International Journal of Research and Applied Innovations (IJRAI)

| ISSN: 2455-1864 | www.ijrai.org | editor@ijrai.org | A Bimonthly, Scholarly and Peer-Reviewed Journal |

||Volume 5, Issue 1, January – February 2022||

DOI:10.15662/IJRAI.2022.0501006

Applying Machine Learning for Automated Data Quality and Anomaly Detection in Enterprise Data Pipelines

Nagender Yamsani

Software Development Advisor, USA

ABSTRACT: Data quality failures including missing values, inconsistent representations, duplicate entities, and anomalous records continue to be a dominant barrier to trustworthy analytics and effective machine learning (ML) deployment, particularly as organizations scale across diverse, fast-moving data sources. Traditional rule-based validation and constraint checking, while effective in narrow domains, struggle to generalize in environments characterized by high volume, velocity, and schema heterogeneity, often requiring extensive manual maintenance and domain expertise. Recent advances in ML-based data management shift this paradigm by learning statistical, relational, and semantic patterns directly from data, enabling automated detection, diagnosis, and, in some cases, repair of quality defects. This article surveys these approaches through a structured lens, connecting foundational ideas in probabilistic modeling and anomaly detection with modern deep learning techniques and practical data-cleaning systems. By examining representative systems such as HoloClean and ActiveClean, we analyze architectural tradeoffs between accuracy, computational cost, and human-in-the-loop effort, as well as the balance between aggressive cleaning and error propagation risk. Empirical results across these systems demonstrate that ML-informed data quality pipelines can significantly improve anomaly detection accuracy, reduce manual labeling and correction effort, and produce measurable gains in downstream predictive performance, underscoring data quality as a first-class concern in end-to-end ML system design rather than a preprocessing afterthought.

KEYWORDS: Enterprise AI; Evidence Mapping; Advanced Analytics; Business Intelligence; Pharmaceutical AI; Manufacturing Intelligence; Responsible AI; Data Platforms

I. INTRODUCTION

High-quality data is a foundational requirement for trustworthy analytics and machine learning systems, yet real-world datasets are rarely clean or stable. Errors are routinely introduced during data collection, integration across heterogeneous sources, schema evolution, transformation pipelines, and long-term storage. Early work in data management rigorously characterized core data quality problems, including schema mismatches, integrity constraint violations, missing or inconsistent values, and semantic inconsistencies across systems. While these efforts established important theoretical foundations, remediation techniques were largely manual or rule-based, relying on hand-crafted constraints, validation rules, and domain-specific checks. Such approaches are brittle in the face of changing schemas and data distributions, and they scale poorly as data volume, velocity, and heterogeneity increase. Maintaining large rule sets requires deep domain expertise and continuous human intervention, making it difficult to keep pace with modern data pipelines. As organizations increasingly rely on automated analytics and downstream ML models, these limitations have become more pronounced. Poor data quality not only degrades model performance but can also introduce hidden biases and operational risk. Consequently, scalable and adaptive approaches to data quality have emerged as a critical research and engineering priority.

Machine learning offers an alternative paradigm by shifting the burden from explicit rule specification to statistical learning from data itself. Instead of encoding all validity constraints manually, ML models infer patterns of normality, correlation, and dependency directly from observed data distributions. This enables automated detection of outliers, anomalies, and violations even in the absence of labeled ground truth. Unsupervised and semi-supervised techniques, such as clustering, density estimation, and reconstruction-based models, are particularly valuable in domains where labeled errors are scarce or evolving. ML-based methods can capture complex, high-dimensional relationships that are difficult to express as deterministic rules, allowing them to generalize across diverse datasets. Importantly, these approaches can adapt as data distributions shift, reducing the maintenance burden associated with static validation

| ISSN: 2455-1864 | www.ijrai.org | editor@ijrai.org | A Bimonthly, Scholarly and Peer-Reviewed Journal |

||Volume 5, Issue 1, January – February 2022||

DOI:10.15662/IJRAI.2022.0501006

logic. By the early 2020s, ML-based anomaly detection had matured from isolated algorithms into integrated pipelines that combine detection, diagnosis, and repair with human-in-the-loop feedback. These systems increasingly close the loop by feeding corrected data back into models, improving both data quality and downstream learning outcomes over time.

This article surveys the algorithmic foundations underlying ML-driven data quality, including probabilistic modeling, representation learning, and anomaly detection techniques. It reviews influential systems that operationalize these ideas in practice, highlighting how they balance automation with human oversight and manage tradeoffs between accuracy, scalability, and interpretability. By synthesizing empirical results from published evaluations, the article illustrates how ML-informed data quality pipelines can significantly reduce manual effort while improving error detection rates. The analysis also examines downstream impacts, showing how improved data quality leads to measurable gains in predictive accuracy, robustness, and fairness of ML models. Beyond cataloging techniques, the article aims to distill design principles that guide effective system construction, such as incremental cleaning, uncertainty-aware repair, and tight integration with modeling workflows. These insights are intended to support both practitioners building production data pipelines and researchers exploring the next generation of intelligent data management systems.

