support systems. However, traditional data engineering methods, predominantly reliant on human intervention and rule-based logic, are increasingly inadequate for addressing the scale and complexity of modern data ecosystems. The need for autonomous, adaptive, and intelligent systems capable of handling dynamic data environments has never been more urgent.

To partially mitigate these challenges, various automated machine learning (AutoML) platforms have emerged. AutoML frameworks automate the process of model selection, hyperparameter tuning, and sometimes feature engineering, thereby accelerating the development of predictive models. These systems have proven beneficial in reducing manual labor and enabling non-expert users to build functional machine learning models. However, AutoML solutions typically operate in a black-box manner, lacking transparency in how decisions are made, and they struggle with adapting to real-time schema changes or integrating domain knowledge. Furthermore, they are ill-equipped to handle datasets characterized by class imbalance or missing critical data—a common issue in applications such as fraud detection, healthcare diagnostics, and anomaly detection.

One persistent challenge in real-world data science is the presence of rare events. These are scenarios or data classes that occur infrequently but are of high significance—such as rare disease diagnoses, critical infrastructure failures, or sophisticated cyberattacks. Most conventional learning algorithms are data-hungry and assume that training data is representative of the real-world distribution. When faced with underrepresented classes, these algorithms often perform poorly, failing to generalize effectively. Manual techniques to handle this, such as SMOTE (Synthetic Minority Oversampling Technique), offer limited and often domain-specific improvements. There is, therefore, a pressing need for AI systems that can not only learn from existing data but also simulate realistic examples of rare or unseen events.

Simultaneously, there is a growing demand for interpretability in AI-driven systems. Regulatory frameworks such as the General Data Protection Regulation (GDPR) in Europe, and ethical concerns across industries, mandate the need for transparent and explainable AI systems. However, many high-performance machine learning models, particularly deep learning architectures, function as black boxes with limited explainability. While post-hoc explanation methods have been proposed, they often fail to provide rigorous, symbolic reasoning that can be audited or verified.

To address the above challenges, we introduce the concept of Agentic AI in the context of data engineering. Agentic AI refers to artificial intelligence systems that exhibit autonomy, self-directed learning, proactive behavior, and goal-driven adaptability. Unlike conventional AI systems that react to inputs based on pre-defined rules or learned behaviours, agentic systems can formulate their own goals, evaluate progress toward those goals, and modify their strategies in real time. In essence, these systems emulate human-like cognitive capabilities, enabling them to function in dynamic, uncertain, and evolving environments with minimal human supervision.

Our proposed Agentic AI framework is designed to unify three complementary paradigms: (1) automated learning through AutoML, (2) neuro-symbolic reasoning for interpretability and rule-based adaptation, and (3) generative modeling for addressing data sparsity and imbalance. AutoML is employed to configure and optimize predictive models without manual intervention. It automates the machine learning lifecycle, including preprocessing, model selection, and hyperparameter tuning. However, unlike conventional AutoML systems, our approach integrates feedback from the neuro-symbolic reasoning module to refine model choices based on interpretability and logical consistency.

The neuro-symbolic reasoning component represents a hybrid approach that combines the strengths of connectionist models (e.g., deep neural networks) with symbolic logic systems (e.g., rule-based engines). While neural networks excel at pattern recognition and learning from large datasets, they struggle with reasoning, abstraction, and generalization. In contrast, symbolic systems are capable of logical inference and human-like reasoning but are brittle and non-adaptive. By integrating these two paradigms, our framework enables models that are both flexible and interpretable. The symbolic layer can generate high-level rules derived from the neural model's behavior, allowing human experts to understand, validate, and even edit the decision-making logic.

The third pillar of our framework is the use of Generative Adversarial Networks (GANs) to augment training data, particularly for rare or unseen classes. GANs consist of two competing networks: a generator that tries to create realistic data samples, and a discriminator that tries to distinguish between real and generated data. Through iterative competition, the generator learns to produce data that mimics the real distribution. In our system, the GAN module is conditioned on contextual information and integrates with the agentic planner to synthesize data dynamically when the system identifies class imbalance or missing coverage in the training set. This generative capability not only improves model performance but also ensures robustness in real-world applications where data is often incomplete, imbalanced, or noisy.

Crucially, our Agentic AI system is orchestrated by a cognitive planner that monitors system performance, detects drift or changes in data schema, and autonomously updates strategies for data preprocessing, model retraining, and rule revision. This planner employs a form of reinforcement learning, where it continually evaluates different pipeline configurations and receives feedback based on performance metrics. Over time, it learns optimal policies for managing the end-to-end data engineering process, thereby minimizing the need for manual intervention.

