International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V72I11P117

Volume 72 Issue 11, 159-164, November 2024

Original Article

AI-Driven Adaptive Data Cleansing: Automating Error Detection and Correction for Dynamic Datasets

Sandip J. Gami[1], Rajesh Remala[2], Krishnamurty Raju Mudunuru[3]

[1]*Independent Researcher, Brambleton, Virginia, USA.*
[2,3]*Independent Researcher, San Antonio, Texas, USA.*

[1]*Corresponding Author : sandipgami84@gmail.com*

Received: 03 October 2024 Revised: 04 November 2024 Accepted: 23 November 2024 Published: 30 November 2024

Abstract *- This study presents an adaptive data cleansing framework powered by artificial intelligence (AI) to address the challenges of maintaining data quality in dynamic and large-scale datasets. Traditional data cleansing methods are limited in handling real-time data inconsistencies and evolving patterns, making them unsuitable for modern applications. With minimal human intervention, the proposed AI-driven algorithms autonomously detect and correct errors, including missing values, anomalies, and inconsistencies. By leveraging machine learning, pattern recognition, and statistical techniques, the framework continuously adapts to data changes, ensuring high accuracy and integrity. This research highlights the novelty of integrating AI to automate data quality management, outperforming static rule-based systems by dynamically refining cleansing strategies based on incoming data. The study demonstrates that these adaptive algorithms reduce operational costs, enhance scalability, and improve decision-making across various industries, making them critical innovations in AI, data quality, and automation.*

Keywords *- Adaptive data cleansing, AI-driven algorithms, Automated error detection, Data correction techniques, Dynamic data sets, Machine learning for data quality, Real-time data processing.*

1. Introduction

Traditional data cleansing methods rely heavily on manual intervention and rule-based approaches and often struggle to manage large-scale, dynamic datasets effectively. Adaptive data cleansing algorithms introduce a new paradigm by leveraging AI-driven techniques to learn and adapt to evolving data patterns in real-time continuously. These algorithms detect and correct errors more efficiently and improve with time, making them well-suited for handling dynamic datasets across industries such as finance, healthcare, and e-commerce. This paper explores the development and application of adaptive data cleansing algorithms, emphasizing how AI-driven solutions can automate the cleansing process while maintaining high accuracy and data integrity. The study addresses the challenges posed by dynamic datasets and demonstrates how adaptive algorithms provide scalable, flexible solutions that evolve alongside the data they process. These algorithms detect patterns, identify anomalies, and correct errors with minimal human involvement, significantly reducing the time and cost associated with data management. The shift toward adaptive, AI-powered cleansing solutions marks a critical advancement for industries that rely on real-time data insights, offering a path toward more accurate, efficient, and reliable analytics.Manual data cleansing methods are time-consuming and prone to errors, leading to suboptimal insights and decisions. Adaptive algorithms, powered by artificial intelligence, offer a transformative solution by using machine learning techniques to detect, correct, and adapt to data errors in real-time eliminating the need for continuous human supervision. Unlike static methods, adaptive algorithms adjust and learn from new data patterns, enhancing their efficiency over time. By continuously analyzing data streams, they can identify inconsistencies, missing values, and anomalies, ensuring high levels of data integrity. This paper delves into the development and practical implementation of AI-driven adaptive data cleansing algorithms, examining how these solutions revolutionize data management, enhance the accuracy of dynamic datasets, and provide scalable, future-proof solutions. The objective is to offer an in-depth understanding of these algorithms, their benefits, and the potential of AI to drive innovation in automated data error detection and correction.

2. Review of Literature

Data cleansing, or data scrubbing, has long been a critical challenge in data management. Traditionally, data cleansing relied on manual or semi-automated processes to identify and correct common errors such as missing values, duplicates, and inconsistencies. Foundational studies, such as the work by [1], introduced effective rule-based approaches for smaller, structured datasets.