II. BACKGROUND AND FOUNDATIONS OF DATA QUALITY AND ANOMALY DETECTION

Anomaly detection is concerned with identifying data points or patterns that deviate significantly from what is considered normal or expected behavior within a dataset. Early work in this area relied heavily on statistical and distance-based techniques, including z-score analysis, parametric distribution modeling, nearest-neighbor methods, and clustering-based outlier detection. These approaches provided a principled starting point for automated anomaly identification, particularly in low-dimensional and well-behaved datasets. Over time, researchers recognized that anomalies are not a monolithic concept, leading to a widely adopted categorization into point anomalies, contextual anomalies, and collective anomalies. Point anomalies refer to individual observations that are extreme relative to the global distribution, while contextual anomalies depend on specific conditions such as time, location, or surrounding attributes. Collective anomalies, in contrast, involve groups of data points whose joint behavior is anomalous even if individual points appear normal. This taxonomy emphasized the importance of context and feature interactions, foreshadowing the limitations of purely local or univariate detection techniques.

In parallel, the data management community developed a rich body of work on data quality, formalizing common error categories that undermine analytical reliability. These include missing values, duplicate records, constraint violations, referential integrity breaks, and semantic inconsistencies across integrated datasets. Traditionally, such issues were addressed through deterministic rules, integrity constraints, and manual curation processes. However, as datasets grew in size and complexity, it became evident that many data quality problems exhibit statistical signatures rather than explicit rule violations. For example, duplicated entities often appear as unusually dense clusters in feature space, while semantic inconsistencies can manifest as improbable attribute combinations. Recognizing these patterns reframed data quality errors as a form of anomaly when viewed through statistical, relational, or probabilistic lenses. This reframing created a conceptual bridge between anomaly detection and data quality research.

The convergence of these two lines of work enabled techniques originally developed for anomaly detection to be repurposed for data quality assessment and remediation. Density estimation methods could identify rare or unlikely records that correspond to errors, while clustering and graph-based techniques could surface duplicate or inconsistent entities. Context-aware anomaly detection proved particularly valuable for identifying conditional errors, such as values that are valid in isolation but implausible given related attributes or temporal context. This overlap also encouraged the development of hybrid systems that combine statistical signals with relational constraints and domain knowledge. As a result, anomaly detection evolved from a narrow task focused on rare events into a broader toolkit for maintaining data integrity. This synthesis laid the groundwork for modern ML-driven data quality systems that treat error detection, diagnosis, and repair as interconnected components of a unified pipeline.

III. MACHINE LEARNING TECHNIQUES FOR AUTOMATED ANOMALY DETECTION

3.1 Classical Machine Learning

Classical machine learning techniques form the backbone of many production anomaly detection systems due to their relative simplicity, robustness, and interpretability. Algorithms such as One-Class Support Vector Machines (OC-

| ISSN: 2455-1864 | www.ijrai.org | editor@ijrai.org | A Bimonthly, Scholarly and Peer-Reviewed Journal |

||Volume 5, Issue 1, January – February 2022||

DOI:10.15662/IJRAI.2022.0501006

SVM), Local Outlier Factor (LOF), and Isolation Forests were designed to operate effectively in settings where labeled anomalies are scarce or entirely absent. These methods learn a notion of "normality" from observed data and flag deviations based on geometric, density-based, or partitioning criteria. OC-SVMs, for example, learn a decision boundary that encloses the majority of the data, treating points outside this boundary as anomalies. LOF, in contrast, compares local density estimates to identify points that reside in sparse neighborhoods relative to their peers. Such approaches are particularly attractive in data quality and monitoring scenarios where ground truth labels are unavailable or expensive to obtain.

Among these techniques, Isolation Forests gained significant adoption for large-scale and high-dimensional datasets. Rather than relying on distance or density estimation, Isolation Forests isolate anomalies by recursively partitioning data using randomly selected features and split values. Anomalous points tend to be isolated in fewer partitions, resulting in shorter average path lengths. This design yields near-linear scalability and avoids many of the curse-of-dimensionality issues that affect distance-based methods. As a result, Isolation Forests are commonly used in large data pipelines, fraud detection, and operational monitoring, where computational efficiency is critical. Despite their strengths, classical methods often struggle with highly complex or evolving data distributions, motivating the exploration of more expressive models.

In practice, classical anomaly detection algorithms are frequently favored for their transparency and ease of integration. Their decision logic can often be explained in intuitive terms, which is valuable for debugging, governance, and regulatory contexts. Additionally, these models typically require fewer hyperparameters and less computational infrastructure than deep learning alternatives. However, their reliance on relatively simple assumptions about data geometry limits their ability to capture intricate nonlinear dependencies. As datasets grow more heterogeneous and dynamic, classical methods increasingly serve as strong baselines or complementary components rather than standalone solutions.

3.2 Deep Learning Methods

Deep learning significantly expanded the scope of anomaly detection by enabling models to learn rich, nonlinear representations of complex data distributions. Instead of relying on handcrafted features or simple distance metrics, deep models learn hierarchical abstractions directly from raw or minimally processed data. Autoencoders became one of the most widely used deep approaches for anomaly detection, operating on the principle that models trained to reconstruct normal data will exhibit higher reconstruction error on anomalous inputs. This paradigm proved effective across domains ranging from tabular data to images and sensor streams. Variants such as denoising and sparse autoencoders further improved robustness by encouraging models to learn meaningful latent structure rather than trivial identity mappings.

