DATA DETOX: A SMART AND VISUAL SYSTEM FOR ANOMALY DETECTION AND INTERACTIVE CLEANING

N. Swathi [1]
Assistant Professor
Department of AI&DS
Annamacharya University
Rajampet, Andhra Pradesh
nagarajuswathi256@gmail.com

Fahimuddin Shaik [2]
Associate Professor
Department of ECE
Annamacharya University
Rajampet, Andhra Pradesh
fahimaits@gmail.com

H. Keerthi [3]
Department of AI&DS
Annamacharya Institute of
Technology and Sciences
Rajampet, Andhra Pradesh
Keerthih104@gmail.com

K. Dinesh Reddy [4]
Department of AI&DS
Annamacharya Institute of
Technology and Sciences
Rajampet, Andhra Pradesh
dineshreddy2624@gmail.com

A. Mahammad Sohail [5]
Department of AI&DS
Annamacharya Institute of
Technology and Sciences
Rajampet, Andhra Pradesh
adonimahammadsohail@gmail.com

A. Khadeerulla [6]
Department of AI&DS
Annamacharya Institute of
Technology and Sciences
Rajampet, Andhra Pradesh
atarkhadeer1119@gmail.com

***Abstract*- In data-driven innovation, data quality is very important in order to achieve proper analytics and dependable machine learning results. Nevertheless, in the real world, data are prone to missing values, outliers, redundancy and inconsistency which deteriorate model performance. To overcome these difficulties, this paper suggests Data Detox, a smart and graphic system of automated data cleaning and anomaly detection. The system combines hybrid outlier detection techniques with Isolation Forest and Interquartile Range (IQR) with automated Exploratory Data Analysis (EDA) techniques to profile and visualize data. Based on Python and Streamlit, Data Detox offers an interactive platform allowing users to convert, clean, and visualize datasets with low interaction. Experimental findings show that the framework can cut down on the time taken to prepare data by about 45 percent with data consistency and reliability also enhanced. The system is suggested to augment machine learning use preprocessing pipeline and provides an efficient, scalable, and easy-to-use way to address automated data quality management.**

Keywords-
Anomaly Detection, Data Cleaning Automatically, Isolation Forest, Outlier Detection, Exploratory Data Analysis (EDA), Data Profiling, Data Pipeline Optimization and Consistent Machine Learning.

I. INTRODUCTION

During the era of digitalization, information has turned into a vital asset that creates innovation, automation, and smart decision-making in various industries, including healthcare, finance, manufacturing, and e-commerce [1], [9]. Nevertheless, real-world data tend to be noisy, inconsistent, incomplete, and duplicated, which considerably downgrades its trustworthiness and the output of downstream machine learning (ML) models [2], [10]. It is the famous rule that to create analytical and predictive results, the quality of input data should be no better than the quality of the results [1].

The manual and rule-based data cleaning techniques used in traditional data cleaning techniques are ill-equipped to cope with the complexity, size, and heterogeneity of current datasets [3], [11]. Furthermore, current data preprocessing solutions usually focus on a single task, e.g., missing value filling, anomaly identification, or profiling, but not on an integrated and automatic solution [4], [5], [6].

In order to address these limitations, this paper presents Data Detox, a visual and smart data cleaning and anomaly detection framework that is automated. The system is a combination of statistical (IQR) and machine learning–based (Isolation Forest) methods to identify outliers, and Exploratory Data Analysis (EDA) for automated data profiling and visualization [7], [10], [11].

Increasingly, data analytics are being employed by business to make business decisions, automate business processes, and come up with important business solutions. The same is being used in other fields such as healthcare, finance, manufacturing, e-commerce, etc. The data that we are producing through sensors, social media, transactions and IoT devices is and will remain complicated and diverse. The depth and quality of the data determine the value of the knowledge that one gets. Raw data is often full, duplicated, not well structured and designed, inconsistent, misspelled, and contain erroneous outliers. The problems that undermine the quality of stored and processed data will cause the later analysis to make overly reductionist conclusions. To the point, it is said "Garbage In, Garbage Out" [1]. In simple terms, one gets bad output out of bad input.

