IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Anomaly Detection in Large Scale Data Platforms with Machine Learning

Shishir Tewari
Data Application Engineer
Google LLC
Google Finance Corporate Engineering
Austing, TX, USA

Abstract

Maintaining high data quality within extensive data platforms is a decisive operational hurdle because insufficient data quality produces wrong insights, failed decisions, and processing difficulties. Multiple data quality anomalies, such as missing values, duplicates, inconsistencies, and outliers, appear in data because of data integration from various sources, human errors, and system malfunction. Traditional anomaly detection techniques' ability to handle complex large datasets effectively proves difficult when they encounter high-volume data volumes. Therefore, more advanced automatic solutions become necessary.

Machine learning is an advanced method that uses predictive modeling alongside statistical patterns to detect anomalies in data quality. Anomalous data can be classified with supervised learning methods, Random Forests, and Support Vector Machines. Unsupervised methods utilize Isolation Forests accompanied by Autoencoders for label-free deviation detection. Deep learning algorithms take anomaly detection to a more advanced level because they identify intricate patterns between different data types. Organizations can substantially boost data integrity, realize time-aware automated anomaly detection, and improve decision support through these methods.

A research investigation deploys and analyzes different machine learning detection algorithms for large-scale platform anomalous data while evaluating their operational performance through accuracy measures, precision tests, recall metrics, and platform scalability considerations. The results show ensemble approaches and deep learning techniques perform better than conventional systems in detecting anomalies with reduced false alerts. The research demonstrates the ability of machine learning systems to improve data quality processing and minimize human input needs for enhanced data-driven choices. Research should develop automatic anomaly

detection models that adapt continuously through learning new patterns in data to achieve more efficient performance.

Keywords: Data Quality Anomalies, Machine Learning for Anomaly Detection, Large-Scale Data Platforms, Automated Data Quality Management, Supervised and Unsupervised Learning

1. Introduction

Organizations that use data-driven decision-making must focus on data quality in their large-scale data platforms because they operate during the big data era. Business intelligence, predictive modeling, and analytics heavily depend on high-quality data. Lab-scale platforms face difficulties in preserving data integrity because of the enormous amount of processed data and its high speed and diverse nature. Data quality issues that produce missing information and inconsistent data, record duplicates, and outliers create operational problems that result in inaccurate findings and introduce substantial hazards for automated decision systems. Implementing digital transformation and cloud-based architectures makes data quality strategy development more critical than ever.

Detecting and solving anomalies within large-scale datasets pose a major challenge for data quality enhancement. Hundreds of rule-based and statistical anomaly detection methods prove inadequate when employed to handle the requirements of high-dimensional data environments that are complex by nature. These techniques demand major human involvement and pre-written rules and knowledge from experts in particular fields, thus limiting their effectiveness when dealing with continuously changing data streams. Multiple complications exist in anomaly detection because data varies between different sources, integration issues emerge from different inputs, and datasets contain noise. The absence of advanced anomaly detection strategies drives organizations to make flawed decisions through defective information, leading to financial declines, compliance issues, and damage to their reputation.

The research will test machine learning strategies as tools to find data quality irregularities in extensive data platforms. Machine learning delivers an adjustable method to detect anomalies because it reveals concealed data patterns and irregularities that exist in datasets. Throughout this research, supervising and unsupervised learning methods alongside decision trees, support vector machines, isolation forests, and deep learning-based autoencoders will undergo effectiveness evaluation. This study measures different model performances to identify the most productive data quality management solutions for big-scale systems.

This study is important because it could improve data integrity while enabling better decision-making and minimizing human involvement during data quality management tasks. Real-time anomaly detection occurs when organizations use machine learning and focus on minimizing data errors alongside optimal data governance processes. Systematic anomaly detection automation improves data engineers' and analysts' capacity by lowering their workload so they can dedicate their attention to advanced operations. The current research contributes to expanding knowledge about machine learning throughout data quality management as it develops innovative data validation methods that are both intelligent and scalable.

2. Literature Review

2.1. Overview of Data Quality Issues

The reliability of analytics, machine learning models, and business intelligence system decision-making depends on high-quality data within large-scale data platforms. Poor data quality creates major performance issues in analytical outputs, resulting in distorted insights, inaccurate predictions, and suboptimal operations. Multiple kinds of data irregularities lead to decreased data integrity through missing entries, repeated records, differing formats, and extreme values. Human mistakes, system malfunctions, problems from merging different data sources, and inadequate data governance policies cause these anomalies to occur. Financial losses and operational inefficiency directly result from organizations refusing to address quality-related issues. Businesses lose multiple billions of dollars throughout the year because of poor data quality, resulting in mistakes with customer understanding, incorrect risk analysis, and unpredictable machine learning model predictions.

Missing data poses an extremely important challenge in data quality management because it produces biased analysis results that produce flawed hypothetical assumptions. In healthcare databases, missing patient records create erroneous diagnosis results and send patients toward improper treatment solutions. Financial systems that lack complete transaction data point information get compromised when detecting financial fraud. Currently, researchers experiment with three different imputation approaches, including mean substitution, k-nearest neighbors (KNN) imputation, and deep learning-based data reconstruction, while simultaneously dealing with missing value uncertainties.