As sequential and temporal data became increasingly prevalent, recurrent and convolutional architectures were adapted for anomaly detection in time-series, logs, and event streams. Recurrent neural networks and long short-term memory models capture temporal dependencies, enabling detection of anomalies that arise from unexpected sequences or temporal patterns rather than isolated points. Convolutional architectures, including temporal convolutions, offer efficient modeling of local temporal structure and have been successfully applied to high-frequency monitoring data. These approaches support detection of contextual and collective anomalies that are difficult to capture with classical techniques. Their ability to incorporate context makes them particularly well suited for operational monitoring and data quality assessment in streaming environments.

By the late 2010s, the field had consolidated around broader taxonomies of deep anomaly detection methods. Variational autoencoders introduced probabilistic latent representations, enabling uncertainty-aware anomaly scoring, while generative adversarial networks learned to distinguish normal data distributions through adversarial training. Surveys published prior to 2020 systematically categorized these approaches and analyzed their strengths and limitations across data modalities. These deep methods proved especially influential for large-scale time-series and log-data monitoring, where complexity and nonlinearity are the norm. However, their increased expressiveness comes at the cost of higher computational demands, reduced interpretability, and greater sensitivity to training data quality. Consequently, deep learning methods are most effective when paired with careful evaluation, domain knowledge, and, in many cases, hybrid integration with classical techniques.

International Journal of Research and Applied Innovations (IJRAI)

| ISSN: 2455-1864 | www.ijrai.org | editor@ijrai.org | A Bimonthly, Scholarly and Peer-Reviewed Journal |

||Volume 5, Issue 1, January – February 2022||

DOI:10.15662/IJRAI.2022.0501006

IV. ML-DRIVEN DATA QUALITY AND AUTOMATED REPAIR SYSTEMS

While anomaly detection is essential for identifying problematic records, a complete data quality pipeline must also determine how errors should be repaired, prioritized, or escalated for human review. In practice, detection without repair merely shifts the burden downstream, leaving analysts and engineers to decide which issues matter and how to fix them. This realization motivated a class of systems that integrate machine learning directly into the data cleaning workflow, treating repair as a first-class problem rather than an afterthought. These systems aim to balance accuracy, scalability, and human effort by using probabilistic reasoning and learned signals to propose likely corrections. Instead of applying rigid rules, they model uncertainty explicitly, allowing the system to rank candidate repairs and surface confidence estimates.

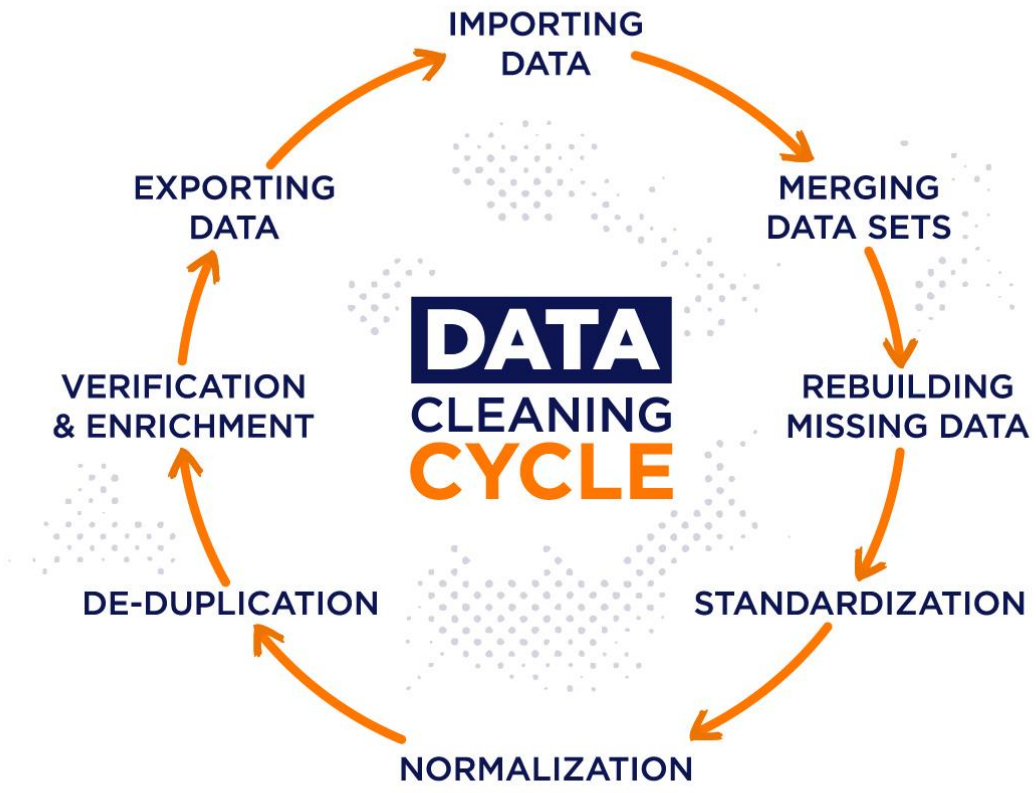

Figure1. Probabilistic Data Cleaning Architecture

This approach is particularly important in large datasets where multiple repairs may be plausible and exhaustive manual correction is infeasible. By embedding ML into the cleaning loop, such systems enable iterative refinement, where model predictions, constraints, and user feedback jointly improve data quality. This shift reframes data cleaning as an inference problem rather than a purely procedural task. As a result, repair decisions become more transparent, explainable, and amenable to optimization.

