SisInfo
JURNAL SISTEM INFORMASI DAN INFORMATIKA

e-ISSN: 2655-867X
p-ISSN: 2655-8661
Volume 8 No. 1 February 2026

Systematic Literature Review: Machine Learning Algorithm Performance Evaluation of Extract-Transform-Load

Muhammad Faisal Ashshidiq, Mohammad Nurkamal Fauzan
Informatics Engineering, Vocational School, Universitas Logistik dan Bisnis Internasional
Email: sidiqfaisal30@gmail.com; m.nurkamal.f@ulbi.ac.id

Accepted:	Accepted After Revision:	Published:
16 July 2025	5 February 2026	27 February 2026

Abstract

The exponential growth of data in the digital era poses significant challenges for effective data utilization. The Extract, Transform, Load (ETL) process is the foundation for preparing large-scale, unstructured data from various sources (NoSQL databases, log files) for analysis in a data warehouse. However, handling complex data structures such as nested arrays in MongoDB is a major obstacle during the transformation phase. In addition, the purpose of the transformation process is to maintain data quality and integrity. This crucial need requires a robust mechanism for anomaly detection to identify unusual patterns or events that indicate data corruption or system errors. The process of handling system errors requires analyzing nested array data structures using relevant machine learning algorithms for anomaly detection. This literature study is expected to provide valuable insights and identify relevant algorithms in data anomaly detection after the ETL process.

Keywords: Machine Learning, Anomaly Detection, Nested Array, SLR

1 INTRODUCTION

The digital age, marked by rapid data growth known as Big Data, poses significant challenges for organizations in managing and using data effectively to support evidence-based decisions. Various industries currently collect large amounts of data from diverse sources, including relational databases and NoSQL systems such as MongoDB. However, raw data is typically unstructured, inconsistent, and contains redundancies and missing values, making it difficult for advanced analysis.

To resolve these issues, organizations implement the Extract–Transform–Load (ETL) process as a core part of modern data architecture. The ETL process involves three main steps: (1) Extract, which retrieves data from sources such as MongoDB in JSON format; (2) Transform, which includes schema modifications, normalization, nested array splitting, and integrating data relationships; and (3) Load, which involves loading the transformed data into a data warehouse for analytics and reporting.

MongoDB, as a document-oriented database, can store complex data structures, including nested arrays and hierarchical documents. Its flexibility offers storage advantages, but the complexity of these structures increases computational load during the ETL transformation stage. Processing errors can affect data quality after ETL, so evaluating data quality using metrics such as null value percentage, attribute inconsistency, duplication, and outliers is necessary. Indicators of poor data quality suggest that the ETL process has failed to meet business analytics needs.

Research and analysis clearly indicate that a mature data management strategy is essential. Unsupervised machine learning (ML) methods are used to automate ETL performance evaluation and detect anomalies in data quality metrics. Techniques such as Isolation Forest, Local Outlier Factor (LOF), and One-Class Support Vector Machine (OC-SVM) are used to identify abnormal patterns in large-scale unlabeled data. Using ML improves anomaly detection accuracy while maintaining ETL system scalability and reducing processing delays in production.

However, existing research shows that the most popular algorithms are not always the most effective in every scenario. Some studies choose certain methods for ease or popularity, while others highlight the superior performance of alternative algorithms under specific conditions, such as high dimensionality or concept drift. Therefore, a systematic approach is needed to differentiate between how often algorithms are used and

how effectively they are based on comprehensive evaluation criteria.

Performance evaluation of anomaly detection during ETL processing cannot rely on a single metric, such as accuracy. This study adopts an integrated approach, considering three key dimensions: (1) detection accuracy, including precision, recall, and F1-score; (2) scalability, involving computation time, ability to handle large data, and compatibility with real-time ETL; and (3) interpretability, which assesses how well data practitioners can understand detection results to support operational decisions.LO

High anomaly-detection accuracy can be misleading, especially with highly imbalanced data where anomalies are rare compared to normal data. Results show models can achieve high accuracy simply by predicting the majority of class but fail to detect critical issues that indicate data quality problems. Hence, the literature increasingly emphasizes recall-based metrics, false positive rates, and the business impact of detection errors.

This journal presents a Systematic Literature Review (SLR) following PRISMA guidelines to examine research published from 2019 to 2025 on applying machine learning to ETL evaluation and optimization, particularly for semi-structured data and nested arrays. The article selection process aligns with the PRISMA flow diagram, covering identification, screening, eligibility, and final inclusion. As part of the review, several selected journal articles are summarized in a table, including authors, publication year, dataset types, algorithms used, and main contributions:

Table 1. Summary table of selected articles

Author	Year	Dataset	Algorithm	Contribution
Marcelli E, Barbariol T, Susto G	2022	Sensor data (Node-RED),	Isolation Forest	High AUC (0.997), F1 (95.29) active learning for anomaly detection [1].
Usman N, Utami E, Hartanto A	2023	Credit transactions.	LOF	Accuracy 97.6%, F1 (97). Density-based outlier scoring [2].
Usman N, Utami E, Hartanto A	2023	High Dimensional Sensor data	OC-SVM	AUC 0.9, TPR 0.9. efficient for high-dim anomalies [2].

