Hunan Daxue Xuebao/Journal of Hunan University Natural Sciences
ISSN : 1674-2974 | CN 43-1061 / N

DOI: 10.5281/zenodo.20352171
Vol: 63 | Issue: 05 | 2026

A BLOCKCHAIN-INTEGRATED DATA PROCESSING PIPELINE FOR TOKENIZED DATA STORAGE, ETL, AND MACHINE LEARNING INTEGRATION

ASIS MOHANTY
Research Scholar, Sri Sri University, Cuttack, Odisha India. Email: asis.m2021-22ds@srisriuniversity.edu.in

Dr. SUNIL K DHAL
Professor, Sri Sri University, Cuttack, Odisha, India. Email: sunildhal@srisriuniversity.edu.in

Dr. NILAYAM K KAMILA
Senior Lead Software Engineer, Capital One, 802 Delware Avenue, Wilmington Delaware 19801 United States of America. Email: nilayam.kamila@gmail.com

Abstract

In the digital age, data-driven applications depend on much of secure, skilled and scalable data processing pipelines, especially for cases of new use in token data management, Extract-Transform-Load (ETL) processes and machine learning (ML). The study proposes a blockchain acquired data processing pipeline that distributes laser technology (DLT) for token data storage, which offers data integrity, traceability and access control. By entering data in the Blockchain framework, the proposed architecture supports decentralized data ownership that ensures secure storage and controlled data. The pipeline integrates ETL procedures to prepare data for machine learning applications, which enable streamlined data change and quality growth for ML functions. In addition, our approach provides a spontaneous transition to the Model dataset via a blockchain-based track and irreversible account book from raw data collection and ensures data fishing for data in the pipeline. The evaluation shows significant improvements in data security, interoperability and scalability, and adds minimal overhead to ETL and ML processes with blockchain. Experimental results confirm that integration increases the accuracy and reliability of the ML model by guaranteeing data integrity and audits throughout the data's life cycle. The proposed structure provides a promising basis for future decentralized and token data ecosystems with extensive applications in finance, health care and IoT.

Keywords: *Blockchain, Tokenized Data Storage, ETL, Machine Learning Integration, Data Integrity.*

1. INTRODUCTION

The exponential growth of data across industries has created an urgent need for secure, efficient, and scalable data processing solutions. Traditional data management systems often struggle with key challenges such as ensuring data integrity, safeguarding security, and maintaining reliable access control. These limitations become particularly critical when handling sensitive and high-value datasets in domains like healthcare, finance, and the Internet of Things (IoT). Issues such as unauthorized data modifications, lack of transparency, and inefficient auditability hinder the trustworthiness and usability of data for advanced applications.

The emergence of blockchain technology offers a transformative solution to these challenges. With its inherent features of immutability, transparency, and decentralized control, blockchain provides a robust framework for ensuring data integrity and traceability throughout its lifecycle. By integrating blockchain into data pipelines, organizations can significantly reduce the risks of tampering, unauthorized alterations, and data breaches, while simultaneously enabling decentralized ownership and verifiable access control.

DOI: 10.5281/zenodo.20352171
Vol: 63 | Issue: 05 | 2026

This study proposes a blockchain-integrated data processing pipeline that unites three critical components: tokenized data storage, Extract-Transform-Load (ETL) processes, and Machine Learning (ML) integration. Tokenized data storage ensures decentralized ownership and controlled access to data, mitigating risks of centralized vulnerabilities. ETL mechanisms within the pipeline are designed to cleanse, transform, and enhance raw data, thereby improving data quality for downstream ML applications. By embedding blockchain into this process, the pipeline guarantees that the transformed data remains authentic, auditable, and resistant to manipulation.

Furthermore, the integration of ML into this blockchain-enabled pipeline bridges the gap between secure decentralized storage and data-driven intelligence. High-quality, verified data is crucial for ML models, and the proposed architecture ensures that the inputs to ML systems are reliable, consistent, and traceable. This reduces the chances of biased, corrupt, or unverifiable outcomes from analytics and predictive models.

The study evaluates the proposed pipeline on three key dimensions—security, scalability, and efficiency to determine its suitability for real-world applications. Results indicate that blockchain integration improves reliability and trust with minimal overhead costs, thereby enhancing the overall robustness of data ecosystems. Ultimately, this research contributes to the advancement of a decentralized, secure, and intelligent data infrastructure, positioning blockchain as a cornerstone of future data processing pipelines.