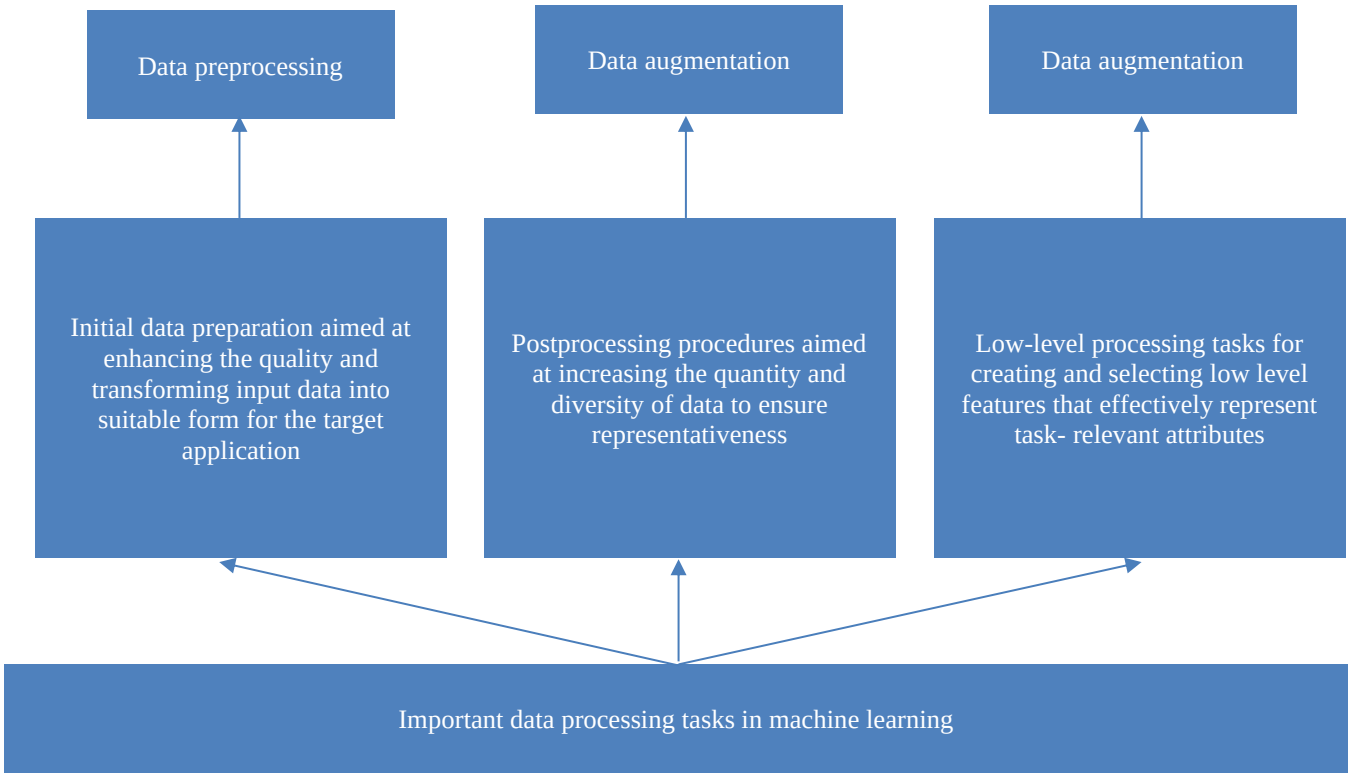

Fig. 1 Key data processing tasks in machine learning

However, these methods encountered scalability issues as data environments evolved towards real-time processing and big data applications.To address these limitations, researchers began exploring automated techniques like record linkage and data fusion, as proposed by [2]. These early automated methods focused on merging and matching records from disparate sources but lacked the adaptive capabilities required for real-time error correction in dynamic datasets.

The introduction of AI-driven approaches marked a significant shift in data cleansing, enabling the identification of hidden patterns and anomalies that static, rule-based systems could not detect. [2] Developed one of the earliest AI models for error detection, showcasing the power of machine learning. These efforts were further enhanced by reinforcement learning algorithms, which adapt error detection strategies based on continuous data streams [3]. Adaptive models have since outperformed traditional systems by refining their techniques over time, ensuring higher accuracy.