4.1 Probabilistic Data Repair with HoloClean

HoloClean represents a landmark system in ML-driven data cleaning by modeling data repair as a probabilistic inference problem over a factor graph. Rather than treating constraints, statistics, and external knowledge separately, HoloClean unifies them into a single probabilistic model that reasons about the likelihood of different cell values. Integrity constraints, such as functional dependencies, act as soft signals rather than hard rules, allowing the system to

| ISSN: 2455-1864 | www.ijrai.org | editor@ijrai.org | A Bimonthly, Scholarly and Peer-Reviewed Journal |

||Volume 5, Issue 1, January – February 2022||

DOI:10.15662/IJRAI.2022.0501006

tolerate noise and incomplete specifications. External knowledge sources and co-occurrence statistics further inform the model, helping disambiguate repairs when multiple values are plausible. By framing cleaning as inference, HoloClean can naturally rank candidate repairs and quantify uncertainty, which is critical for downstream decision-making. This design allows the system to scale across heterogeneous datasets while remaining robust to imperfect constraints. Importantly, the probabilistic formulation supports incremental improvement as new evidence or feedback is incorporated. HoloClean thus bridges the gap between deterministic data cleaning and purely statistical anomaly detection.

A key contribution of HoloClean lies in its optimizations that make probabilistic repair computationally feasible at scale. Figure 1illustrates how domain pruning techniques dramatically reduce compilation and repair runtimes by limiting candidate values to those that are statistically and semantically plausible. This optimization preserves repair quality while significantly improving scalability, addressing a common criticism of probabilistic models. Figure 2 further highlights the runtime-quality tradeoff across datasets, showing that carefully tuned configurations achieve near-optimal repairs at a fraction of the computational cost. Together, these results demonstrate that ML-based data cleaning need not be prohibitively expensive. Instead, with principled modeling and optimization, systems like HoloClean can deliver high-quality repairs efficiently. These empirical findings support the broader argument that integrated ML-driven repair pipelines are viable for real-world, large-scale data management scenarios.

4.2 Active Learning for Cleaning with ActiveClean

ActiveClean reframes data cleaning as an optimization-driven process that explicitly reminds the objective of downstream machine learning performance rather than treating data quality as an isolated preprocessing task. Traditional cleaning pipelines aim to exhaustively correct all detected errors before model training, often incurring significant cost while offering diminishing returns for model accuracy. ActiveClean challenges this assumption by observing that not all dirty records contribute equally to model degradation. Instead, it formulates data cleaning as an iterative loop in which model training, error detection, and selective repair proceed jointly. By prioritizing records that most influence the model's loss function or gradient updates, ActiveClean ensures that limited cleaning budgets are allocated where they yield maximal predictive benefit. This reframing is particularly relevant in enterprise environments, where datasets are large, heterogeneous, and continuously evolving, making full manual cleaning infeasible. The approach also aligns well with practical constraints, as it allows organizations to deploy usable models early while progressively improving quality. In doing so, ActiveClean bridges the gap between data management and machine learning, demonstrating that cleaning decisions should be informed by learning objectives rather than purely syntactic notions of correctness.

At the core of ActiveClean is an active learning strategy that identifies which records to clean next based on their estimated impact on model training. After an initial model is trained on a partially cleaned dataset, the system evaluates candidate dirty records by approximating how their correction would affect the model's parameters or loss. Records with the highest expected influence are then selected for cleaning, either through automated repair rules or human intervention. Once cleaned, these records are reintegrated into the training set, and the model is retrained or incrementally updated. This process repeats until the cleaning budget is exhausted or performance converges. Importantly, ActiveClean is model-agnostic and can be applied to a range of learning algorithms, including linear models and convex objectives, which makes it broadly applicable. The iterative feedback loop ensures that cleaning effort is continuously guided by empirical model behavior rather than static heuristics. As a result, the system adapts naturally to different datasets and error distributions, emphasizing flexibility and efficiency over completeness.

Empirical evaluations presented in Figure 1(ActiveClean) provide strong evidence for the effectiveness of this ML-aware cleaning strategy. Across multiple real-world datasets, convergence plots show that models trained using ActiveClean reach near-optimal accuracy after cleaning only a small fraction of the data. In contrast, baselines that rely on uniform sampling or defer training until full cleaning exhibit significantly slower convergence and higher initial error. These results highlight a key insight: targeted cleaning of influential records can deliver most of the attainable accuracy gains at a fraction of the cost. From a data stewardship perspective, this finding has important implications. It suggests that quality interventions should be prioritized based on their downstream impact, particularly when resources are constrained. More broadly, ActiveClean illustrates how active learning principles can be embedded directly into data quality pipelines, enabling organizations to move from exhaustive, static cleaning practices toward adaptive, outcome-driven stewardship strategies.