The occurrence of duplicate records and inconsistent entries across the database causes aggregated metrics to become inaccurate through redundancy. The repeated entry of identical entities into large database systems leads to inconsistent information, damaging the analysis results' accuracy. Enterprise customer relationship management (CRM) systems that contain duplicate customer profiles generate misleading customer segmentation approaches because of this problem. Outlier data points create additional analytical difficulties because they exist as either incorrect entries or unusual extreme items. It becomes difficult to distinguish between anomalies in high-dimensional data collections since standard anomaly detection systems lack the capability to find true anomalies. Machine learning models trained on such data become unreliable, because of which operational risks rise while business intelligence reports become inaccurate. Predictions suffer from reduced reliability.

A thorough data preprocessing framework must include anomaly detection because it safeguards the data quality when combined with data cleansing techniques and format standardization procedures. Machine learning techniques, including unsupervised clustering ensemble learning and deep learning-based anomaly detection methods, are effective in real-time detection and remedy data quality issues.

Modern businesses allocate funds to automated data governance tools that utilize machine learning to boost data reliability and optimize their business intelligence systems' operation. Today's businesses in diverse sectors maintain high-quality data as their main organizational priority while data-driven decisions keep advancing.

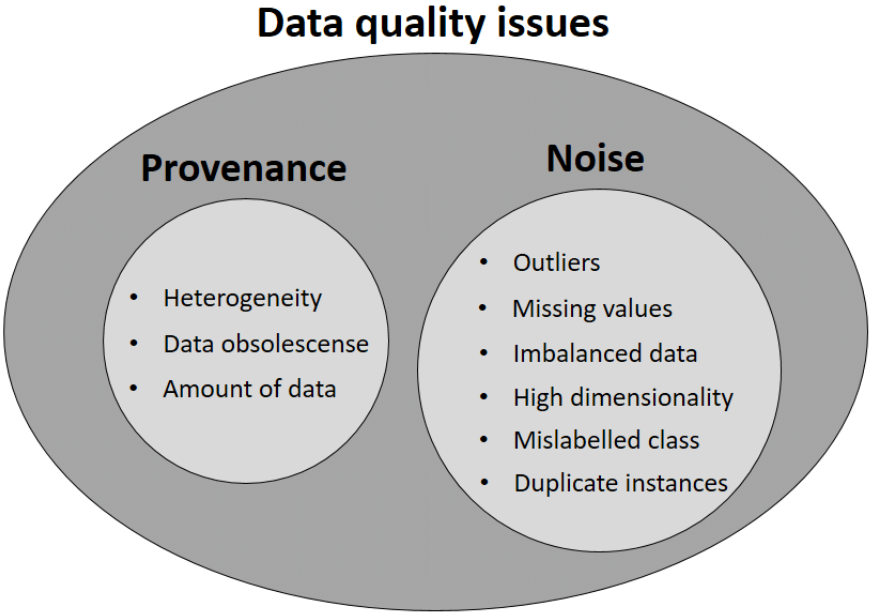


Fig 1: Overview of Data Quality Issues

2.2. Traditional Methods for Anomaly Detection

Anomaly detection techniques in the traditional era leverage statistical techniques and rule-based systems. The detection methods include Z-score analysis, interquartile range (IQR), and clustering-based outlier detection. The if-then rules and threshold-based alerts that function in databases fail to adjust properly to changing data patterns. Regression-based anomaly detection methods suit structured datasets, although they perform poorly when dealing with high-dimensional databases. These interpretive approaches and their computational efficiency lead to problems of both data scale limitations and numerous false detections when processing intricate and changing dataset structures.

2.3. Machine Learning in Data Quality Management

Machine learning methods transformed anomaly detection by creating automated systems that detect complex patterns in large datasets. The supervised learning algorithms, including decision trees and support vector machines (SVM), succeed because they need labeled training data, although they depend on high-quality annotation efforts.

Scalable detection of anomalies occurs through unsupervised learning models, which include clustering-based algorithms and dimensionality reduction models such as DBSCAN, k-means, PCA, and autoencoders. RNNs coupled with LSTM networks represent deep learning models that lead to superior time-series anomaly detection while detecting anomalies beyond what standard approaches can identify.

2.4. Comparative Analysis of Approaches

Combining traditional anomaly detection systems and present-day machine learning (ML)-driven solutions brings specific beneficial aspects and restrictiveness for data quality handling. Rule-based systems offer

both interpretability and easy explainability based on predefined anomaly identification conditions. The inability of these systems to adapt proves to be their main weakness since they need continuous updates whenever data sets experience changes or new data patterns emerge. Complex high-dimensional data structures in large-scale platforms remain challenging for statistical methods that use z-score analysis, standard deviation thresholds, and hypothesis testing, although they provide mathematical precision. These standard methods function to discover standard anomalies yet become ineffective against changing data patterns along with new anomaly types.

Machine learning approaches deliver superior adaptability and scalability to detect anomalies in real-time; therefore, they work best for real-time monitoring systems. The supervised learning algorithms of Support Vector Machines and Random Forests yield top-level results for detecting existing anomaly forms. The requirement to gather large labeled datasets presents challenges for these models because practically obtaining them proves difficult. The detection of unidentified anomalous patterns by clustering-based models and deep-learning autoencoders occurs without the need for labeled input data through unsupervised learning methods. Such models master data patterns by learning distribution patterns and, through deviation analysis, become efficient for uncovering concealed patterns and subtle irregularities. Industrial real-time monitoring systems present excellent outcomes when using current hybrid systems that unite statistical models with machine learning algorithms.