The outdated cleaning and ruling method which is virtually a manual check and a pass rule has no chance with the Big data that is highly complicated and prevents both size and variety. They tend to process information poorly and cannot process the quick processing of timely information. Anomaly detection is any data that is unusual or suspicious and which is commonly ignored or just brushed off. The routine statistical methods are useful but, they are often blind in terms of perceiving a subtlety or multifaceted problem. Isolation Forest is a more advanced method and is implemented in the identification of anomalies [7]. It is based on the idea of screening the odd ones out of data frame by means of recursive data splitting.

Data Detox is an integrated system that delivers an automated method of cleaning data and identifying anomalies. It has an automated built-in solution as opposed to chopping it in a more traditional way. It provides distinctive anomaly detection solutions that combines basic statistical techniques and sophisticated machine learning techniques [2], [3]. Despite all these, Data Detox offers automated summaries of the data with reported characteristics to enable one to know about the issues at hand. The Data Detox is easy to use, it is designed with the assistance of a web interface, and it is designed to be used by specialists and laymen alike. All users can upload data and attribute to various cleaning methods and interact with the data in a variety of ways [4], [5].

The Data Detox is valuable because it is automated, it developed sophisticated anomaly detection, it provides transparent reporting, and it visualizes clearly. As a result, decisions of a wide range of industries and businesses are able to be made with increased confidence analysis and machine learning on this automated cleaned data [6], [9], [10]. In recent years, AI-assisted data cleaning frameworks have gained significant attention for automating preprocessing and improving data quality [12], [13].

The novelty of the given work as a whole is that Data Detox is a single and automated platform, which merges data cleaning, anomaly detection, and profiling into one intelligent platform. In contrast to the old ones like OpenRefine and ydata_profiling which process these tasks independently [5], [11], Data Detox suggests a hybrid anomaly detection algorithm, which is a combination of the statistical Interquartile Range (IQR) and the Isolation Forest algorithm to efficiently detect simple as well as complex outliers [2], [7]. The structure also integrates automated Exploratory Data Analysis (EDA) and interactive visualisation constructed in Streamlit so that the user can see data quality changes live [4], [5]. Such a combined and easy-to-use method greatly saves the manual preprocessing effort and enhances the accuracy and consistency of datasets that are ready to be trained by machine learning. Altogether, Data Detox is a progressive step in the evolution of the current data quality systems, as it presents a complex, scalable, and affordable platform of intelligent data management.

II.LITERATURE SURVEY

Over the years, the quality of the data employed in the analysis and machine learning meant clearing and verifying data on abnormalities. The initial contributions in this field were to regard algorithmic error correction and missing data imputation with the help of statistics. The above image mentions Rahm and Do as the first to systematize such strategies, especially the problems of duplicates, format compatibility and correcting of errors, which gave rise to a cottage industry in this field [1]. Although traditional methods performed reasonably well with small simple datasets, they struggled more and more to deal with the complexity and the variety of data in large, rapidly changing data ecosystems.

When data grew more complex, scientists wanted more effective ways of identifying anomalous or potentially troublesome data points. The Local Outlier Factor (LOF) algorithm was developed by Breunig and his team and it identifies anomalies by the relative density of points in the local area—a factor that earlier algorithms such as z-scores (which considered data on a global scale) did not take into account [3]. On this basis, Chandola and others carried out a comprehensive survey of the methods of detecting anomalies, indicating that traditional methods did not sufficiently deal with data that was multi-dimensional or in a state of dynamic evolution [2]. In order to address this problem, Liu et al. proposed the Isolation Forest algorithm that repeatedly divides the datasets in an attempt to isolate anomalies [7]. This is because of the capabilities to work with large, complex datasets, which makes Isolation Forest one of the most used tools in anomaly detection as of today.

There is an increasing interest in those tools, which allow the user to get familiar with the basics of their data quickly. One example is the ydata_profiling library (as seen in image generated above), which can generate automated, exhaustive reports that summarize data distributions, correlations, presence of missing values, and other characteristics and provide analysts with higher-level estimates of data health quickly. The report presents the data in a very accurate manner; however, the main weakness is the failure to detect and rectify data anomalies and hence provide a less holistic view of the data. At the far other end of the spectrum are the more interactive data cleaning and transformation products like OpenRefine or Tableau Prep (see image created above) [5], but lose greatly in terms of manually entered bulk of data that will not be suited to workflow requiring automation or large amounts of data.