International Journal of Research and Applied Innovations (IJRAI)

| ISSN: 2455-1864 | www.ijrai.org | editor@ijrai.org | A Bimonthly, Scholarly and Peer-Reviewed Journal |

||Volume 5, Issue 1, January – February 2022||

DOI:10.15662/IJRAI.2022.0501006

V. EVALUATION METHODOLOGIES AND PERFORMANCE TRADEOFFS

Evaluating anomaly detection and data quality systems presents distinctive methodological challenges that set them apart from conventional supervised learning tasks. Ground-truth labels for anomalies are often scarce, noisy, or subjective, and true anomalies typically represent a tiny fraction of the data, leading to extreme class imbalance. As a result, naïve accuracy-based metrics can be misleading, rewarding models that simply predict "normal" behavior. Early evaluation practices therefore struggled to compare methods consistently or to reflect real operational priorities. Benchmarks such as the **Numenta Anomaly Benchmark** addressed this gap by reframing evaluation around timeliness, robustness, and tolerance to noise in time-series settings. Instead of penalizing late detections equally with missed detections, NAB emphasized early warning capability, which is often more valuable in operational contexts. This shift highlighted that anomaly detection quality is multidimensional, encompassing detection speed, stability, and resilience to changing conditions. Consequently, evaluation frameworks increasingly reflect the realities of deployment rather than idealized laboratory settings.

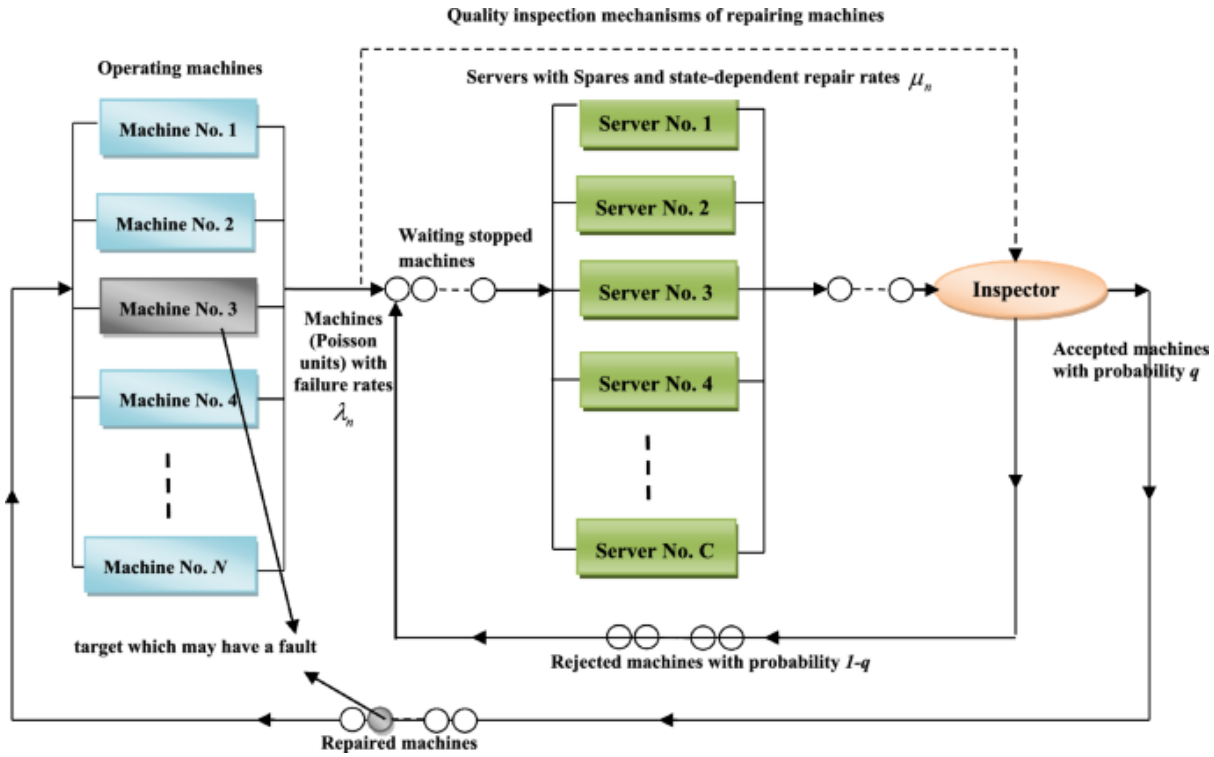

Figure2. Runtime-Quality Tradeoff in Automated Data Repair

Beyond benchmarks, practical evaluation in production environments must account for data drift, evolving schemas, and changing definitions of "normal" behavior. An anomaly detection model that performs well at deployment time may degrade silently as upstream data pipelines change or business processes evolve. Continuous evaluation strategies therefore incorporate distribution monitoring, alert review rates, and feedback from downstream users. Precision-recall tradeoffs are often tuned dynamically based on operational tolerance for false positives versus missed issues. Importantly, evaluation is not limited to detection accuracy but extends to the cost of investigation and remediation triggered by alerts. Systems that generate frequent low-quality alerts impose cognitive and operational burdens that can outweigh their theoretical benefits. As a result, practical metrics increasingly include alert fatigue indicators, mean time to resolution, and the proportion of alerts that lead to confirmed data issues.