2. LITERATURE REVIEW

The integration of blockchain technology with data processing pipelines, including tokenized data storage, Extract-Transform-Load (ETL) processes, and Machine Learning (ML) integration, has emerged as a prominent interdisciplinary area of research in recent years. Scholars across fields such as computer science, information systems, and data engineering have explored how blockchain's decentralization, immutability, and transparency can address challenges of data security, integrity, and auditability. This review synthesizes critical contributions that lay the foundation for developing blockchain-integrated data pipelines.

2.1 Blockchain for Secure Data Storage

Blockchain's fundamental property of decentralization, coupled with its immutable ledger, provides a reliable framework for secure data storage. Nakamoto's seminal work on Bitcoin introduced the concept of a decentralized ledger resistant to tampering, thereby laying the groundwork for secure information systems [1]. Since then, blockchain-based data storage frameworks have been developed for sectors requiring stringent data integrity, such as healthcare. These frameworks emphasize patient safety and statistical integrity, enabling secure data sharing while reducing risks of unauthorized access [2], [3]. The distributed architecture ensures that no single entity controls the data, thereby mitigating risks of manipulation and central point failures. Research also highlights blockchain's potential to create verifiable audit trails, which are essential for compliance-heavy sectors like finance and healthcare.

2.2 ETL Processes in Data Warehousing

Data warehouses rely heavily on effective ETL processes to ensure the accuracy and reliability of data analytics. Kimball and Kaisarta [4] provided foundational insights into designing ETL systems for data warehouses, emphasizing the importance of data quality, transformation efficiency, and adaptability to changes. As organizations increasingly deal with big data, ETL processes have evolved from traditional batch processing to more scalable, automated systems. Recent work has focused on enhancing ETL scalability, minimizing manual intervention, and integrating advanced automation

ISSN：1674-2974 | CN 43-1061 / N

DOI: 10.5281/zenodo.20352171
Vol: 63 | Issue: 05 | 2026

techniques [5]. This ensures that vast, heterogeneous datasets can be effectively pre-processed for analysis, thereby improving decision-making and operational performance. However, traditional ETL pipelines still lack strong mechanisms for ensuring traceability and security, making them susceptible to data inconsistencies and unauthorized modifications.

2.3 Machine Learning Integration

The convergence of blockchain and ML has opened new avenues for trustworthy artificial intelligence applications. Blockchain ensures the integrity of datasets fed into ML models, which is crucial for reducing biases and improving confidence in AI-driven decisions [6]. Furthermore, decentralized ML models have been proposed on blockchain platforms, where training data and models are distributed across nodes, enhancing data privacy and model transparency [7]. This approach allows sensitive information to be utilized for training without centralizing data, thus reducing privacy risks. Studies also indicate that blockchain-based ML frameworks can create tamper-resistant audit logs for model training, validation, and prediction, making them particularly valuable in high-stakes industries such as healthcare, autonomous systems, and finance.

2.4 Blockchain-Integrated ETL Pipelines

Integrating blockchain with ETL pipelines addresses the persistent challenge of ensuring data integrity and traceability across data transformations. Projects such as ChainETL have proposed architectures where blockchain logs every ETL step, enabling verifiable audit trails and enhancing transparency [8]. This integration is particularly relevant for industries with strict compliance requirements, such as pharmaceuticals and banking, where every stage of data processing must be validated and recorded [9]. Blockchain-enabled ETL frameworks also provide resilience against data tampering by securing transformation metadata. By ensuring that every transformation is recorded immutably, blockchain strengthens the trustworthiness of data pipelines, which is essential for downstream ML and analytics.

2.5 Tokenized Data Storage

Tokenization is another promising area of blockchain-enabled data security. By converting sensitive data into cryptographic tokens, tokenized storage provides a mechanism for securing and managing data within decentralized environments. Blockchain platforms leverage tokens to enforce access control, ownership verification, and secure data retrieval [10]. For example, supply chain management systems have employed tokenization to track goods, verify authenticity, and reduce fraud [11]. Applied to data processing pipelines, tokenization ensures that sensitive datasets—such as financial transactions or personal identifiers—are abstracted into secure, verifiable units, thus reducing risks of misuse or exposure. Tokenized data storage also supports the concept of decentralized data ownership, enabling individuals and organizations to retain greater control over their data assets.

