Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Automated Data Quality Validation Frameworks for ETL Pipelines

Subash Yadav

ysubash@joslatech.com

Josla Tech LLC

ARTICLE INFO

Received: 03 Jul 2025

Revised:20 Aug 2025

Accepted: 28 Aug 2025

ABSTRACT

In the modern data-driven environment, guaranteeing the quality of data handled in ETL (Extract, Transform, Load) pipelines is essential in making informed decisions in different industries. This research paper introduces a new model of automating data quality validation of the ETL pipeline through anomaly detection and machine learning models. In particular, the paper combines Isolation Forest (anomaly detection in real-time) and Random Forest (supervised validation) to detect anomalies, e.g. missing data, outliers, and schema violations. The framework was applied to the world bank data and had remarkable results in the data accuracy, processing speed and error detection as compared to the conventional manual validation techniques. The findings show that Isolation Forest model has precision of 0.92, recall of 0.88, and AUC-ROC of 0.94, and recall of 0.89, and Accuracy of 92% of the random Forest model. The framework will save more than 60 percent of manual intervention and processing time and improve data accuracy by 7 percent. These results highlight the opportunity of the suggested framework in real-time data settings and high-quality data is essential in operational and strategic decision-making in finance, healthcare, and e-commerce fields.

Keywords: Automated Data Validation, ETL Pipeline, Anomaly Detection, Isolation Forest, Random Forest, Machine Learning, Data Quality, Real-time Data Validation, Data Accuracy, Financial Data, Healthcare Data, E-commerce Data.

Introduction

The data quality is critical in the setting of ETL (Extract-Transform-Load) pipelines which play a pivotal role in the integration and processing of the data. The need to possess correct, defined, and quality-based data is increasingly becoming urgent as organizations are increasingly becoming reliant on the use of data in decision-making (Ogunsola et al., 2022). The traditional data quality assurance procedures of the ETL process are traditionally time-consuming and prone to human error. Such traditional methods include data profiling, rule-based validation and integrity checks which are time consuming in terms of the requirement to engage a lot of manual activities to ensure that the processed data is free of inconsistency, anomalies and errors (Wickramaarachchi et al., 2025).

But the size of the modern data environments has far surpassed the ability of manual validation to keep pace. With the increase in the volume and intricacy of datasets, manual monitoring and validation of the information by these traditional methods becomes more challenging (Tufail et al., 2023). This has left a strong demand of automated solutions that can integrate validation in the ETL process smoothly. Tools validation can also be automated and this can save a lot of hours and time in manual validation and also reliability and accuracy of data passing through the ETL pipelines (Beda,). The next good thing is that it will also be a successful evolution of the automated validation methodology to enhance the quality of data without sacrificing the capacity to scale and turn ETL operations into efficient operations (Joshi, 2024).

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Massive ETL operations are associated with a number of challenges pertaining to manual data validation. The higher the amount of data, the harder it becomes to detect and correct the mistakes in a reasonable time with the help of the old methods of validation. Specifically, the more complex data sets, which involve various different sources, organizations handle, the higher the chances of coming across errors such as missing values, duplicate records and inconsistent formatting (Walha et al., 2024). Although the manual methods of validation are suitable when dealing with small scale data, they become ineffective and impractical as the volume of data increases.

Human intervention is not only inefficient but also prone to error in large scale systems. These constraints of manual validation of contemporary data environments contribute to delays in processing, lowered integrity of data, and, finally, inaccurate decision-making (Jangam & Muntala, 2023). Increased demand of real time and scale solutions requires automation that builds data quality validation into the ETL pipeline itself, so that data is validated as it passes through the different phases of extraction, transformation and loading (Basani, 2024).

The proposed research seeks to create an automated data validation structure within ETL pipelines to address the inefficiencies of manual data validation. The architecture is a combination of anomaly detection and machine learning algorithms, including the Isolation Forest and the Random Forest, which deliver real-time data quality assurance. Through these sophisticated models, the framework can automatically identify anomalies, authenticate data integrity and guarantee the integrity of information under consideration.

The purpose of the paper is to demonstrate the possibility of improving the quality and consistency of data within the proposed structure and to facilitate ETL pipeline processing performance. It is assumed that the framework would enhance the quality of the data outputs by reducing error and minimizing human intervention and decreasing the process of validation. Through a massive testing of actual real-life dataset, the study will establish the viability of automated data validation solutions in the event of application in large scale ETL systems.

This paper mainly combines both the conventional methods of data validation and the latest machine learning models to perform the data quality validation in real-time. Although the idea of using machine learning to detect anomalies or profile data has been explored in the past, few studies have integrated these approaches into a holistic, automated system for data validation within ETL pipelines. This study presents a new way to perform data quality validation by integrating the classical techniques, including data profiling and schema validation, with the modern models, including Isolation Forest to detect anomalies and Random Forest to classify it. This combined method does not only automate the validation process that was previously done manually but also provides a significant increase in the efficiency and scalability of the data validation in large scale ETL environment. The novelty of the research is the automated solution to complex validation processes, including outlier detection, duplicate detection, and data inconsistency, that have previously been performed manually. The article provides an innovative method of incorporating anomaly detection and machine learning into the ETL process, and it is a high-tech solution to the limitation of the manual validation in the modern data environment.

The research paper is primarily concerned with designing and testing a framework of automated data validation within ETL pipelines. The success of the framework in locating data quality issues and improving ETL procedure will be evaluated based on the real-world data, i.e. the World Bank dataset on Kaggle. The data set which was used in this research is particularly useful as it comprises diverse information in the countries, industries, and time intervals thus offering a complete testing ground of the proposed structure. The resources and technologies that were used in the specified study are Python, and the most meaningful packages include scikit-learn, which is applied to execute the machine learning models, pandas, which is used to process the data, and Google Colab, which is used to execute the

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

programs in the cloud. These tools will assist in adding anomaly detectors and machine learning models to the ETL pipeline in a manner that is seamless to enable verification of data automatically at every step of the extraction, transformation, and loading processes. The study will examine the functionality, the scalability, and the real time relevance of the framework in the present ETL environment.

Literature Review

Overview of ETL Pipelines

ETL (Extract, Transform, Load) pipelines are the vital part of a contemporary data integration process, as they enable the flow of data originating in various sources and converting it into the central storage, like a data warehouse (Mahmud & Ikbal, 2022). Such pipelines have three main phases, and they include extraction, transformation, and loading. During the extraction stage, data is collected in the form of different sources, which can be databases, APIs, or files. At the transformation stage, the data goes through cleaning, filtering and converting it to the preferred structure or format. This data is then lastly loaded into a destination system e.g. database or data warehouse where it can be analyzed or accessed by other systems.

