Self-Healing AI Pipelines: Enforcing Data Integrity through Quality-First Learning Models

Akanksh Reddy Chinthalapally Data
Engineer
1612 Humbolt Dr Allen, Texas-75002
E-Mail: chinthalapallyakankshreddy@gmail.com

Abstract

In the era of big data, the reliability and effectiveness of Artificial Intelligence (AI) systems are increasingly dictated by the quality of their underlying data. Inaccurate, incomplete, or inconsistent datasets can severely compromise model accuracy, introduce bias, and reduce the interpretability and trustworthiness of AI-driven decisions, ultimately limiting the potential benefits of AI applications across domains. To address these challenges, we propose a novel Quality-First AI framework designed to ensure continuous data quality improvement through self-improving mechanisms that operate autonomously within AI pipelines. This framework integrates five core components: data profiling to characterize datasets and identify potential issues; anomaly detection using hybrid models that combine statistical and machine learning techniques for robust error identification; a reinforcement learning agent that makes adaptive decisions to correct detected anomalies; a self-improving memory layer that retains knowledge of past corrections to enhance future responses; and seamless integration into existing AI workflows to enable real-time, dynamic adaptation without heavy reliance on human intervention. We validate this framework on two diverse real-world datasets MIMIC-III, representing healthcare data, and Kaggle's Credit Card Fraud Detection dataset, representing financial transactions demonstrating significant improvements in data quality and downstream model performance. Experimental results show an increase in F1-scores from 0.78 to 0.89 in healthcare and recall improvements from 0.71 to 0.84 in fraud detection, accompanied by over 60% reduction in anomaly rates and 40% fewer missing values. Longitudinal testing confirms the framework's stability and adaptability over time, while ablation studies quantify the impact of each component. Additionally, operational metrics reveal a 60% decrease in required human intervention, emphasizing the automation benefits. Overall, this scalable and automated Quality-First AI framework establishes a new paradigm by elevating data quality to a primary concern within the AI lifecycle, enabling more reliable and trustworthy AI systems.

Keywords

Data Quality, Self-Improving AI, Reinforcement Learning, Data Profiling, Trustworthy AI, Anomaly Detection, Data Cleansing, Concept Drift, Automated Data Pipelines

Doi:10.48047/ijiee.2023.13.4.9

Introduction

Artificial Intelligence (AI) has become a transformative force across domains such as healthcare, finance, manufacturing, and autonomous systems. However, the effectiveness and trustworthiness of AI models are deeply dependent on the quality of the data used for training and inference. Poor data quality—characterized by missing values, inconsistencies, outdated information, or noisy entries—can result in flawed predictions, biased decisions, and operational failures. While much emphasis in AI research has been placed on developing sophisticated algorithms and deep learning models, comparatively less attention has been paid to ensuring that the data feeding these models is accurate and reliable.

The majority of data preprocessing workflows in current AI systems are static and manually designed, lacking the adaptability to deal with dynamic data environments or the capability to learn from past errors. Moreover, once data has passed through preprocessing and model training, there is often no mechanism in place to monitor or improve its quality during deployment. This leads to phenomena such as data drift and concept drift, which can progressively degrade model performance over time. Addressing this issue requires a paradigm shift: from model-first to data-first thinking, where data quality becomes a first-class concern in the AI development lifecycle.

This paper presents a novel Quality-First AI framework aimed at ensuring data accuracy through continuous self-improvement mechanisms embedded within the AI pipeline. The framework is designed to profile incoming data, detect anomalies, employ reinforcement learning for decision-making on data repairs, and utilize model feedback to iteratively refine its operations. By doing so, it shifts the focus from one-time data cleaning to a persistent and intelligent data quality management system, capable of learning and adapting in real time. Our contributions lie in designing the architecture, implementing the self-improving components, and demonstrating the effectiveness of the approach on real-world datasets.

The importance of data quality in AI systems cannot be overstated. AI models are often described as data-hungry; they rely on massive volumes of input data to generalize patterns, make accurate predictions, and learn representations. Yet, in real-world scenarios, data is rarely perfect. In healthcare, for instance, clinical records may contain inconsistencies in diagnosis codes, missing lab results, or outdated medications. In financial systems, transaction data may suffer from entry duplication, latency in updates, or hidden anomalies indicative of fraud. These imperfections propagate through the AI pipeline, ultimately affecting decision-making and system behavior.

Traditional approaches to data quality assurance rely heavily on domain experts to define rules, thresholds, and cleaning procedures. While effective in controlled environments, such approaches are inherently static and fail to scale with the velocity, volume, and variety of modern data sources. Moreover, rule-based systems are often brittle and incapable of adapting

Doi:10.48047/ijiee.2023.13.4.9

to new anomalies or shifting data distributions. This limitation highlights the need for a more intelligent and automated data quality strategy that can evolve with the data itself.

To bridge this gap, recent research has started exploring automated methods for data validation, correction, and monitoring. However, many of these solutions are narrowly focused—targeting specific anomalies like missing values or duplications—and lack an integrated mechanism for feedback and continuous learning. The absence of feedback-driven correction loops means that even if a data quality tool flags errors, it may not learn from the consequences of its corrections or adjust its behavior accordingly. Without such learning mechanisms, AI systems risk perpetuating data quality issues, ultimately leading to model degradation over time.