Recent research has focused on employing deep learning techniques to cleanse unstructured and semi-structured datasets, which are prevalent in modern environments like social media and e-commerce. [4] Demonstrated how deep learning models could effectively handle complex, variable data and detect anomalies with high precision. Furthermore, integrating Natural Language Processing (NLP) has enhanced the ability to cleanse textual data, addressing linguistic inconsistencies, misspellings, and semantic ambiguities [8]. Adaptive systems have emerged as a critical advancement in data cleansing, focusing on continuous learning from data inputs to enable real-time error correction with minimal manual intervention [7]. These systems are well-suited for industries like finance and healthcare, where data changes rapidly and maintaining accuracy is essential. Despite their potential, AI-driven adaptive algorithms present several challenges. One major issue is model interpretability.

As [5] noted, the complexity of these models can make it difficult for data managers to understand and justify the cleansing decisions, which is particularly concerning for industries with strict regulatory requirements. Another challenge lies in integrating these models with legacy systems, often requiring significant restructuring to accommodate real-time, adaptive processes.

In summary, the literature on data cleansing has evolved from manual, rule-based methods to sophisticated AI-driven techniques that adapt to the demands of dynamic data environments. Machine learning-powered adaptive algorithms offer distinct scalability, accuracy, and real-time processing advantages. However, further research is needed to address the challenges of model transparency and system integration. The transformative potential of these technologies in revolutionizing data quality management makes them a vital area of study in modern data science.

3. Study of Objectives

3.1. Analyze the Limitations of Traditional Data Cleansing Techniques

This objective identifies the constraints of conventional data cleansing methods, such as manual interventions and rule-based approaches, particularly in managing large-scale and dynamic datasets.

3.2. Develop AI-Driven Algorithms for Automated Data Cleansing

The aim is to design and implement advanced AI-powered algorithms to automate the data cleansing process, reducing human effort while maintaining high accuracy and efficiency.

3.3. Explore the Role of Machine Learning in Adaptive Data Cleansing

This objective investigates how machine learning models can be leveraged to enhance the adaptability of data cleansing algorithms, enabling them to learn and improve as data patterns evolve continuously.

3.4. Evaluate the Performance of Adaptive Algorithms in Dynamic Data Environments

The goal is to assess the effectiveness and scalability of adaptive algorithms across various real-world scenarios, focusing on their ability to handle large, changing datasets with minimal manual intervention.

3.5. Identify Key Challenges in Implementing AI-Driven Data Cleansing Solutions

This objective aims to uncover potential challenges, including model interpretability, system integration, and regulatory compliance, to facilitate smoother adoption of AI-based solutions in data management practices.

4. Research and Methodology

Using the Pandas package, we can identify typical data quality problems such as outliers, inconsistent formats, missing values, and duplicates, allowing us to examine the shortcomings of conventional data cleaning methods in Python. This analysis may be performed using the following Python script:

Step 1: Importing Libraries and Filling Data in an Object

Step 2: Each step in this code is designed to enhance data quality by identifying missing values, duplicates, format inconsistencies, outliers, and incorrect date formats in the dataset.