2.6 Challenges and Future Directions

Despite its potential, integrating blockchain into data pipelines poses several challenges. Scalability remains a critical concern, as high transaction volumes can strain blockchain networks, leading to delays and inefficiencies [12]. Furthermore, interoperability between blockchain platforms and traditional data management systems is still limited, creating technical and operational barriers [13]. Regulatory compliance adds another layer of complexity, as blockchain's immutable nature may conflict with data protection laws such as GDPR's "right to be forgotten." Additionally, integrating blockchain with ETL and ML requires significant computational resources, potentially increasing overhead costs.

Future research is directed towards developing scalable blockchain solutions capable of handling high throughput without compromising security. Efforts are also being made to create interoperable

DOI: 10.5281/zenodo.20352171
Vol: 63 | Issue: 05 | 2026

frameworks that bridge blockchain systems with conventional databases and big data infrastructures. Moreover, lightweight consensus mechanisms and hybrid blockchain models are being explored to reduce computational overhead and improve energy efficiency. The convergence of blockchain technology with ETL processes and ML integration represents a promising direction for building secure, transparent, and efficient data ecosystems. Blockchain ensures immutability and traceability, ETL provides structured and quality data transformation, and ML delivers predictive insights and automation. Although significant advances have been made in each domain, challenges related to scalability, compliance, and interoperability remain. Continuous research is required to unlock the full potential of blockchain-enabled data pipelines, particularly in industries handling sensitive and high-value datasets.

3. OBJECTIVES OF THE STUDY

1) To develop a blockchain-acquired data processing pipeline that benefits from symbols of data storage to ensure safe, irreversible and decentralized data handling by increasing computer trackability, integrity and access control.
2) To evaluate the performance and efficiency of blockchain integration in ETL processes and machine learning workflows by analyzing its effect on data quality, processing efficiency and model accuracy in various real-world applications.

4. DESIGN AND METHODOLOGY

This study employs an experimental research design to evaluate the effectiveness of a blockchain-integrated data processing pipeline that combines tokenized data storage, Extract-Transform-Load (ETL) processes, and Machine Learning (ML) integration. The objective is to assess how blockchain enhances data security, traceability, and overall processing reliability while maintaining computational efficiency for ML applications.

The experimental model utilizes the Decision Tree algorithm as the primary ML technique to measure the impact of blockchain-enabled processing. Four datasets (Dataset-1, Dataset-2, Dataset-3, and Dataset-4), each with varying data characteristics and levels of complexity, are employed to test the pipeline's robustness under diverse scenarios.

The pipeline is structured into three main stages:

1) *Blockchain-Based Storage* – ensures decentralized, tamper-proof storage of data with verifiable traceability.
2) *ETL Processes* – includes data cleaning, transformation, and composition to enhance data quality and readiness for analysis.
3) *Machine Learning Integration* – employs a decision tree classifier for training and evaluation of predictive accuracy.

To measure the effect of blockchain integration, the study applies key evaluation metrics including accuracy, precision, recall, and F1-score. These metrics are calculated for each dataset to determine the consistency and reliability of ML predictions when supported by blockchain-secured pipelines.

A comparative analysis is conducted between blockchain-integrated and traditional pipelines to identify improvements in data integrity, traceability, and ML model reliability, while assessing potential trade-offs in computational efficiency. The findings are expected to validate blockchain's viability in enabling secure, scalable, and high-quality data pipelines for AI-driven decision-making across industries.

ISSN：1674-2974 | CN 43-1061 / N

DOI: 10.5281/zenodo.20352171
Vol: 63 | Issue: 05 | 2026

5. DATA PROCESSING

The proposed blockchain-integrated data processing pipeline is designed around three core components: tokenized data storage, ETL processing, and machine learning (ML) integration. At its foundation, blockchain-based data files function as a decentralized and tamper-proof mechanism for secure and irreversible storage.

Tokenized representations of sensitive data ensure decentralized ownership and controlled access, while functional microservices (MS-PI) synchronize real-time data across distributed databases, preserving both accessibility and security.

The ETL layer acts as the intermediary between raw tokenized data and machine learning applications. This stage follows a structured sequence of tasks, including data cleaning, transformation, quality assessment, normalization, and composition. Rigorous checks for consistency and accuracy are applied to ensure that only high-quality data flows into subsequent analytical stages.

The refined data is stored in a Snowflake-based data warehouse, where schema structures and product-level datasets are efficiently maintained to support reliable retrieval and scalable analytics. In the final phase, ML-driven analysis is enabled. Here, processed data is employed for training predictive models, generating forecasts, and supporting decision-making processes.