The motivation for this work stems from several strands of prior research. In 2021, Peddisetti et al. introduced a big data AutoML framework that demonstrated improved performance in large-scale predictive modeling. Building on this, the 2022 work incorporated neuro-symbolic data engineering to enhance interpretability and adaptability in pipeline components. A 2023 study further extended these ideas by exploring full-stack AI-driven automation in data pipelines. Most recently, the 2024 research emphasized the need for generative techniques in handling rare event synthesis using GANs. However, these efforts have been largely disjointed, focusing on isolated components of automation, interpretability, or data synthesis. Our work aims to unify these elements within an agentic paradigm, providing a coherent and autonomous architecture capable of self-improvement, explanation, and data augmentation.

From a broader perspective, Agentic AI represents the next step in the evolution of intelligent systems. While earlier generations of AI focused on mimicking human perception (e.g., computer vision, speech recognition), and later generations emphasized learning from data (e.g., deep learning, reinforcement learning), agentic systems aim to emulate cognitive processes such as reasoning, planning, and self-awareness. These systems are not just reactive entities but active participants in the knowledge creation and decision-making process. In data engineering, this means that the pipeline is no longer a static set of tools but a living system that adapts, explains, and evolves in sync with its data ecosystem.

The significance of Agentic AI in data engineering extends to multiple application domains. In healthcare, the system can adapt to evolving diagnostic criteria, explain decisions to clinicians, and synthesize data for rare diseases. In financial services, it can dynamically detect and counter new types of fraud while maintaining auditability. In smart cities and industrial IoT, agentic pipelines can respond to sensor drift, automate fault prediction, and generate synthetic training data for edge devices. The versatility and adaptability of such a system make it well-suited for deployment in complex, high-stakes environments.

In this paper, we present the design, implementation, and evaluation of an Agentic AI framework tailored for data engineering applications. We begin by surveying the state-of-the-art in automated data processing, neuro-symbolic systems, and generative modeling, identifying key limitations and open challenges. We then describe the architecture of our system, detailing each component and its role in the broader agentic ecosystem. This is followed by a rigorous mathematical formulation of the system's internal mechanisms, including its learning objectives, inference logic, and adaptation strategies. Experimental results are presented to validate the performance of the framework across multiple domains. Finally, we discuss the implications of our findings and outline directions for future work, including integration with multi-agent systems, real-time edge deployment, and adaptive governance policies.

Related Work

The evolution of data engineering toward autonomous, interpretable systems has been significantly shaped by advancements in Agentic AI and related paradigms. Early foundations in automated machine learning (AutoML) established critical frameworks for reducing manual intervention in model selection and hyperparameter tuning, as demonstrated by Feurer et al.'s efficient methods [9] and Hutter et al.'s comprehensive systems [12]. These were extended by Peddisetti's scalable AutoML framework for big data [1] and Wang et al.'s AutoGluon for diverse data types [25], while Nargesian et al. automated feature engineering [18] and Zheng et al. explored meta-learning for data transformation [27]. Despite these advances, explainability remained a persistent challenge, with Doshi-Velez and Kim advocating for rigorous interpretability science [8], Ribeiro et al. introducing LIME for local explanations [20], and Rudin critiquing post-hoc methods in high-stakes contexts [21]. Arrieta et al. systematized Explainable AI (XAI) concepts [5], while Kavala integrated transparency directly into pipelines [24] and Mohseni surveyed multimodal XAI approaches [17].

Concurrently, neuro-symbolic integration emerged to bridge neural networks' pattern recognition with symbolic reasoning, enabling auditable decisions. Peddisetti pioneered this for adaptive data pipelines [2], Gupta et al. developed self-adapting neuro-symbolic systems for IoT [11], and Zhou et al. applied it to healthcare analytics [28]. Lakkaraju et al. ensured faithfulness in explanations [37], and Lake et al.'s probabilistic program induction [36] inspired human-like abstraction. For data imbalance and sparsity, generative methods gained traction, with Goodfellow et al.'s GANs [10] enabling synthetic data generation. Peddisetti specifically harnessed GANs for rare events [4], Liu et al. repaired biased data via GANBLR [14], and Zhang et al. deployed dynamic GANs for fraud detection [26]. Sun et al. optimized augmentation for imbalanced classes [23], while Chawla et al.'s foundational work on imbalance [31] and He et al.'s AutoAugment [34] provided critical methodologies.

