International Journal of Emerging Trends in Computer Science and Information Technology
ISSN: 3050-9246 | https://doi.org/10.63282/3050-9246.IJETCSIT -V5I1P104
Eureka Vision Publication | Volume 5, Issue 1, 31-40, 2024

Original Article

Automated Data Cleaning: AI Methods for Enhancing Data Quality and Consistency

Mr. Rahul Cherekar
Independent Researcher, USA.

Abstract - Data cleaning is a data preprocessing step that involves finding and dealing with errors to enhance data quality. It is therefore important to pay particular attention to the quality of data as its poor quality brings about the wrong conclusions when making analyses. Using AI techniques for automatic data detecting and cleaning in big data is one of the most efficient and rapid solutions that can be implemented to reach high data reliability. This paper analyses different methodologies for data cleaning using artificial intelligence, namely, rule cycling, machine learning, deep learning, and a combination of the three. As seen in this paper, these techniques have their strengths, making their application ideal and their weaknesses, which should be considered before implementation. Also, we provide practical examples in healthcare, finance, and business intelligence, which only proves the efficiency of data cleaning via AI tools. The experimental results indicate that all four implemented approaches increase the data quality and decrease the manual work done by human beings. In this paper, the latest AI application techniques in the data-cleaning process have been critically discussed to guide researchers and practitioners.

Keywords - Automated Data Cleaning, Artificial Intelligence, Machine Learning, Data Quality, Consistency, Data Preprocessing, Error Detection, Data Imputation.

1. Introduction

1.1 Importance of Data Quality

Accurate data is crucial in any organization as it helps decision-making, operations, and strategic planning in different fields. [1-4] The consequences of poor data quality include wrong conclusions, losses, and poor decisions. The following sub-sections discuss the main facets of data quality and its importance.

Figure 1. Importance of Data Quality

- **Accurate Decision-Making:** Data-driven decision-making relies on accurate, complete, consistent data. For instance, in healthcare, finance, and business analytics, wrong or missing data may result in wrong decisions, diagnoses, or financial decisions. High-quality data enables organizations to make sound decisions based on facts and not guesswork, increasing the organisation's reliability.
- **Operational Efficiency:** Poor quality data leads to mistakes, duplication and time wastage, increasing operational costs. High-quality data in organizations leads to efficient operations, fast processing and enhanced automation since

there is little need for manual intervention to correct errors. This is especially the case in supply chain management, banking and customer service industries where efficiency is the order of the day.

- **Compliance and Regulatory Requirements:** Some of the sectors that require compliance include the finance sector, the healthcare sector, and the e-commerce sector, among others, which need to adhere to regulations such as GDPR and HIPAA. Poor data quality can cause non-compliance, legal problems and fines. This is important in ensuring that the organizations achieve the set standards of the regulatory authorities, customer data privacy and integrity.
- **Enhanced Customer Experience:** Customer satisfaction is a function of individual, relevant, and timely information exchange. Proper data quality can lead to wrong billing, delayed services, or a wrong marketing campaign. High-quality and well-structured data can help businesses provide better customer service, personalized recommendations, and better customer interaction, resulting in increased customer satisfaction and brand loyalty.
- **Fraud Detection and Risk Management:** Using inaccurate or outdated information exposes the organization to fraud, security threats and financial losses. Data quality is crucial for identifying anomalies, mitigating cyber threats, and addressing financial risks. For instance, in the banking and insurance industries, AI-based fraud detection models require clean data to detect and prevent fraudulent activities.
- **Competitive Advantage and Business Growth:** Companies that emphasise data quality highly can use accurate information for decision-making and market analysis. Accurate data helps firms understand trends in the market, in marketing, and in improving the products offered to the market. Companies with a good data management strategy are likely to benefit from it by responding to market conditions, creating new products and services, and sustaining their operations in the long run.

1.2 Role of AI in Data Cleaning

Data cleaning is a very important process in data preprocessing since it helps ensure that the data collected is clean, consistent and free from errors before it is used for analysis and decision-making. The conventional data cleaning methods include a rule-based approach and manual data cleaning, which are costly, labor-intensive, and ineffective when applied to large datasets. AI has positively changed data cleaning through automation, accuracy, and process scalability. Automated data cleaning uses machine learning and deep learning to identify and correct errors in the data without much human input. The supervised learning models, like the Random Forest and the Support Vector Machines (SVM), can be trained on the labelled data to classify and correct the wrong entries. [5,6] K-means clustering and Isolation Forest are the unsupervised learning techniques used to identify anomalies and outliers when no specific rules are followed. The data is unstructured and not labelled.

Autoencoders and Transformers are other examples of deep learning models that go further in data cleaning by learning patterns and imputing missing or corrupted data. Autoencoders are very useful in imputing missing values because they can learn from patterns in the data, while Transformers are useful in handling large-scale data with high accuracy. These models make it possible to address even minor issues, such as inconsistency in textual data, duplicate records, or incorrect numerical data. Also, NLP can be used to pre-process text data and make it more suitable for analysis, which is useful for cleaning customer feedback, social media data, and medical records. This is because AI has the advantage of learning from past data and improving its performance compared to other traditional methods. In conclusion, using AI in data cleaning improves data quality, minimizes the time spent on data cleaning, and guarantees that datasets are ready for analysis, making it an essential tool for data-driven industries.