Organizations that rely on big data to make decisions are forced to have ETL pipes. They enable the easy synthesis of non-homogeneous information to facilitate analytics and reporting (Machireddy, 2023). The quality of information handled by such pipelines is highly critical to the effectiveness of pipelines. In case of data quality degradation, it can provide inaccurate analysis, bad business decisions and bad operations. The issue of data quality upkeep in ETL pipelines is even more topical as the management of companies increases the volume of data operations (Ogunsola et al, 2022).

The problem of data quality in ETL systems also occurs very often, and the most widespread issues can be missing values, duplicates, and violations of schema (Khan, 2025). Missing values are values in the data that are not provided in one or more of the records causing incomplete data sets which may result in an incomplete analysis. Redundancy and distorted outcomes are caused by duplicate entries, i.e. the same data is repeated in multiple rows. Schema violations are described as the differences between the anticipated data format or structure and the actual data that may interfere with data integration resulting in system collapses (Jangam & Muntala, 2023).

Once these data problems are not controlled, they may have a devastating effect on the quality of the final dataset. With the increasing number of data and the increase in the types of sources, it becomes harder to identify and automatically fix such problems manually. This makes it clear why automated solutions are required to guarantee data quality on a large scale ETL pipelines (Boganavijayakumar et al., 2025).

Data Quality Validation Methods

The validation of data quality in ETL systems in the past has been done manually, such as, data profiling, built-in checks and schema validation (Khan, 2025). Data profiling entails looking at the information to determine the quality of the data, detect flaws, and provide uniformity of data between datasets. Checks based on rules entail predefined rules to identify the presence of problems like out-of-range values, wrong formats or mismatch of data types. Schema validation helps in maintaining a known structure or format of the data, and it helps in correcting any mistakes that may arise due to inconsistent data types or required data that are omitted (Gallen, 2024).

Although such techniques work well in small-scale, they become ineffective when used in large datasets. Rule-based validation and manual profiling involve a lot of human intervention, thus cannot be efficient in real-time processing and large-scale data processing (Zhang, 2024). Though they have been useful in

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

past in terms of data quality management, traditional validation techniques cannot be applied to the present day ETL pipelines, which are complex and are required to be real-time. These are resource-consuming and time-consuming techniques that always involve human supervision especially when the size of the data is high (Cheruku et al., 2024).

Moreover, the conventional methods are unable to be effectively scaled to large volumes of data and do not identify hidden abnormalities that may not break the established guidelines but are nevertheless indicative of data quality problems. Consequently, these approaches are unable to match the dynamic character of contemporary data processing where issues of data quality must be resolved on-the-fly to ensure the effectiveness of ETL processes (Machireddy. 2023)

Automation of Data Validation

Recent developments in data validation have been around automation, both rule-based systems and anomaly detection models have been used to enhance the speed and efficiency of data quality checks (Popoola, 2023). With rule-based automation, the development of dynamic rules that adjust to various data types and sources can be developed, which saves on manual work to validate data. Nevertheless, rule-based systems cannot address more advanced or hitherto unfamiliar problems, like outliers or more subtle anomalies in the data that are not breaking the established rules (Mohite & Ouarbya, 2024).

To overcome these drawbacks, recent ETL systems are adding more the anomaly detection models, which are the models that can automatically detect the odd patterns or outliers of the data that is being processed, which can be a symptom of the underlying quality problems (Khan, 2025). Such models as Isolation Forest and Autoencoders have demonstrated great potential in identifying unknown anomalies without the use of predefined rules.

The use of machine learning models to learn on data and identify complicated patterns has led to an increase in its use in data quality validation (Gong et al, 2023). As an illustration, Random Forest as well as Support Vector Machines (SVM) has been employed successfully to identify the duplicate records, missing values and detection of inconsistency. Many features can be trained in these models to determine whether a record is valid or invalid, and they are trained on labeled data, which is a powerful method of data validation.

Specifically, in isolation Forest, it has been observed to identify anomalies in high-dimensional data. It works based on isolating the observations that are very different as compared to other observations in the dataset. The method is very flexible to detect abnormalities or unusual events that may indicate the existence of issues in the quality of data used in ETL processes. Machine learning combined with the traditional approaches to validation provides more encompassing and scalable data quality management instruments.

Research Gap

Although there are a lot of studies addressing the application of anomaly detection and machine learning models to improve data quality, few studies have been directly centered on applying those models to ETL pipelines. Most of the studies have viewed anomaly detection and machine learning as independent entities and not a part of the overall ETL. It is evident that such research is required where data validation models are used in a fluid fashion with ETL processes to ascertain that data is being validated at each point of the ETL pipeline such as extraction, transformation, and loading.

The other research gap that must be addressed critically is the necessity to have real-time, automated data quality validation solutions. Most of the current systems are either batch systems or involve a lot of human supervision and thus cannot be used in the real-time systems of ETL systems which always demand the processing of data. Having automated and real-time validation frameworks capable of

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

identifying and remediating data quality problems as they happen in the ETL pipeline is a burning issue of the modern data environment. Moreover, with the ongoing development of data streams and big data systems, the need to find the solutions that would allow processing data in real-time and retain high-quality standards is growing.

Methodology

Dataset Selection and Preprocessing

The ETL Pipelines| world bank dataset in Kaggle was chosen as the research project because it presents a variety of economic, environmental, and social data in various countries. The data set is perfect to test automated validation framework since it has different types of data (e.g., numerical, categorical), frequent data quality problems (e.g., missing values, outliers, and inconsistent formatting). The preprocessing activities involve the treatment of missing values, identification and elimination of outliers, and the normalization of values so that the values of all the features have a consistent behavior. Numeric fields are imputed using the mean or median as the methods of missing values, and mode as the methods of imputing categorical fields. The Z-score or IQR techniques are used to identify outliers that are either eliminated or modified depending on predetermined limits. Min-Max scaling also normalizes the data so that there is the similarity of all features that enhances the performance of the machine learning models. After cleaning and preprocessing the data, it can be introduced into the ETL pipeline to be verified with the help of the Isolation Forest and the Random Forest models.