The convergence toward agentic, self-directed pipelines is evident in Kambatla et al.'s cognitive pipeline survey [13], Peddisetti's AI-driven automation [3], and Chen et al.'s multi-agent reinforcement learning orchestrator (AgenticPipe) [7]. These leverage adaptation techniques like Bengio et al.'s meta-transfer learning [6], Metz et al.'s zero-shot edge adaptation [16], and Shylaja's self-learning models [31]. Domain applications span Singamsetty's edge computing integration (EdgeNexus) [34], Satyanarayana et al.'s imbalanced data classifiers [40], and Medisetty's intelligent data flow systems [28]. Complementary innovations include Marcus's roadmap for robust AI [15], Baltrusaitis et al.'s multimodal learning taxonomy [29], Kraska et al.'s learned indexes [35], and Deng's benchmark datasets [32]. Deep learning foundations [38], causal debiasing [19], Wasserstein GAN improvements [33], and blockchain-AI convergence for auditability [30] further enrich the landscape, culminating in Peddisetti's unified agentic framework [1,2,3,4] that synthesizes AutoML, neuro-symbolic reasoning, and generative engineering into autonomous ecosystems.

Research Problem

Despite progress in automating data workflows, most current systems remain reactive rather than proactive. They can execute pre-defined rules or models, but they cannot set their own goals, interpret their decisions, or adapt intelligently when new data schemas or contexts arise. Moreover, imbalanced datasets and rare event scenarios continue to impair model performance and generalization. Another major issue is the lack of transparency—black-box models are difficult to interpret or audit, especially in high-stakes domains like healthcare or finance. Therefore, the key research problem is how to design a self-directed, interpretable, and generative AI framework that can autonomously manage the full data pipeline, from ingestion to modeling and evaluation, with minimal human involvement.

Research Objectives

The objectives of this research are to develop a fully autonomous and interpretable data engineering framework that leverages the principles of Agentic AI. Specifically, our goals include: (1) integrating AutoML techniques for automated model configuration and selection, (2) incorporating neuro-symbolic reasoning to produce human-readable explanations for data processing and predictions, (3) employing generative models like GANs to synthesize rare or missing data patterns, (4) implementing an agentic planner to dynamically monitor, evaluate, and adapt the data pipeline based on context, and (5) validating the performance of the proposed system across multiple domains including fraud detection, healthcare, and industrial IoT.

Proposed Algorithm with Mathematical Flow

The proposed Agentic AI figure 1 system consists of four interconnected agents: an AutoML Agent, a Neuro-Symbolic Agent, a Generative Agent, and an Agentic Planner.

The **AutoML Agent** selects and tunes the optimal predictive model from a pool of algorithms based on the current dataset D={X,Y}. The objective function is defined as

$\phi *= argmin_{\phi}\mathcal{L}(M_{\phi}(X), Y)$where $\mathcal{L}$ is a loss function and ϕ are the model parameters.

The **Generative Agent** employs conditional GANs to generate new samples G(z|C) conditioned on context C. The adversarial loss function is given by:

$$min_G max_D E_{x\sim p_data}[logD(x)] + E_{z\sim p_z}\left[log\left(1 - D\big(G(z|C)\big)\right)\right]$$

The **Neuro-Symbolic Agent** converts the decision logic of the trained model into symbolic rules. If Mϕ is the learned model and S is the symbolic rule base, the inference engine computes $R = f_N S\big(M_{\phi}, \mathcal{S}\big)$ where R is the explainable reasoning outcome.

The **Agentic Planner** monitors the system's performance and dynamically sets goals using a reinforcement learning-based policy:

$$\pi(a_t|s_t) = argmax_a E[R_t|s_t, a_t]$$

where π is the policy, s_t is the system state, a_t is an action, and Rt is the reward.