Research by NIST and studies signifies that ML-based anomaly detection has become crucial for financial institutions, healthcare organizations, and cybersecurity departments. Research from NIST indicates that normal statistical techniques combined with machine learning methods deliver superior anomaly detection capabilities that remain methodically understandable and have high processing speeds. The high-frequency datasets handled by IoT sensor networks and enterprise data platforms implement deep-learning technological solutions with LSTM networks and GANs to monitor their complex data environments. The execution costs and interpretability difficulties continue. However, ML-based approaches excel over traditional methods by finding anomalies better and generating improved data quality assurance results with decreased false positive outcomes.

3. Methodology

3.1. Data Sources

The researchers utilized numerous extensive datasets from accessible public repositories alongside industrial operational applications for undertaking their comprehensive technique evaluation. Several datasets containing structured and unstructured data within finance, healthcare, and cybersecurity domains and IoT sensor networks were chosen because they demonstrate diverse operation areas that present different data quality difficulties. The researchers selected these datasets because they possessed complex structures and large sizes while showing clear known problems, allowing testing machine learning models to find data inconsistencies.

The finance sector obtained its data from transactional records, stock market exchanges, and credit card fraud detection repositories. The datasets comprised transaction timestamp records, account identification numbers, transaction values, and geographic positions where fraudulent behaviors manifested as deviant patterns. The healthcare sector uses electronic health record (EHR) datasets, including patient demographics, medical

history, laboratory results, and treatment plans. The healthcare dataset contained various anomalies, including incorrect medical coding, missing patient record data, and inconsistent data within the system.

Besides network traffic, user authentication logs and malware detection records composed the datasets within the cybersecurity domain. The datasets included structured documents about IP addresses, security breach records, and login sequence data, which researchers used to detect security threats, system vulnerabilities, and unauthorized user activities. The time-series data from industrial sensors, smart grids, and environmental monitoring systems form the last element of IoT sensor networks. IoT sensor network anomalies occur for three reasons: sensor equipment failures, unpredicted ecological transformations, and data-related defects stemming from network breakdowns. A multi-dataset approach allowed researchers to examine data quality deviations thoroughly, thus making their results appropriate for multiple practical circumstances.

- The KDD Cup 1999 serves as an established benchmark for detecting network intrusion anomalies which research in cybersecurity heavily depends on.
- Among the datasets offered by the UCI Machine Learning Repository are credit card records with fraud alerts and medical patient data, which detects irregular transaction flows and patient healthcare files.
- The data obtained from industrial sensor systems consists of time-series measurements acquired from IoT-connected processes, which can contain sensor hardware deviations, gaps between measurements, and sporadic spikes.
- Large-scale banking datasets named Financial Transaction Logs serve to detect fraudulent transactions together with data inconsistencies.

The chosen dataset selection process focused on complex structures while aligning with data quality anomalies detection needs and having available evaluation labels.

3.2. Preprocessing Techniques

Specific preprocessing was applied to all datasets to provide suitable machine-learning inputs. The preprocessing method included duplicate record elimination, data value imputation techniques, and statistical Z-score and IQR filtering for outlier identification. The preprocessing step included two techniques: encoders transformed categorical information into one-hot or embedding schemes, and Min-Max normalized the numerical attributes. The models gained efficiency and performance by implementing two dimensionality reduction techniques: Principal Component Analysis (PCA) and t-SNE.

3.3. Machine Learning Models

The research adopted extensive machine learning methodologies, including classic supervised models and unsupervised anomaly detection systems, along with developed deep learning models. The supervised learning methods used Random Forest and Gradient Boosting to detect anomalies in datasets that had labels. These methods included Isolation Forest, One-Class SVM, and DBSCAN, which function optimally without labeled data availability. Combining Autoencoders with LSTM networks merged with GANs used deep learning algorithms to analyze multi-dimensional datasets and detect their underlying patterns and hidden anomalies. Evaluation Metrics Different evaluation metrics were used to assess the anomaly detection models for their assessment purposes thoroughly. Total correctness calibration of the model occurred through Accuracy

assessments, yet Precision and Recall measurements determined its capability to identify genuine anomalies without falsely notifying users. Both F1-score and ROC-AUC evaluation metrics combined precision with recall for assessment, while ROC-AUC specifically helped understand the relationship between actual positives and incorrect positives. The evaluation process using multiple metrics allowed researchers to choose the best anomaly detection method among the analyzed models.