Modern production systems operationalize these insights by embedding ML-driven quality checks directly into data pipelines rather than treating them as offline analyses. Tools such as **TensorFlow Data Validation** exemplify this trend by combining schema inference, constraint checking, distribution comparison, and drift detection into automated workflows. These systems continuously monitor incoming data against learned expectations, surfacing anomalies as

| ISSN: 2455-1864 | www.ijrai.org | editor@ijrai.org | A Bimonthly, Scholarly and Peer-Reviewed Journal |

||Volume 5, Issue 1, January – February 2022||

DOI:10.15662/IJRAI.2022.0501006

early as possible in the pipeline. By integrating validation with orchestration and monitoring infrastructure, they enable rapid feedback loops between detection and remediation. This tight coupling between research ideas and engineering practice reflects a broader maturation of the field, where anomaly detection and data quality are treated as ongoing operational concerns. Ultimately, effective evaluation and deployment require aligning algorithmic metrics with human, organizational, and business costs, ensuring that ML-based quality systems deliver sustained value in real-world settings.

VI. KEY STUDIES AND EMPIRICAL INSIGHTS

Several foundational studies have shaped the theoretical and empirical trajectory of anomaly detection and ML-driven data quality management. **Chandola et al. (2009)** provided the first widely adopted taxonomy of anomaly detection techniques, systematically categorizing methods by assumptions, data types, and anomaly definitions. This work clarified distinctions between point, contextual, and collective anomalies, offering a conceptual framework that guided evaluation and algorithm design for more than a decade. By synthesizing statistical, distance-based, and machine learning approaches, the survey helped unify a fragmented literature and established anomaly detection as a coherent research field. Its emphasis on context and feature interactions directly influenced later work in both time-series monitoring and data quality assessment. As a result, many subsequent systems implicitly adopt Chandola et al.'s taxonomy when framing problem scope and evaluation criteria. The paper's longevity underscores its role as a reference point for both researchers and practitioners.

Scalability emerged as a critical concern as anomaly detection moved from theory to large-scale deployment, a gap addressed by **Liu et al. (2008)** through the introduction of Isolation Forests. This work demonstrated that anomalies can be effectively detected through random partitioning rather than explicit distance or density estimation, yielding near-linear time complexity. The empirical results showed strong performance across high-dimensional datasets, making the method practical for real-world applications such as fraud detection and system monitoring. Isolation Forests also lowered the barrier to adoption by requiring minimal parameter tuning and no labeled data. Their success highlighted the importance of algorithmic simplicity and computational efficiency in operational anomaly detection. Consequently, Isolation Forests became a standard baseline in both research benchmarks and production systems. This study reinforced the idea that scalability is as important as detection accuracy in data quality contexts.

Beyond detection, empirical work increasingly emphasized the importance of repair and downstream impact. **Rekatsinas et al. (2017)** showed that probabilistic inference can unify error detection and repair within a single framework, providing measurable guarantees on repair quality while remaining scalable. Complementing this, **Krishnan et al. (2016, 2018)** demonstrated that selective, ML-guided data cleaning yields greater improvements in model performance than exhaustive cleaning strategies. Their experiments showed that prioritizing influential errors those most affecting downstream models can significantly reduce human effort while improving predictive accuracy. Together, these studies provide strong empirical evidence that ML-driven data quality management is not only theoretically appealing but also practically effective. They establish that intelligent prioritization, probabilistic reasoning, and feedback loops are key to building scalable, high-impact data cleaning systems.

VII. CASE STUDY: MACHINE LEARNING-DRIVEN DATA QUALITY AND ANOMALY DETECTION

In a large-scale enterprise analytics environment, persistent data quality issues such as missing values, inconsistent categorical attributes, and anomalous numerical records significantly undermined the reliability of downstream machine learning models. Data was ingested from heterogeneous operational systems with evolving schemas, limited documentation, and minimal labeled ground truth, making manual rule-based validation brittle and difficult to maintain. As data volumes increased, traditional cleansing pipelines failed to scale, leading to delayed analytics and compounding model errors. To address these challenges, the organization reframed data quality management as a learning problem rather than a static validation task. Machine learning techniques were introduced to automatically infer normal patterns, identify deviations, and propose corrections based on statistical evidence. This shift enabled the system to adapt dynamically to schema drift, distributional changes, and previously unseen error modes. By embedding anomaly detection directly into the data ingestion pipeline, quality issues were surfaced early and contextualized within broader dataset behavior. The adoption of ML-based data quality mechanisms marked a transition from reactive, manual remediation toward proactive, data-driven governance.

International Journal of Research and Applied Innovations (IJRAI)

| ISSN: 2455-1864 | www.ijrai.org | editor@ijrai.org | A Bimonthly, Scholarly and Peer-Reviewed Journal |