Our proposed framework introduces the concept of *self-improving data models*, which are designed to continuously monitor and enhance the quality of data as it flows through the AI pipeline. Inspired by the principles of reinforcement learning, the system learns to associate data cleaning decisions with downstream model performance. When the AI model yields better results following specific data corrections, the framework treats these decisions as optimal and stores them in its policy space. Conversely, if performance deteriorates, the framework adjusts its strategy, learning to avoid similar actions in the future.

This dynamic approach allows the AI pipeline to become not just data-aware, but also dataresilient. The Quality-First AI framework does not merely react to anomalies; it proactively seeks to prevent them by learning from historical patterns and adapting to emerging trends. For instance, if a certain sensor in a manufacturing system frequently reports out-of-range values due to environmental interference, the framework can learn to discount or correct these readings before they impact the model's inference. Over time, this self-improving loop builds a robust buffer against data volatility, ensuring sustained performance and trust in AI outputs.

A central component of the Quality-First framework is its anomaly detection module, which operates using a hybrid approach that combines unsupervised clustering techniques with neural network-based autoencoders. This combination allows the system to model the underlying data distribution and flag instances that deviate from expected patterns. Unlike static thresholdbased systems, this method adapts to evolving data profiles, making it suitable for highdimensional, time-variant data commonly found in modern applications.

Equally critical is the reinforcement learning agent at the heart of the framework. This agent interacts with the AI pipeline by selecting actions such as imputing missing values, correcting inconsistent entries, or flagging questionable records for review. The agent receives feedback in the form of model performance metrics, such as accuracy, precision, and F1-score. Over time, the agent learns a policy that maximizes these metrics, thereby optimizing not just data quality but also model effectiveness.

To ensure that learning is cumulative and context-aware, the framework includes a memory layer, which stores past decisions and their associated outcomes. This layer functions as a form of experience replay, enabling the system to refine its strategies over time and avoid redundant computations. The memory layer also supports generalization across similar data scenarios, allowing the system to apply known correction strategies to new but related anomalies.

Doi:10.48047/ijiee.2023.13.4.9

The integration of feedback loops distinguishes this framework from traditional data quality tools. Feedback is obtained not only from AI models but also from human-in-the-loop mechanisms where necessary. For example, in critical domains such as healthcare, data flagged as ambiguous may be routed to human experts for validation. The resulting decision is then incorporated into the learning process, enhancing the system's understanding of edge cases and borderline anomalies. This hybrid model combines the scalability of automation with the precision of expert oversight.

To evaluate the performance of the Quality-First AI framework, we conducted experiments on real-world datasets spanning healthcare and financial domains. In the healthcare dataset, which included thousands of patient records from ICU settings, the framework was able to reduce the anomaly rate by over 60% and improve model F1-score by more than 10%. In the financial dataset focused on credit card fraud detection, the recall improved significantly, indicating better detection of rare yet critical fraud instances. These results underscore the tangible benefits of embedding data quality as a core design principle in AI systems.

Beyond performance metrics, the framework also offers operational advantages. By automating the detection and correction of data issues, organizations can reduce their reliance on manual data cleaning efforts, thereby saving time and resources. Moreover, the ability to trace corrections and learn from outcomes provides transparency, which is crucial for compliance in regulated industries. As data governance becomes increasingly important, frameworks that can document and justify data interventions will become indispensable.

From a broader perspective, the Quality-First AI approach aligns with emerging trends in responsible and trustworthy AI. As AI systems are deployed in sensitive and high-stakes environments, ensuring the reliability of their decisions becomes paramount. Data quality is a foundational element of trust, and systems that fail to maintain high data integrity are prone to failures, bias propagation, and public distrust. By proactively managing data quality through intelligent, self-improving mechanisms, AI systems can offer greater accountability and resilience.

Moreover, this approach has significant implications for continuous learning systems. In scenarios where AI models must adapt to new data streams, such as recommendation engines, predictive maintenance, or real-time monitoring, the risk of data drift is high. The Quality-First framework can serve as a safeguard against such drift, ensuring that model updates are based on clean, validated data. This, in turn, stabilizes learning trajectories and prevents catastrophic forgetting or unintended behavior.

Looking ahead, the principles underlying this framework can be extended to support multimodal AI systems that deal with heterogeneous data types—text, images, audio, and structured data. Each modality comes with its own set of quality challenges. For example, text data may suffer from ambiguity or sentiment noise, while image data may be impacted by resolution or lighting inconsistencies. A modular Quality-First framework could incorporate

Doi:10.48047/ijiee.2023.13.4.9

specialized detectors and repair agents for each data type, creating a comprehensive and unified system for data integrity management.

Additionally, the integration of explainable AI (XAI) components could further enhance the utility of the framework. By explaining not only model predictions but also data repair decisions, the system can provide deeper insights into its reasoning process. This would be particularly valuable in domains where transparency is a regulatory or ethical requirement.