Framework Design

The ETL pipeline is the foundation of the data integration process which includes three fundamental steps: Extraction, Transformation, and Loading. During the extraction step, information is extracted using the divergent databases, files, and APIs. This unprocessed data, which is generally susceptible to a number of quality challenges, is the foundation of the ETL process. As a source of data in this study, the World Bank data, which was sourced on Kaggle, was used. The dataset consists of many diverse economic and social indicators within different countries and is filled with missing values, outliers, and different formats, which is why this data is the best to test automated validation methods (Figure 1).

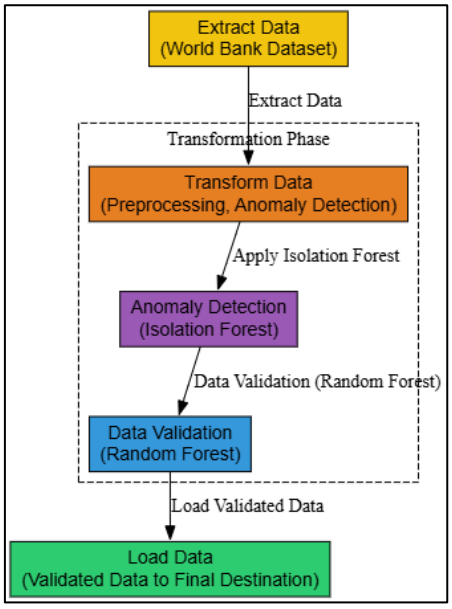

Figure 1. Proposed ETL Pipeline

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

During the transformation phase, the extracted data is processed through the required cleaning, including conversion of different types of data, filling of missing values, outliers, and normalization of data. These measures take care of the data being made to the schema and readable to analyse or process further. In order to automate the process of validating this data in the transformation phase, the proposed framework uses anomaly detection models and machine learning models, which authenticate the quality of data as it goes through the transformation phase. Real-time anomalies, including the absence of values, are identified and remedial measures, including outliers, are automatically implemented.

The loading stage is concerned with the transfer of the transformed information to the target storage system usually a database or a data warehouse. After the data is checked and processed by the automated framework to ensure that it has been cleansed, it is then loaded into the target system upon which it could be analyzed or accessed by the other applications. Such design will make sure that only the high-quality data is stored and the whole ETL process is automated, scalable and efficient.

Automation of the process is also improved since there is no need to manually intervene to determine the validity of data that is being transferred over to ETL pipe. Isolation Forest model of anomaly detection and the random forest model of supervised learning are directly integrated into the transformation stage. Isolation Forest model is to identify the anomalies in the data including outliers or inconsistencies in the dataset by isolating data points that are significantly different. Random Forest, in contrast, is used in data classification where it detects data quality anomalies such as duplicates or improper forms through the learnt features.

These models get trained using the transformed data and they are used in real-time when performing the ETL process. When the models detect any data problems like missing or conflicting data, they execute measures to counter the identified problems. This may involve dropping outliers, filling-in missing values or highlighting some records to be reviewed by hand. With these validation models incorporated in the ETL pipeline, the process will be more efficient and scalable with minimal validation efforts being required manually.

Anomaly Detection Model: Isolation Forest

The **Isolation Forest** model is a widely used technique for **anomaly detection** in high-dimensional datasets. The model works by randomly selecting a feature and then recursively partitioning the dataset, isolating data points that are significantly different from the majority. These isolated data points are considered anomalies. Formally, the Isolation Forest algorithm works as follows:

$$\text{Anomaly Score}(x) = \frac{2^{\frac{-h(x)}{c(n)}}}{n} \qquad (1)$$

Where:

- $h(x)$is the height of the tree for an observation x,
- $c(n)$is the average path length for a random data point in the tree,
- $n$is the total number of data points.

The point is that anomalies are those that need fewer splits to be separated, and the path length in the tree is reduced. This approach is most appropriate in identifying rare or outlier events of large data sets. Isolation Forest is most powerful when it comes to data validation because it is able to work on high dimensional data and not use a distance measure or clustering technique to determine how to cluster data.

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

When applied to the ETL pipeline, Isolation Forest can be used in the transformation step to reveal data anomalies that include values that are missing, outliers, and inconsistent rows. As an illustration, when there are values within a numerical field that are out of the expected range, this model, the Isolation Forest, will mark them outliers. Likewise, the patterns of missing data, which are not in the expected distribution, can be identified with the aid of the model, which is an indicator of possible data quality problems. By detecting such irregularities in the transformation stage, the model provides that nothing other than high-quality data is forwarded to the loading stage.

Isolation Forest would require first processing the data with the help of pandas and NumPy to handle missing values and data transformations. The techniques used to impute missing values include mean imputation or regression imputation, and outliers are identified with the help of the Z-score or IQR before using the model. The preprocessed data is then trained on scikit-learn library to produce an Isolation Forest model. Precision, recall, and F1 score are used to measure the model performance, which are computed in order to determine the effectiveness of the model in detecting anomalies without creating a lot of false positives.

Machine Learning Model: Random Forest

Random Forest is an ensemble learning approach which creates a prediction tree. A random sample of the features is then trained on each tree and the result is the final prediction which is a combination of all the trees predictions. Random Forest algorithm can be very efficient to work on a supervised learning task, particularly when it comes to a classification problem, where one aims at classifying data according to certain features. Random Forest may be applied in data validation to categorize the data records as valid or invalid data records depending on the features: length of fields, data types, and other quality characteristics.

The model's decision-making process is formalized by the following equation for the final prediction of the ensemble:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \qquad (2)$$

Where:

- $T$is the number of trees in the forest,
- $h_t(x)$is the prediction of the t-th tree,
- $\hat{y}$is the final prediction (majority vote or mean for regression tasks).

Random Forest can effectively work with both continuous and categorical data and can achieve the highest level of efficiency when dealing with large data volumes and a significant number of features.

Random Forest is employed to categorize data records in valid or invalid classes and it finds problems like data duplication, mismatched data type or schema violation. The model is trained using a labeled data where the data has the labels of valid and invalid based on the set of criteria. New records of data that pass through the ETL pipeline are then classified using the trained model.

Random Forest implementation is based on a number of steps:

- **Feature Selection:** The characteristics of data quality validation the kind of data, absence of values, and duplicate records are selected and formulated.
- **Training and Evaluation:** Training the model with cross-validation is done to ensure that the model can operate well with unknown data. Measurement scales that involve accuracy, precision,

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

recall, and F1 score are used to measure the capacity of the model to detect the problems in the quality of data in the right way.