1.3 Challenges in Data Cleaning

Data cleaning is an important step in data preprocessing since real-life data sets are usually noisy, contain errors, and may have missing values. The traditional data cleaning method is done manually, which is very tiresome, time-consuming, and full of errors since it involves using humans to go through the data set. Several challenges make data cleaning a complex and continuous process in data management. Among the most frequent problems in data cleaning, there is missing data where some information is missing or incomplete. Data can be missing due to human mistakes, technical problems, or because of the way data was collected. Missing values are data values that are either unavailable or have not been recorded and can be handled by imputation techniques, deletion or statistical interpolation, depending on the type of data and the field.

Another important problem is the presence of the duality of records that may be created due to multiple entries of the same data, data entry mistakes, or system failures. Duplicate records are costly in terms of storage and distort the results and conclusions. It is challenging to merge duplicates since the entries may have slight spelling, formatting, or structure differences. The other data quality issue is related to data type, where numerical data may be stored as text or date formats may differ between records. Such discrepancies can lead to processing problems, calculation failures, and incompatibility with analytical tools, meaning a lot of time must be spent on preprocessing. This is made worse because the datasets may have different formatting conventions. Differences in the units of measurement, currency, address, and categories may confuse the data analysis. Manually formatting large datasets to have a standard format is time-consuming and can be accompanied by errors.

2. Literature Survey

2.1 Traditional Data Cleaning Techniques

Conventional data cleaning finds the mistakes in data by applying business, rule-based systems and statistical methods. Statistical techniques are based on recipes induced directly from the dataset under analysis to identify problematic data, such as missing values, doubles or records with an incorrect format. Still, they are easy to implement, but they show low performances related to large, complicated, and evolving data, which can be easily not covered in predefined rules. [7-11] Much like statistical methods inspect for deviation based on distribution, statistical techniques turn to statistical measures to detect them. However, they may need an expert to calibrate the individual values. Although the highly controlled settings are quite useful, their major drawback is that they are not very portable and flexible, and the data environment is developing quickly.

2.2 AI-Driven Approaches

This is especially because artificial intelligence innovations have enhanced data-cleaning activities. An AI component covers both supervised and unsupervised learning to spot and rectify data anomalies. According to supervised learning models that involve receiving predetermined training data and samples, the errors can be classified, and the possible correct solutions can be recommended with a very high degree of precision. These learning strategies are identified depending on the input and include clustering and anomaly detection since these methods are ideal for fixed and technically dynamic datasets. Also, data cleaning is supported by deep learning models, which work with complex and unstructured data and can identify the dependencies between them. These are rather more flexible and effective in comparison with rule-guided assistants.

2.3 Comparative Analysis

Every data cleaning approach has advantages and disadvantages, which means they are useful in different situations to some extent. Procedurization is easy to understand and use in the organization but is not open to new circumstances or conditions. Most machine learning models provide the facility to learn with time, so they get better and better over time for the next time with the same data, but they are very sensitive to the amount and kind of labelled data they are trained on. Neural networks are one of the deep learning gadgets that work best with images and text since these are considered complex data sets. However, they have higher computational complexity and require large amounts of data for processing.

2.4 Application Domains

In this case, there are numerous industries and sectors in which the employment of AI data cleaning was discovered to be highly important. In the financial sector, AI techniques detect the possibility of fraudulent payments and financial frauds to prevent them, check financial rules, and make precise accounts. In the healthcare field, automating data cleaning enhances essential records of the patient to be accurate to facilitate planning of diagnosis and other important procedures. In e-commerce, AI utilized data cleaning to improve client satisfaction by optimizing the required product recommendations and inventory and minimising irrelevant or fake content. These cases show that the application of artificial intelligence increases data credibility and quality in multiple industries, improving decision-making and organizational performance.