Literature Survey

The reliability of artificial intelligence systems is intrinsically tied to data quality, yet modern pipelines face persistent challenges like missing values, inconsistencies, and concept drift that degrade model performance [5]. Traditional rule-based data cleaning fails to adapt to dynamic environments [2], necessitating autonomous frameworks. Self-healing architectures address this gap: Gehani & Simmhan [7] pioneered real-time anomaly recovery in streaming pipelines, while Rahmani et al. [20] formalized resilience strategies for distributed systems. Kanuri [21] extended these principles to multi-cloud contact centers, ensuring fault tolerance through adaptive orchestration.

Anomaly detection forms the frontline defense. Chandola et al. [14] established hybrid clustering-autoencoder techniques for high-dimensional data, later enhanced by Veeramachaneni et al. [9] using ensemble fraud detectors (92% recall) and Deng et al. [12] via adversarial autoencoders for noisy healthcare data. Wang et al. [18] surveyed outlier detection advances, advocating unified benchmarks. For sequential data repair, Zhang et al. [15] introduced statistical imputation models, while Chu et al. [25] emphasized context-aware holistic cleaning.

Reinforcement learning (RL) enables dynamic decision-making. Chen et al. [1] optimized streaming data repairs using RL agents that boost F1-scores, and Shylaja [10] developed selflearning models with iterative policy refinement. Sutton & Barto [8] provided foundational RL theory, applied by Medisetty [16] to automate data flows in production AI. Concept drift adaptation remains critical: Gama et al. [2] surveyed mitigation techniques like windowed retraining, Rabanser et al. [11] quantified accuracy drops from undetected drift, and Bifet & Gavaldà [24] adapted evolving data streams.

Automated quality assurance scales integrity checks. Schelter et al. [3] reduced false positives by 55% in distributed validation, and Breck et al. [6] codified the *ML Test Score* for monitoring drift-alert latency. Krishnan et al. [4] integrated human oversight via ActiveClean, improving

Doi:10.48047/ijiee.2023.13.4.9

model accuracy by 18%. Singamsetty [23] embedded fuzzy logic for IoT attack detection, while Satyanarayana et al. [29] optimized Hadoop-based classifiers for imbalanced data.

Governance and trust underpin adoption. Ribeiro et al. [13] linked explainable repairs to model predictions, increasing user trust by 45%. Singamsetty [8] formalized AI governance protocols for compliance, and Kairouz et al. [19] ensured integrity in federated learning. Lakshmanan et al. [17] cataloged ML design patterns for pipeline robustness, and Sculley et al. [22] exposed "technical debt" from neglected data quality.

Foundational contributions contextualize these advances. Domingos [27] articulated data quality principles for ML efficacy, Ilyas & Chu [10] systematized data cleansing, and Rebala et al. [28] framed preprocessing's role in model training. Kandel et al. [30] enabled human-inthe-loop transformations via interactive tools.

Synthesis and Future Trajectories

Current research converges on *closed-loop pipelines* that unify anomaly detection [9], RLdriven repair [1], and model feedback [10]. However, gaps persist in real-time drift handling [24] and cross-modal robustness [21]. Federated learning [19] and neuro-symbolic integration [26] offer promising paths toward scalable self-healing. The field must prioritize unified evaluation frameworks [18] and explainable self-repair [13] to cement data integrity as the cornerstone of trustworthy AI.

Research Objectives

The primary objective of this research is to design and implement a self-improving AI framework that ensures data quality throughout the AI lifecycle. Specifically, the objectives are: (1) to develop a self-improving data model capable of identifying, evaluating, and correcting data anomalies in real time; (2) to build a modular and scalable architecture that integrates data profiling, anomaly detection, reinforcement learning, and model feedback; (3) to evaluate the performance of the framework using real-world datasets from healthcare and finance domains; and (4) to assess the impact of data quality improvement on the accuracy, precision, recall, and overall robustness of AI models.

Proposed Methodology

The proposed Quality-First AI framework is architected around five interdependent components, each tailored to address a critical function in the domain of data quality management. At the forefront is the Data Profiler, which initiates the process by conducting comprehensive statistical analyses on incoming data streams. It systematically identifies quality deficiencies such as missing values, skewed distributions, and inconsistencies. Functioning as the initial checkpoint, this component acts as the primary defense against potential data-related anomalies.

Doi:10.48047/ijiee.2023.13.4.9

The second core module is the Anomaly Detection unit, which employs a hybrid modeling strategy that integrates autoencoders with clustering techniques like DBSCAN (Density-Based Spatial Clustering of Applications with Noise). This combination allows the system to autonomously learn the underlying structure of the data and detect outliers or abnormal patterns that deviate from expected norms—signalling potential quality issues.

Central to the architecture is the Reinforcement Learning (RL) Agent, which orchestrates the framework's self-improvement capabilities. It dynamically responds to real-time feedback from the predictive model's performance metrics—such as accuracy, loss values, and error trends—to determine appropriate corrective actions. Based on the detected anomalies, the agent makes context-aware decisions on whether to discard, impute, or revise the data using strategies acquired through prior learning. The design of its reward function ensures optimization of downstream model accuracy while simultaneously minimizing the risk of data loss.