The SLR also includes a dedicated subsection on quality assessment of the studies, considering clarity of methodology, reproducibility, dataset relevance, and transparency in reporting evaluation metrics. The approach ensures that the synthesis is both quantitative and methodologically solid.

Based on the literature review, the study identifies open research gaps such as limited exploration of hybrid ML-rule-based ETL approaches that combine machine learning flexibility with deterministic rule reliability, the scarcity of studies on online anomaly detection in streaming ETL environments especially for real-time settings with concept drift and the lack of integration of Explainable Artificial Intelligence (XAI) methods to clarify the causes of anomalies in ETL processes, helping practitioners to follow up effectively.

Therefore, this systematic review aims to provide an overview of research trends, algorithm effectiveness, evaluation criteria, and future directions for applying machine learning to improve data quality and Extract-Transform-Load (ETL) system performance.

1.1 Research Questions

There are several research questions in this study, among others:

1. RQ1: What is the most effective Machine Learning method for anomaly detection in nested array data after the Extract-Transform-Load process?
2. RQ2: How is the implementation and performance evaluation of machine learning algorithms in optimizing the Extract-Transform-Load process on large-scale data?

2 LITERATURE REVIEW

2.1 ETL Process (Extract, Transform, Load)

The ETL process is a fundamental methodology in data warehousing that consists of three stages: extracting data from MongoDB in JSON format (Extract), transforming the data schema, and modifying some related data for analysis purposes (Transform), and the transformed data is loaded into the data warehouse (Load) for further analysis as reporting [3]. Data with large volumes is known as “Big Data”. To handle such

large amounts of data, an ETL process is needed to make data processing easier, faster, and automated, using Apache Airflow [4].

2.2 Machine Learning

Machine Learning (ML) is a technology to address various challenges, especially in the final output of the ETL process. ML algorithms can automate several aspects of data transformation, proactively identify anomaly patterns, and predict potential issues [5]. Machine learning has two types of approaches: supervised and unsupervised. Supervised machine learning requires labeled data to train the model, whereas unsupervised machine learning does not, as its approach focuses on discovering patterns within the data independently [6].

2.3 MongoDB

MongoDB is a leading NoSQL database and a choice for storing structured and semi-structured data. MongoDB is one of the sources used for analysis to obtain business information within an organization [7] MongoDB's unique ability to store nested arrays provides flexibility but also presents challenges in the ETL process.

Research conducted by [8] shows that the schema-less MongoDB data structure requires an adaptive ETL approach to handle the variations in dynamic data structures [9] A comparative study comparing ETL performance between relational databases. The main challenges in the ETL process using data from MongoDB include denormalization, handling nested objects, and query optimization for unstructured data [10] Proposing that MongoDB's complex array improves ETL performance by up to 40% [11], [12].

2.4 Anomaly Detection

Anomaly detection is the process of identifying patterns that do not conform to normal expectations. Data anomalies can include unreasonable values, unusual patterns, or data inconsistencies that can affect the quality of analysis [12]. Anomaly detection after the ETL process becomes crucial to maintain the integrity of the processed data [13]

3 RESEARCH METHODS

3.1 Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)

The Preferred Reporting Items for Systematic Review and Meta-Analyses (PRISMA) method is a systematic and structured framework for conducting literature reviews [14]. The process includes the identification, evaluation, and synthesis of results from several previous studies in a systematic and structured manner [14]. The approach is chosen as a review of Machine Learning algorithms in optimizing the Extract-Transform-Load (ETL) process, especially in handling complex data such as nested arrays and anomaly detection [3]. In this study, we followed the PRISMA guidelines to review various machine learning algorithms, reducing potential bias and ensuring the objectivity of the literature review process through transparent and systematic reporting [14].

3.2 Stages of Systematic Literature Review

This research involves several stages that must be carried out when studying and summarizing previous research to provide an overview and guide the process from planning to final reporting of results.