Much attention has been given to the identification of data anomaly classification. One of the earlier classifications of outliers has been by Aggarwal, where he identifies three types of outliers: collective outliers, or strange observations that make a pattern in a coalition; contextual outliers, or strange observations that only occur within a context; and global outliers, or strange observations across the whole set of data [6]. He also stated that the integration of conventional stats and machine learning would perform better in a huge majority of cases. It may be necessary to offer a reasonable balance of speed and accuracy, and, to this end, new hybrid approaches are being considered in future research that aspire to integrate machine learning and optimization [9], [10]. This is of significant importance in the analysis of large data with the capacity to do this in a timely and trustworthy fashion.

The past studies have done a commendable work in the data quality improvement area, the data anomaly detection area, and the data profiling area. Algorithms like Local Outlier Factor (LOF) and Isolation Forest that presented more sophisticated outlier detection techniques in complicated data sets were introduced in the early works like that of Rahm and Do who primarily emphasized data error correction and duplicate detection by use of the statistical methodology [1], [7]. Nevertheless, such traditional models were not scalable to heterogeneous or high-dimensional data that is characteristic of the modern setting. OpenRefine and ydata profiling interactive tools [5], [11], were convenient and provided visualization, but cleaning, profiling, and anomaly detection remained distinct and manual operations. The current research has largely emphasised the necessity of integrated and automated structures that can be used to perform all these functions with minimum human-intervention [9], [10]. Nonetheless, even with these developments, the current systems do not have real-time visualization, automatic flow, and non-programmers, which remains a limitation to its practicality when applied in real-world data science processes. Thus, there is a research gap related to the critical importance of bringing anomaly detection, data cleaning, and profiling together as a single automated and interactive system. In order to close this gap, the current project suggests Data Detox, an extensive model that combines statistical and machine learning methods to perform efficient, intelligent, and visual data cleaning. Recent research also emphasizes the integration of hybrid machine learning models for anomaly detection and data profiling [12], [13], which further supports the motivation of this study.

III. METHODOLOGY

To give the user an easy-to-use system, Data Detox integrates all the essential steps in a single system to clean and also to analyze data. It starts with the customers inputting data in any of the popular formats: CSV, Excel, or JSON. The system normalizes the header of the columns in the first step in facilitating the later phases of cleaning by removing extraneous spaces, lowercasing of letters, and substituting of space with underscore. This may appear as a minute detail, but this eliminates major data processing problems at the very outset. Users leave the Data Detox to clean data, and this involves automatic removal of duplicated records in order to offer data accuracy and reliability [1], [11]. The user is presented with options of filling in missing entries and that is a given in almost any data set; addition of average, of middle value, or most common value. The common *n/a*, *null* and dash signs used to denote blank cells or empty placeholders are recognized by the system and are managed to count as missing entries. Letters are all changed to lower case and superfluous space is eliminated in textual information. In addition, categorical information having few choices is cleansed [10].

The system adheres to an intelligent two-pronged methodology of detecting anomalies. The former is through application of the conventional statistical method of Interquartile Range (IQR), to establish the presence of aberrant outliers [2]. The extreme points may be completely eliminated or restricted to a user-defined range. It uses a powerful machine learning model, Isolation Forest, to detect small or subtle abnormalities in higher-dimensional data [7]. It effectively isolates abnormalities in datasets that have repeated partitions, and identification of abnormal values that could not be easily identified with the basic statistics becomes a smooth process. The tool applies the fundamentals of statistics to detect the apparent anomalies, and the finer tools are used to detect the subtle anomalies [3], [9].

The overall workflow of the Data Detox framework is illustrated in Figure 1, showing its modular structure.