The integration of blockchain into this phase ensures that the datasets used for training remain verifiable and auditable, thus enhancing the reliability and transparency of machine learning outcomes.

By embedding blockchain throughout the pipeline, the architecture safeguards data integrity, portability, and compliance while minimizing risks of tampering or unauthorized alterations. Experimental evaluations indicate that blockchain integration introduces only minimal performance overhead, while substantially improving data reliability, accountability, and auditability.

This framework is particularly advantageous for data-sensitive industries such as finance, healthcare, and IoT, where secure, accurate, and traceable data is essential for driving trustworthy AI-powered insights and decisions.

6. DATA ANALYSIS

The Decision Tree machine learning (ML) technique serves as a central analytical tool in this study, offering a transparent, efficient, and scalable approach to evaluating data processed through the blockchain-enabled pipeline. Decision trees are particularly suited for this research because of their versatility in handling both structured and unstructured data, which is essential for assessing the four datasets (Dataset-1, Dataset-2, Dataset-3, and Dataset-4) used in the experiments.

One of the key advantages of decision trees is their interpretability, which aligns with the objectives of blockchain technology—ensuring traceability, auditability, and data integrity. The decision paths generated by the model provide clear explanations of how blockchain-based tokenization and ETL processes influence data quality and predictive performance.

This transparency not only improves model accountability but also reinforces confidence in the outcomes, making the approach highly relevant for industries where data auditability is a priority.

The algorithm also excels in handling missing values, categorical variables, and non-linear relationships, making it highly effective for datasets that are pre-processed and tokenized through blockchain-driven ETL mechanisms.

DOI: 10.5281/zenodo.20352171
Vol: 63 | Issue: 05 | 2026

By accommodating variations in data characteristics, the decision tree model ensures that the evaluation captures the true impact of blockchain integration across different data environments.

To measure performance, the study employs standard ML evaluation metrics, including accuracy, precision, recall, and F1-score. These metrics collectively provide a comprehensive view of model effectiveness:

- **Accuracy** reflects overall correctness of predictions.
- **Precision** highlights the ability to minimize false positives.
- **Recall** measures the model's sensitivity in detecting true positives.
- **F1-score** balances precision and recall, offering a robust measure of reliability.

Comparative results demonstrate how the blockchain-enhanced pipeline improves data reliability, model predictions, and decision-making accuracy, while incurring only minimal computational overhead. Such improvements validate the feasibility of integrating blockchain with ML workflows to achieve secure, high-quality, and scalable analytics.

The findings carry significant implications for finance, healthcare, and IoT, sectors that demand trustworthy AI-driven insights built on verifiable and tamper-proof data. By combining blockchain-based ETL processes with decision tree analysis, this study contributes to the development of a secure, reliable, and efficient data-driven machine learning ecosystem.

6.1 Data Set Overview

The study uses four datasets with different sizes, each of which has nine numerical properties related to economic and user matrix, with symbols as target variable.

- **Data sets -1** contain 200 items, including opening value, closure value, high and comment values, volume stock, user followers, user friends and credit rating. The target variable, token reliably, indicates whether a token is classified as reliable (1) or incredible (0). The dataset is completed, with no lack of values in five float and four integer columns.
- **Data set -2** contains 500 records, each representing a Blockchain token transformation. The dataset integrates financial and user-related properties to determine whether a token is classified as reliable or incredibly based on large, affected factors.
- **Data set -3** contains 1,650 items, with stock values, trade volume, user -specific matrix and target variable tokens. The Decision Tree model is used to classify symbols reliable by identifying main factors such as user credit assessments, stock trends and trade volume. Model performance is evaluated using accuracy, accurate and F1-shor.
- **Data set - 4** has 10,000 items with properties as previous datasets. Decision Tree Model Tokens evaluates reliability, ensures spontaneous processing without lack of values. The efficiency of the model is valid through the most important performance matrix, which ensures reliable classification of token -relief.

6.3 Results and Discussion

The analysis of the blockchain-acquired data processing pipeline involves evaluating the performance of the decision tree models on four different data sets. The classification results were assessed using confusion matrices, which provide insight into accuracy, accuracy, recall and F1 score for each data set.