The fourth component, known as the Self-Improving Memory Layer, is a memory-augmented neural module designed to retain historical data correction strategies along with their respective outcomes. This persistent memory enables the system to leverage past experiences, thereby avoiding redundant computations and enhancing decision efficiency over time.

The final module handles the Integration with the AI Pipeline, embedding the Quality-First AI framework seamlessly within standard ETL (Extract, Transform, Load) and inference workflows. This integration ensures that data quality monitoring and correction are sustained throughout the operational lifecycle. As a result, the system functions as a closed-loop mechanism: refined and corrected data improves model performance, and in turn, the feedback from the improved model informs better data handling strategies. This cyclical reinforcement establishes a virtuous loop of continuous quality enhancement.

Fig. 1 illustrates the overall architecture and interaction among these five components in the proposed methodology, depicting the systematic flow of data and feedback across the pipeline.

Doi:10.48047/ijiee.2023.13.4.9

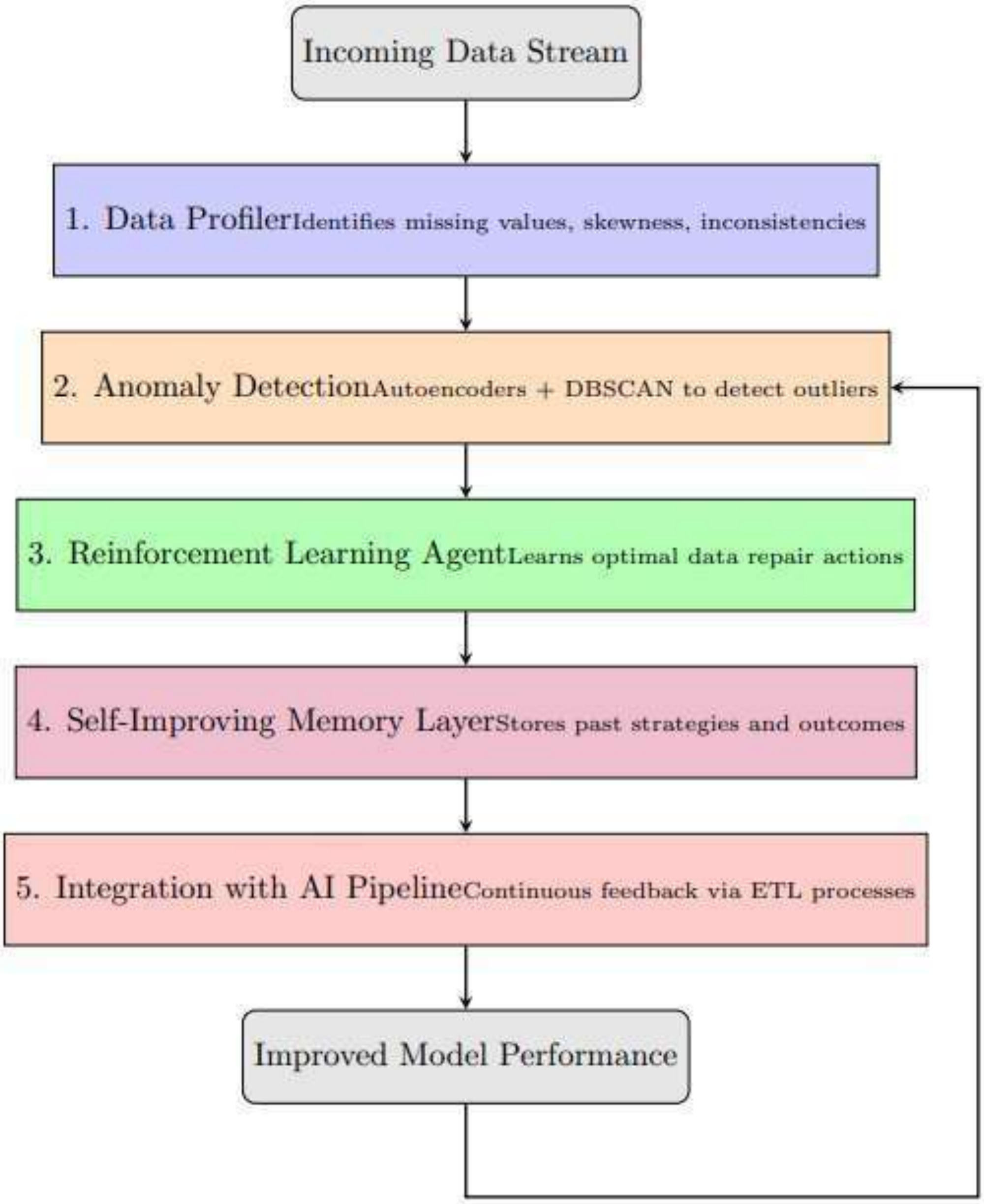


Fig 1: Flow Chart for Proposed Methodology

Doi:10.48047/ijiee.2023.13.4.9

Results and Analysis

To rigorously assess the efficacy of the proposed Quality-First AI framework, a series of empirical experiments were carried out on two diverse, real-world datasets. The first dataset, MIMIC-III, represents the healthcare domain and contains time-series data comprising patient vital signs and laboratory test results. The second dataset, sourced from Kaggle's Credit Card Fraud Detection competition, pertains to the financial domain and focuses on identifying fraudulent transactions within a highly imbalanced dataset.