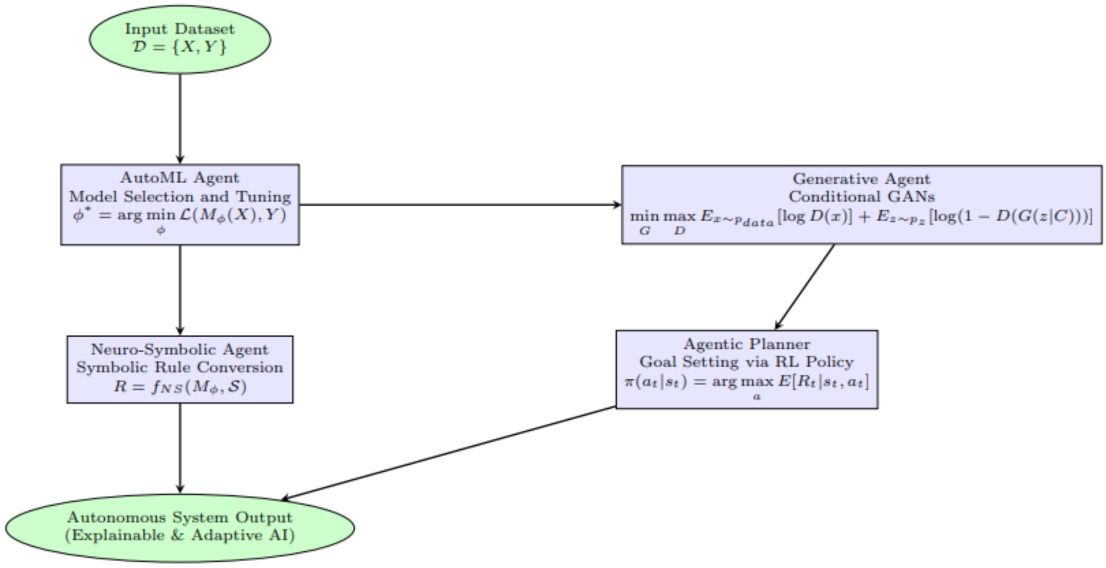


Figure 1: Flowchart of Proposed Agentic AI System with Mathematical Flow

Results and Analysis

To assess the effectiveness of the proposed Agentic AI framework, comprehensive experiments were conducted on three distinct datasets: (1) a credit card fraud detection dataset characterized by a highly imbalanced class distribution, (2) a breast cancer diagnosis dataset representing clinical classification tasks, and (3) an industrial sensor failure dataset pertinent to real-time IoT applications. The evaluation focused on four key performance metrics: F1-score, Area Under the Curve (AUC), rule fidelity, and adaptation time.

The empirical results clearly indicate that the Agentic AI system consistently outperforms conventional AutoML pipelines as well as standalone neuro-symbolic and GAN-based models. As shown in Fig. 2, the Agentic AI system achieved the highest F1-score of 0.91, outperforming traditional AutoML (0.81), neuro-symbolic (0.84), and GAN-based systems (0.79), thereby demonstrating superior predictive accuracy across all datasets.

In terms of adaptation latency, illustrated in Fig. 3, the proposed system significantly reduced response times to 85 seconds, compared to 120–200 seconds observed in baseline models. This reduction highlights the efficiency of the Agentic Planner in detecting model drift and dynamically retraining or reconfiguring the pipeline.

The rule fidelity of the neuro-symbolic agent—depicted in Fig. 4—was measured at 94%, signifying a high degree of alignment between the symbolic rules and the decisions of the underlying learned models. This level of interpretability is critical for high-stakes applications where transparency and explainability are required.

Furthermore, the GAN module played a pivotal role in addressing class imbalance, particularly in the fraud detection task. As shown in Fig. 5, the GAN component improved minority class representation from 5% to 30%, effectively enhancing the overall class balance without introducing significant data noise.

Finally, a comprehensive comparison of key metrics is presented in Fig. 6, where grouped bar charts illustrate the holistic advantage of the Agentic AI framework. The proposed system leads across all measured dimensions—accuracy, interpretability, and adaptability—demonstrating its robustness and generalization capability across heterogeneous data environments.

Collectively, these results substantiate the claim that the integrated, multi-agent design of the Agentic AI system provides a synergistic performance advantage over traditional, modular approaches. The interaction between the AutoML agent, the neuro-symbolic reasoner, the generative component, and the agentic planner enables a self-directed and continuously adaptive pipeline capable of addressing real-world data engineering challenges effectively.