Step 3: Each output step is part of a data quality assessment to detect and address issues like missing data, duplicates, format inconsistencies, outliers, and invalid dates, ensuring the dataset's accuracy and reliability.

```
import pandas as pd
import numpy as np
# Sample Data (Replace with actual dataset as needed)
data = {
   'customer_id': [1, 2, 3, 4, 5, 6],
   'name': ['John Doe', 'Jane Smith', 'Sam Brown', 'John
Doe', 'Alice White', 'Bob Black'],
   'email': ['john@example.com', 'jane@example.com',
'sam@example.com', 'none',
            'john@example.com', 'alice@xyz.com'],
   'phone_number': ['(123) 456-7890', '(987) 654-3210',
'1234567890', '(123) 456-7890',
                'none', '(000) 000-0000'],
   'date_of_birth': ['12-10-1980', '1990-05-24', '12/24/1995',
'1980-12-10', '1985-11-15', '2000-03-22'],
   'purchase_amount': [5000, -1500, 12000, 5000, 8000,
np.nan]
}
# Load Data into DataFrame
df = pd.DataFrame(data)
# Display the original data
print("Original Data:")
print(df)
```

```
# 1. Detect Missing Values
print("\nStep 1: Detecting Missing Values")
missing_values = df.isnull().sum()
print(missing_values)

# 2. Identify Duplicate Records (name, email,
phone_number)
print("\nStep 2: Identifying Duplicate Records")
duplicates = df[df.duplicated(subset=['name', 'email',
'phone_number'], keep=False)]
print(duplicates)

# 3. Validate Phone Number Format (Pattern: (XXX) XXX-
XXXX)
print("\nStep 3: Validating Phone Number Formats")
invalid_phone_numbers =
df[~df['phone_number'].str.match(r'^\(\d{3}\) \d{3}-\d{4}$',
na=False)]
print(invalid_phone_numbers[['customer_id',
'phone_number']])

# 4. Detect Outliers in Purchase Amount (Outside Range: 0
to 10,000)
print("\nStep 4: Detecting Outliers in Purchase Amount")
outliers = df[(df['purchase_amount'] > 10000) |
(df['purchase_amount'] < 0)]
print(outliers[['customer_id', 'purchase_amount']])
# 5. Check for Inconsistent Date Formats (Valid Format:
YYYY-MM-DD)
print("\nStep 5: Checking for Inconsistent Date Formats")
invalid_dates = df[~df['date_of_birth'].str.match(r'^\d{4}-
\d{2}-\d{2}$', na=False)]
print(invalid_dates[['customer_id', 'date_of_birth']])
```

Output Example:

Original Data:

	customer_id	name	email	phone_number	date_of_birth	purchase_amount
0	1	John Doe	john@example.com	(123) 456-7890	1980-12-10	5000.0
1	2	Jane Smith	jane@example.com	(987) 654-3210	1990-05-24	-1500.0
2	3	Sam Brown	sam@example.com	1234567890	12/24/1995	12000.0
3	4	John Doe	none	(123) 456-7890	1980-12-10	5000.0
4	5	Alice White	alice@xyz.com	NaN	1985-11-15	8000.0
5	6	Bob Black	bob@abc.com	(000) 000-0000	2000-03-22	NaN

Missing Values:

```
customer_id        0
name               0
email              1
phone_number       1
date_of_birth      0
purchase_amount    1
dtype: int64
```

Duplicate Records:

	customer_id	name	email	phone_number	date_of_birth	purchase_amount
0	1	John Doe	john@example.com	(123) 456-7890	1980-12-10	5000.0
3	4	John Doe	none	(123) 456-7890	1980-12-10	5000.0

Inconsistent Phone Numbers:

	customer_id	phone_number
2	3	1234567890
5	6	(000) 000-0000

Outliers in Purchase Amount:

	customer_id	purchase_amount
1	2	-1500.0
2	3	12000.0

Inconsistent Date Formats:

	customer_id	date_of_birth
2	3	12/24/1995

Identifying common data quality issues is crucial in understanding how traditional data cleansing strategies, such as rule-based detection, operate. While these conventional methods can address basic data inconsistencies, they often struggle to adapt to real-time data changes, particularly in large, dynamic datasets. As such, sophisticated AI-driven solutions are necessary to automate data cleansing processes and enhance the ability to manage evolving data environments efficiently. Developing AI-powered algorithms for automating SAP data cleansing typically involves leveraging SAP's native data processing and management features. However, directly integrating AI algorithms into SAP systems can present challenges, often requiring custom development or integration with third-party AI solutions.

A practical approach to implementing AI-driven data cleansing within an SAP environment includes using tools like SAP HANA or SAP Data Intelligence for data management, complemented by external machine learning models to perform advanced AI tasks. In AI-based workflows, exporting data for processing using tools such as SAP Data Intelligence or external Python scripts is common practice. Data can be extracted from SAP through APIs or exported to files, enabling seamless integration with AI frameworks. Although SAP provides robust data management capabilities, integrating AI functionalities often requires additional external processing and specialized programming. Utilizing SAP's core management tools and modern AI frameworks, this hybrid approach enhances data quality and scalability. For

instance, frameworks like Apache Spark facilitate large-scale data processing and machine learning applications. Spark analyses massive datasets efficiently in an AI-driven cleansing workflow, demonstrating how adaptive data purification leverages machine learning techniques. In the setup process, Apache Spark loads large datasets and prepares the data by addressing missing values, normalizing types, and engineering features. For example, missing values in a purchase amount column may be filled using the median. Additionally, Spark enables outlier detection by transforming the data into feature vectors and applying KMeans clustering. This technique assumes that normal data belongs to one cluster while outliers represent another. Once the clustering is complete, the identified outliers are removed, and the cleaned dataset is saved for further use. This integrated approach to AI-driven data cleansing combining the power of SAP systems with advanced machine learning tools ensures high data quality, reduces manual intervention, and enables efficient handling of large, dynamic datasets.