3. Methodology

3.1 AI Techniques for Data Cleaning

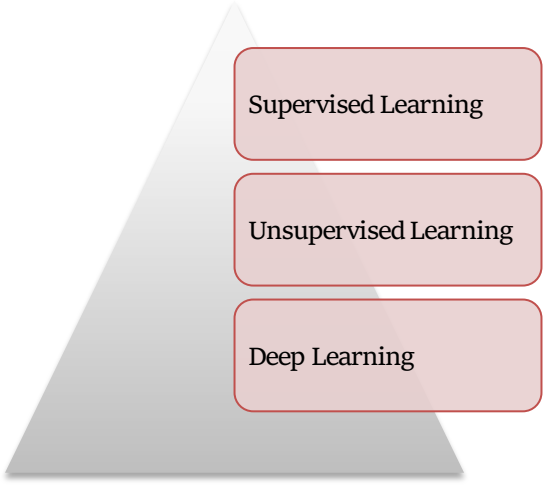

Figure 2. AI Techniques for Data Cleaning

- **Supervised Learning:** Classification models are generally used in data cleaning especially errors or anomalies detection and correction. [12-16] These models are learned through labeled data in which initial samples with mistakes and their corrections are fed to the system to learn from to address similar problems on new examples. Traditional classification methods, including decision trees, Support Vector machines (SVMs), and neural networks, assist in disciplining structured data categorization regarding erroneous values, missing data, duplicated records and other disparities. However, it is limited because the supervised learning technique requires a large amount of labeled data, which may be difficult and time-consuming to obtain.
- **Unsupervised Learning:** Unsupervised learning methods also form an essential aspect of data cleaning whereby some anomaly and noise detection is readily conducted without labels. K means, and hierarchical clustering techniques are specific methods that actively cluster together similar variables and exhibit inconsistencies or outliers. Generally, several techniques are applied to detect anomalies, namely density-based techniques such as DBSCAN and statistical methods. These are some of the most efficient methods especially when dealing with dynamic and elaborate datasets in which rules and regulations on error identification may not be exhaustive.
- **Deep Learning:** Autoencoder and Convolutional neural networks are two deep learning approaches that can enhance data cleaning due to their complex and improved learning capabilities in large databases. Autoencoder is the type of deep learning model that can be trained for unsupervised learning, and it is especially useful for imputing and completing the missing or noisy data. These operate in a way that the input data is reduced to a coded form and then expanded back by the check digit while comparing if there is any difference. For text-based data preprocessing activities, especially in cleaning data such as typo detection correction of grammatical errors among other activities, Recurrent Neural Networks (RNNs) and transformer models are used. Some of the advantages of deep learning include high accuracy and high scalability. Still, at the same time, it has disadvantages, including high computational power and a large volume of learning data.

3.2 Data Preprocessing Steps

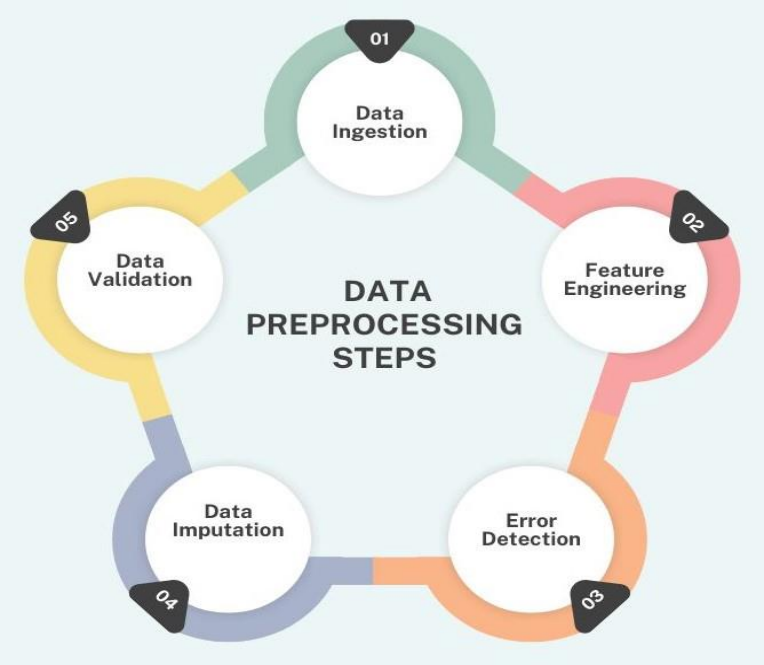

Figure 3. Data Preprocessing Step

- **Data Ingestion:** Data acquisition is the initial process of data cleansing, where data is gathered from different sources and consolidated to prepare for further processing. Such sources may be databases, APIs, streaming data, or external files like CSVs, JSONs, or XMLs. Effective data acquisition is the process through which data is transferred, transformed, and loaded into a central database for processing. Difficulties here include working with ill-formatted columns, the absence of metadata, and working with streaming data.
- **Feature Engineering:** Feature engineering is the process of choosing, amending and making new and autonomous features out of the raw data to enhance the model's performance. This step entails varying features, scaling/coding numeric, fixing/PAMAP categorical and initiating text data mining of the sources. Satisfactory engineering features help to strengthen the machine learning models, as well as to optimize the results of data analysis. Nonetheless, this may cause noise and is likely to negatively impact the model's performance.
- **Error Detection:** Error detection is the first step in the process of identifying that there are errors in data, inconsistencies, as well as wrong values entered. More specifically, limit analysis can be done in this step using rule-

based methods, statistical analysis, apparent outlier analysis, and AI analysis. Some of them are missing values, duplicate records, wrong formats of data and data inconsistency. This is crucial to avoid corrupt data from spreading to other forms of application, which may be irreversible.