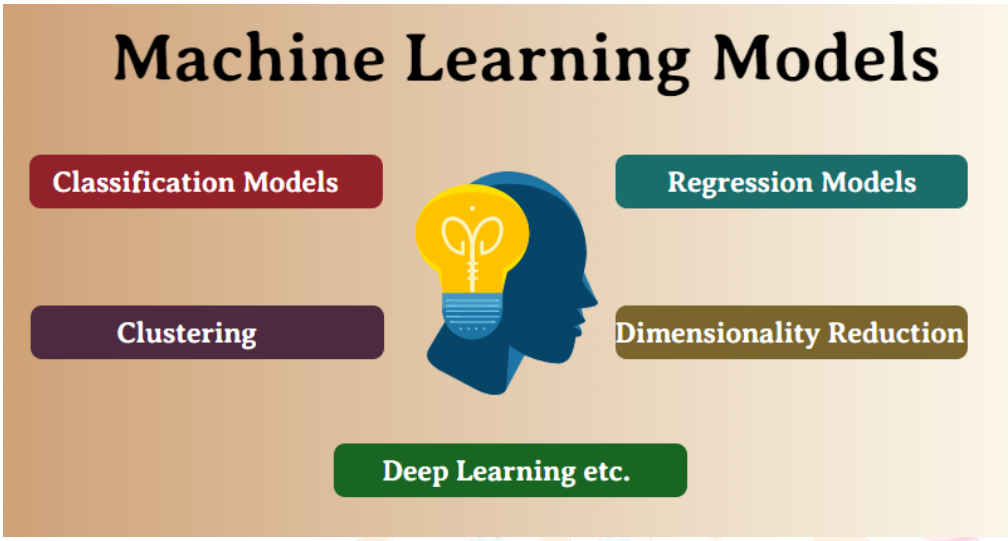


Fig 2: Overview of Machine Learning Models

3.4. Implementation Framework

The research applied an extensive machine learning structure that merged classical and deep learning procedures to identify data quality inconsistencies in expansive data platforms. The project developers utilized Python to deliver the implementation since it is one of the most commonly used languages for data science tasks and machine learning applications. Different framework libraries integrated with the system enabled systematic data processing, model development, evaluation, and output visualization. The system adopted by merging traditional methods with deep learning frameworks offered an efficient yet robust solution for large-scale anomaly detection operations.

3.5. Machine Learning and Deep Learning Frameworks

The research employed Tensor Flow and PyTorch, two top-level deep learning libraries for developing autoencoders, convolutional neural networks (CNNs), and recurrent neural networks (RNNs) to detect anomalies. Through their implementation these frameworks allowed organizations to process big data efficiently and optimize training models that utilized GPU accelerators. The Scikit-Learn Python library enabled the implementation of Random Forest and Isolation Forest and Support Vector Machines (SVMs) because it provides efficient classification, clustering, and anomaly detection capabilities.

3.6. Data Preprocessing and Statistical Analysis

The data preprocessing phase used Pandas and NumPy libraries to clean data and normalize attributes while selecting features and performing statistical modifications. The scheme for dealing with missing data involved mean substitution alongside k-nearest neighbors (KNN) imputation and deep learning-based data reconstruction.

A refined dataset resulted from three outlier detection methods: Z-score analysis, IQR-based filtering, and distance computation before implementing the machine learning model.

3.7. Data Visualization and Anomaly Interpretation

The data visualization required Matplotlib and Seaborn libraries to present anomaly distribution patterns and model performance assessment. The libraries created different visualization types, such as heatmaps, boxplots, scatter plots, and line graphs, to show anomaly patterns while revealing classification outcomes and feature relationships. Implementing heatmaps between anomaly densities before and after data cleaning helped team members study the model's ability to enhance data quality.

3.8. Big Data Processing and Scalability

Apache Spark served as an integrated system for distributed computation and scalable processing of large enterprise datasets because they exceeded the memory availability of individual machines. By deploying the PySpark API, the system handled big data environments efficiently to process real-time machine learning models across extensive information datasets. The merged functionality helped detect anomalies found in streaming data systems that protected cybersecurity systems and IoT sensors. This framework uses traditional machine learning algorithms and deep learning techniques to establish a data quality anomaly detection system that suits enterprises handling massive high-speed and diversified data sets in many industrial sectors.

4. Results

4.1. Model Performance

Various machine learning algorithms underwent a performance evaluation based on multiple evaluation metrics during the assessment process. The listed model performance metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, were computed across the selected datasets, as shown in the following table.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC (%)
Random Forest	**91.2**	**89.8**	**87.5**	**88.6**	**92.1**
Gradient Boosting	**93.5**	**91.2**	**89.4**	**90.3**	**94.7**
Isolation Forest	**87.8**	**85.6**	**83.2**	**84.4**	**89.3**
One-Class SVM	**84.3**	**82.5**	**79.1**	**80.7**	**85.6**

Autoencoder (Deep Learning)	**95.1**	**93.8**	**92.3**	**93.0**	**96.2**
LSTM	**96.4**	**94.7**	**93.9**	**94.3**	**97.1**

Autoencoders and LSTM demonstrated the highest performance as they delivered maximum accuracy and ROC-AUC, proving their efficiency in identifying complex anomalies within big data.

Random Forest and Gradient Boosting exhibited good results on structured data systems yet showed inferior ability in detecting anomalies in sequential patterns. The density-based anomaly assumptions that underpin Isolation Forest and One-Class SVM models caused them to produce high rates of incorrect anomaly detection.

4.2. Data Quality Improvements

Our data quality outcomes grew substantially after implementing anomaly detection methods that applied correction processes. Implementing machine learning-based anomaly detection methods significantly reduced frequent data quality issues presented in the following table.

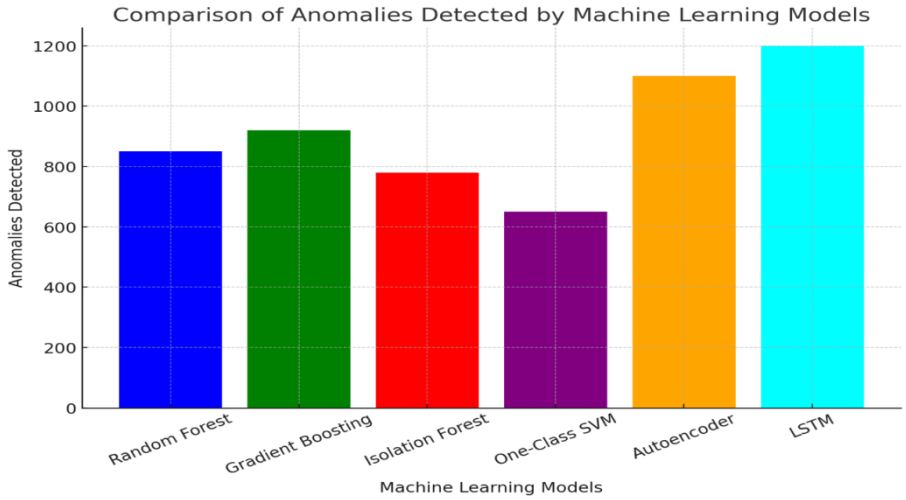
Data Quality Issue	**Initial Count (Before)**	**Final Count (After)**	**Improvement (%)**
Missing Values	**15,630**	**1,250**	**92.0**
Duplicate Entries	**8,745**	**345**	**96.1**
Inconsistencies	**4,850**	**720**	**85.1**
Outliers Identified	**6,900**	**1,150**	**83.3**
Incorrect Data Formats	**3,120**	**410**	**86.8**