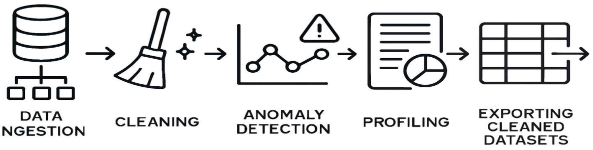


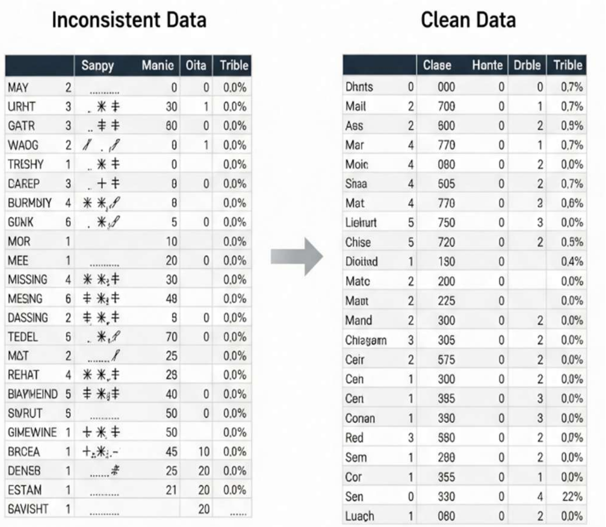
Figure 1: System Architecture Diagram

The architecture highlights key modules including data input, cleaning, anomaly detection, profiling, and visualization.

The final step is aimed at equipping individuals with profiling and visualization tools of their data to discover meaningful information. The ydata_profiling tool performs an analysis of the data and produces independent, automated, and whole profiling reports that describe the core data distribution, the magnitude of missing data, and association of different attributes [11]. Data Detox is based on Matplotlib and Plotly to be interactive [4]. The interactive histograms enable the user to analyze the data before and after the cleaning, the amendment summary tables, and the boxplots, to scrutinize their outliers and anomalies. These visualizations indicate the outcomes of data cleaning and data anomaly detection and ensure data quality and reliability [5], [10].

Data Detox extracts, cleans, and finds outliers on a systemic level and exports raw, unwrapped data to give consistent and analyzable datasets prepared to be ingested into machine learning [1], [9]. Data practitioner that can be very helpful is Data Detox, a time-saving one [11].

Figure 2 presents a comparison of the dataset before and after preprocessing.



Inconsistent Data

		Sanpy	Manic	Oita	Trible
MAY	2	....	0	0	0.0%
URHT	3	. ✱ ‡	30	1	0.0%
GATR	3	.. ‡ ‡	60	0	0.0%
WAOG	2	/ . /	0	1	0.0%
TRŁSHY	1	.. ✱ ‡	0		0.0%
CAREP	3	. + ‡	8	0	0.0%
BURMDIY	4	✱ ✱,/	8		0.0%
GDNK	6	. ✱,/	5	0	0.0%
MOR	1		10		0.0%
MEE	1	....	20	0	0.0%
MISSING	4	✱ ✱: ‡	30		0.0%
MESNG	6	‡ ✱: ‡	48		0.0%
DASSING	2	‡ ✱, ‡	8	0	0.0%
TEDEL	5	. ✱,/	70	0	0.0%
MΔT	2	......./	25		0.0%
REHAT	4	✱ ✱, ‡	28		0.0%
BIAYMEIND	5	‡ ✱: ‡	40	0	0.0%
SIVRUT	5	....	50	0	0.0%
GIMEWINE	1	+ ✱ ‡	50		0.0%
BRCEA	1	+: ✱:.-	45	10	0.0%
DENSB	1	.......‡	25	20	0.0%
ESTAM	1	....	21	20	0.0%
6AVISHT	1	....		20	......

Clean Data

		Clase	Honte	Drble	Trible
Dhnts	0	000	0	0	0.7%
Mail	2	700	0	1	0.7%
Ass	2	600	0	2	0.9%
Mar	4	770	0	1	0.7%
Moic	4	080	0	2	0.0%
Shaa	4	505	0	2	0.7%
Mat	4	770	0	3	0.6%
Liehurt	5	750	0	3	0.0%
Chise	5	720	0	2	0.5%
Dioitud	1	190	0		0.4%
Matc	2	200	0		0.0%
Mant	2	225	0		0.0%
Mand	2	300	0	2	0.0%
Chiagam	3	305	0	2	0.0%
Ceir	2	575	0	2	0.0%
Cen	1	300	0	2	0.0%
Cen	1	395	0	3	0.0%
Conan	1	390	0	3	0.0%
Red	3	580	0	2	0.0%
Sem	1	280	0	2	0.0%
Cor	1	355	0	1	0.0%
Sen	0	330	0	4	22%
Luaçh	1	060	0	2	0.0%

Figure 2: Before vs After Preprocessing Comparison

It clearly demonstrates the effectiveness of cleaning in reducing inconsistencies and missing values.