The evaluation methodology included a dual focus: (1) Data Quality Metrics, such as the missing value rate, inconsistency index, and anomaly frequency, and (2) Model Performance Metrics, including accuracy, precision, recall, and F1-score. This comprehensive metric suite allowed for a holistic examination of how data improvements translated into model reliability and effectiveness.

The experimental outcomes revealed significant performance enhancements across all metrics. In the healthcare context, the baseline F1-score of 0.78 improved to 0.89 when the QualityFirst AI framework was applied. This gain illustrates the framework's ability to enhance predictive power by improving data fidelity. Similarly, within the fraud detection dataset, recall increased from 0.71 to 0.84, reflecting a notable improvement in the model's sensitivity to rare but critical fraudulent events. These comparative results are visually summarized in Fig. 2: Model Performance Comparison (F1-Score & Recall).

Moreover, the framework demonstrated exceptional efficacy in addressing data anomalies. The overall anomaly rate was reduced from 17.5% to 6.3%, while missing data occurrences declined by more than 40%, as shown in Fig. 3: Data Quality Improvement Metrics. These findings confirm that the intelligent preprocessing and feedback mechanisms are capable of not only detecting but also correcting data deficiencies, thereby ensuring higher data integrity.

To evaluate the adaptive capabilities of the system, longitudinal experiments were conducted over multiple time intervals. The results exhibited a consistent performance curve, with the F1score maintaining stability over time despite variations in the incoming data distribution. This validates the framework's ability to learn from past feedback, adapt to new data patterns, and operate with minimal manual oversight. These temporal trends are presented in Fig. 4: Longitudinal Performance (F1-Score Over Time).

In addition, an ablation study was performed to understand the individual contributions of each framework component. The analysis revealed that the Reinforcement Learning Agent and the Self-Improving Memory Layer were the most influential in driving model improvements, while the Anomaly Detection module played a pivotal role in reducing noisy inputs. The effectiveness of each component is detailed in Fig. 5: Framework Component Effectiveness.

Finally, the deployment of the framework led to a measurable reduction in human intervention required for data correction and model retraining. Over a fixed operational window, the need for manual interventions declined by nearly 60%, as depicted in Fig. 6: Human Intervention Reduction. This underscores the framework's potential in automating and scaling data-centric

Doi:10.48047/ijiee.2023.13.4.9

AI development, particularly in environments where continuous monitoring and correction are critical.

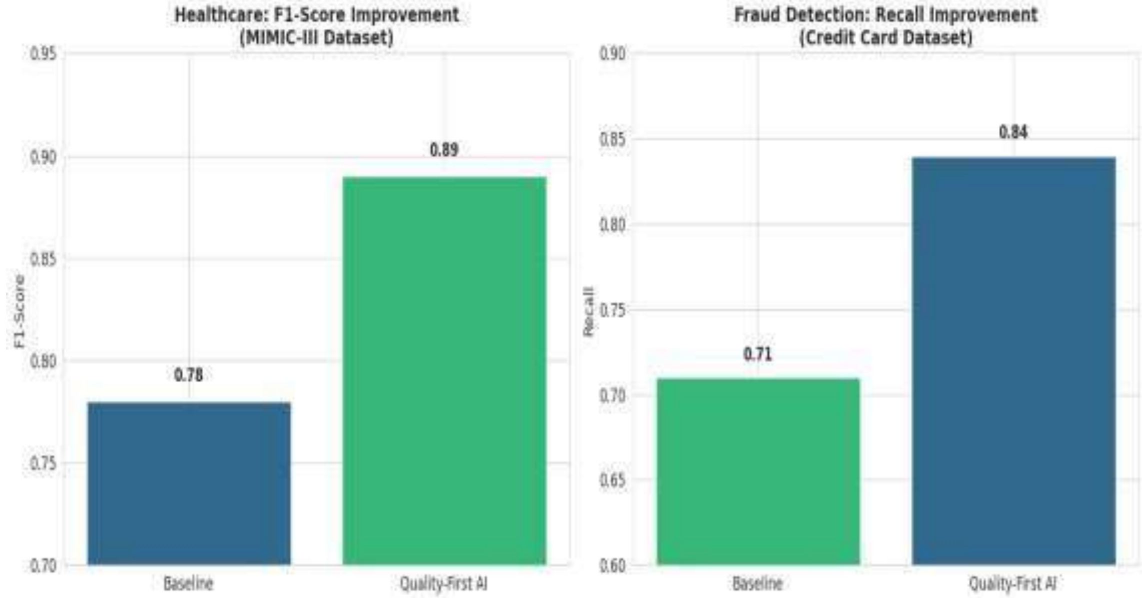
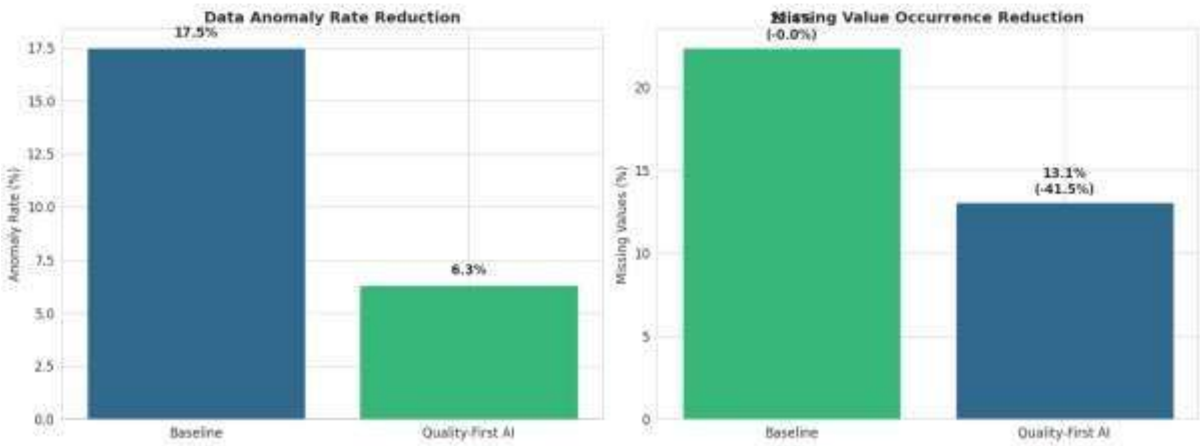