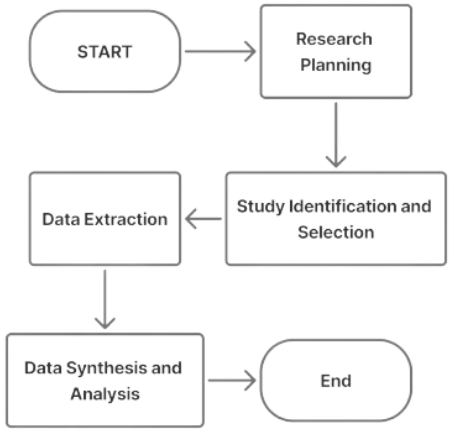

Figure 1. Flowchart Systematic Literature Review

Figure 1 shows the planning diagram for the systematic literature review conducted to examine and summarize several journal articles from 2019 to 2025. There are some discussions in the systematic literature review chart to understand the process carried out. Here is the discussion of the flowchart:

1. Formulation of research questions
   This research is guided by two Research Questions (RQs) designed to explore in depth the role of machine learning algorithms in evaluating the performance of the Extract-Transform-Load (ETL) process. Specifically related to anomaly detection in complex data structures.

2. Inclusion and Exclusion Criteria
   This process ensures that the results are relevant and that the quality of the studies to be analyzed is high. A set of inclusion and exclusion criteria is established. The inclusion and exclusion criteria serve as filters in the journal article selection

process. Additionally, the researcher will focus on the main objective, which is literature relevant to the predetermined main question. The following are the objectives of the research, among others:

a. Inclusion Criteria
Relevant journal articles on the topic are published by reputable publishers such as IEEE, ACM, Springer, Elsevier, MDPI, and others. The journal articles specifically discuss the application of machine learning algorithms to evaluate Extract-Transform-Load performance, anomaly detection, data processing for storage in data warehouses, data schema transformation, and several processes to address issues with nested array data types, all with a modular structure.

b. Exclusion Criteria
Identified duplicate studies, deleting irrelevant journal articles based on a review of titles and abstracts, and studies that do not directly discuss the application of machine learning for anomaly detection or the performance of Extract-Transform-Load on nested array data [15].

3. Data Sources and Search Strategies
Literature research is conducted through various methods, using a variety of sources, including reputable academic databases, to ensure coverage of relevant studies [16]. The literature will be processed using well-known academic databases, such as Scopus.

This search strategy uses keywords, then compiles several relevant journal articles. The results obtained are approximately 40 journal articles. These results are based on a systematic literature review process.

3.3 Study Identification and Selection

This stage involves identifying journal article data and screening studies according to the established planning criteria. Based on the process carried out, the PRISMA guidelines need to be followed to ensure transparency and reproducibility. Here are some of the steps taken, among others:

1. The specified database search is for journal articles indexed academically, one of the well-known examples being Scopus.
2. Data deletion of irrelevant journal articles.
3. Filtering titles and abstracts.
4. Full text screening for journal articles that pass the relevance verification process and comply with inclusion and exclusion criteria.

3.4 Data Extractions

This stage follows the analysis of relevant journal articles, in which input data are selected to classify the articles that will serve as references for this research. This process will produce information on several important aspects, including the publication, the research methodology used, the characteristics of the dataset employed, the performance metrics used for evaluation, and the solutions proposed by previous studies to assist future research on similar topics.

3.5 Data Synthesis and Analysis

In the final stage, the extraction and processing stage, the final visualization synthesis results are produced based on the publication years of the journal articles, which are set from 2019 to 2025. From the approximately 7 years of selected publication years, the data will be analyzed to answer the research questions.

4 RESULTS AND DISCUSSION

The results of the process carried out through a systematic literature review present the final outcomes of the identification and selection of journal articles, based on well-known academic studies. Based on the processing conducted using the PRISMA method, the following will be discussed:

4.1 Keyword Search Results

The stages of identifying and selecting studies are the initial steps in collecting relevant literature. Processing begins with keyword searches related to the journal article topics referenced. There are several well-known journal publishers, and then the next step is screening, aimed at ensuring that journal articles meeting the inclusion criteria are selected as the final results. The selection of criteria is based on the results of the journal article screening process, summarized in Table 1, which outlines a systematic procedure for sorting and selecting relevant studies.

Table 2. Stages of a systematic literature review

Stages	Explanation
Research	The process of removing duplicate literature is limited by search processors, resulting in the database in Watase or the selection of literature that does not match the topic.

Selection Based on Title and Abstract	Filtering journal articles based on titles and abstracts. Selecting relevant and irrelevant studies.
Selection Based on the method used for research topic adjustment	Filtering relevant topics based on the method or model used.
Selection based on full text	A comprehensive analysis for the extraction process of relevant journal article selection results will be automatically saved in the full text menu.
Documentation of the Selection Process	The entire selection process is documented using PRISMA to show the journal articles reviewed at each stage.

The discussion based on the table visualization above has resulted in various keyword searches. Here are the final results of the keyword search.

No	Keyword	Raw
1	performance evaluation, big data	47
2	Machine Learning Algorithms, big data	64
3	anomaly detection, isolation forest	56
4	algorithm data warehouse	30
5	extract transform load	16

Figure 2. Identify and select keywords

Note that Figure 2 shows the final result of keyword identification and selection processing. The process aims to identify relevant journal articles for the most general topic. The search was conducted using general keywords and more specific research keywords. Below are some of the final results obtained from the search for relevant journal articles.