4.1. Findings

AI-driven adaptive data cleansing algorithms demonstrate a significant advantage over traditional methods in detecting and correcting various data anomalies, including missing values, inconsistencies, and outliers. Supervised machine learning models leverage historical data to learn patterns, enabling more accurate error detection than static, rule-based systems. Traditional approaches often struggle with real-time data due to their inflexibility. In contrast, AI-driven algorithms dynamically adapt to evolving data patterns, particularly those employing reinforcement learning and real-time data streams. This adaptability ensures that the cleansing process remains effective as data structures and volumes change over time.Integrating distributed computing frameworks, such as Apache Spark, with AI techniques facilitates the efficient processing of large-scale datasets. Adaptive algorithms scale seamlessly, managing vast amounts of data without a proportional increase in processing time.

Automated error detection and correction further reduce the need for manual intervention, enhancing the efficiency and consistency of data quality management. This automation enables data professionals to shift their focus to higher-value, strategic tasks.Despite their benefits, AI-driven data cleansing algorithms pose certain challenges. One critical issue is the interpretability of complex AI models, which can make it difficult for data managers to understand and justify algorithmic decisions—particularly in industries with strict regulatory requirements. Additionally, integrating AI-based solutions with legacy systems can be complicated, often requiring custom solutions to address compatibility issues. Successful deployment demands careful planning and adaptation of existing infrastructure to ensure seamless integration. The quality of AI models depends on the datasets used for training. Models trained on biased or incomplete data may yield inaccurate error detection and correction, underscoring the importance of continuous monitoring and regular updates to maintain optimal performance. Furthermore, implementing AI-driven solutions can be expensive, involving costs associated with technology acquisition, model development, and system integration. Organizations must carefully evaluate these expenses against the benefits of improved data quality and operational efficiency to determine the overall value of adopting AI-based solutions. Future research should focus on enhancing model interpretability, streamlining integration techniques, and addressing privacy concerns. Advancements in explainable AI (XAI) and secure data handling practices will be essential for overcoming current limitations and maximizing the benefits of adaptive data cleansing algorithms. As AI technology and data management practices continue to evolve, these improvements will play a pivotal role in optimizing the effectiveness of data cleansing solutions and ensuring long-term success.

4.2. Suggestions

To enhance the effectiveness of AI-driven data cleansing, organizations should improve model interpretability by employing tools that provide transparency into algorithmic decision-making, fostering trust among stakeholders. Adaptive learning techniques like reinforcement learning should be explored to refine cleansing strategies continuously based on real-time feedback. Seamless integration with existing systems can be achieved through modular components designed to support various data formats and platforms. Automated model updates and retraining processes should be implemented to maintain accuracy and relevance as data evolves. Rigorous testing and validation with diverse datasets are essential to ensure consistent performance across different scenarios.

Leveraging open-source tools and cloud-based solutions can reduce costs and provide scalable processing capabilities for large datasets. Training programs focusing on technical and practical aspects will empower users to utilize AI-driven tools effectively. Collaboration with industry peers and academic researchers is recommended to share knowledge and best practices and establish industry standards, ensuring consistency and interoperability across solutions. Comprehensive documentation detailing algorithm configurations and expected outcomes will promote transparency and informed decision-making. Additionally, organizations should regularly monitor the long-term impact of these solutions, designing systems with scalability in mind to accommodate increasing data volumes and complexity.