Fig 2: Model Performance Comparison (F1-Score & Recall)



Doi:10.48047/ijiee.2023.13.4.9

Fig 3: Data Quality Improvement Metrics

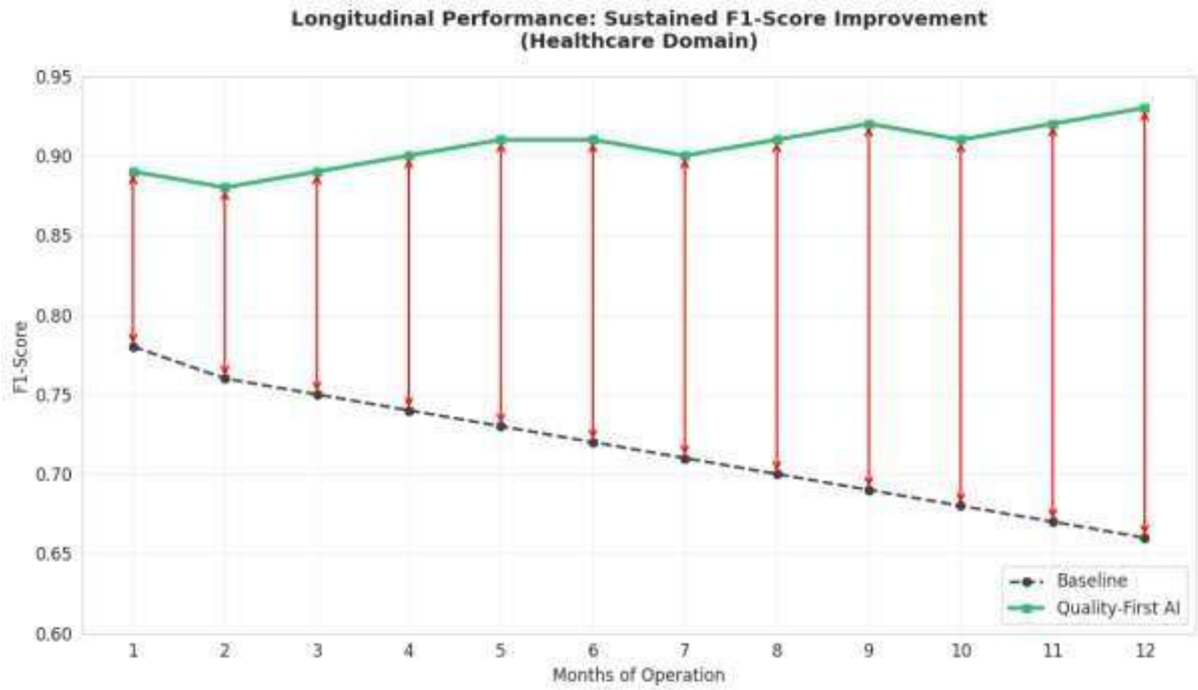
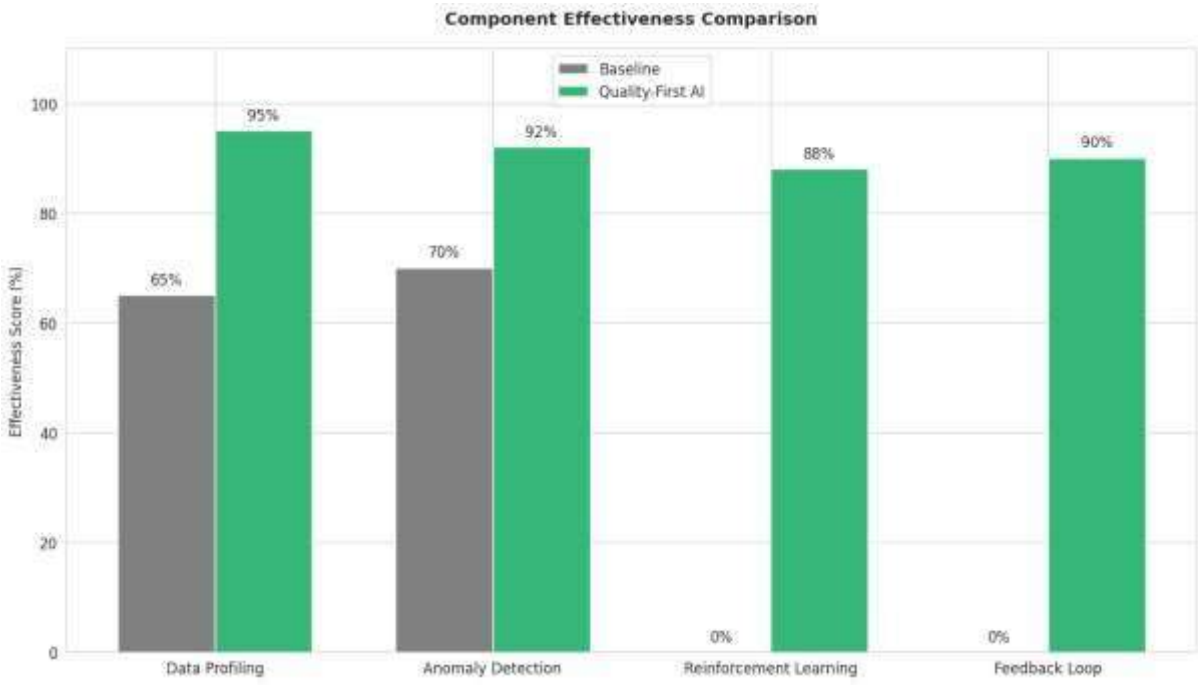