1. "Performance evaluation, big data." This keyword retrieves journal data. A total of 47 data points were filtered down to about 14.
2. "Machine Learning Algorithm, big data" This keyword retrieves 64 journal data entries, which are then filtered down from 64 to approximately 24 entries.
3. "Anomaly detection, isolation forest." This keyword retrieved a total of 56 journal data entries, which were then filtered down to 32.
4. "Algorithm data warehouse." This keyword has 30 journal data entries, which are then filtered down to 7.
5. "Extract, transform, load." For this keyword, there are 16 journal data entries, which were then filtered down to about 5.

The keyword search above filters journal article data in the Watase database. The publication year follows a 7-year cycle, from 2019 to 2025, and is indexed by reputable publishers such as Scopus. After successful processing, the next step is to produce a PRISMA diagram.

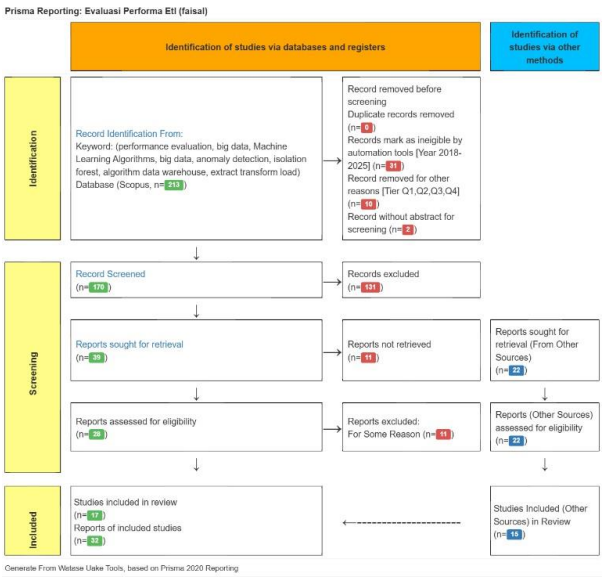

Figure 3 PRISMA Result

Please refer to Figure 3, which shows the final results of the keyword search process, also known as the identification and selection of studies. The search process was conducted using databases from various reputable publishers. A total of 213 journal article data entries were found. Of these, 31 irrelevant and duplicate journal articles were removed. Thus, 182 data entries remained. Next, approximately 10 journal articles that did not meet the criteria were deleted, leaving approximately 172. The remaining data were reduced by 2 journal articles that lacked abstracts, resulting in a final PRISMA output of 170 journal articles needed for the study.

The processing of these 170 journal articles will be repeated, with some excluded, resulting in 39 articles suitable for the research. Next, a retrieval process was carried out to select relevant articles. The selection was done manually, resulting in 11 reports that could not be obtained, while the remaining 28 reports were successfully retrieved and deemed suitable as research references. Additionally, the final results of the keyword search for journal articles produced a visualization of a synthesis graph by year, showing trends over time, in accordance with external data from publishers.

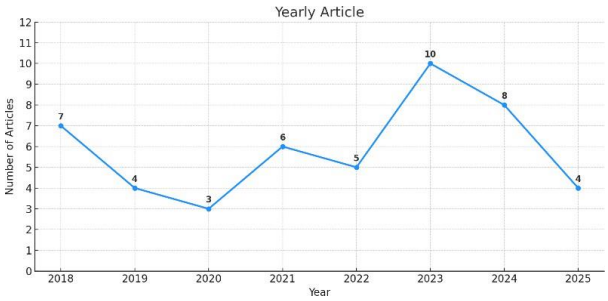

Figure 4. Exclusive results based on the year of journal article publication

Note that Figure 4 shows the graph of the number of journal articles selected based on the publication year, from 2018 to 2025. The publication year 2018 successfully obtained 7 journal articles, the year 2021 obtained 6 journal articles, the year 2022 obtained 5 journal articles, and the year 2024 obtained 8 journal articles. The most journal articles were obtained from the publication year 2023, with 10 articles, while the fewest were in 2020, with 3 articles. Based on the line graph, the average is 4 journal articles for 2019 and 2025.