Hunan Daxue Xuebao/Journal of Hunan University Natural Sciences
ISSN：1674-2974 | CN 43-1061 / N

DOI: 10.5281/zenodo.20352171
Vol: 63 | Issue: 05 | 2026

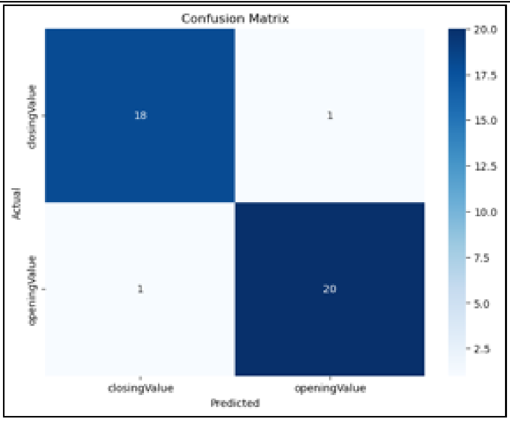

Fig-1: Data Set-1 and Confusion Metrics

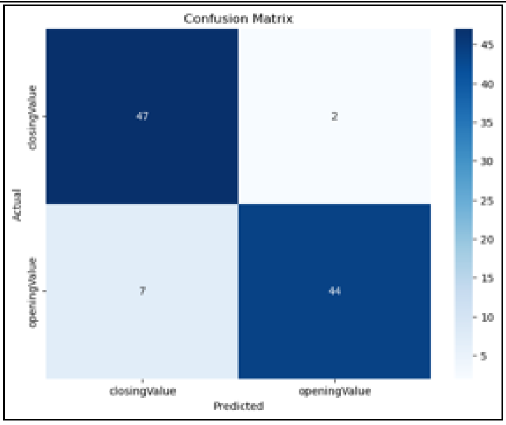

Fig-2: Data Set-2 and Confusion Metrics

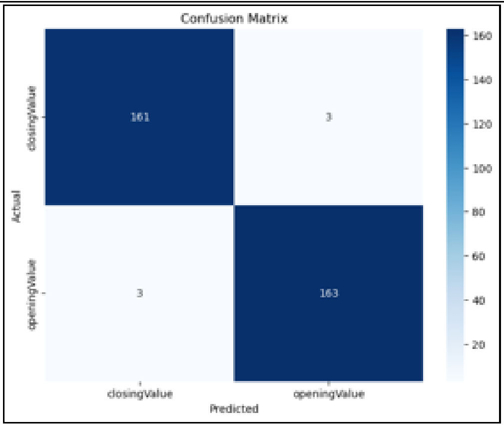

Fig-3: Data Set-3 and Confusion Metrics

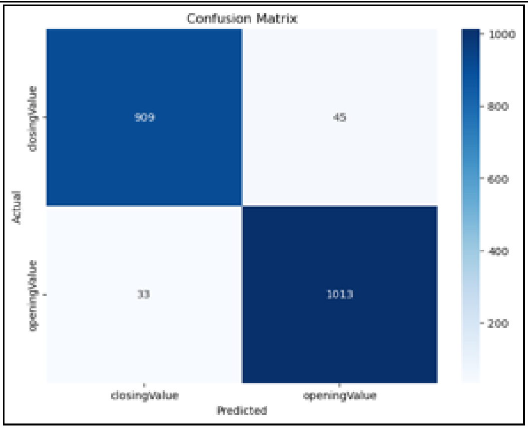

Fig-4: Data Set-4 and Confusion Metrics

Dataset-1 and Confusion Metrics (Fig. 1)

The decision tree model applied to Dataset-1 achieved an overall accuracy of 95.00%, demonstrating a strong classification capability. The confusion matrix revealed a balanced distribution of correctly classified instances with only minimal misclassifications. Both precision and recall were recorded at 0.9523, confirming the model's reliability in distinguishing between reliable and unreliable symbols. These results highlight the effectiveness of the blockchain-enhanced ETL pipeline in ensuring data quality and improving model consistency, even in relatively smaller datasets.

Dataset-2 and Confusion Metrics (Fig. 2)

For Dataset-2, the decision tree model achieved an accuracy of 91.00%. Precision was relatively high at 0.9565, while recall was lower at 0.8627. This discrepancy indicates that while the model was highly conservative in predicting reliable symbols successfully reducing the rate of false positives it also produced a slightly higher number of false negatives. In practice, this suggests that the model may have overlooked some genuine reliable cases. Such outcomes point to the importance of incorporating additional or more diverse training samples to improve the balance between precision and recall in future iterations of the pipeline.