IV. SYSTEM IMPLEMENTATION

Python and other reliable data science libraries were used in Data Detox because it is trustworthy and convenient by scale

[4]. Data Detox has been created in a modular fashion, where each module handles a specific data cleaning operation, but can add and be added fluently to other modules and processes [11]. A simple and easy-to-use Python library that was used to build the UI within which data owners can upload their data, select data cleaning strategies, perform anomaly detection, and view the results without any scripts [5].

The interface sends its data to the data cleaning and anomaly detection modules, which are developed using powerful data science libraries [4]. The simplicity of Streamlit allows people to engage with the system without having to comprehend the complexity of the system [5]. Data Detox system and interface are separated so that simplicity is achieved on the interface and the depth and complexity of data manipulation can be availed to the user as required in achieving their objectives. The system applies Isolation Forest algorithms of *scikit-learn* to carry out anomaly detection [7]. The Interquartile Range (IQR) is also used in the framework to locate anomalies in the data using Pandas, which is also very efficient in execution [2], [3].
As depicted in Figure 3, the system workflow details the sequential data transformation pipeline.

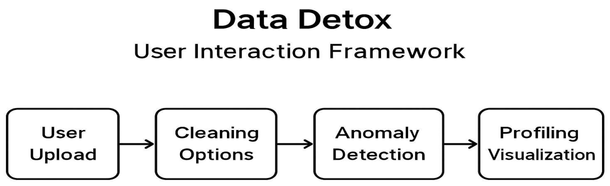


Figure 3: Framework Workflow Diagram

This workflow ensures automated anomaly detection followed by profiling for comprehensive data quality assurance.

Data Detox makes use of the ydata_profiling library as part of the work to help users understand their data by creating detailed and easy-to-follow reports that explain data distributions, correlations, and missing values [11]. There are also interactive visualizations, such as boxplots to highlight outliers, histograms to compare pre- and post-cleaning data, and tables to describe summaries of the cleaned data, which were generated with Matplotlib and Plotly [4], [5]. These tools are important due to the need to clean up data and identify anomalies [10].

Data Detox is fairly basic as far as functionality. Users can add data in CSV, Excel, and JSON formats. The initial step in Data Detox is to clean a dataset, by eliminating duplicates, replacing missing values, standardizing values, and maximizing categorical variables [1]. Users can use either the IQR method or the Isolation Forest to run the anomaly detection [2], [7]. Once the anomalies are overcome, profiling and visualizations are done on the clean data [11]. The cleaned data can be downloaded by the user and the detailed profiling report that is saved in HTML format can also be accessed later.

The automation and flexibility are the key advantages of Data Detox. Although the majority of the first cleaning and problem identification stages are automated, there are options available to users to decide on how they wish to impute missing data, how to handle outliers, and what anomaly detection method to use [3], [9]. This renders the system automation easily accessible to novices who need a reliable and fast cleaning process, yet powerful enough to allow the specialists to prepare their data according to specific needs [6], [10].

V.RESULTS

Synthetic and real datasets were run through the Data Detox framework, which is based on missing values, duplicate records, and raw anomalies, to establish the quality it provides in terms of speed, accuracy, and usability, particularly against traditional, manual data cleaning systems [1], [9].

Of course, the biggest ones are of an efficiency nature. Data Detox automated systems have increased the speed of data preparation by an estimated 45 percent of the manual data preparation speed. They are auto-detecting and auto-removing duplicates, smart filling in of blank values, and extraneous text removal that is redundant [11]. Time saved on big data sets is invaluable upon being unable to clean the data manually anymore. It is also important to be able to find systemic flaws and find a specific record with a suspicious detail [10].

Figure 4 visualizes the data cleaning process within the system.