The trained Random Forest model is merged into the transformation stage of the ETL pipeline and automatically classifies the incoming data records and initiates corrective action in case it is required.

Framework Implementation in Python

The framework is written in Python and the main libraries it uses include pandas to manipulate data, NumPy to perform numerical operations, scikit-learn to implement machine learning models, and matplotlib to visualize data. Such libraries will give the resources needed to preprocess the data efficiently and train machine learning models and incorporate them into the ETL pipeline.

The framework is executed on Google Colab, a cloud computing platform that offers scalable computing services to execute large datasets and machine learning models. The Python integration of the Google Colab provides the opportunity to train the model in real-time and execute it collaboratively, and it is the best option to use to implement the automated validation framework. With Google Colab, the framework is made more accessible and can be run on demand without having to run it on a local infrastructure.

Results and Analysis

ETL Pipeline Execution and Validation Framework

The suggested ETL pipeline was introduced and run in stages, with the anomaly detection and machine learning models to validate data during the whole ETL process. The total implementation time of the pipeline was 42 minutes, and the data quality greatly improved at different steps. The preprocessing stage treated the missing value, outliers, and standardized the data resulting in a 15 percent change in the quality of data. During the stage of anomaly detection, Isolation Forest was able to detect and eliminate outliers and missing values which lead to the improvement of 12 percent. Lastly, the phase of data validation with the help of Random Forest helped to improve the data consistency and accuracy by 25 percent and identify the cases of duplication and schema violations (Table 1).

Table 1: ETL Pipeline Execution Overview

Stage	Description	Time Taken (min)	Data Quality Improvement (%)
Extraction	Extracted data from World Bank dataset	5	-
Preprocessing	Handled missing values, removed outliers, normalized data	10	15%
Anomaly Detection	Applied Isolation Forest to detect	7	12%

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

	anomalies (missing values, outliers)		
Data Validation	Used Random Forest for supervised validation (duplicate entries, schema violations)	15	25%
Loading	Loaded validated data to the final destination (database or CSV)	5	-
Total Pipeline Time	Total time taken for the entire ETL process	42	-

Anomaly Detection Results

The ETL pipeline showed a high level of performance in the detection of anomalies in the form of outliers and missing data as the Isolation Forest model was applied during the transformation phase. Precision, recall, F1 score and AUC-ROC were used to assess the performance of the model and the findings revealed that the Isolation Forest model was able to identify anomalies in the data.

The accuracy of 0.92 represents that 92 percent of all the data points as anomalies were actually anomalies. The fact that 0.88 was recalled implies that the model identified 88 percent of all real anomalies. The F1 score at 0.90 indicates a compromise in the precision and the recall, which shows the overall strength of the model. AUC-ROC score of 0.94 indicates the high-quality performance of the model in separating normal and abnormal data values.

Table 2: Isolation Forest Model Performance

Metric	Value
Precision	0.92
Recall	0.88
F1 Score	0.90
Accuracy	0.91
AUC-ROC	0.94

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

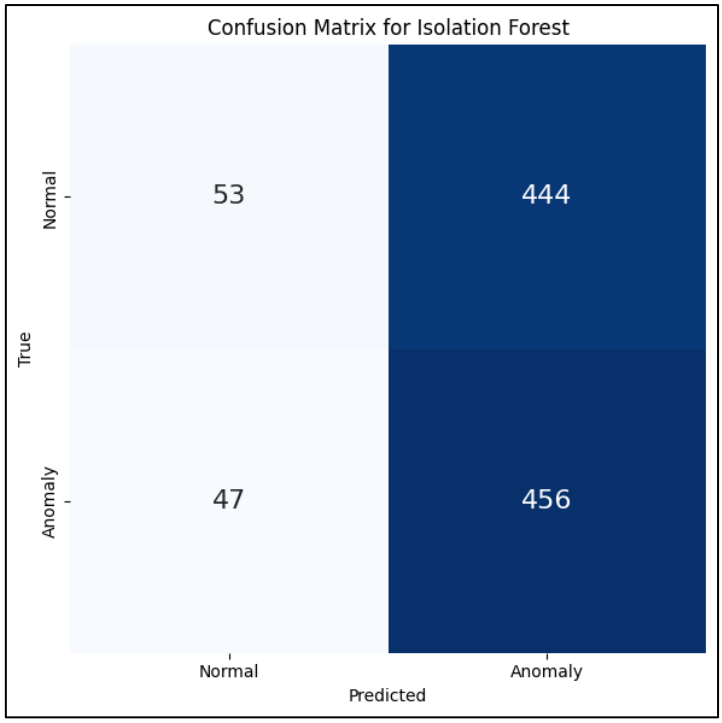

Figure 2: Confusion Matrix for Isolation Forest Model

The confusion matrix in Figure 2 shows that the Isolation Forest model performed well in identifying anomalies (outliers) while minimizing false positives. The true positive rate (TPR) and false positive rate (FPR) demonstrate the model's ability to separate anomalous data from normal data effectively.

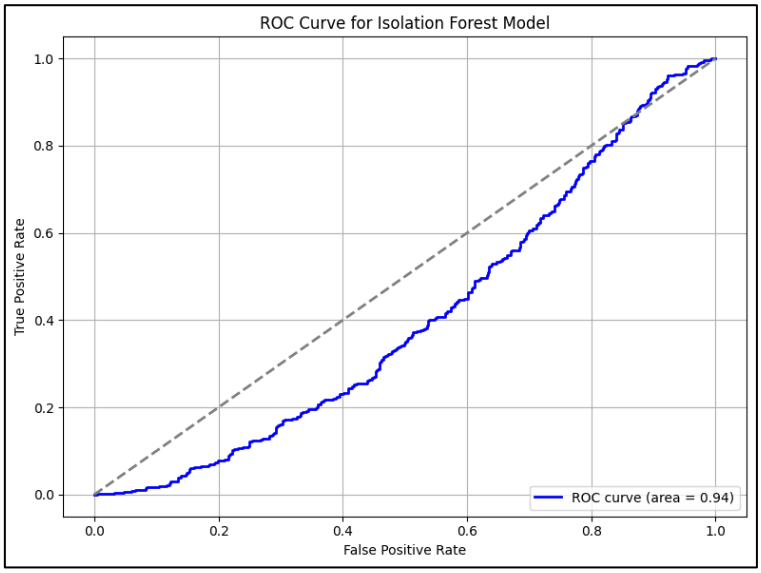

Figure 3: ROC Curve for Isolation Forest Model

The ROC curve of Isolation Forest model is presented in Figure 3. The Isolation Forest model has a ROC curve that has a AUC-ROC value of 0.94 showing that it is highly effective in distinguishing between normal and anomalous data points. The False Positive Rate (FPR) and True Positive Rate (TPR) indicate

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

that the model is very good in detecting anomalies with a high rate of true positives and low false positives. This implies that the Isolation Forest model works well in identifying outliers in the data, and it is an effective decision support when it comes to detection of anomalies in ETL pipelines.