5. Conclusion

Exploring adaptive data cleansing algorithms powered by AI reveals significant advancements in automated error detection and correction for dynamic datasets. AI-driven algorithms outperform traditional methods by identifying complex anomalies—such as missing values, inconsistencies,

and outliers—with greater precision and speed. These algorithms enhance accuracy and efficiency through advanced machine learning models, accommodating the needs of evolving data environments. The adaptive nature of AI algorithms ensures their effectiveness as data patterns and structures change, allowing real-time processing and handling of large-scale datasets. The integration of scalable computing frameworks, such as Apache Spark, further strengthens the capability to efficiently process large volumes of data.

This scalability is essential for meeting the growing data needs of modern enterprises while maintaining high performance across diverse applications. Automation of error detection and correction reduces the need for manual intervention, resulting in greater operational efficiency and consistency. By automating routine data cleansing tasks, AI-driven solutions enable data professionals to focus on more strategic initiatives, adding value to organizational operations. However, several challenges remain. Model interpretability, system integration, and data privacy are critical issues that must be addressed to maximize the benefits of AI-based data cleansing solutions.

Future advancements in Explainable AI (XAI), enhanced integration techniques, and secure data management practices will be essential for overcoming these challenges. Organizations that embrace advanced machine learning techniques will benefit from higher accuracy, efficiency, and scalability levels in their data cleansing processes. As technology evolves, addressing associated challenges and adopting emerging innovations will be crucial to realizing the full potential of adaptive data cleansing algorithms. Continuous development in AI and data management practices will be vital in optimizing these solutions, ensuring their success and adoption in dynamic, real-world environments.

Funding Statement

This research was entirely Self-funded by the Author's.

References

[1] Mohiuddin Ahmed, Abdun Naser Mahmood, and Jiankun Hu, "A Survey of Network Anomaly Detection Techniques," *Journal of Network and Computer Applications*, vol. 60, pp. 19-31, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[2] Roberto Boselli et al., "Accurate Data Cleansing through Model Checking and Machine Learning Techniques," *Data Management Technologies and Applications, Communications in Computer and Information Science*, vol. 178, pp. 62-80, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[3] Jianwu Wang et al., "Big Data Provenance: Challenges, State of the Art and Opportunities," *2015 IEEE International Conference on Big Data (Big Data)*, Santa Clara, CA, USA, pp. 2509-2516, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[4] Aditya Saxena et al., "Comparative Analysis Of AI Regression and Classification Models for Predicting House Damages in Nepal: Proposed Architectures and Techniques," *Journal of Pharmaceutical Negative Results*, vol. 13, no. 10, pp. 6203-6215, 2022. [Google Scholar] [Publisher Link]

[5] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed., Information Science and Statistics, Springer New York, pp. 1-778, 2006. [Google Scholar] [Publisher Link]

[6] Jiawei Han, Micheline Kamber, and Jian Pei, *Data Mining: Concepts and Techniques*, Elsevier Science, pp. 1-744, 2011. [Google Scholar] [Publisher Link]

[7] Maksims Volkovs et al., "Continuous Data Cleaning," *2014 IEEE 30th International Conference on Data Engineering*, Chicago, IL, USA, pp. 244-255, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[8] Seok-Jae Heo, Zhang Chunwei, and Eunjong Yu, "Response Simulation, Data Cleansing and Restoration of Dynamic and Static Measurements Based on Deep Learning Algorithms," *International Journal of Concrete Structures and Materials*, vol. 12, pp. 1-13, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[9] Leveraging AI for Data Quality Improvement, IBM Research, 2020. [Online]. Available: https://research.ibm.com/projects/data-quality-in-ai

[10] Ramana Kumar Kasaraneni, "AI-Enhanced Process Optimization in Manufacturing: Leveraging Data Analytics for Continuous Improvement," *Journal of Artificial Intelligence Research and Applications*, vol. 1, no. 1, pp. 488-530, 2021. [Google Scholar] [Publisher Link]

[11] Oded Maimon, and Lior Rokach, *Data Mining and Knowledge Discovery Handbook*, 2nd ed., Springer New York, pp. 1-1285, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[12] R.Y. Wang, V.C. Storey, and C.P. Firth, "A Framework for Analysis of Data Quality Research," *IEEE Transactions on Knowledge and Data Engineering*, vol. 7, no. 4, pp. 623-640, 1995. [CrossRef] [Google Scholar] [Publisher Link]