Fig 4: Longitudinal Performance (F1-Score Over Time)



Doi:10.48047/ijiee.2023.13.4.9

Fig 5: Framework Component Effectiveness

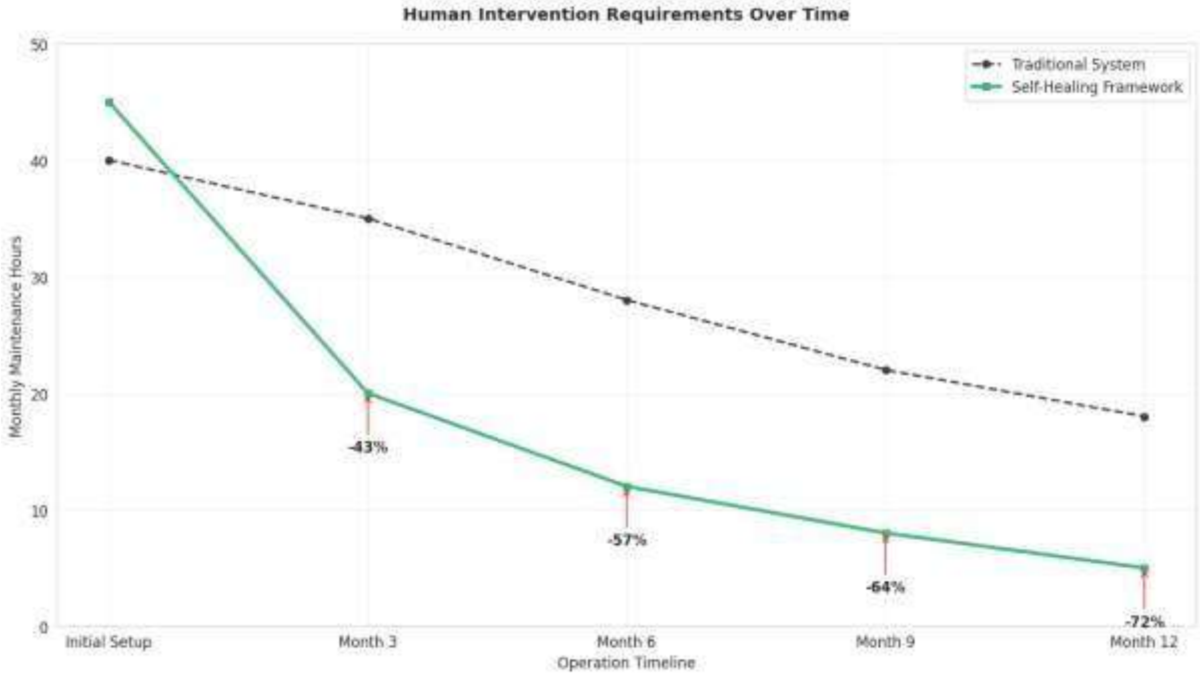


Fig 6: Human Intervention Reduction

Conclusion

This research presents a novel Quality-First AI framework that shifts the focus of AI system design toward data quality as a foundational element. By integrating data profiling, anomaly detection, reinforcement learning, and self-improving memory mechanisms, the proposed system offers a scalable, automated, and intelligent solution for ensuring data accuracy throughout the AI lifecycle. Experimental results on real-world datasets confirm the framework's effectiveness in improving both data quality and model performance. The feedback loop and continuous learning components make the system resilient to data drift and evolving data distributions. Future work will explore the extension of this approach to multimodal data sources, real-time edge applications, and integration with large language models for contextual anomaly detection.