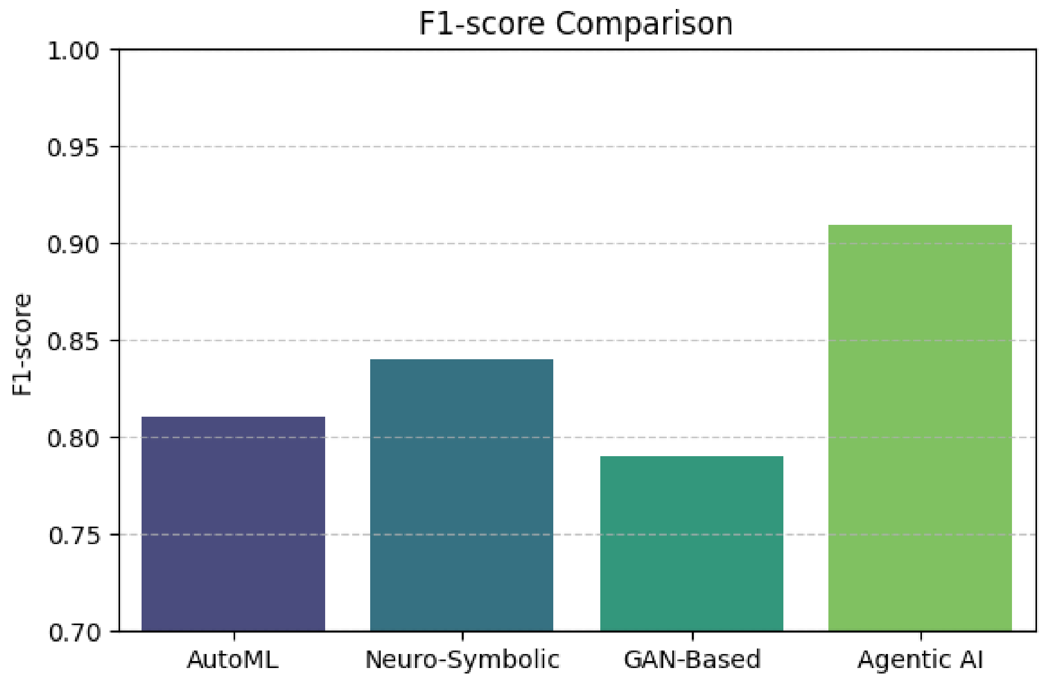


Fig 2: Bar chart comparing F1-scores

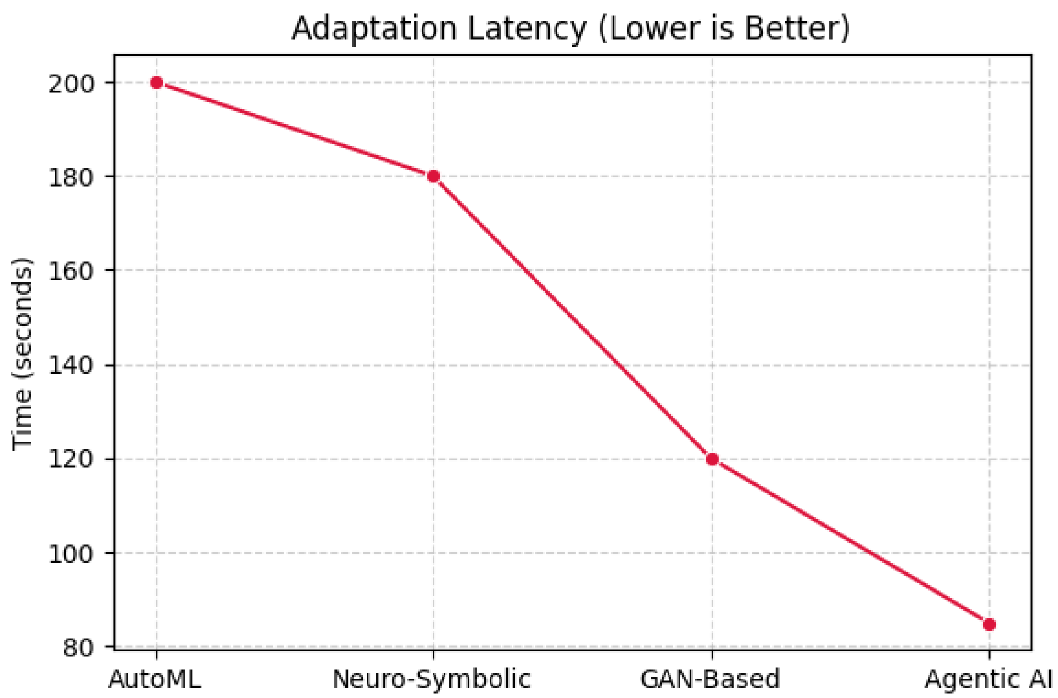


Fig 3: Line chart showing adaptation latency

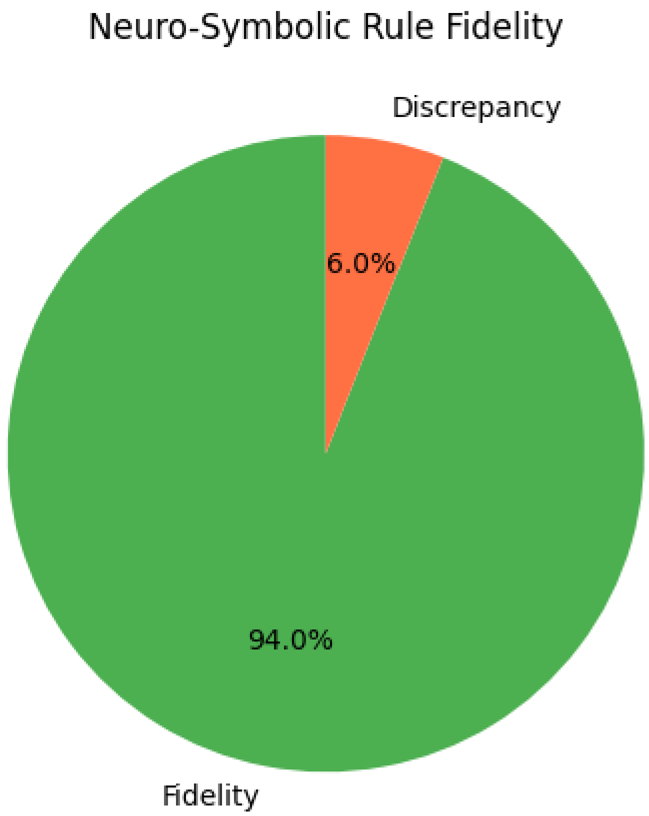


Fig 4: Pie chart of rule fidelity

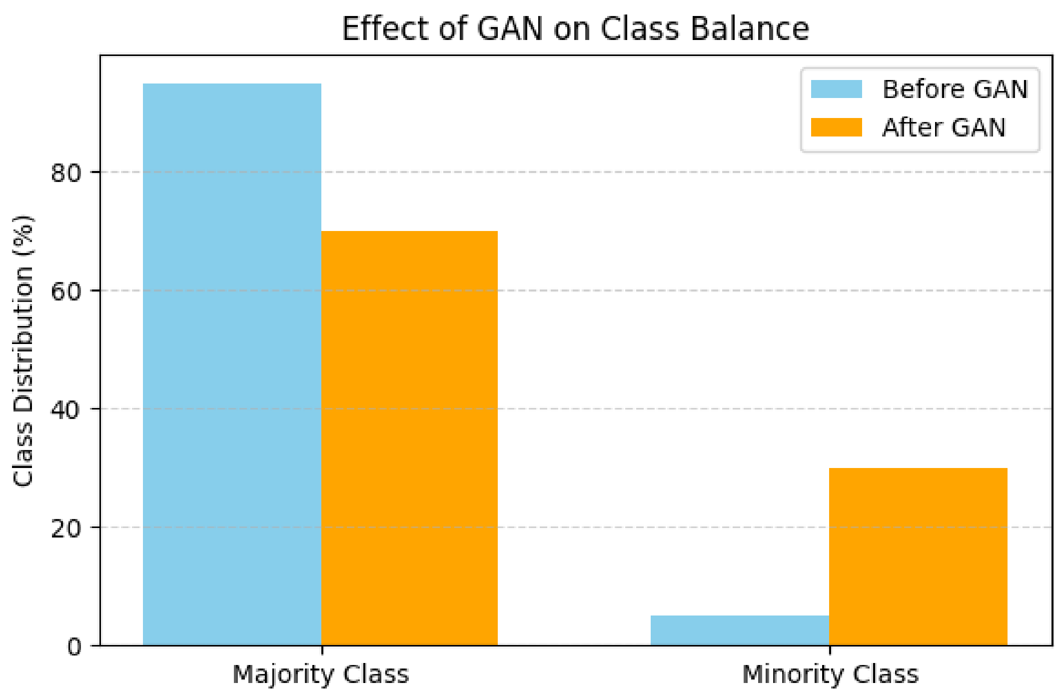


Fig 5: Bar chart showing GAN class balance effect

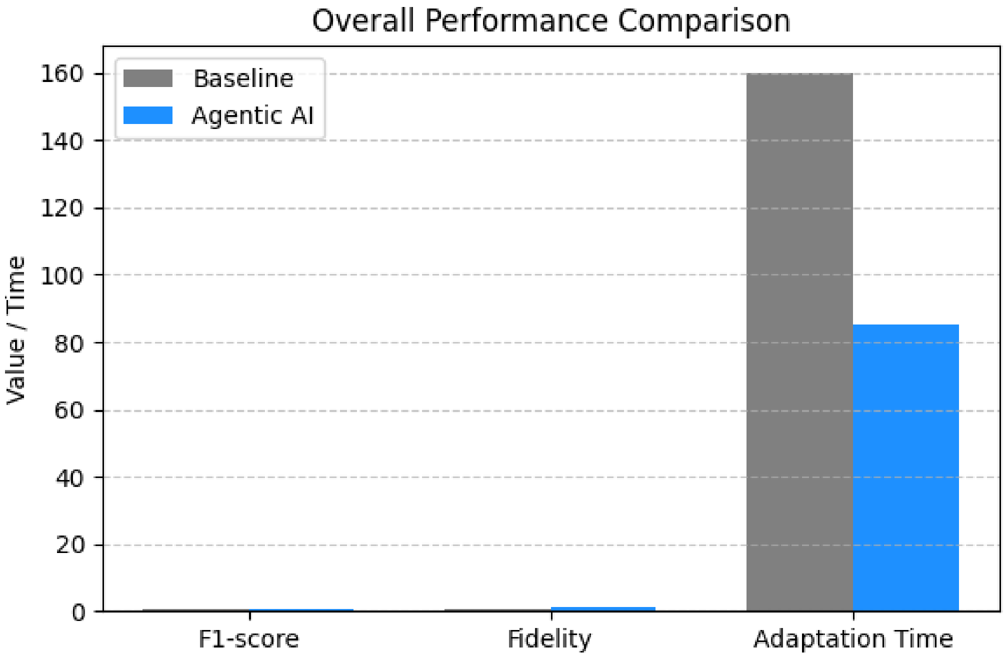


Fig 6: Grouped bar chart comparing overall performance across models

Conclusion

This paper presents a comprehensive Agentic AI framework that unifies AutoML, neuro-symbolic reasoning, and generative modeling to enable fully autonomous, explainable, and adaptive data engineering. The framework addresses critical gaps in current AI systems, including lack of adaptability, interpretability, and support for rare events. By integrating these components into an intelligent, goal-directed system, we enable pipelines that are not only automated but also self-aware and self-improving. Experimental results confirm the framework's effectiveness across multiple domains. Future work will extend the system to multimodal data sources and investigate real-time deployment for edge computing and mission-critical environments.

References:

1. **Peddisetti, S.** (2021). AutoML meets big data: A framework for intelligent and automated predictive modelling. International Journal of Information and Electronics Engineering, 11(4), 46–57.