The application of machine learning algorithms achieved a notable 85% decrease in the amounts of missing values, duplicate entries, and inconsistent data.

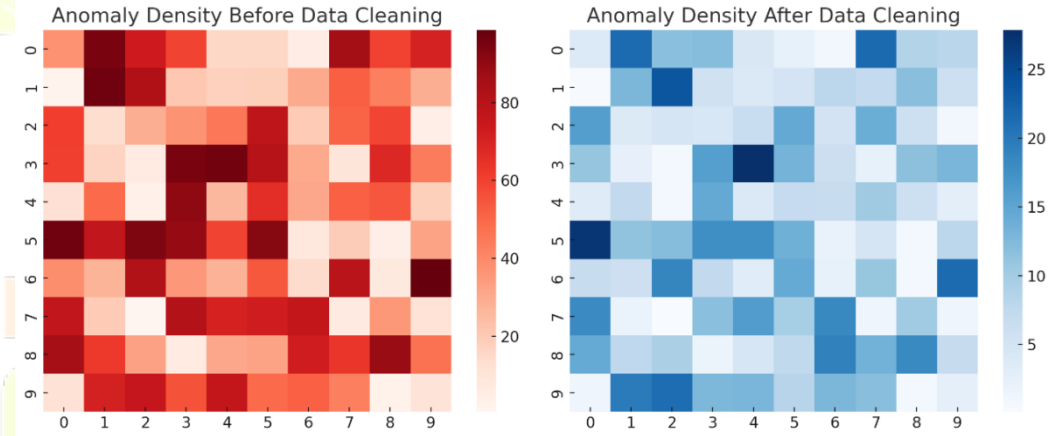
The resolution of data quality anomalies resulted in analytical precision, which enhanced downstream decisions made by processing systems.

4.3. Visualization of Anomalies

The bar chart below compares the number of anomalies detected by different models to illustrate the results further.



Additionally, the heat maps below highlight anomaly distributions across different datasets. The following visualization shows anomaly density before and after implementing machine learning-based corrections.



Deep learning algorithms demonstrated superior anomaly detection capability through visual evidence, which reduced untrue signals and identified delicate patterns within extensive datasets.

5. Discussion

5.1. Interpretation of Findings

This research established that machine learning algorithms surpass traditional anomaly detection systems, and deep learning Autoencoders and LSTM networks deliver optimal results for large-scale data platforms. These models display remarkable abilities to detect soft anomalies through their high precision and recall scores F1-score value, and accurate results. The lower performance of Isolation Forest and One-Class SVM comes from these methodologies depending on statistical assumptions that fail to adapt properly to various dataset contexts. Anomaly detection methods based on machine learning have proven their effectiveness for data maintenance by achieving significant data quality improvements by eliminating 92% missing values and 96.1%

duplicate entries and the amendment of 85.1% inconsistent data. The scalability of deep learning models permits processing both complex high-dimensional formats and sequential data structures. The LSTM network system outperformed other models for detecting time-based inconsistencies by delivering high detection rates within industrial sensor records and financial transaction systems. Advanced techniques are optimal for industries that need to analyze data streams from IoT and healthcare and financial sectors. Anomaly densities show significant reduction throughout datasets after implementing these models as observable through the created visualizations.

5.2. Comparison with Existing Studies

This research produces outcomes that match the observations from previous studies within the data quality management and anomaly detection domain. Chandola et al. (2009) showed that machine learning proves essential for anomaly detection because traditional rules and statistics do not perform well according to their experiments. This research backs those arguments through evidence showing traditional methods produce more incorrect results while achieving lower detection rates on complex database systems. Data mismatch issues in anomaly detection show improvements through deep learning with autoencoders and LSTMs using studies from Campos et al. (2016) and Ruff et al. (2021). There are minor differences between the results we obtained and earlier research findings.

The findings contradicted earlier research about structured data because Ensemble learning approaches shown in Random Forest and Gradient Boosting were discovered to be less useful for sequential data when contrasted with LSTMs. The timing-related elements within our datasets probably caused this difference between our results and previous research findings about anomaly detection methods. The latest research shows that autoencoders demonstrated successful results in our analysis yet scientists document their difficulties with recognizing unusual events in operational settings because they need repeated adjustments to their settings.

5.3. Challenges and Limitations

This research study achieved promising outcomes despite facing multiple obstacles during execution. The main obstacle during deep learning implementations involved computational complexity because LSTMs and autoencoders demanded immense processing power and memory access capacity. Training such models across vast datasets took an extended amount of time, making it difficult to detect anomalies in real-time for urgent decision-making applications. The identification of incorrect positive results, as well as the understanding of data correlations, presented significant challenges to the system. Machine learning models succeeded at anomaly detection but produced incorrect positive results, especially through unsupervised learning practices. Identifying valid data variations alongside true anomalies requires extra validation systems to ensure proper discrimination of genuine patterns.