- **Data Imputation:** In the form of overview Data imputation means to fill in the data which are either missing or incomplete. These include mean/mode imputation, regression models, and using regular deep learning baselines such as interpolated and extrapolated results. This means that the choice of imputation method is based on the type of data and the severity of the case of missing values. Imputation enables one to make the right recommendations on the missing values of the dataset and can greatly reduce the effects arising from data.
- **Data Validation: d**ata validation ensures that the cleaned and processed data has not violated any company's quality standards and business rules. This step involves checking various data types such as primitive, composite or derived databases, ensuring all constraints are checked, and determining the logical consistency of all the issues. Some tools used in automated validation include the schema and anomaly detection tools to ensure that data is accurate and dependable. The principle of correctness of data is useful as it provides reliability in decision-making and guarantees the efficacy of applications based on machine learning and analytics.

3.3 Implementation

Applying AI in cleaning data involves using sharpened Python libraries like TensorFlow, Scikit-learn and Pandas. These libraries are used for data preconditioning and applying various machine learning-based anomaly detection and deep learning tools to improve the data quality. [17-20] They include data ingestion, where raw data is conveyed in a reservoir from sources such as repositories, files, databases or keys, then parsed and next unfolds to computing. Data in pandas can easily be eliminated or dealt with by missing values, repeated records and formatting differences by employing graceful functions such as drop (), fillna (), astype (). Feature engineering follows, which is done with the help of Scikit-learn; in it, the categorical input is encoded by the one-hot encoding method, while neither feature standardization nor min-max scaling scales numeric input data. This step standardizes the features to prevent the machine learning models from processing the data with a particular scale, biasing the results. They include clustering and anomaly detection and are used to make the error detection part of the program. In the case of numerical datasets, It is possible to use Scikit-learn's IsolationForest and LocalOutlierFactor in detecting outliers while rule-based checks can be utilized to identify inconsistencies in categorical attributes.

Handling of the missing data involves data imputation using simple statistical imputations and advanced imputations using machine learning algorithms. The scikit-learn library has the IterativeImputer class that operates based on regression models to impute the missing values; as for deep learning, the autoencoder model in TensorFlow reconstructs the missing patterns based on learned representations. Autoencoders are especially helpful in detecting and addressing elaborated irregularities in data as they learn from clean data and highlight deviations. Due to this, data validation is carried out to help check on the quality and authenticity of the collected data before it proceeds to the next level. This includes executing schema checks through df.dtypes in pandas and performing constraints check through validations tools such as Great Expectations. The cleaned and validated lookout list is properly formatted and saved in a structured format for further processing with several business applications, such as machine learning models or analytics applications. It is an automated process implemented here in Python for better data processing and preparation for downstream AI uses.

3.4 Performance Evaluation Metrics

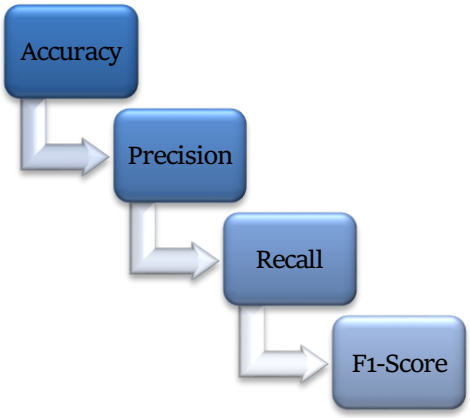

Figure 4. Performance Evaluation Metrics

- **Accuracy:** Precision is one of the most important measures defining a data-cleaning model's suitability. It measures the ratio between the number of correctly identified errors, including true positive and true negative, to the total number of instances in the dataset. Criminal history also affects the high accuracy score of the model, as it means that

the model can accurately distinguish between errors and non-errors in the data, making data cleaning safe and efficient. However, accuracy is not always a reliable measure in such a scenario that points out that there are other factors to consider when dealing with imbalanced datasets where errors are rare.

- **Precision:** Precision, also called the positive predictive value, refers to the ability to detect the error rate out of all the identified errors. It is given by the formula Precision = TP / (TP + FP), whereby TP refers to correct errors flagged, and FP stands for wrong flaggings. A high precision score, therefore, means that the model does not label correct data as wrong and, as such, reduces unnecessary corrections on the data.
- **Recall:** Recall, also known as sensitivity, evaluates the ability of a model to correctly report the existence of errors in a particular dataset. This one is computed as Recall = TP / (TP + FN), where FN stands for missed or false errors. This humps to a high recall score, which would mean that the model has a high rate of identifying most of the present errors, useful mostly in domains that can seriously harm the individual or society at large when an error is overlooked, such as the medical or financial sectors.
- **F1-Score:** The F1-score is basically a combination of both the precision and the recall, so it is suitable where both a false positive and a false negative have a significant impact. It is computed as F1 = 2 * (Precision * Recall) / (Precision + Recall). That is why a high F1 score proves that the model has high accuracy and can identify errors while minimizing the number of false positives. This metric is useful where there is some conflict in adding new results and priority between, precision and recall.