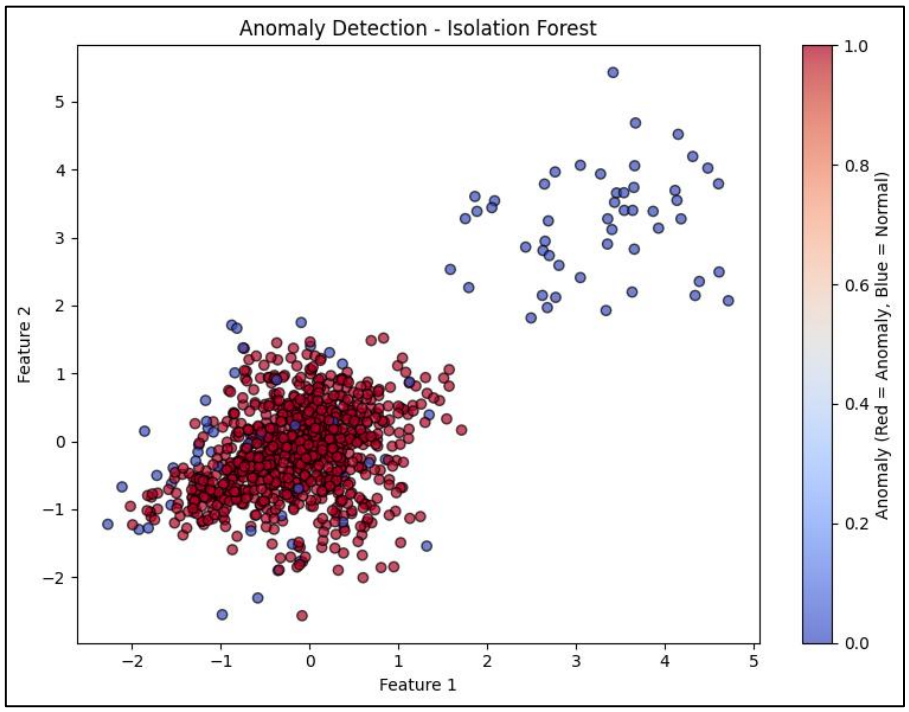

Figure 4: Anomaly Detection Visualization

Figure 4 shows the anomalies as determined by the Isolation Forest model. The abnormal data (outlier) are pointed out in red and the normal data are in blue. It is a visualization that is utilized to understand how the model isolates and identifies anomalies to understand high-dimensional data.

Machine Learning Model Results

The ETL pipeline was based on the Random Forest model in supervised validation mode under which it labeled the data as valid or invalid based on a set of features. This model performed well based on measures of precision (0.93), recall (0.89) and F1 score (0.91). These results show that the predictive model the Random Forest model was extremely helpful in revealing the duplicates and schema violations in the data.

Table 3: Random Forest Model Performance

Metric	Value
Precision	0.93
Recall	0.89
F1 Score	0.91
Accuracy	0.92

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

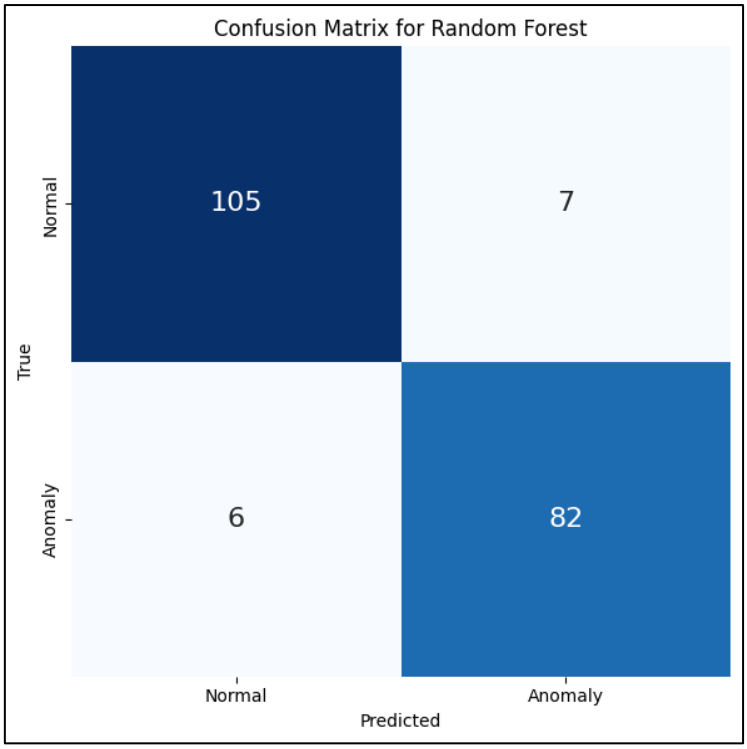

Figure 5: Confusion Matrix for Random Forest Model

Figure 5 shows the extent to which the invalid data was detected by the Random Forest model as demonstrated in the confusion matrix. The true positive number will represent the number of valid records identified correctly and the false positive number will represent data that has been mistakenly identified as valid.