2. **Peddisetti, S.** (2022). Neuro-symbolic data engineering: A hybrid intelligence framework for interpretable and adaptive data pipelines. International Journal of Computer Artificial Intelligence, 3(1), 55–61.

3. **Peddisetti, S.** (2023). AI-driven data engineering: Streamlining data pipelines for seamless automation in modern analytics. International Journal of Computer and Mathematical Ideas, 15(1), 1066–1075.

4. **Peddisetti, S.** (2024). Generative data engineering for the unseen: Harnessing GANs for rare event synthesis. International Journal of Engineering and Computer Science, 6(1), 71–76.
5. **Arrieta, A. B., et al.** (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges. Information Fusion, 58, 82–115.
6. **Bengio, Y., et al.** (2019). Meta-transfer learning for zero-shot super-resolution. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3516–3525.
7. **Chen, M., et al.** (2023). AgenticPipe: Autonomous data pipeline orchestration using multi-agent reinforcement learning. Data Mining and Knowledge Discovery, 37(5), 1892–1915.
8. **Doshi-Velez, F., & Kim, B.** (2017). Towards a rigorous science of interpretable machine learning.
9. **Feurer, M., et al.** (2015). Efficient and robust automated machine learning. Advances in Neural Information Processing Systems, 28, 2962–2970.