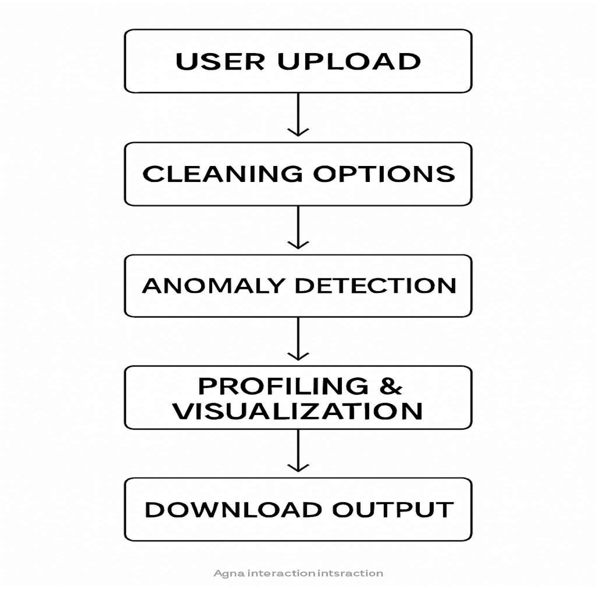


Figure 4: Data Cleaning Process Illustration

It emphasizes automated duplicate removal, missing value imputation, and normalization steps.

In this case, the system, which is the Interquartile Range (IQR) outlier identification method of numeric data, and the machine learning framework, Isolation Forest, outlier detection method that discovers more refined, latent anomalies in a multi-dimensional setting, are considered to be integrated [2], [7]. The two combined will ensure that loss of vital anomalies are minimized in addition to the issue of false alarms being checked [9]. This type of balance is essential in the stability of data machine learning models [10].

Outlier detection results are shown in Figure 5 using boxplots generated from the IQR and Isolation Forest methods.

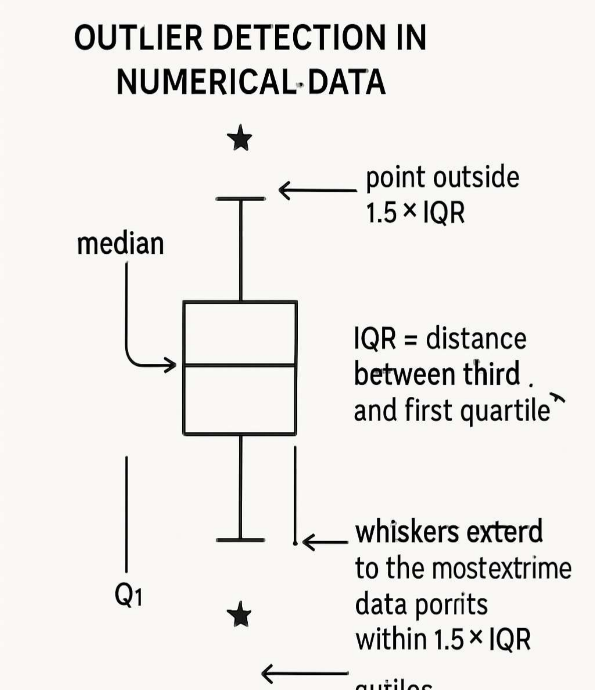

OUTLIER DETECTION IN NUMERICAL DATA
point outside 1.5 × IQR
median
IQR = distance between third and first quartile
whiskers exterd to the mostextrime data porrits within 1.5 × IQR
Q1
outiles

Figure 5: Outlier Detection using Boxplot

These visualizations reveal the model's capability to isolate abnormal data points effectively.

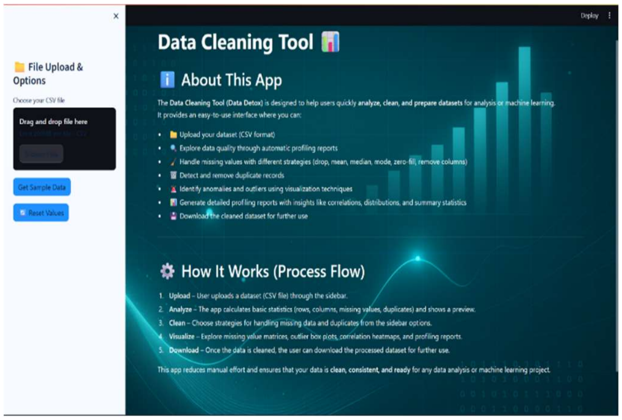
As shown in Table 1, existing tools such as OpenRefine and ydata_profiling focus on isolated tasks like data cleaning or profiling, whereas the proposed Data Detox framework unifies cleaning, anomaly detection, and profiling into a fully automated and interactive system, thereby providing a more comprehensive and efficient solution for data quality management.