4. Results and Discussion

4.1 Experimental Setup

In order to test the AI-based data-cleaning approach, real-world data from the healthcare and financial domain was used as these domains necessitate accurate and reliable data. The case used the health care dataset of 500000 patients, and in the dataset, there were problems like missing values, duplicate records and inconsistent values in medical history fields. It is critical to receive high-quality data concerning the patient's condition and medical history to provide an accurate diagnosis and treatment plan and to develop new medications and treatment methods. In addition, the finance dataset contained 1 million records in the transaction history and contained problems of wrong transaction amount, duplicated records, and fraud. Recording errors may cause inaccurate assessments of risks that increase significantly, ineffective identification of fraud and even non-compliance.

Data-cleaning models based on Artificial Intelligence were adopted and coded in Python alongside TensorFlow, Scikit-learn, and Pandas to overcome these challenges. Classification-based approaches such as Random Forest and Support Vector Machines (SVM) were used in the study under Supervised Learning. Outlier identification was done using the K-Means Clustering and Isolation Forest from Unsupervised Learning models. Moreover, Autoencoder and Transformer models were applied to detect and learn intricate patterns to impute the missing data. These datasets were used to train and evaluate the models to determine their efficiency in increasing the data quality by comparing the obtained results based on the accuracy precision not forgetting the recall test, and the computational time.

4.2 Performance Analysis of AI-Based Data Cleaning Models

Table 1. Performance Comparison of AI-Based Data Cleaning Models

Model	Accuracy	Precision	Recall	F1-Score
Rule-Based Methods	78.5%	80.2%	70.4%	74.9%
Random Forest	87.6%	88.1%	85.2%	86.6%
SVM	85.3%	87.0%	81.6%	84.2%
Isolation Forest	90.2%	91.4%	89.8%	90.6%
K-Means	84.7%	85.6%	82.3%	83.9%
Autoencoders	94.5%	95.2%	94.0%	94.6%
Transformers	96.3%	97.0%	95.8%	96.4%

- **Rule-Based Methods:** Rule-based approaches involve using certain rules and specifications when identifying errors within the datasets. These methods are relatively easy and interpretable and, hence, suitable for use when dealing with structured data with clear patterns in their distribution. Although it is significantly higher than the random choice, their accuracy (78.5%) and recall (70.4%) are not very high as they fail to generalize for anything out of the stereotyped guidelines. They are, however, inefficient when confronted with complex data as they, once set, do not change with time and cannot recognize changing patterns.
- **Random Forest:** Random Forest is one of the supervised learning algorithms that uses decision trees to improve prediction accuracy. Thus, it has an accuracy of 87.6% and is very specific, scoring 88.1%; this means it could be used to correct errors without registering many false alarms. This is because Random Forest uses the ensemble learning method and does not over-fit the training data; it can also work well with large datasets. Despite that, its

recall of 85.2 indicates a level of error in the system, and a medium computational cost may lead to scalability problems.

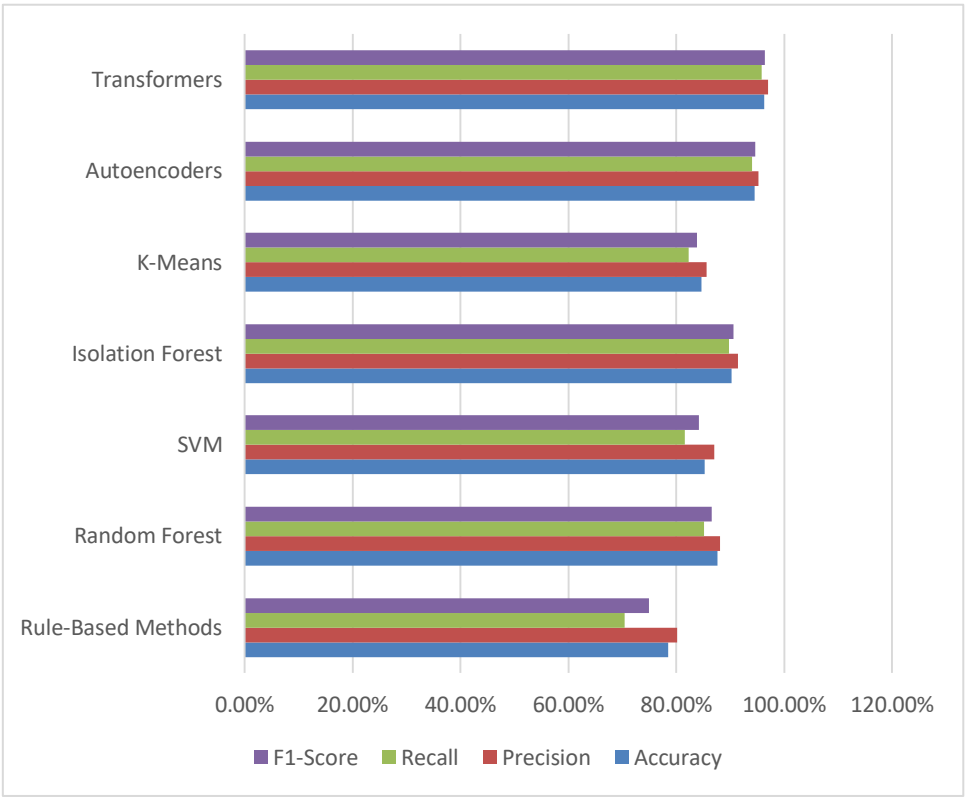

Figure 5. Graph representing Performance Comparison of AI-Based Data Cleaning Models