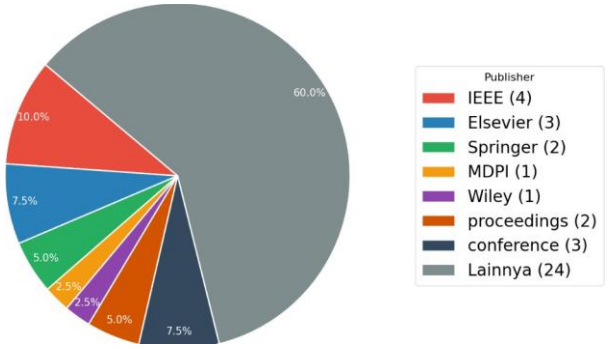

Figure 5. Results of the illusion based on a reputable publisher

Note that Figure 5 is a visualization of the final inclusion results based on various reputable publishers. The visualization was conducted to analyze the distribution of relevant journal articles successfully found by different publishers. Based on the visualization, the source with the most articles is IEEE, with around 4 journal articles, while the source with the fewest is MDPI, and Wiley managed 1 journal article. The final results, displayed as a pie chart, will proceed to the next stage: processing to answer the research questions.

RQ1: What is the most effective Machine Learning method for anomaly detection in nested array data after the Extract-Transform-Load process?

Anomaly detection is the process of identifying unusual patterns or events, such as results from data transformation processes that experience discrepancies or are considered null [17]. Anomaly is also known as outlier, deviation, or exception [18]. Anomalies are data points or entire patterns that deviate from normal behavior. Therefore, anomaly detection is performed to tidy up data with nested array types by converting them into a modular format so the system can understand them. Managing nested data types using the Extract-Transform-Load strategy presents complex challenges that require machine learning algorithms to handle hierarchical and unstructured data structures. Thus, the systematic literature review conducted reveals various machine learning methods that demonstrate the effectiveness of anomaly detection for nested array data patterns [19].

Figure 6. Types of machine learning algorithms

Figure 6 is a visualization of the mind mapping, based on the analysis process that identified two types of machine learning algorithms. From among these algorithms, the researcher will select the most efficient for anomaly detection. The selection is based on the results of the analysis, through visualization and understanding, that have been conducted. The most commonly used algorithm for anomaly detection is unsupervised, and the metric most commonly used to assess whether the data is anomalous or normal is accuracy.

Figure 7. Types of Algorithms for Anomaly Detection

Based on scientific analysis, machine learning algorithms are divided into two types: supervised and unsupervised. The researcher further analyzed and found that anomaly detection processing is often performed using unsupervised algorithms [6], [20]. Therefore, it has been concluded that the results of unsupervised algorithm analysis are frequently used, including by other researchers. Unsupervised algorithms branch into many methods, so researchers choose efficient methods for anomaly detection, data management, and error checking in data. The following are:

1. Isolation Forest
   This algorithm is designed to detect anomalies. Its approach is to isolate data observations that are considered anomalies from those considered normal. [21], [22] The Isolation Forest method demonstrates strong performance, with an AUC of 0.997, an F1-Score of 95.29%, a Precision of 95.73%, and a recall of 94.85%.

2. Local Outlier Factor (LOF)
   This algorithm is designed to measure the 'anomaly' value, such as data points based on data density compared to neighbor looseness. Data points with a significantly higher Local Outlier Factor than their neighbors are considered outliers [23].

3. One-Class Support Vector Machine (OC-SVM)
   This algorithm is designed to classify normal data, identifying data points that are 'extraordinarily' normal as anomalies [24] stated that the OC-SVM algorithm shows performance with an AUC value of 0.9, FPR of 0.39, and TPR of 0.9.

Table 3. Total Data

Algorithm	Citation	Low Metrics	High Metrics	Data Transformation
Isolation Forest	[1]	AUC: 0.935 F1-Score: 93.7 Precision: 93.86 Recall: 93.08	AUC: 0.997 F1-Score: 95.29 Precision: 95.73 Recall: 94.85	Historical machine sensor data from Node-RED to tabular.
Local Outlier Factor (LOF)	[2]	AUC: 0 Accuracy: 77.8 F1-Score: 50 F2-Score: 14 Precision: 50.3 Recall: 50.4 AUROC: 0.5465	AUC: 0.956 Accuracy: 97.6 F1-Score: 97 F2-Score: 95.7 Precision: 97.7 Recall: 97.6 AUROC: 0.4859	Credit card transaction data into tabular.
One-Class Support Vector Machine (OC-SVM)	[2]	AUC: 0.64 FPR: 0.0 TPR: 0.4	AUC: 0.9 FPR: 0.39 TPR: 0.9	Sensor data dimensions are high for feature representation.

Based on the analysis of the table visualization above, the most frequently used algorithm is Isolation Forest to detect anomalies in data that has been processed using the Extract-Transform-Load strategy. [17], [19].

After managing data using the Extract-Transform-Load (ETL) strategy with nested array data from a company in Indonesia, the processing needs to consider several methods as key to the success of implementing machine learning algorithms. Here are some effective data preprocessing steps, among others:

a. Flattening/De-nesting is the process of transforming a nested array structure into a tabular representation.
b. Feature Engineering is a process of extracting various relevant features from nested data array patterns that can be described through the data characteristics comprehensively, such as the number of elements in the array, the data type of elements, or specific patterns.
c. Missing Value Imputation is the process of removing or eliminating unnecessary or potentially appearing values during the Extract-Transform-Load process.
d. Encoding categorical data is the process of converting categorical data into a numerical format.