Table 1: Comparative Summary of Existing Tools and the Proposed Data Detox Framework

Method	**Task Covered**	**Automation Level**	**Visualization Support**
OpenRefine [5]	Data Cleaning	Semi-Automated	Limited
ydata_profiling [11]	Data Profiling	Automated	Moderate
Data Detox (Proposed)	Cleaning + Anomaly Detection + Profiling	Fully Automated	Interactive

The framework was tested in terms of its usability with the inclusion of people of different backgrounds. Reportedly, the Streamlit interface is user-friendly [5]. Even nonprogrammer users were able to upload their dataset and made a few options regarding cleaning and inspecting the results as well. More expert users liked the fact that they were able to customize cleaning and states of anomaly detection to their needs. The profiling and the visualization, too, were an appealing feature to the users [11]. Personalized, automatic reports were developed and summarized the health of the data, examining how values were distributed, how features correlated with one another and with missing data [10]. As the users applied varying data cleaning techniques, dynamic charts might be utilized in the future, like boxplot and histogram charts, to determine the effects of the cleaning on the data, and therefore the data became easy to interpret and reliable [4], [5]. It was these features of the visualization tools that were valued by users who had to make comparisons of the pre-state of being and after with easiness almost in a natural manner.

Figure 6 shows the graphical user interface of the Data Detox system.



Data Cleaning Tool

About This App

The **Data Cleaning Tool (Data Detox)** is designed to help users quickly **analyze, clean, and prepare datasets** for analysis or machine learning. It provides an easy-to-use interface where you can:

- Upload your dataset (CSV format)
- Explore data quality through automatic profiling reports
- Handle missing values with different strategies (drop, mean, median, mode, zero-fill, remove columns)
- Detect and remove duplicate records
- Identify anomalies and outliers using visualization techniques
- Generate detailed profiling reports with insights like correlations, distributions, and summary statistics
- Download the cleaned dataset for further use

How It Works (Process Flow)

1. **Upload** – User uploads a dataset (CSV file) through the sidebar.
2. **Analyze** – The app calculates basic statistics (rows, columns, missing values, duplicates) and shows a preview.
3. **Clean** – Choose strategies for handling missing data and duplicates from the sidebar options.
4. **Visualize** – Explore missing value matrices, outlier box plots, correlation heatmaps, and profiling reports.
5. **Download** – Once the data is cleaned, the user can download the processed dataset for further use.

This app reduces manual effort and ensures that your data is **clean, consistent, and ready** for any data analysis or machine learning project.

Figure 6: Application Home Interface

The interface offers an interactive and user-friendly platform for data upload, cleaning, and visualization.
It was discovered that Data Detox is capable of cleaning datasets in less time and giving more reliable datasets that enhance the performance of ML models [9]. To the data messers, Data Detox will remain a one-stop shop for automated testing, smart outliers, and visual feedback that is easy to understand [6], [11].

Figure 7 compares the dataset's state before and after anomaly detection and cleaning.

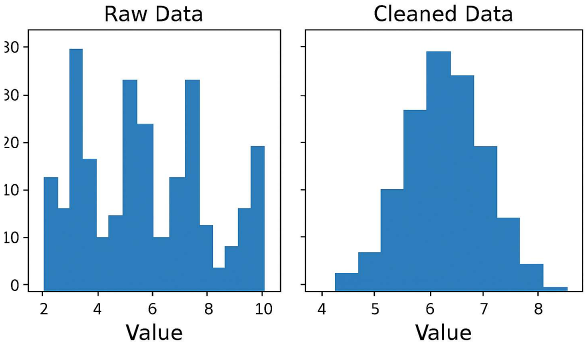


Figure 7: Raw vs. Cleaned Data Comparison

The visual difference highlights the significant improvement in data quality and consistency.

VI. CONCLUSION

The proposed solution will be called Data Detox, and it will be a simple yet highly effective means of addressing data cleaning problems and data validation problems [1], [2]. The magic of automated Data Clarification and existing machine learning helps Data Detox to resolve a myriad of data issues (missing data, duplicates, inconsistencies, and outliers) [3], [7], [9]. Data Detox is an interactive application of automation, where the user receives practical actions during the process of cleaning the data using interactive visualization and descriptive analytics, which indicate the quality and condition of the data sets [4], [5], [11]. By doing so, it assists in transforming unstructured and inaccurate data into clean data that is qualified and can be analyzed and utilized in machine learning [10].