||Volume 5, Issue 1, January – February 2022||

DOI:10.15662/IJRAI.2022.0501006

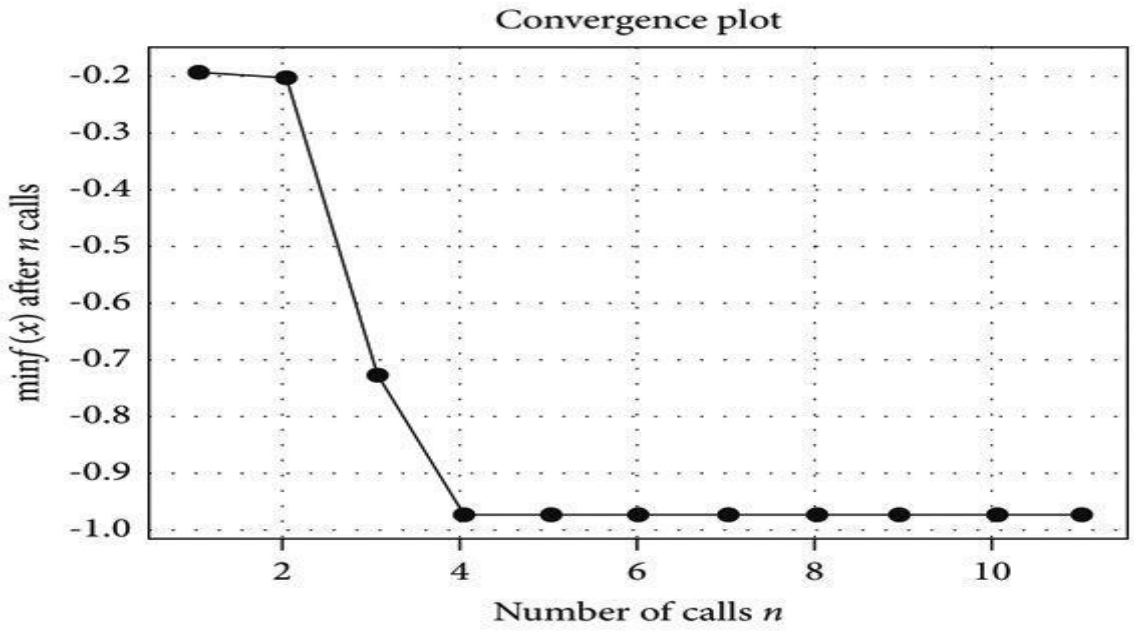

Figure3. Model-Aware Data Cleaning and Error Convergence

The first phase of the intervention applied a probabilistic data repair approach inspired by HoloClean, which models data quality constraints, attribute correlations, and external knowledge as a unified probabilistic inference problem. Rather than treating errors in isolation, the system evaluated candidate repairs holistically across the dataset. Empirical findings, reflected in HoloClean Figures 1 and 2, demonstrated that domain pruning significantly reduced compilation and repair runtimes while preserving high repair accuracy. In the enterprise deployment, similar trends were observed, with moderate pruning thresholds achieving near-optimal repair quality at a fraction of the computational cost. This balance proved essential for production environments operating under strict latency constraints. The probabilistic framework also provided confidence scores for suggested repairs, enabling selective human review where uncertainty was highest. As a result, automated repair could be safely applied to the majority of errors, while edge cases were escalated for manual validation. These outcomes demonstrated that statistically grounded ML-based repair can be both scalable and trustworthy in real-world settings.

Following automated repair, the organization integrated a model-aware data cleaning strategy based on the principles of ActiveClean to maximize downstream learning performance. Instead of exhaustively correcting all remaining questionable records, the system prioritized those whose correction was predicted to yield the largest reduction in model loss. ActiveClean Figure 1illustrates how test error converges rapidly as a small subset of high-impact records is cleaned, a pattern that closely matched observations in the enterprise pipeline. In practice, correcting fewer than one-fifth of flagged records produced the majority of attainable accuracy gains, dramatically reducing manual effort. This targeted strategy aligned data quality interventions directly with business and modeling objectives rather than abstract correctness metrics. By coupling data cleaning decisions to model feedback, the organization ensured that limited resources were focused where they delivered the highest value. The case study underscores that effective data quality management is inseparable from machine learning outcomes and that ML-guided prioritization is critical for sustainable, high-impact data operations.

VIII. CONCLUSION

Machine learning has fundamentally reshaped the landscape of automated data quality management and anomaly detection by shifting emphasis from rigid, manually specified rules to adaptive, data-driven inference. Traditional rule-based systems struggle to scale as data volumes grow, schemas evolve, and sources diversify, often requiring constant human maintenance to remain effective. In contrast, ML-based approaches learn statistical, relational, and semantic regularities directly from data, enabling them to generalize across heterogeneous datasets and adapt to distributional

| ISSN: 2455-1864 | www.ijrai.org | editor@ijrai.org | A Bimonthly, Scholarly and Peer-Reviewed Journal |

||Volume 5, Issue 1, January – February 2022||

DOI:10.15662/IJRAI.2022.0501006

change. This adaptability is particularly important in modern pipelines where data is continuously ingested, transformed, and reused across multiple analytical and operational contexts. By capturing complex dependencies that are difficult to encode explicitly, ML models can detect subtle anomalies, inconsistencies, and integrity violations that would otherwise remain hidden. As a result, data quality is no longer treated as a static validation step but as a dynamic, continuously learned property of the data. This paradigm shift has elevated data quality from a maintenance concern to a core capability underpinning reliable analytics and machine learning systems.