Based on the analysis, it was found that the most frequently used algorithm is Isolation Forest for anomaly detection, and that the processing takes a long time. Therefore, it is necessary to find out information so that the processing time is kept under about 10 minutes to avoid taking too long in data management.

RQ2: How is the implementation and performance evaluation of machine learning algorithms in optimizing the Extract-Transform-Load process on large-scale data?

The application of machine learning (ML) in the Extract-Transform-Load process offers great potential to improve efficiency, accuracy, and automation. However, the implementation faced challenges, such as difficulties in making accurate and careful decisions. The implementation involved a transformation process by adjusting the data schema or data warehouse structure (BigQuery). Then, an implementation issue occurred, so an evaluation of data performance was necessary by

applying machine learning to identify the root cause and determine which data needs to be followed up on.

Here are some metrics that can be used as considerations [25]:

1. Anomaly detection accuracy
   This metric is used to measure precision, recall, and F1-Score.

2. Efficiency of time and computational resources.
   This metric is applied to process data quickly, within approximately 1 minute, without burdening the device's resources.

3. Quality of transformed data
   This metric is used to measure the consistency and completeness of data from the Extract-Transform-Load process.

Among the three materials, the researcher took a wise step by conducting further analysis to understand the meaning of the information obtained. As a result, several machine learning application methods were identified that can assist in many aspects, including:

Table 4. Name of the machine learning method for evaluating ETL performance

Machine Learning Method	**Explanation**	**Citation**
Improvement of Data Quality	This method is used to identify and mark inconsistent, duplicate, or functional data to ensure the quality of data that has been uploaded to the data warehouse.	[26], [27]
Data Transformation Automation	This method is used to handle nested arrays. Additionally, the application of machine learning can learn transformation patterns efficiently and reduce the need for complex manual rules.	[5], [28]
Proactive Anomaly Detection	This Machine Learning method is capable of performing real-time or near-real-time anomaly pattern detection during the ETL process.	[13], [23], [24], [29]
Performance Optimization of Extract-Transform-Load	This Machine Learning method can analyze logs and performance metrics of Extract-Transform-Load processes to identify bottlenecks or areas that need optimization. Examples include job scheduling or resource allocation.	[19], [25], [30]

Through the Extract-Transform-Load (ETL) workflow, the design is processed to prepare data for effective use [31]. Data is collected and then processed according to the needs of analysis, reporting, and ultimately used to make wise business decisions. [32] With the three stages completed, the results are accurate, consistent, and relevant data [3].

The processing is carried out, and the analysis has found that the understanding that can be drawn is a transformation process by changing raw data into a format ready for analysis. An example is managing data using transaction data. [33] The initial data was in a nested array pattern, then its structure was transformed into a tabular format that can be understood by analytics, so it can be analyzed into detailed explanations. Processing was carried out using an unsupervised method so that the design could explain activities, scope, objects, materials, tools, locations, and data collection techniques, and operational definitions of research variables as an effective analysis technique.

5 CONCLUSION

The processing conducted using a systematic literature review strategy can be summarized that there are two types of machine learning algorithms, namely supervised and unsupervised, which are then further narrowed down according to the research needs. Based on the results of narrowing down various types of unsupervised algorithms, three were identified: isolation

forest, local outlier factor, and one-class support vector, all of which have good effectiveness in detecting anomalies in nested array data patterns after processing data with the Extract-Transform-Load strategy. The algorithms obtained were then selected based on their frequent use because they demonstrated some of the best performance, with isolation forest being the most prominent. The results of the isolation forest algorithm include data evaluation metrics with the implementation of optimized Extract-Transform-Load processing to improve efficiency, accuracy, and automation. However, there are challenges in processing and managing data, so careful consideration of the evaluation metrics results is necessary, such as recalculating anomaly detection accuracy, time efficiency, and data quality from the data transformation results.

6 SUGGESTION AND RECOMMENDATION

This research provides a solid foundation for developing data management using a more effective Extract-Transform-Load (ETL) strategy, integrating machine learning algorithms to detect anomalies in nested array data structures.

REFERENCES

[1] E. Marcelli, T. Barbariol, and G. A. Susto, "Active Learning-based Isolation Forest (ALIF): Enhancing Anomaly Detection in Decision Support Systems," Jul. 2022.

[2] N. Usman, E. Utami, and A. D. Hartanto, "Comparative Analysis of Elliptic Envelope, Isolation Forest, One-Class SVM, and Local Outlier Factor in Detecting Earthquakes with Status Anomaly using Outlier," in *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*, IEEE, Feb. 2023, pp. 673–678. doi: 10.1109/ICCoSITE57641.2023.10127748.