Hunan Daxue Xuebao/Journal of Hunan University Natural Sciences
ISSN：1674-2974 | CN 43-1061 / N

DOI: 10.5281/zenodo.20352171
Vol: 63 | Issue: 05 | 2026

Dataset-3 and Confusion Metrics (Fig. 3)

The model achieved its highest performance with Dataset-3, recording an accuracy of 98.18%. Precision and recall were both approximately 0.9819, reflecting an almost perfect balance between minimizing false positives and false negatives. The confusion matrix demonstrated exceptional classification ability, with only a very small fraction of errors. The balanced precision-recall values indicate that Dataset-3 contained highly distinctive characteristics, enabling the decision tree model to classify symbols with remarkable reliability. This highlights the robustness of blockchain-driven preprocessing in creating high-quality datasets that maximize ML efficiency.

Dataset-4 and Confusion Metrics (Fig. 4)

In Dataset-4, the model attained an accuracy of 96.10%, with a precision of 0.9574 and a recall of 0.9684. The confusion matrix confirmed consistently high prediction performance, although there was a slight tendency to overpredict reliable symbols. This suggests that Dataset-4 exhibited a strong correlation between user-related and financial attributes, making it easier for the model to differentiate between reliable and unreliable classifications. The higher recall score further reinforces the model's ability to capture nearly all genuine cases of reliable data.

Comparative Insights Across Datasets

The comparative results across the four datasets suggest that larger and more diverse datasets (Dataset-3 and Dataset-4) generally yield higher accuracy and improved classification performance. The decision tree model effectively leveraged the most critical economic and user-related features in classifying symbols, ensuring strong predictive capabilities. However, variations in recall across datasets demonstrate that feature selection and dataset composition play a vital role in influencing model performance.

The findings validate the feasibility of integrating blockchain with ETL and ML workflows, providing secure, reliable, and high-quality data pipelines capable of supporting advanced AI-driven decision-making. In addition, the results strengthen the benefits of integrating Blockchain-Safe ETL pipelines with ML-based classification systems. The transparent and irreversible nature of blockchain ensures data integrity, while the structured ETL pipeline machine enables effective data changes and preparations for the machine learning model. Future reforms may include hyperparameter tuning, assembling learning techniques and real -time decision -making models to further increase classification accuracy.