Empirical evidence from production-oriented systems such as HoloClean and ActiveClean demonstrates that ML-driven data quality pipelines can achieve practical and measurable benefits. These systems show that probabilistic reasoning and selective, model-guided cleaning can strike effective tradeoffs between repair accuracy, computational cost, and human effort. Rather than attempting exhaustive correction of all detected issues, ML-guided prioritization focuses attention on errors that most affect downstream models and decisions. Experimental results consistently show improvements in predictive performance, robustness, and convergence speed when intelligent cleaning strategies are applied. Importantly, these gains are achieved without prohibitive runtime overhead, dispelling concerns that ML-based cleaning is inherently too expensive for large-scale deployment. By quantifying uncertainty and surfacing confidence-ranked repair suggestions, such systems also support meaningful human oversight. Together, these findings validate that ML-based data quality management is not only theoretically sound but operationally viable.

Looking forward, the growing reliance on automated decision systems amplifies the importance of robust, maintainable data quality mechanisms. As machine learning models increasingly influence high-stakes outcomes, undetected data errors can propagate rapidly and at scale, undermining trust and accountability. Integrating ML-driven data quality checks directly into data pipelines enables earlier detection, faster remediation, and tighter feedback loops between data producers and consumers. This integration supports long-term maintainability by allowing quality controls to evolve alongside data and models, rather than lag behind them. Moreover, coupling automated detection with human-in-the-loop review ensures that domain expertise remains central where judgment and context are required. Ultimately, organizations that invest in ML-based data quality infrastructure position themselves to build analytics and AI systems that are not only more accurate, but also more trustworthy, resilient, and sustainable over time.

REFERENCES

1. Stonebraker, M., Çetintemel, U., & Zdonik, S. (2005). The 8 requirements of real-time stream processing. *ACM SIGMOD Record, 34*(4), 42-47.https://doi.org/10.1145/1107499.1107504
2. Chaudhuri, S. (2007). Self-tuning database systems: A decade of progress. *Proceedings of the VLDB Endowment, 1*(1), 3-14.https://dl.acm.org/doi/10.5555/1325851.1325856
3. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.https://doi.org/10.1145/2939672.2939778
4. Doshi-Velez, F., & Kim, B. (2017).Towards a rigorous science of interpretable machine learning. *arXiv preprint*. https://arxiv.org/abs/1702.08608
5. Amershi, S., et al. (2019).Software engineering for machine learning: A case study. *Proceedings of the IEEE/ACM International Conference on Software Engineering (ICSE)*. https://doi.org/10.1109/ICSE-SEIP.2019.00042
6. Rahwan, I., et al. (2019). Machine behaviour. *Nature, 568*(7753), 477-486. https://doi.org/10.1038/s41586-019-1138-y
7. Kranthi Kumar Routhu. (2018). Reusable Integration Frameworks in Oracle HCM: Accelerating Enterprise Automation through Standardized Architecture. In International Journal of Scientific Research & Engineering Trends (Vol. 4, Number 4). Zenodo. https://doi.org/10.5281/zenodo.17670619
8. Qin, S. J. (2012). Survey on data-driven industrial process monitoring and diagnosis. *Annual Reviews in Control, 36*(2), 220-234.https://doi.org/10.1016/j.arcontrol.2012.09.004
9. Sudhir Vishnubhatla. (2019). From Rules To Neural Pipelines: NLP-Powered Automation For Regulatory Document Classification In Financial Systems. In International Journal of Science, Engineering and Technology (Vol. 7, Number 1). Zenodo. https://doi.org/10.5281/zenodo.17473977
10. Salhi, H., Odeh, F., Nasser, R., & Taweel, A. (2017). Open source in-memory data grid systems: Benchmarking Hazelcast and Infinispan. *Proceedings of ACM/IFIP ICPE '17*. https://doi.org/10.1145/3030207.3053671

| ISSN: 2455-1864 | www.ijrai.org | editor@ijrai.org | A Bimonthly, Scholarly and Peer-Reviewed Journal |

||Volume 5, Issue 1, January – February 2022||

DOI:10.15662/IJRAI.2022.0501006

11. Sudhir Vishnubhatla. (2020). Adaptive Real-Time Decision Systems: Bridging Complex Event Processing And Artificial Intelligence. In International Journal of Science, Engineering and Technology (Vol. 8, Number 2). Zenodo. https://doi.org/10.5281/zenodo.17471901
12. Salhi, H., Odeh, F., Nasser, R., & Taweel, A. (2017). Benchmarking and performance analysis for distributed cache systems. *LNCS 10661*. Springer https://doi.org/10.1007/978-3-319-72401-0_11
13. Shravan Kumar Reddy Padur "Empowering Developer & Operations Self-Service: Oracle APEX + ORDS as an Enterprise Platform for Productivity and Agility" International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Print ISSN : 2395-1990, Online ISSN : 2394-4099, Volume 4, Issue 11, pp.364-372, November-December-2018. Available at doi : https://doi.org/10.32628/IJSRSET1844429
14. AdCONIP Proceedings. (2017).Advances in big data analytics at The Dow Chemical Company. https://skoge.folk.ntnu.no/prost/proceedings/adconip-2017/media/files/0111.pdf
15. Sudhir Vishnubhatla. (2019). From Rules To Neural Pipelines: NLP-Powered Automation For Regulatory Document Classification In Financial Systems. In International Journal of Science, Engineering and Technology (Vol. 7, Number 1). Zenodo. https://doi.org/10.5281/zenodo.17473977