[3] J. C. Nwokeji and R. Matovu, "A Systematic Literature Review on Big Data Extraction, Transformation and Loading (ETL)," 2021, pp. 308–324. doi: 10.1007/978-3-030-80126-7_24.

[4] J. Nwokeji, F. Aqlan, A. Anugu, and A. Olagunju, "Big Data ETL Implementation Approaches: A Systematic Literature Review (P)," Jul. 2018, pp. 714–721. doi: 10.18293/SEKE2018-152.

[5] F. Raymand, B. Najafi, A. Haghighat Mamaghani, A. Moazami, and F. Rinaldi, "Machine learning-based estimation of buildings' characteristics employing electrical and chilled water consumption data: Pipeline optimization," *Energy Build.*, vol. 295, p. 113327, Sep. 2023, doi: 10.1016/j.enbuild.2023.113327.

[6] Y. Gong, F. Gu, K. Chen, and F. Wang, "The Architecture of Micro-services and the Separation of Frond-end and Back-end Applied in a Campus Information System," in *2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications(AEECA)*, IEEE, Aug. 2020, pp. 321–324. doi: 10.1109/AEECA49918.2020.9213662.

[7] A. A. Yulianto, "Extract Transform Load (ETL) Process in Distributed Database Academic Data Warehouse," *APTIKOM Journal on Computer Science and Information Technologies*, vol. 4, no. 2, pp. 61–68, Jul. 2019, doi: 10.11591/APTIKOM.J.CSIT.36.

[8] M. Gorawski, K. Pasterak, A. Gorawska, and M. Gorawski, "The stream data warehouse: Page replacement algorithms and quality of service metrics," *Future Generation Computer Systems*, vol. 142, pp. 212–227, May 2023, doi: 10.1016/j.future.2023.01.003.

[9] S. R. Cheruku, S. Jain, and A. Aggarwal, "Managing Data Warehouses in Cloud Environments: Challenges and Solutions," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 6, no. 8, Sep. 2024, doi: 10.56726/IRJMETS61249.

[10] F. F. Hasan and M. S. A. Bakar, "Data Transformation from SQL to NoSQL MongoDB Based on R Programming Language," in *2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 2021, pp. 399–403. doi: 10.1109/ISMSIT52890.2021.9604548.

[11] B. R. Chang, H.-F. Tsai, and Y.-D. Lee, "Integrated High-Performance Platform for Fast Query Response in Big Data with Hive, Impala, and SparkSQL: A Performance Evaluation," *Applied Sciences*, vol. 8, no. 9, p. 1514, Sep. 2018, doi: 10.3390/app8091514.

[12] A. Herreros-Martínez, R. Magdalena-Benedicto, J. Vila-Francés, A. J. Serrano-López, S. Pérez-Díaz, and J. J. Martínez-Herráiz, "Applied Machine Learning to Anomaly Detection in Enterprise Purchase Processes: A Hybrid Approach Using Clustering and Isolation Forest," *Information*, vol. 16, no. 3, p. 177, Feb. 2025, doi: 10.3390/info16030177.

[13] M. Nalini, B. Yamini, C. Ambhika, and R. Siva Subramanian, "Enhancing early attack detection: novel hybrid density-based isolation forest for improved anomaly detection," *International Journal of Machine Learning and Cybernetics*, vol. 16, no. 5–6, pp. 3429–3447, Jun. 2025, doi: 10.1007/s13042-024-02460-5.

[14] A. R. Fadillah and M. N. Fauzan, "Systematic Literature Review: Identifying Key Variables and Measuring Maximum Loan Limits," *Jurnal ELTIKOM : Jurnal Teknik Elektro, Teknologi Informasi dan Komputer*, vol. 8, no. 2, pp. 100–110, Dec. 2024, doi: 10.31961/eltikom.v8i2.1156.

[15] S. Mishra, S. Konidala, and J. Manda, "Improving the ETL process through declarative transformation languages ," *Distributed Learning and Broad Applications in Scientific Research*, vol. 5, Jun. 2019.

[16] B. Oliveira, Ó. Oliveira, T. Matos, V. Santos, and O. Belo, "AN ETL PATTERN FOR LOG CONFIGURATION AND ANALYSIS," in *Proceedings of the International Conferences Big Data Analytics, Data Mining and Computational Intelligence 2019; and Theory and Practice in Modern Computing 2019*, IADIS Press, Jul. 2019, pp. 39–46. doi: 10.33965/bigdaci2019_201907L005.

[17] A. Herreros-Martínez, R. Magdalena-Benedicto, J. Vila-Francés, A. J. Serrano-López, S. Pérez-Díaz, and J. J. Martínez-Herráiz, "Applied Machine Learning to Anomaly Detection in Enterprise Purchase Processes: A Hybrid Approach Using Clustering and Isolation Forest," *Information*, vol. 16, no. 3, p. 177, Feb. 2025, doi: 10.3390/info16030177.