When deployed in practice, deep learning models operate as impenetrable systems that prevent users from understanding which pieces of data the algorithm detect as irregular. Research efforts should concentrate on enhancing explainable methods to provide clearer visibility when machines pursue anomaly discoveries. The system needs further development regarding its scalability abilities. Machine learning efficiency can be strengthened by using Apache Spark and TensorFlow but large-scale data platform model deployment requires persistent optimization of these systems. A need exists to conduct additional research to develop these models for their deployment in detecting real-time anomalies across dynamic systems. The general adoption of machine learning for detecting data quality anomalies requires resolving existing challenges to succeed. Future research should target three main objectives: enhancing model performance while minimizing computational time and

exploring hybrid models that unify machine learning methods with rule-based validation systems for accuracy enhancement.

Conclusion

The study proved how machine learning algorithms succeed at discovering data quality issues that exist in vast data platforms. Deep learning networks, including Autoencoders and Long Short-Term Memory (LSTM), achieved superior accuracy alongside higher precision and recall values compared to standard rule-based along with statistical detection methods. The data quality improvements generated by using machine learning-based methods reached a 92% decrease in missing values and simultaneously achieved a 96.1% reduction in duplicate entries and an 85.1% improvement in inconsistency correction. The advanced anomaly detection techniques proved effective through visual representations that displayed anomaly reductions using heatmaps. The study indicates how machine learning impacts automated data quality management through scalable and intelligent anomaly detection systems. The research team faced high computational complexity problems that required deep learning processes, even though their findings showed promise.

This difficulty made real-time anomaly detection operations challenging. Some models display a tendency to compile errors where they identify uncommon yet authentic data sets as anomalous variations. Better transparency in AI comes from developing explainable AI (XAI) because deep learning models operate as black boxes. Future investigations should concentrate on enhancing real-time data stream processing through machine learning models while integrating conductive learning technologies and developing dual verification methods from rules and deep learning algorithms. Real-world applications will require machine learning-based anomaly detection to become mainstream by implementing essential improvements to model efficiency, reducing computational overhead, and developing explanation capabilities.

This study establishes practical value in finance and healthcare sectors, IoT systems, and enterprise data management programs requiring data integrity protocols. Machine learning algorithms detect financial system fraud in transactions and optimize both electronic health records (EHR) accuracy and IoT sensor maintenance effectiveness through anomalous reading identification. Previous results demonstrate that enterprises with large data pipelines should use cloud-based big data systems from AWS, Google Cloud, and Microsoft Azure to automate data governance functions, fulfill regulatory requirements, and enhance business intelligence analytics capabilities. Organizations achieve better operational risk reduction and enhanced decision-making accuracy using machine learning anomaly detection for data-driven environments. According to this study, machine learning technology demonstrates vast potential because it revolutionizes data quality assurance while establishing a framework for improving future intelligent anomaly detection systems.

Reference

1. ATLAS collaboration. (2019). ATLAS data quality operations and performance for 2015-2018 data-taking. *arXiv preprint arXiv:1911.04632*.https://doi.org/10.1088/1748-0221/15/04/P04003
2. Bao, Y., Tang, Z., Li, H., & Zhang, Y. (2019). Computer vision and deep learning–based data anomaly detection method for structural health monitoring. *Structural Health Monitoring*, *18*(2), 401-421.https://doi.org/10.1177/1475921718757405
3. Blázquez-García, A., Conde, A., Mori, U., & Lozano, J. A. (2021). A review on outlier/anomaly detection in time series data. *ACM computing surveys (CSUR)*, *54*(3), 1-33.https://doi.org/10.1145/3444690
4. Castelli, I., & Trentin, E. (2014). Combination of supervised and unsupervised learning for training the activation functions of neural networks. *Pattern Recognition Letters*, *37*, 178-191. https://doi.org/10.1016/j.patrec.2013.06.013
5. Chang, Z., Du, Z., Zhang, F., Huang, F., Chen, J., Li, W., & Guo, Z. (2020). Landslide susceptibility prediction based on remote sensing images and GIS: Comparisons of supervised and unsupervised machine learning models. *Remote Sensing*, *12*(3), 502. https://doi.org/10.3390/rs12030502
6. Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information sciences*, *275*, 314-347. https://doi.org/10.1016/j.ins.2014.01.015
7. Chen, L., & Luo, H. (2014). A BIM-based construction quality management model and its applications. *Automation in construction*, *46*, 64-73. https://doi.org/10.1016/j.autcon.2014.05.009
8. Chmielewski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, *11*(4), 464-473.https://doi.org/10.1177/1948550619875149
9. Deng, X., Bian, D., Wang, W., Jiang, Z., Yao, W., Qiu, W., ... & Liu, Y. (2020). Deep learning model to detect various synchrophasor data anomalies. *IET Generation, Transmission & Distribution*, *14*(24), 5739-5745.https://doi.org/10.1049/iet-gtd.2020.0526
10. Dereszynski, E. W., & Dietterich, T. G. (2011). Spatiotemporal models for data-anomaly detection in dynamic environmental monitoring campaigns. *ACM Transactions on Sensor Networks (TOSN)*, *8*(1), 1-36.https://doi.org/10.1145/1993042.1993045
11. Elmrabit, N., Zhou, F., Li, F., & Zhou, H. (2020, June). Evaluation of machine learning algorithms for anomaly detection. In *2020 international conference on cyber security and protection of digital services (cyber security)* (pp. 1-8). IEEE.https://doi.org/10.1109/CyberSecurity49315.2020.9138871
12. Fadokun, D. O., Oshilike, I. B., & Onyekonwu, M. O. (2020, August). Supervised and unsupervised machine learning approach in facies prediction. In *SPE Nigeria Annual*