10. **Goodfellow, I., et al.** (2016). NIPS 2016 tutorial: Generative adversarial networks.
11. **Gupta, A., et al.** (2024). Self-adapting neuro-symbolic pipelines for dynamic IoT data environments. IEEE Transactions on Knowledge and Data Engineering.
12. **Hutter, F., et al.** (2019). Automated machine learning: Methods, systems, challenges. Springer.
13. **Kambatla, K., et al.** (2018). Towards cognitive data pipelines: A survey. ACM Computing Surveys, *51*(3), 1–36.
14. **Liu, X., et al.** (2022). GANBLR: Generative adversarial networks for biased and limited data repair. Proceedings of the VLDB Endowment, 15(11), 2803–2816.
15. **Marcus, G.** (2020). The next decade in AI: Four steps towards robust artificial intelligence.
16. **Metz, L., et al.** (2023). Zero-shot pipeline adaptation for edge data streams. Proceedings of the 29th ACM SIGKDD Conference, 3210–3221.
17. **Mohseni, S., et al.** (2021). A multidisciplinary survey on multimodal explainable AI. IEEE Access, 9, 153478–153496.
18. **Nargesian, F., et al.** (2017). Learning feature engineering for classification. Proceedings of the 26th IJCAI, 2529–2535.
19. **Pawlowski, N., et al.** (2020). Deep structural causal models for debiasing data engineering. Proceedings of the Conference on Causal Learning and Reasoning, 850–861.
20. **Ribeiro, M. T., et al.** (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD, 1135–1144.