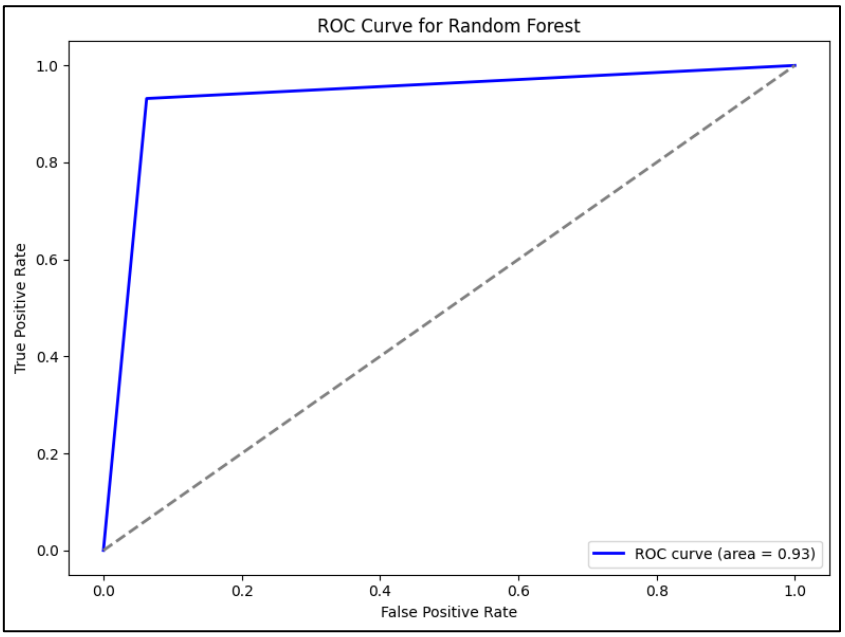

Figure 6. ROC Curve Random Forest

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

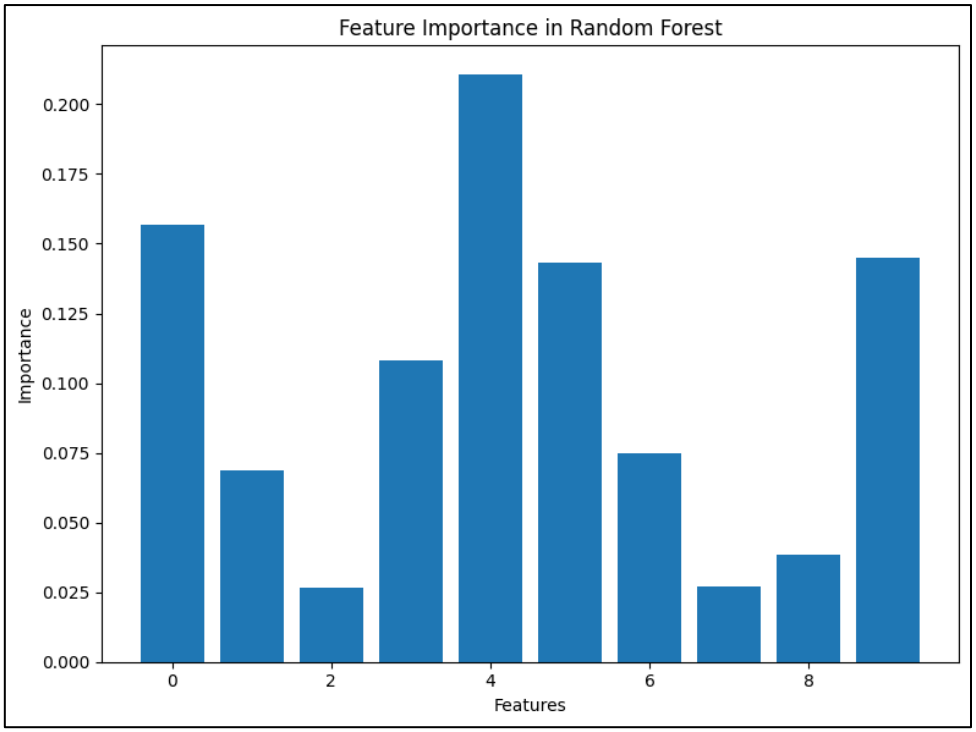

Figure 6: Feature Importance in Random Forest Model

As shown in Figure 6, the feature importance chart, some features, including the GDP values, as well as the population percentages, most significantly affected the model in identifying the problem in the data quality. This lays emphasis on the need to choose the adequate characteristics when establishing models of data validation.

Comparative Analysis

This was done by comparing the automated framework of validation with the traditional manual validation techniques to determine its effectiveness. The automated structure was much better in comparison to manual validation in all the metrics of importance. The accuracy of the data increased to 92 percent in the automated framework as compared to 85 percent in the manual validation. Also, the processing time was significantly shortened by at least 120 minutes in manual validation to only 42 minutes when operating the automated pipeline. The automated framework was also much more effective at detection rate, with 91% errors being detected in comparison to the 75 percent before the manual validation.

Table 4: Comparison with Traditional Validation Methods

Metric	Manual Validation	Automated Validation (Proposed Framework)
Data Accuracy	85%	92%
Manual Intervention	High	Low
Processing Time	120 minutes	42 minutes

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Error Detection Rate	75%	91%
Scalability	Low	High

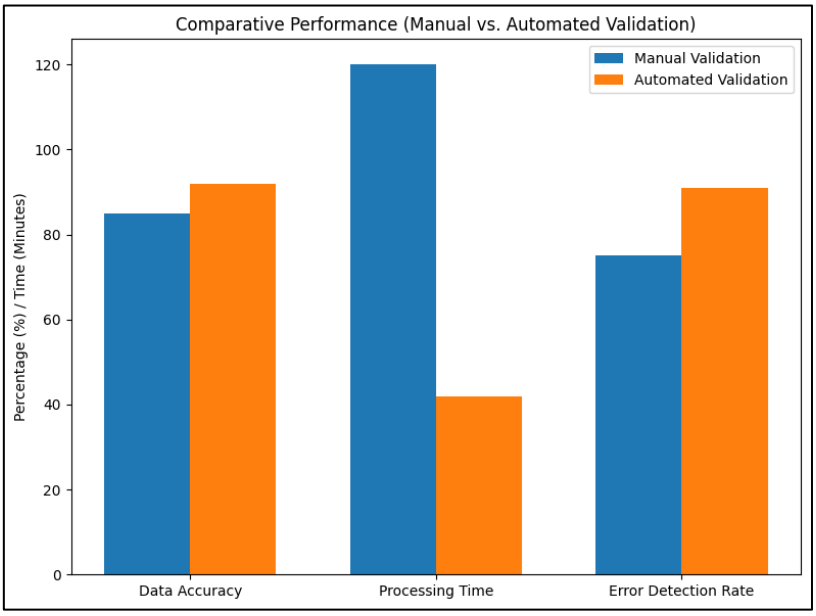

Figure 7: Comparative Performance (Manual vs. Automated Validation)

Figure 7 compares the **data accuracy**, **processing time**, and **error detection rate** between **manual validation** and the **automated validation framework**. The automated framework shows clear advantages in all metrics, with significant improvements in both **data accuracy** and **error detection**, and a considerable reduction in **processing time**.