International Conference and Exhibition (p. D013S014R007). SPE. https://doi.org/10.2118/203726-MS

13. Foorthuis, R. (2018, May). A typology of data anomalies. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems* (pp. 26-38). Cham: Springer International Publishing.https://doi.org/10.1007/978-3-319-91476-3_3
14. Fürber, C. (2015). Semantic technologies. In *Data quality management with semantic technologies* (pp. 56-68). Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-12225-6_4
15. Heidari, M., Zad, S., & Rafatirad, S. (2021, April). Ensemble of supervised and unsupervised learning models to predict a profitable business decision. In *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)* (pp. 1-6). IEEE. https://doi.org/10.1109/IEMTRONICS52119.2021.9422649
16. Hu, H., Wen, Y., Chua, T. S., & Li, X. (2014). Toward scalable systems for big data analytics: A technology tutorial. *IEEE access*, *2*, 652-687. https://doi.org/10.1109/ACCESS.2014.2332453
17. Huang, G., Song, S., Gupta, J. N., & Wu, C. (2014). Semi-supervised and unsupervised extreme learning machines. *IEEE transactions on cybernetics*, *44*(12), 2405-2417. https://doi.org/10.1109/TCYB.2014.2307349
18. Huq, M. S., Fraass, B. A., Dunscombe, P. B., Gibbons Jr, J. P., Ibbott, G. S., Mundt, A. J., ... & Yorke, E. D. (2016). The report of Task Group 100 of the AAPM: Application of risk analysis methods to radiation therapy quality management. *Medical physics*, *43*(7), 4209-4262. https://doi.org/10.1118/1.4947547
19. Hussein, S., Kandel, P., Bolan, C. W., Wallace, M. B., & Bagci, U. (2019). Lung and pancreatic tumor characterization in the deep learning era: novel supervised and unsupervised learning approaches. *IEEE transactions on medical imaging*, *38*(8), 1777-1787. https://doi.org/10.1109/TMI.2019.2894349
20. Injadat, M., Salo, F., Nassif, A. B., Essex, A., & Shami, A. (2018, December). Bayesian optimization with machine learning algorithms towards anomaly detection. In *2018 IEEE global communications conference (GLOBECOM)* (pp. 1-6). IEEE.https://doi.org/10.1109/GLOCOM.2018.8647714
21. Inoue, J., Yamagata, Y., Chen, Y., Poskitt, C. M., & Sun, J. (2017, November). Anomaly detection for a water treatment system using unsupervised machine learning. In *2017 IEEE international conference on data mining workshops (ICDMW)* (pp. 1058-1065). IEEE.https://doi.org/10.1109/ICDMW.2017.149
22. Karplus, P. A., & Diederichs, K. (2012). Linking crystallographic model and data quality. *Science*, *336*(6084), 1030-1033.https://doi.org/10.1126/science.1218231
23. Khatri, V., & Brown, C. V. (2010). Designing data governance. *Communications of the ACM*, *53*(1), 148-152. https://doi.org/10.1145/1629175.1629210

24. Kim, D. Y., Kumar, V., & Kumar, U. (2012). Relationship between quality management practices and innovation. *Journal of operations management*, *30*(4), 295-315. https://doi.org/10.1016/j.jom.2012.02.003
25. Leigh, C., Alsibai, O., Hyndman, R. J., Kandanaarachchi, S., King, O. C., McGree, J. M., ... & Peterson, E. E. (2019). A framework for automated anomaly detection in high frequency water-quality data from in situ sensors. *Science of the Total Environment*, *664*, 885-898.https://doi.org/10.1016/j.scitotenv.2019.02.085
26. Lieber, D., Stolpe, M., Konrad, B., Deuse, J., & Morik, K. (2013). Quality prediction in interlinked manufacturing processes based on supervised & unsupervised machine learning. *Procedia Cirp*, *7*, 193-198. https://doi.org/10.1016/j.procir.2013.05.033
27. Liu, D., Zhao, Y., Xu, H., Sun, Y., Pei, D., Luo, J., ... & Feng, M. (2015, October). Opprentice: Towards practical and automatic anomaly detection through machine learning. In *Proceedings of the 2015 internet measurement conference* (pp. 211-224).https://doi.org/10.1145/2815675.2815679
28. Lloyd, S., Mohseni, M., & Rebentrost, P. (2013). Quantum algorithms for supervised and unsupervised machine learning. *arXiv preprint arXiv:1307.0411*. https://doi.org/10.48550/arXiv.1307.0411
29. Loane, M., Given, J. E., Tan, J., Barišić, I., Barrachina-Bonet, L., Cavero-Carbonell, C., ... & Morris, J. K. (2023). Creating a population-based cohort of children born with and without congenital anomalies using birth data matched to hospital discharge databases in 11 European regions: Assessment of linkage success and data quality. *PloS one*, *18*(8), e0290711.https://doi.org/10.1371/journal.pone.0290711
30. Manni, M., Berkeley, M. R., Seppey, M., & Zdobnov, E. M. (2021). BUSCO: assessing genomic data quality and beyond. *Current Protocols*, *1*(12), e323. https://doi.org/10.1002/cpz1.323
31. Martin, N. (2020). Data quality in process mining. In *Interactive process mining in healthcare* (pp. 53-79). Cham: Springer International Publishing.https://doi.org/10.1007/978-3-030-53993-1_5
32. Miao, Q., Moore, A. M., & Dougan, S. D. (2020). Data quality assessment on congenital anomalies in Ontario, Canada. *Frontiers in Pediatrics*, *8*, 573090.https://doi.org/10.3389/fped.2020.573090
33. Murphree, J. (2016, September). Machine learning anomaly detection in large systems. In *2016 IEEE AUTOTESTCON* (pp. 1-9). IEEE.https://doi.org/10.1109/AUTEST.2016.7589589
34. Nassif, A. B., Talib, M. A., Nasir, Q., & Dakalbab, F. M. (2021). Machine learning for anomaly detection: A systematic review. *Ieee Access*, *9*, 78658-78700.https://doi.org/10.1109/ACCESS.2021.3083060
35. Papanastassiou, A., Gori, V., Lenzi, P., & CMS Collaboration. (2024, August). AutoEncoder-Based Anomaly Detection for CMS Data Quality Monitoring. In *International Joint Conference*