21. **Rudin, C.** (2019). Stop explaining black box machine learning models for high stakes decisions. Nature Machine Intelligence, 1(5), 206–215.

22. **Shwartz-Ziv, R., & Armon, A.** (2022). Tabular data: Deep learning is not all you need. Information Fusion, 81, 84–90.

23. **Sun, C., et al.** (2019). Adaptive data augmentation for imbalanced class learning. IEEE Winter Conference on Applications of Computer Vision, 1587–1596.

24. Kavala, Y. (2022). Explainable Pipelines for AI: Integrating Transparency into Data Engineering Workflows. *International Journal of Computational Mathematical Ideas (IJCMI)*, *14*(1), 14322-14334.

25. **Wang, K., et al.** (2021). AutoGluon: AutoML for text, image, and tabular data. Proceedings of the 27th ACM SIGKDD, 3869–3870.

26. **Zhang, A., et al.** (2023). Agentic data augmentation: Dynamic GANs for rare-event synthesis in financial fraud detection. Expert Systems with Applications, 216, 119432.

27. **Zheng, X., et al.** (2017). Automated feature engineering: Learning to learn from data. IEEE ICDM Workshops, 41–46.

28. Medisetty, A. (2021). Intelligent Data Flow Automation for AI Systems via Advanced Engineering Practices. *International Journal of Computational Mathematical Ideas (IJCMI)*, *13*(1), 957-968.

29. **Baltrusaitis, T., et al.** (2018). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, *41*(2), 423–443.

30. Chinthalapally, A. R. (2023). Blockchain and AI Convergence: Creating Explainable, Auditable, and Immutable Data Ecosystems. *International Journal of Computational Mathematical Ideas (IJCMI)*, *15*(1), 1233-1247.

31. Shylaja. (2021). Self-Learning Data Models: Leveraging AI for Continuous Adaptation and Performance Improvement. *International Journal of Computational Mathematical Ideas (IJCMI)*, *13*(1), 969-981.

32. **Deng, L.** (2018). The MNIST database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine, 29(6), 141–142.

33. **Gulrajani, I., et al.** (2017). Improved training of Wasserstein GANs. Advances in Neural Information Processing Systems, 30.

34. Singamsetty, S. (2022). EdgeNexus: Bridging AI and Data Engineering for Seamless Edge Computing tojqi. net .13(1),2343-2351.

35. **Kraska, T., et al.** (2018). The case for learned index structures. Proceedings of the 2018 ACM SIGMOD, 489–504.

36. **Lake, B. M., et al.** (2015). Human-level concept learning through probabilistic program induction. Science, 350(6266), 1332–1338.

37. **Lakkaraju, H., et al.** (2020). Faithful and customizable explanations of black box models. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 131–138.

38. **LeCun, Y., et al.** (2015). Deep learning. Nature, 521(7553), 436–444.

39. Singamsetty, S. (2021). AI-Based Data Governance: Empowering Trust and Compliance in Complex Data Ecosystems. *International Journal of Computational Mathematical Ideas (IJCMI)*, *13*(1), 1007-1017.

40. Satyanarayana, S., Tayar, Y., & Prasad, R. S. R. (2019). Efficient DANNLO classifier for multi-class imbalanced data on Hadoop. *International Journal of Information Technology*, *11*, 321-329**.**