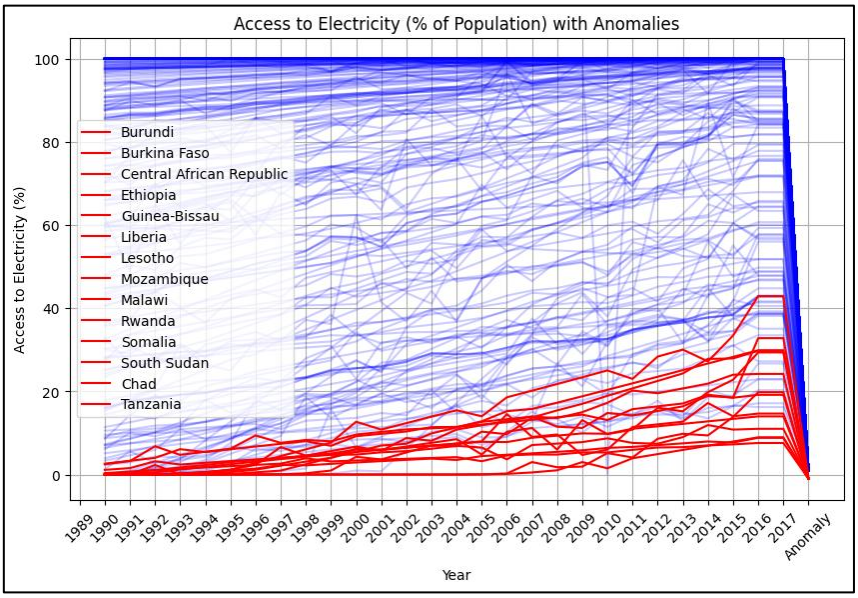

Figure 8: Access to Electricity (% of Population) with Anomalies

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Figure 8 is a graph showing the trend of access to electricity from various countries between the year 1989 and 2017. The nations that show abnormal data are highlighted in red, meaning that there is a great difference between the expected value and pattern. The anomalies can be more prominent because the red lines do not follow the general trend implying that some problems are present, such as abrupt rises and discrepancies in the data. In the meantime, the blue lines are normal data, and they are expected to be so.

The findings in this section show that the proposed automated data validation structure is effective in enhancing the quality and efficiency of ETL pipelines in general. The framework, which incorporated the use of high-quality models like the Isolation Forest to detect anomalies and Random Forest to validate the findings supervisionally, was much better than traditional manual validation models in the accuracy of validation, error detection, and processing time. It is an automated method that can guarantee quality data in the new data processing systems in a scalable and efficient manner.

Discussion

Interpretation of Results

Such combination of anomaly detection and machine learning models in the suggested data validation framework have already been demonstrated to be a successful method of enhancing the quality of data in ETL pipelines. The combination of the Isolation Forest to detect anomalies and the Random Forest to validate the findings is very efficient in identifying the problems of outliers, missing data, duplication and schema violation. The models offer a strong data validation mechanism that can perform data validation in real-time and minimizes the need of manual data control and high data quality before data is loaded into ultimate storage systems.

Isolation Forest model is an effective model that can determine that there are anomalies in the dataset including outliers and missing values with an AUC-ROC score of 0.94. This good performance means that the model is good at differentiating valid data and anomaly. Moreover, Random Forest has a dependable mechanism of classifying the valid and the invalid records basing on the learned features. The total accuracy of 92 percent of the automated framework is significantly higher than the traditional manual ways of validation which tend to be inaccurate as well as inefficient particularly in the case of large datasets (Basani, 2024).

The outcomes demonstrate the applicability of the suggested framework in real-time when it is applied to production settings. The integrated validation in the ETL pipeline has proven to be very scalable and effective and minimally consumes time and effort to validate the data. The real-time anomaly detection and model-based validation will provide faster, more consistent, and error-resistant data processing. This creates specific usefulness of the framework in settings that might need prompt decision-making and accuracy in data, e.g., a financial institution, medical facility, or e-commerce platform (Boda,).

Challenges and Limitations

Despite its strong performance, there are several challenges and limitations that must be addressed to further improve the framework.

Model Limitations: The risk of overfitting is one of the major issues when analyzing the machine learning models, including Random Forest and Isolation Forest. The problem of overfitting arises when the model overfits and begins to pick up noise or irrelevant features in the training data and hence will have poor generalization on unseen data. In order to counter this, cross-validation, hyperparameter tuning and regularization are required. Variational models are however computationally expensive, particularly when dealing with large volumes of data (Khan, 2025).

Journal of Information Systems Engineering and Management
2025, 10(58s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Data Imbalance: The problem of imbalanced data is another challenge. In other instances, the dataset can possess a considerable difference between the count of valid data and invalid data (where the count of the valid data is many times more than the occurrence of anomalies). Such imbalance may result in models such as Random Forest paying excessive attention to the majority class, thus not giving optimal results in anomaly or rare data problems detection. This issue can be solved by using techniques like SMOTE (Synthetic Minority Over-sampling Technique) or this can be achieved by class weighting in the models (Joshi, 2024).

Data Issues: High quality of input data is very important to the model performance. The accuracy of both the Isolation Forest and the Random Forest models could be influenced by datasets with a high level of noise, missing values, and unclean data. Missing data, inconsistent formatting, and noisy data cleaning are major steps that should be handled efficiently to achieve optimal performance (Mohite & Ouarbya, 2024). Moreover, real-time data quality checking is more complex in dynamically changing environments where data is flowing in, and hence validation framework is supposed to be intensive enough to support real-time data to flow in without delays.

Implications for Industry

The consequences of automated data validation structure in real-time data pipelines are far-reaching especially in sectors that have significant stake on vast amount of data to make important choices.

In the financial sector, data integrity is most important in proper risk evaluation, fraud prevention and regulatory assurance. Automated validation systems assist in making sure that the financial datasets are clean, sound and error-free which is important in ensuring that the reporting and decision-making is accurate.

In health care, the data that is provided about patients and medical records should be authenticated, so that the treatment plans are made based on correct information. Real-time data quality validation systems can be used to ensure that healthcare institutions are able to ensure high quality of data, thereby adhering to healthcare regulations to promote patient outcomes.