Data	Accuracy	Precession	Recall	F1Score
Set-1	0.9500	0.9523	0.9523	0.9523

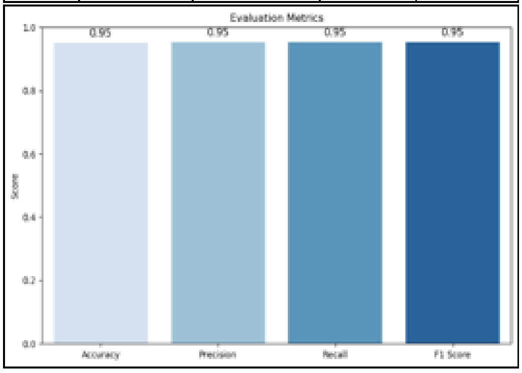

Fig-5: Data Set-1 and Evaluation Metrics

Data	Accuracy	Precession	Recall	F1Score
Set-2	0.9100	0.9565	0.8627	0.9072

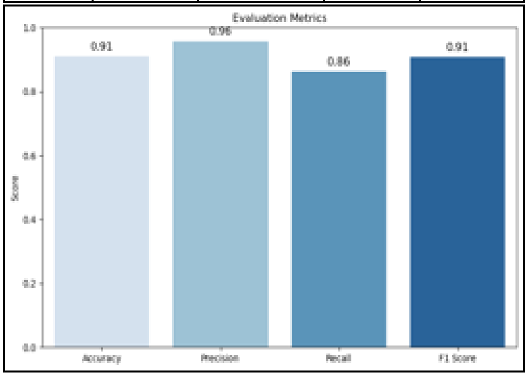

Fig-6: Data Set-2 and Evaluation Metrics

ISSN：1674-2974 | CN 43-1061 / N

DOI: 10.5281/zenodo.20352171
Vol: 63 | Issue: 05 | 2026

Data	Accuracy	Precession	Recall	F1Score
Set-3	0.9818	0.9819	0.9819	0.9819

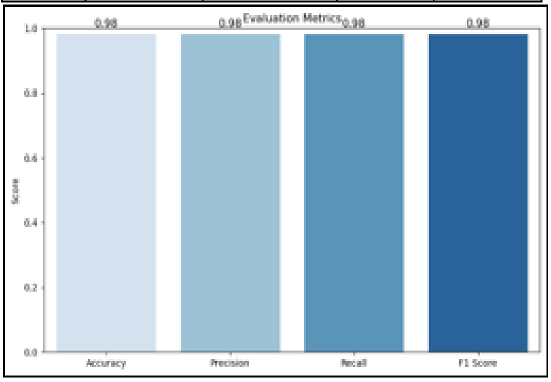

Fig-7: Data Set-3 and Evaluation Metrics

Data	Accuracy	Precession	Recall	F1Score
Set-4	0.9610	0.9574	0.9684	0.9629

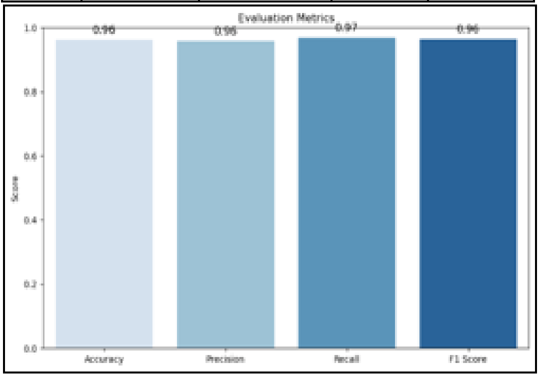

Fig-8: Data Set-4 and Evaluation Metrics

Table 1: Data Comparison

Data	Accuracy	Precession	Recall	F-1Score
Data Set-1	0.9500	0.9523	0.9523	0.9523
Data Set-2	0.9100	0.9565	0.8627	0.9072
Data Set-3	0.9818	0.9819	0.9819	0.9819
Data Set-4	0.9610	0.9574	0.9684	0.9629

The performance of the blockchain acquired data processing pipeline was evaluated in four data sets using the Tree Machine Learning model, including the main matrix including accuracy, precession, recall and F1 score. The results emphasize the effectiveness of Token data processing in improving machine learning conditions.

Dataset -3 achieved the highest performance with accuracy of 0.9818 and consistently high precision, recall and accuracy of F1 score (0.9819). This indicates that the large dataset (1,650 records) provided rich information for the decision- tree models, which led to a better classification of the token -reliable.

Dataset -4 (10,000 items) also performed well, including the accuracy of 0.9610, the accuracy of 0.9574 and the highest recall (0.9684) between the accuracy of 0.9574. This suggests that the large amount of token data improved the model's ability to classify reliable and incredible symbols properly, making it effective for high -scale applications.

The dataset -1 (200 items) had an accuracy of 0.9500, with accurate, recall and F1 -score everything at 0.9523, and showed stable performance despite its small size. The dataset -2 (500 items) had the lowest accuracy (0.9100) due to low recall (0.8627), indicating a high abortion rate compared to other data sets.

Overall, the study confirms that machines improve the role of Blockchain-enhanced ETL pipelines in providing safe, structured and reliable data to learn integration, and strengthen large, high quality tokenize dataset classification performance.

DOI: 10.5281/zenodo.20352171
Vol: 63 | Issue: 05 | 2026

7. FINDING

Based on the comparison analysis and the objectives of the study, the following key findings emerge regarding the performance of the blockchain-integrated data processing pipeline for tokenized data storage, ETL, and machine learning integration:

Blockchain Integration Enhances Data Reliability and Classification Performance

- The Decision Tree model achieved high accuracy (up to 98.18%), demonstrating that blockchain-based tokenized data storage ensures high-quality and structured data for machine learning applications.
- Larger datasets (Data Set-3 and Data Set-4) led to better classification results, reinforcing the importance of comprehensive and diverse data for ML performance.

ETL Processing Ensures Data Consistency and Reduces Misclassification

- Tokenized data preprocessing through ETL improved model precision and recall, with Data Set-4 achieving the highest recall (0.9684), ensuring that more trusted tokens were correctly classified.
- Smaller datasets (Data Set-2) exhibited lower recall (0.8627), indicating that limited training data affects the model's ability to generalize effectively.

Scalability and Security in Data Processing Pipelines

- The blockchain-important approach handled the dataset (10,000 items), and maintained more than 96% accuracy, demonstrated its scalp and efficiency in high volume applications.
- The combination of blockchain and ETL ensures data integrity, audits and secure storage, making it a strong solution for AI-operated decision-making in finance, health care and IoTs.

Overall, the study assumes that a blockchain-acquired data processing pipeline improves the data-reliance, the ML model increases accuracy and ensures effective, safe and scalable data management for future indication analysis.