[18] B. J. Wheeler and H. A. Karimi, "Enhancing Hyperspectral Anomaly Detection Algorithm Comparisons: Leveraging Dataset and Algorithm Characteristics," *Remote Sens. (Basel).*, vol. 16, no. 20, p. 3879, Oct. 2024, doi: 10.3390/rs16203879.

[19] A. Gautama Putrada, I. Dian Oktaviani, M. Nurkamal Fauzan, and N. Alamsyah, "CNN Pruning for Edge Computing-Based Corn Disease Detection with a Novel NG-Mean Accuracy Loss Optimization," *Telematika*, vol. 17, no. 2, pp. 68–83, Aug. 2024, doi: 10.35671/telematika.v17i2.2899.

[20] E. F. Agyemang, "Anomaly detection using unsupervised machine learning algorithms: A simulation study," *Sci. Afr.*, vol. 26, p. e02386, Dec. 2024, doi: 10.1016/j.sciaf.2024.e02386.

[21] D. Sartor, T. Barbariol, and G. A. Susto, "Bayesian active learning isolation forest (B-ALIF): A weakly supervised strategy for anomaly detection," *Eng. Appl. Artif. Intell.*, vol. 130, p. 107671, Apr. 2024, doi: 10.1016/j.engappai.2023.107671.

[22] G. Hannák, G. Horváth, A. Kádár, and M. D. Szalai, "<scp>Bilateral-Weighted</scp> Online Adaptive Isolation Forest for anomaly detection in streaming data," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 16, no. 3, pp. 215–223, Jun. 2023, doi: 10.1002/sam.11612.

[23] M. S. Hossain and H. Mahmood, "Short-Term Load Forecasting Using an LSTM Neural Network," in *2020 IEEE Power and Energy Conference at Illinois (PECI)*, IEEE, Feb. 2020, pp. 1–6. doi: 10.1109/PECI48348.2020.9064654.

[24] Y. Qiao, K. Wu, and P. Jin, "Efficient Anomaly Detection for High-Dimensional Sensing Data With One-Class Support Vector Machine," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 404–417, Jan. 2023, doi: 10.1109/TKDE.2021.3077046.

[25] J. C. Quiroz, T. Chard, Z. Sa, A. Ritchie, L. Jorm, and B. Gallego, "Extract, transform, load framework for the conversion of health databases to OMOP," *PLoS One*, vol. 17, no. 4, p. e0266911, Apr. 2022, doi: 10.1371/journal.pone.0266911.

[26] E. Widad, E. Saida, and Y. Gahi, "Quality Anomaly Detection Using Predictive Techniques: An Extensive Big Data Quality Framework for Reliable Data Analysis," *IEEE Access*, vol. 11, pp. 103306–103318, 2023, doi: 10.1109/ACCESS.2023.3317354.

[27] T. Nguyen, H.-T. Nguyen, and T.-A. Nguyen-Hoang, "Data quality management in big data: Strategies, tools, and educational implications," *J. Parallel Distrib. Comput.*, vol. 200, p. 105067, Jun. 2025, doi: 10.1016/j.jpdc.2025.105067.

[28] F. F. Hasan and M. S. A. Bakar, "Data Transformation from SQL to NoSQL MongoDB Based on R Programming Language," in *2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, IEEE, Oct. 2021, pp.

399–403. doi: 10.1109/ISMSIT52890.2021.9604548.

[29] S. Alam, S. K. Sonbhadra, S. Agarwal, and P. Nagabhushan, "One-class support vector classifiers: A survey," *Knowl. Based. Syst.*, vol. 196, p. 105754, May 2020, doi: 10.1016/j.knosys.2020.105754.

[30] Q. Yang and Y. Tang, "Big Data-based Human Resource Performance Evaluation Model Using Bayesian Network of Deep Learning," *Applied Artificial Intelligence*, vol. 37, no. 1, Dec. 2023, doi: 10.1080/08839514.2023.2198897.

[31] J. Awiti, "Algorithms and Architecture for Managing Evolving ETL Workflows," 2019, pp. 539–545. doi: 10.1007/978-3-030-30278-8_51.

[32] D. Andriansyah, "Implementasi Extract-Transform-Load (ETL) Data Warehouse Laporan Harian Pool," *Jurnal Teknik Informatika*, vol. 8, no. 2, pp. 45–49, Aug. 2022, doi: 10.51998/jti.v8i2.486.

[33] M. Hendayun, E. Yulianto, J. F. Rusdi, A. Setiawan, and B. Ilman, "Extract transform load process in banking reporting system," *MethodsX*, vol. 8, p. 101260, 2021, doi: 10.1016/j.mex.2021.101260.