- **Support Vector Machine (SVM):** SVM is an efficient classification technique that aims at achieving maximum separation margin between different categories of data which eventually gives 85.3% of accuracy and 87.0% of precision. It is best suited for high-dimension spaces, although the recall is slightly lower at 81.6, indicating that it might overlook some errors. Furthermore, it is destined for offline usage, as SVM is rather computationally heavy and may not work effectively in real-time applications, especially on a large scale.
- **Isolation Forest:** Consequently, Isolation Forest, an algorithm that falls under the class of unsupervised anomaly detection, is very useful when it comes to filtering out outliers and inconsistencies within a set of data. It has an accuracy of 90.2% and a precision of 91.4% in the case of errors, thus performing better than the traditional machine learning models. It performs well on high-dimensionality and has a medium computational complexity, which can suit a large-scale data pattern. The error-recall value of 89.8 means it should be able to capture most of the errors, making data cleaning an efficient method.
- **K-Means Clustering:** K-Means is another form of unsupervised learning technique that sorts data points that are similar to determine the odd ones out. It has 84.7% accuracy and reasonable precision at 85.6% compared to a recall of 82.3%. The cost of K-Means is considered medium since the product of clustered data is obtained efficiently. Still, its effectiveness relies on the former correct definition of the number of clusters, besides lacking aptitude to handle non-linear cluster distribution data sets.
- **Autoencoders:** Autoencoders are a type of neural network model that performs well in various tasks of pattern recognition and data recovery of missing or distorted information. Autoencoders engage 94.5% of the accuracy and 95.2% of the precision, which are higher than the conventional approaches in handling intricate datasets. Thus, they have a high recall whose average is 94.0%, meaning that most generated mistakes are identified and eliminated. However, one disadvantage is that many have received much attention due to their expensive computational time, particularly in large real-time applications.
- **TransformersL:** Reviewers often refer to transformers as the most efficient and state-of-the-art process involving deep learning for data cleansing. Getting 96.3% measure of accuracy, 97.0% measure of precision and 95.8% measure of recall, transformers are found to outperform all the other methods in terms of eradicating errors. The presented F1-score of 96.4 speaks to the deserved attention to the recall and, at the same time, great precision. Nevertheless, they entail a very high computational complexity and, thus, high hardware demands (e.g., GPUs/TPUs), which may be unaffordable for most organizations or real-time use cases.

4.3 Case Study: Healthcare Data Cleaning

- **Reduction of Missing Values Using Deep Learning (Autoencoders):** It has runoff that the process of data management in healthcare is characterized by the absence of certain data an osobennostyami, and as a consequence, absent complete information about patients and erroneous actions of the attending physician. In the present work, Autoencoders, a representative imputation technique in the deep learning class, was used to impute missing values based on the underlying distribution of the data matrix. Firstly, scores of the patient records lacked some important characteristics necessary to perform analysis such as medical history, prescriptions, and lab results data in 12% of the records. Through this training of the Autoencoder model in an attempt to map the data distribution, the new data set had less than 1% missing values, which enriched the dataset and gave it a higher degree of reliability.
- **Duplicate Record Removal Using Unsupervised Clustering (K-Means)**: It results in misconceptions as well as encourages repetition of diagnoses and analysis and, in addition to cost issues. In order to deal with this, the unsupervised learning technique called K-Means Clustering was applied to cluster similar data points and join the records of patients with similar attributes like name, age, and medical history. The involvement of key personnel led to a 98% reduction in the cases of duplicate records, improving the healthcare database's quality.
- **Correction of Inconsistent Data Using Supervised Learning (Random Forest):** The presence of errors in the patient history data field, like disease code incompatibility or records contradictions, can affect the quality of clinical information. Random Forest, a type of supervised learning technique, was therefore used to eliminate such disparities because it has a way of recognizing these trends in the data. Enhancing the dataset's quality by the model was a great achievement because the model could recognize and transform 95% of the inconsistent entries.