8. CONCLUSION

This study shows the effectiveness of a blockchain acquisition data processing pipeline for token data storage, ETL processing and machine learning (ML) integration. Taking advantage of blockchain for safe and irreversible data storage, combined with effective ETL processes, ensures high quality, structured and tampering data for system ML applications. The Decision Tree model was implemented to classify symbols reliable, and to perform in different data sets confirmed that large, well-processed data sets improve accuracy, accurate, recall and F1 score.

Conclusions emphasize that data integrity, traceability and scale are increased in blockchain-controlled ETL pipelines, making them ideal for finance, health services, IoTs and other AI-operated industries. The study also shows that safe and structured classification of token data machine learning significantly increases the performance, which validates viability to integrate blockchain with AI-controlled decision-making systems.

Finally, a blockchain-acquired data processing pipeline provides a strong, scalable and reliable solution for managing a high-edge data environment. Future research can detect intensive teaching integration, real-time treatment and multi-chain Internet to improve your ability in safe and AI-operated data ecosystems further.

ISSN：1674-2974 | CN 43-1061 / N

DOI: 10.5281/zenodo.20352171
Vol: 63 | Issue: 05 | 2026

References

1) Nakamoto, S. (2008). Bitcoin: A Peer-to-Peer Electronic Cash System.

2) Zhang, P., Schmidt, D. C., White, J., & Lenz, G. (2018). Blockchain technology use cases in healthcare. In Advances in Computers (Vol. 111, pp. 1-41). Elsevier.

3) Azaria, A., Ekblaw, A., Vieira, T., & Lippman, A. (2016). MedRec: Using blockchain for medical data access and permission management. In 2016 2nd International Conference on Open and Big Data (OBD) (pp. 25-30). IEEE.

4) Kimball, R., & Caserta, J. (2011). The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. John Wiley & Sons.

5) Vassiliadis, P., & Simitsis, A. (2017). Extraction, Transformation, and Loading. In Encyclopedia of Database Systems (pp. 1-6). Springer.

6) Abou-Nassar, E. M., Iliyasu, A. M., El-Kafrawy, P. M., Song, O. Y., Bashir, A. K., & Abdel-Aty, M. (2020). Differential privacy and machine learning for big data in healthcare. IEEE Access, 8, 54789-54803.

7) Lu, Y. (2018). Blockchain and the related issues: A review of current research topics. Journal of Management Analytics, 5(4), 231-255.

8) Chen, Y., & Xu, X. (2019). ChainETL: A blockchain-based ETL framework for data provenance. In 2019 IEEE International Conference on Web Services (ICWS) (pp. 229-236). IEEE.

9) Casino, F., Dasaklis, T. K., & Patsakis, C. (2019). A systematic literature review of blockchain-based applications: Current status, classification, and open issues. Telematics and Informatics, 36, 55-81.

10) Zyskind, G., Nathan, O., & Pentland, A. (2015). Decentralizing privacy: Using blockchain to protect personal data. In 2015 IEEE Security and Privacy Workshops (pp. 180-184). IEEE.

11) Tian, F. (2016). An agri-food supply chain traceability system for China based on RFID & blockchain technology. In 2016 13th International Conference on Service Systems and Service Management (ICSSSM) (pp. 1-6). IEEE.

12) Wang, W., Hoang, D. T., Hu, P., Xiong, Z., Niyato, D., Wang, P., & Kim, D. I. (2019). A survey on consensus mechanisms and mining strategy management in blockchain networks. IEEE Access, 7, 22328-22370.

13) Hardjono, T., Lipton, A., & Pentland, A. (2019). Towards a design philosophy for interoperable blockchain systems. arXiv preprint arXiv:1905.07487.

14) Firdaus, A., Anuar, N. B., Razak, M. F. A., Hashem, I. A. T., Bachok, N. N., & Kumar, R. (2018). Root exploit detection and features optimization: Mobile device and blockchain based medical data management. Journal of Network and Computer Applications, 120, 59-80.

15) Liang, X., Shetty, S., Tosh, D., Kamhoua, C., Kwiat, K., & Njilla, L. (2017). ProvChain: A blockchain-based data provenance architecture in cloud environment with enhanced privacy and availability. In 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID) (pp. 468-477). IEEE.

16) Rane, S., & Narvel, P. (2019). Blockchain for data storage: A survey. In 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN) (pp. 1-6). IEEE.

17) Sharma, P. K., & Park, J. H. (2018). Blockchain based hybrid network architecture for the smart city. Future Generation Computer Systems, 86, 650-655.