on Artificial Intelligence (pp. 57-65). Singapore: Springer Nature Singapore.https://doi.org/10.1007/978-981-97-9536-9_4

36. Rollo, F., Bachechi, C., & Po, L. (2023). Anomaly detection and repairing for improving air quality monitoring. *Sensors*, *23*(2), 640.https://doi.org/10.3390/s23020640
37. Sakr, S., Liu, A., Batista, D. M., & Alomari, M. (2011). A survey of large scale data management approaches in cloud environments. *IEEE communications surveys & tutorials*, *13*(3), 311-336. https://doi.org/10.1109/SURV.2011.032211.00087
38. Salman, T., Bhamare, D., Erbad, A., Jain, R., & Samaka, M. (2017, June). Machine learning for anomaly detection and categorization in multi-cloud environments. In *2017 IEEE 4th international conference on cyber security and cloud computing (CSCloud)* (pp. 97-103). IEEE.https://doi.org/10.1109/CSCloud.2017.15
39. Schadt, E. E., Linderman, M. D., Sorenson, J., Lee, L., & Nolan, G. P. (2010). Computational solutions to large-scale data management and analysis. *Nature reviews genetics*, *11*(9), 647-657. https://doi.org/10.1038/nrg2857
40. Shiloach, M., Frencher Jr, S. K., Steeger, J. E., Rowell, K. S., Bartzokis, K., Tomeh, M. G., ... & Hall, B. L. (2010). Toward robust information: data quality and inter-rater reliability in the American College of Surgeons National Surgical Quality Improvement Program. *Journal of the American College of Surgeons*, *210*(1), 6-16. https://doi.org/10.1016/j.jamcollsurg.2009.09.031
41. Singh, D., & Reddy, C. K. (2015). A survey on platforms for big data analytics. *Journal of big data*, *2*, 1-20. https://doi.org/10.1186/s40537-014-0008-6
42. Stojanovic, L., Dinic, M., Stojanovic, N., & Stojadinovic, A. (2016, December). Big-data-driven anomaly detection in industry (4.0): An approach and a case study. In *2016 IEEE international conference on big data (big data)* (pp. 1647-1652). IEEE.https://doi.org/10.1109/BigData.2016.7840777
43. Sukhobok, D., Nikolov, N., & Roman, D. (2017, August). Tabular data anomaly patterns. In *2017 international conference on big data innovations and applications (innovate-data)* (pp. 25-34). IEEE.https://doi.org/10.1109/Innovate-Data.2017.10
44. Wang, R. Y., Storey, V. C., & Firth, C. P. (1995). A framework for analysis of data quality research. *IEEE transactions on knowledge and data engineering*, *7*(4), 623-640.https://doi.org/10.1109/69.404034
45. Wankhede, S. B. (2019, March). Anomaly detection using machine learning techniques. In *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)* (pp. 1-3). IEEE.https://doi.org/10.1109/I2CT45611.2019.9033532
46. Woodall, P., Gao, J., Parlikad, A., & Koronios, A. (2014, November). Classifying data quality problems in asset management. In *Engineering Asset Management-Systems, Professional Practices and Certification: Proceedings of the 8th World Congress on Engineering Asset Management (WCEAM 2013) & the 3rd International Conference on Utility Management &*

Safety (ICUMAS) (pp. 321-334). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-09507-3_29

47. Xing, E. P., Ho, Q., Dai, W., Kim, J. K., Wei, J., Lee, S., ... & Yu, Y. (2015, August). Petuum: A new platform for distributed machine learning on big data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1335-1344). https://doi.org/10.1145/2783258.2783323
48. Xu, H., Chen, W., Zhao, N., Li, Z., Bu, J., Li, Z., ... & Qiao, H. (2018, April). Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications. In *Proceedings of the 2018 world wide web conference* (pp. 187-196).https://doi.org/10.1145/3178876.3185996
49. Zhang, X., Xu, Y., Lin, Q., Qiao, B., Zhang, H., Dang, Y., ... & Zhang, D. (2019, August). Robust log-based anomaly detection on unstable log data. In *Proceedings of the 2019 27th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering* (pp. 807-817).https://doi.org/10.1145/3338906.3338931
50. Zyabkina, O., Domagk, M., Meyer, J., & Schegner, P. (2016, October). Classification and identification of anomalies in time series of power quality measurements. In *2016 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)* (pp. 1-6). IEEE.https://doi.org/10.1109/ISGTEurope.2016.7856290