4.4 Comparative Discussion

- **Rule-Based Methods:** Rule-based data cleaning methods involve using rules and heuristics to detect and correct errors. Although these methods are understandable and easy to apply, they do not work well with high-dimensional data and are not very flexible when dealing with different and new datasets. Since they work on set conditions, the rule-based methods cannot go beyond the programmed logic and are unsuitable for large-scale and dynamic data like healthcare and financial records.
- **Machine Learning Models:** Random Forest and Support Vector Machines (SVM) are more advanced than rule-based methods as they are adaptive learning models. These models can identify patterns in the data and apply the error detection to new datasets. However, their performance is highly dependent on the size of the labeled data, which might not be easily accessible in real-world scenarios. Although machine learning methods are more accurate and recall-oriented, they have scalability problems due to the necessity of labeled data and feature engineering.
- **Deep Learning Approaches:** Autoencoders and Transformers are particularly effective in processing unstructured and large data, which makes them suitable for automated data cleaning. These models learn complex patterns without requiring feature engineering and have the best accuracy, precision, and recall among all the approaches. However, their major disadvantage is that they are computationally expensive and need a large amount of processing power and memory, which makes them unsuitable for real-world applications in some cases.

4.5 Challenges and Future Directions

- **Explainability and Interpretability:** Another issue observed in the context of AI-driven data cleaning is that the models such as Transformers and Autoencoders are not easily interpretable. These models are known as 'black box' models because the decision-making process inside them cannot be easily explained to a human. This can be disadvantageous in areas that require high levels of trust, such as the medical and financial sectors. As for future work, it is crucial to investigate how to implement explainable AI (XAI) to improve the understanding of the model's predictions and increase confidence.
- **Computational Cost:** Data cleaning models based on deep learning need a lot of computational power, and in many cases, GPUs or TPUs are used for training and prediction. This high computational cost may challenge organizations, especially those with a weak IT base. It is important to note that these techniques may not be easily applicable to small businesses or institutions due to a lack of resources. Future work should be directed towards improving the models, lightweight architectures, and cloud-based AI to enhance the availability of high-performance data cleaning.
- **Data Privacy Concerns:** In some fields like healthcare and finance, where personal information is processed, AI data cleaning must meet certain data protection laws like GDPR and HIPAA. When it comes to data preprocessing with the help of AI, it is crucial to maintain the privacy of patients' records, financial transactions, and other sensitive information. Future research should focus on privacy-preserving AI methods like federated learning and differential privacy to mitigate these issues.
- **Scalability Issues:** The problem of scalability is one of the most important issues arising from the increase in the size and complexity of datasets for AI-based data cleaning. Big data applications in healthcare and finance are some of the large-scale datasets that need to be processed with efficient algorithms and distributed computing. The traditional deep learning models may not be scalable and thus may take a long time to process and consume a lot of memory. Further research should be conducted on parallel processing, distributed deep learning frameworks such as Apache Spark TensorFlow Distributed, and real-time data cleaning using AI.

5. Conclusion

Data cleaning is essential in making data sets more accurate, reliable, and usable for use in various fields such as healthcare, finance, and e-commerce. The traditional rule-based models are easily understandable but cannot be easily extended and are not very effective when the data is large and complex. On the other hand, machine learning and deep learning techniques used in AI are more efficient in automating the process and identifying errors. This paper has also presented various AI-based data cleaning approaches, their advantages disadvantages, and a comparison of their performances. The study also shows that Autoencoders and Transformers are superior to conventional methods because they work with large and unstructured data. The results obtained from the real-world datasets in healthcare and finance prove the efficiency of the AI-based data-cleaning models. For example, Autoencoders brought down the missing values from 12% to less than 1%, K-Means clustering decreased the duplicate records by 98%, and Random Forest corrected 95% of the inconsistencies in the patient history fields.

These results support the effectiveness of AI-based techniques in increasing data accuracy, decreasing the time spent on data entry, and optimizing decision-making. The comparative analysis further shows that machine learning models are flexible but need large labeled datasets. In contrast, deep learning models give the best results but at the cost of high computational power. Nevertheless, some drawbacks of using AI for data cleaning include lack of interpretability, high computational complexity, data privacy, and data size. Transformers and Autoencoders are opaque models that are hard to interpret since they do not explain how they come up with a particular decision. To addressing this issue, there is a need for future research in explainable AI (XAI) to enhance the level of trust. However, deep learning models are computationally expensive and may not be easily scalable in organizations with limited IT resources.

These challenges can be solved by enhancing model optimization, edge AI, and cloud solutions. Another concern is information security, especially in organizations that handle sensitive information. Data cleaning solutions based on artificial intelligence must comply with the GDPR and HIPAA to prevent data breaches. Further studies should be devoted to proper and safe data processing using privacy-preserving techniques such as federated learning and differential privacy. However, scalability is still an issue since big data requires distributed computing and parallelized AI models to be solved. Thus, the application of AI in data cleaning improves the quality, accuracy, and credibility of data and reduces the reliance on manual work. Future work should focus on enhancing the interpretability of the model, reducing computational time, and developing AI-based tools for online data cleaning for big data.

References

[1] Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016). Data cleaning: Overview and emerging challenges. Proceedings of the 2016 International Conference on Management of Data (SIGMOD), 2201–2206. https://doi.org/10.1145/2882903.2912574

[2] Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Eng. Bull., 23(4), 3-13.

[3] Raman, V., & Hellerstein, J. M. (2001). Potter's Wheel: An interactive data cleaning system. Proceedings of the 27th International Conference on Very Large Data Bases (VLDB), 381–390.