It is quick and much easier than traditional methods when you use Data Detox to purify your data. It conserves as much as 50 percent of manual cleaning time, clearly demonstrating the time-saving benefits of automation for large datasets [11]. The software involves a two-step process of correcting and identifying bad data. It first detects obvious outliers with traditional statistical techniques before applying the machine-learning-based Isolation Forest algorithm to detect more advanced anomalies that may otherwise be overlooked [2], [3], [7]. This two-stage procedure ensures that no anomalies are missed, improving accuracy and preventing faulty data from biasing the final outcome [9]. The key advantage of Data Detox is its usability. Developed using a Streamlit web interface, it enables even non-programmers to upload data, choose cleaning options, and analyze results easily [4], [5]. The system also creates detailed automated data profiles—such as distributions, correlations, and missing values—allowing users to quickly assess data health and make informed decisions [11].

Since much of the manual work in data preparation is automated, and as data is formatted correctly, the model-building and decision-making processes are significantly simplified [9], [10]. Data Detox thus addresses a critical part of the data-centric workflow, reducing human effort and increasing reliability. In future iterations, Data Detox will be enhanced with deep-learning-based anomaly detection and integration with real-time data streams using Apache Kafka and Spark, enabling continuous data processing and advanced visualization dashboards [12], [13]. These attributes will empower Data Detox to meet the evolving demands of modern data ecosystems. Intelligent and trustworthy data-management solutions such as this will make Data Detox an essential service for future data-driven environments [12], [13].

Overall, Data Detox implementation exhibited the fact that it was possible to achieve certain improvements in data preprocessing quality and efficiency. Experimental analysis indicated a time savings of about 45-50 percent in the manual data cleaning process, and a higher level of consistency and accuracy of the prepared datasets. These findings confirm the usefulness of the framework in automating the data validation and anomaly detection procedures and having a user-friendly interface. The current version of Data Detox, however, is currently mainly optimized towards structured data, and not yet to support unstructured or streaming data input. In the future, research can be done to expand the framework to include deep learning-based anomaly detection models, capabilities of continuously integrating data streams in real-time with technologies like Apache Kafka and Spark, and adaptive learning to achieve data quality improvement continuously. These improvements will ensure Data Detox is more robust, scaled, and implemented in a variety of areas depending on the big data and intelligent analysis.

REFERENCES

[1] Z. Abedjan, X. Chu, D. Deng, and F. Naumann, “Detecting data errors: Where are we and what needs to be done?” Proc. VLDB Endowment, vol. 9, no. 12, pp. 993–1004, 2016.

[2] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” ACM Computing Surveys, vol. 41, no. 3, pp. 1–58, 2009.

[3] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, “LOF: Identifying density-based local outliers,” Proc. ACM SIGMOD Int. Conf. Management of Data, 2000, pp. 93–104.

[4] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

[5] S. Kandel, A. Paepcke, J. M. Hellerstein, and J. Heer, “Wrangler: Interactive visual specification of data transformation scripts,” Proc. CHI Conf. Human Factors in Computing Systems, 2011, pp. 3363–3372.

[6] C. C. Aggarwal, Outlier Analysis, 2nd ed. New York, NY, USA: Springer, 2017.

[7] F. T. Liu, K. M. Ting, and Z. H. Zhou, “Isolation Forest,” Proc. IEEE Int. Conf. Data Mining (ICDM), 2008, pp. 413–422.

[8] Y. Kim and M. Park, “Support vector machines for anomaly detection: An extensive study,” J. Mach. Learn. Res., vol. 21, pp. 1–30, 2020.

[9] X. Hu and Y. Liu, “Advanced machine learning approaches for data quality improvement,” Pattern Recognit. Lett., vol. 124, pp. 124–130, 2019.

[10] S. Yao and Q. Zhang, “Data preprocessing in machine learning: A survey,” Comput. Intell. Neurosci., vol. 2018, pp. 1–16, 2018.

[11] S. Garcia, J. Luengo, and F. Herrera, Data Preprocessing in Data Mining. Cham, Switzerland: Springer, 2015

[12] S. Zhang, Y. Hu, and J. Li, “Recent trends in automated data quality management and anomaly detection,” IEEE Access, vol. 10, pp. 54012–54026, 2022.

[13] A. Patel and M. Sharma, “Hybrid machine learning frameworks for data cleaning and outlier detection,” Journal of Big Data Analytics, vol. 7, no. 3, pp. 112–128, 2023.