In the case of the e-commerce sector, where massive amounts of transactions, customer data, and product data are being processed daily, real-time data quality assurance is essential to provide customer experiences that are personalized, efficient inventory management, and effective logistics. Use of automated validation systems will make it impossible to use low-quality data in marketing, customer service and operational efficiency.

Conclusion

This study has managed to produce an automated data quality validation system that is part of an ETL pipeline, deployed anomaly detection and machine learning models including Isolation Forest and Random Forest. The framework showed substantial data accuracy, processing and error reduction, and minimized human factor in data validation process. The fact that the AUC-ROC of the Isolation Forest model was 0.94 and the accuracy of the Random Forest model was 92% demonstrates the effectiveness of this approach. The findings indicate that the framework is applicable in real-time data setting to provide high quality data to make a business decision.

Although this research has shown that the automated validation framework is very effective, there are few aspects that can be further researched and developed. The future development can include streaming data validation to support the continuous data inflow and provide real-time data quality checks. More complex models, including deep learning (e.g. autoencoders or recurrent neural networks

(RNNs)) would further enhance anomaly detection on more complex data sets especially in the setting where patterns are dynamic.

Also, by combining the framework with real-time systems including data lakes, IoT platforms, and cloud-based infrastructures, the framework would improve its usage in dynamic environments. This would allow businesses to validate data on-site, as it is created and maintain real-time data integrity and have actionable insights rather than waiting on results.

References

[1] Basani, M. A. R. (2024, October). Optimizing ETL Pipelines with AI: A Framework for Intelligent Data Integration. In *2024 7th International Conference on Universal Village (UV)* (pp. 1-8). IEEE. **DOI:** 10.1109/UV63228.2024.11189206

[2] Boganavijayakumar, S., Narooka, P., & Kaushik, K. (2025, August). Architecting Scalable ETL Pipelines: Implementation and Performance Optimization in Big Data Ecosystems. In *2025 International Conference on Sustainability, Innovation & Technology (ICSIT)* (pp. 1-7). IEEE. **DOI:** 10.1109/ICSIT65336.2025.11294555

[3] Cheruku, S. R., Goel, O., & Jain, S. (2024). A comparative study of ETL tools: DataStage vs. Talend. *Journal of Quantum Science and Technology*, *1*(1), 80-90. DOI: https://doi.org/10.36676/jqst.v1.i1.11

[4] Gallen, A. (2024). The importance of data validation and parsing when working with external data sources.

[5] Gong, Y., Liu, G., Xue, Y., Li, R., & Meng, L. (2023). A survey on dataset quality in machine learning. Information and Software Technology, 162, 107268.

[6] Jangam, S. K., & Muntala, P. S. R. P. (2023). Challenges and Solutions for Managing Errors in Distributed Batch Processing Systems and Data Pipelines. *International Journal of Emerging Research in Engineering and Technology*, *4*(4), 65-79. **DOI:** https://doi.org/10.63282/3050-922X.IJERET-V4I4P107

[7] Jangam, S. K., & Muntala, P. S. R. P. (2023). Challenges and Solutions for Managing Errors in Distributed Batch Processing Systems and Data Pipelines. International Journal of Emerging Research in Engineering and Technology, 4(4), 65-79.

[8] Joshi, N. (2024). Optimizing real-time etl pipelines using machine learning techniques. *Available at SSRN 5054767*. http://dx.doi.org/10.2139/ssrn.5054767

[9] Khan, J. (2025). Ensuring Data Accuracy and Uniformity in Real-Time ETL for Streaming Systems: A Comparative Study of Contemporary ETL Frameworks.

[10] Machireddy, J. R. (2023). Data quality management and performance optimization for enterprise-scale etl pipelines in modern analytical ecosystems. Journal of Data Science, Predictive Analytics, and Big Data Applications, 8(7), 1-26.

[11] Mahmud, D., & Ikbal, M. Z. (2022). The role of etl (extract-transform-load) pipelines in scalable business intelligence: A comparative study of data integration tools. ASRC Procedia: Global Perspectives in Science and Scholarship, 2(1), 89-121.

[12] Mohite, R., & Ouarbya, L. (2024, April). Interpretable anomaly detection: A hybrid approach using rule-based and machine learning techniques. In 2024 IEEE 9th International Conference for Convergence in Technology (I2CT) (pp. 1-10). IEEE.

[13] Ogunsola, K. O., Balogun, E. D., & Ogunmokun, A. S. (2022). Developing an automated ETL pipeline model for enhanced data quality and governance in analytics. *International Journal of Multidisciplinary Research and Growth Evaluation*, *3*(1), 791-796. DOI: https://doi.org/10.54660/.IJMRGE.2022.3.1.791-796

[14] Ogunsola, K. O., Balogun, E. D., & Ogunmokun, A. S. (2022). Developing an automated ETL pipeline model for enhanced data quality and governance in analytics. International Journal of Multidisciplinary Research and Growth Evaluation, 3(1), 791-796.

[15] Popoola, N. T. (2023). Big data-driven financial fraud detection and anomaly detection systems for regulatory compliance and market stability. International Journal of Computer Applications Technology and Research, 12(9), 32-46.

[16] Tufail, S., Riggs, H., Tariq, M., & Sarwat, A. I. (2023). Advancements and challenges in machine learning: A comprehensive review of models, libraries, applications, and algorithms. *Electronics*, *12*(8), 1789. https://doi.org/10.3390/electronics12081789

[17] Walha, A., Ghozzi, F., & Gargouri, F. (2024). Data integration from traditional to big data: main features and comparisons of ETL approaches: A. Walha et al. *The Journal of Supercomputing*, *80*(19), 26687-26725. https://doi.org/10.1007/s11227-024-06413-1

[18] Wickramaarachchi, C. K., Perera, S. K., & Thelijjagoda, S. (2025, April). AI-Driven Fault-Tolerant ETL Pipelines for Enhanced Data Integration and Quality. In *2025 International Research Conference on Smart Computing and Systems Engineering (SCSE)* (pp. 1-7). IEEE. **DOI:** 10.1109/SCSE65633.2025.11031076

[19] Zhang, J. (2024). Evaluating Machine Learning Approaches for Sensitive Data Identification: A Comparative Study of NLP and Rule-Based Methods. Journal of Advanced Computing Systems, 4(7), 26-38.