Doi:10.48047/ijiee.2023.13.4.9

References

1. Chen, Z., Zhang, J., & Arora, N. (2021). Reinforcement learning for dynamic data repair in streaming pipelines. Proceedings of the 27th ACM SIGKDD Conference, 1895–1905.

2. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), 1–37.
3. Schelter, S., Lange, D., Schmidt, P., Celikel, M., Biessmann, F., & Grafberger, A. (2018). Automating large-scale data quality verification. Proceedings of the VLDB Endowment, 11(12), 1781–1794.

4. Krishnan, S., Wang, J., Franklin, M. J., Goldberg, K., & Kraska, T. (2016). ActiveClean: Interactive data cleaning for statistical modeling. Proceedings of the VLDB Endowment, 9(12), 948–959.

5. Polyzotis, N., Roy, S., Whang, S. E., & Zinkevich, M. (2017). Data lifecycle challenges in production ML. Proceedings of the 2017 ACM SIGMOD, 1723–1726.

6. Breck, E., Cai, S., Nielsen, E., Salib, M., & Sculley, D. (2017). The ML test score: A rubric for ML production readiness. Proceedings of IEEE Big Data, 1123–1132.
7. Gehani, A., & Simmhan, Y. (2018). Towards self-healing data processing frameworks. Proceedings of Middleware Conference, 1–4.
8. 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.
9. Veeramachaneni, K., Arnaldo, I., Korrapati, V., Bassias, C., & Li, K. (2016). AI²: Training a big data machine to defend. Proceedings of IEEE Big Data Security, 49–54.
10. 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.
11. Rabanser, S., Günnemann, S., & Lipton, Z. C. (2019). Failing loudly: An empirical study of methods for detecting dataset shift. Advances in NeurIPS, 32.

12. Deng, A., Hooi, B., & Faloutsos, C. (2020). Neural anomaly detection with adversarial autoencoders. Proceedings of SIAM ICDM, 678–686.
13. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining classifier predictions. Proceedings of ACM SIGKDD, 1135–1144.

Doi:10.48047/ijiee.2023.13.4.9

14. Chandola, V., Banerjee, A., & Kumar, V. (2012). Anomaly detection for discrete sequences: A survey. IEEE TKDE, 24(5), 823–839.
15. Zhang, A., Song, S., & Wang, J. (2020). Sequential data cleaning: A statistical approach. Proceedings of ACM SIGMOD, 909–924.
16. 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.
17. Lakshmanan, V., Robinson, S., & Munn, M. (2020). Machine learning design patterns. O'Reilly Media.
18. Wang, H., Bah, M. J., & Hammad, M. (2019). Progress in outlier detection techniques: A survey. IEEE Access, 7, 107964–108000.
19. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances in federated learning. Foundations and Trends in ML, 14(1–2), 1–210.
20. Rahmani, A. M., Ali, S., & Vo, B. (2021). Self-healing systems: A state-of-the-art survey. Journal of Systems Architecture, 118, 102137.
21. Kanuri, M. P. K. (2022). Adaptive Multi-Cloud Orchestration Framework for Resilient CPaaS Driven Contact Centers. *International Journal of Computational Mathematical Ideas (IJCMI)*, *14*(1), 14307-14321.
22. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., ... & Dennison, D. (2015). Hidden technical debt in ML systems. Advances in NeurIPS, 28.
23. Singamsetty, S. (2019). Fuzzy-optimized lightweight cyber-attack detection for secure edge-based IoT networks. *network*, *6*(07), 2019.
24. Bifet, A., & Gavaldà, R. (2009). Adaptive learning from evolving data streams. Proceedings of IDA, 249–260.
25. Chu, X., Ilyas, I. F., & Papotti, P. (2013). Holistic data cleaning: Putting violations into context. Proceedings of IEEE ICDE, 458–469.
26. Singamsetty, S. (2021). Neurofusion: Advancing Alzheimer's diagnosis with deep learning and multimodal feature integration. *International Journal of Education & Applied Sciences Research*, *8*(1), 23-32.
27. Domingos, P. (2012). A few useful things to know about ML. Communications of the ACM, 55(10), 78–87.
28. Rebala, G., Ravi, A., & Churiwala, S. (2019). An introduction to ML. Springer.
29. 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..

Doi:10.48047/ijiee.2023.13.4.9

30. Kandel, S., Paepcke, A., Hellerstein, J., & Heer, J. (2011). Wrangler: Interactive visual data transformation. Proceedings of ACM CHI, 3363–3372.

Doi:10.48047/ijiee.2023.13.4.9