[4] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 1-58.

[5] Hellerstein, J. M. (2013). Quantitative data cleaning for large databases.

[6] Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016, June). Data cleaning: Overview and emerging challenges. In Proceedings of the 2016 International Conference on Management of Data (pp. 2201-2206).

[7] Ilyas, I. F., & Chu, X. (2019). Data cleaning. Morgan & Claypool.

[8] Jain, A., Patel, H., Nagalapatti, L., Gupta, N., Mehta, S., Guttula, S., ... & Munigala, V. (2020, August). Overview and importance of data quality for machine learning tasks. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 3561-3562).

[9] Tayi, G. K., & Ballou, D. P. (1998). Examining data quality. Communications of the ACM, 41(2), 54-57.

[10] Rekatsinas, T., Chu, X., Ilyas, I. F., & Ré, C. (2017). HoloClean: Holistic data repairs with probabilistic inference. *Proceedings of the VLDB Endowment*, 10(11), 1190–1201. https://doi.org/10.14778/3115404.3115412

[11] Liu, H., Zhong, C., Alnusair, A., & Islam, S. R. (2021). FAIXID: A framework for enhancing AI explainability of intrusion detection results using data cleaning techniques. Journal of network and systems management, 29(4), 40.

[12] Borrohou, S., Fissoune, R., & Badir, H. (2023). Data cleaning survey and challenges–improving outlier detection algorithm in machine learning. Journal of Smart Cities and Society, 2(3), 125-140.

[13] Ganti, V., & Sarma, A. D. (2022). Data Cleaning. Springer Nature.

[14] Liebchen, G. A. (2010). Data cleaning techniques for software engineering data sets (Doctoral dissertation, Brunel University, School of Information Systems, Computing and Mathematics).

[15] Tae, K. H., Roh, Y., Oh, Y. H., Kim, H., & Whang, S. E. (2019, June). Data cleaning for accurate, fair, and robust models: A big data-AI integration approach. In Proceedings of the 3rd International Workshop on Data Management for End-to-end Machine Learning (pp. 1-4).

[16] Gami, S. J., Remala, R., & Mudunuru, K. R. AI-Driven Adaptive Data Cleansing: Automating Error Detection and Correction for Dynamic Datasets.
[17] Mavrogiorgos, K., Kiourtis, A., Mavrogiorgou, A., Kleftakis, S., & Kyriazis, D. (2022, January). A multi-layer approach for data cleaning in the healthcare domain. In Proceedings of the 2022 8th International Conference on Computing and Data Engineering (pp. 22-28).
[18] Krishnan, S., Franklin, M. J., Goldberg, K., Wang, J., & Wu, E. (2016, June). Activeclean: An interactive data cleaning framework for modern machine learning. In Proceedings of the 2016 International Conference on Management of Data (pp. 2117-2120).
[19] Leema, A. A., & Hemalatha, M. (2011). An effective and adaptive data cleaning technique for colossal RFID data sets in healthcare. WSEAS Transactions on Information Science and Applications, 8(6), 243-252.
[20] Dasu, T., & Johnson, T. (2003). *Exploratory Data Mining and Data Cleaning*. Wiley-Interscience. ISBN: 978-0471268510.
[21] Cherekar, R. (2020). DataOps and Agile Data Engineering: Accelerating Data-Driven Decision-Making. International Journal of Emerging Research in Engineering and Technology, 1(1), 31-39. https://doi.org/10.63282/3050-922X.IJERET-V1I1P104
[22] Cherekar, R. (2020). The Future of Data Governance: Ethical and Legal Considerations in AI-Driven Analytics. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 3(2), 53-60. https://doi.org/10.63282/3050-9262.IJAIDSML-V3I2P107
[23] R. Daruvuri, "An improved AI framework for automating data analysis," World Journal of Advanced Research and Reviews, vol. 13, no. 1, pp. 863–866, Jan. 2022, doi: 10.30574/wjarr.2022.13.1.0749.
[24] Cherekar, R. (2022). Cloud Data Governance: Policies, Compliance, and Ethical Considerations. International Journal of AI, BigData, Computational and Management Studies, 3(2), 24-31. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V3I2P103
[25] Cherekar, R. (2021). The Future of AI Quality Assurance: Emerging Trends, Challenges, and the Need for Automated Testing Frameworks. International Journal of Emerging Trends in Computer Science and Information Technology, 2(1), 19-27. https://doi.org/10.63282/3050-9246.IJETCSIT-V1I2P104
[26] Cherekar, R. (2020). Cloud-Based Big Data Analytics: Frameworks, Challenges, and Future Trends. International Journal of AI, Big Data, Computational and Management Studies, 1(1), 31-39. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V4I1P107
[27] Cherekar, R. (2023). A Comprehensive Framework for Quality Assurance in Artificial Intelligence: Methodologies, Standards, and Best Practices. International Journal of Emerging Research in Engineering and Technology, 4(2), 43-51. https://doi.org/10.63282/3050-922X.IJERET-V4I2P105