PAPER NAME

Systematic Literature Review of Data Distribution in Preprocessing Stage with Focus on Outliers.pdf

AUTHOR

Ahmad Rofiqul Muslikh

WORD COUNT

4869 Words

CHARACTER COUNT

27246 Characters

PAGE COUNT

6 Pages

FILE SIZE

296.8KB

SUBMISSION DATE

Feb 23, 2024 10:26 AM GMT+7

REPORT DATE

Feb 23, 2024 10:26 AM GMT+7

9% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 9% Internet database
- 0% Publications database
- Crossref database
- Crossref Posted Content database

Excluded from Similarity Report

- Submitted Works database
- Bibliographic material

Systematic Literature Review of Data Distribution in Preprocessing Stage with Focus on Outliers

Ahmad Rofiqul Muslikh
Faculty of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
p41202100033@mhs.dinus.ac.id

Pulung Nurtantio Andono
Faculty of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
pulung.nurtantio.andono@dsn.dinus.ac.id

Aris Marjuni
Faculty of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
aris.marjuni@dsn.dinus.ac.id

Heru Agus Santoso
Faculty of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
heru.agus.santoso@dsn.dinus.ac.id

***Abstract*— Data Preprocessing refers to the steps and techniques applied to raw data before it is ready to be analyzed or modeled as a substantive part of the data flow and aims to transform, clean and organize data in a revised way for the quality, relevance and efficiency of subsequent data analysis tasks. Handling outliers in the N2O Emissions Dataset, Fertilizer Prediction and Crop Yield Prediction Dataset is an important step in the data analysis process. The approach taken will depend on the specific context and purpose of the analysis, and it is important to carefully consider the impact of outliers on the results. Using the methods discussed researchers and analysts can effectively identify and treat outliers in the N2O Emissions Dataset, Fertilizer Prediction, Crop Yield Prediction Dataset, and produce more accurate and reliable results. Implemented a systematic literature that involved searching for articles published from 2015 to 2023 for review. The quality of the existing studies used the assessment criteria of 50 relevant studies identified as having been conducted following systematic literature guidelines.**

***Keywords*—SLR, Data Preprocessing, Data Distribution, Outliers**

I. INTRODUCTION

Data preprocessing refers to the steps and techniques applied to raw data before it is ready for analysis or modelling. It is a substantive part of the data pipeline, and it aims to transform, clean and organize data in a revised manner to the quality, relevance and efficiency of subsequent data analysis tasks [1]. There are many ways to improve the quality of data obtained through data preprocessing. In the context of data preprocessing, data distribution refers to the distribution of the data features or variables within a dataset. In a data set, outlier detection can significantly affect the distribution of data, especially in the presence of an abnormal distribution[2].

Data distribution plays a crucial role in the preprocessing phase of data analysis. It helps in outlier detection, guides the selection of preprocessing techniques, informs feature engineering decisions, ensures model assumptions are met, and aids in data normalization [3]. Outliers can have a significant impact on subsequent analysis or modelling tasks. We can detect outliers through statistical measures or visualizations, enabling their proper handling during preprocessing by examining the data distribution. Different preprocessing techniques are suitable for different types of data distributions. For example, normalization techniques like Z-score or Min-Max scaling can be applied if the data follows a normal distribution. Otherwise, data transformation techniques such as logarithmic or Box – Cox transformations may be appropriate if the data is skewed or non – normal distribution [4]. Understanding the data distribution guides the selection of the most suitable preprocessing techniques to improve data quality and prepare it for analysis.

The discrepancy between data distribution and preprocessing can arise due to several reasons like handling skewed data, outliers and extreme values. Many preprocessing techniques, such as normalization or standardization, assume a symmetric distribution. When dealing with skewed data, the chosen preprocessing techniques may need to be more suitable and effective in capturing the underlying patterns or relationships in the data. Outliers can distort the data distribution, making it non – normal or non – linier. Preprocessing techniques often involve handling outliers, such as removing or transforming them. However, if the outliers are not appropriately addressed or if they are treated as errors when they represent valid information, it can lead to a gap between the data distribution and the applied preprocessing techniques. In process to analyze data, we need a dataset to perform the preprocessing techniques[5].

We will examine the importance of identifying and addressing outliers and the impact that these observations can have on the results of statistical analyses and machine learning models by using The N2O Emissions Dataset, Fertilizer Prediction and Crop Yield Prediction Dataset. The N2O Emissions dataset can be used to understand the sources and levels of N2O emissions in different countries, and to identify trends and patterns over time[6]. Some common methods for dealing with outliers in datasets include identifying and correcting errors in the data, excluding outliers from the analysis, or transforming the data to make it more normal. Ultimately, the approach taken will depend on the specific context and goals of the analysis[7].

In our research we propose, for the first time, a validation process for each feature data set. We will validate every process that start with a raw data, then checking the outliers using boxplot, remove data containing small outliers, checking the data distribution using several techniques algorithm among others Histogram Analyzing, Percentage Immuter Data, Anderson and Saphiro Test. We will come up with the best algorithm for checking the data distribution,

then we need to remove outliers to perform the output of normal data distribution. Each of the processes described above will be validated one by one[8].

In conclusion, handling outliers in the N2O Emissions Dataset, Fertilizer Prediction and Crop Yield Prediction Dataset is a solution in the data preprocessing process. The approach taken will depend on the specific context and goals of the analysis, and it is important to carefully consider the impact that outliers can have on the results. By using the methods discussed in this chapter, researchers and analysts can effectively identify and address outliers in the N2O Emissions Dataset, Fertilizer Prediction, Crop Yield Prediction Dataset, and produce more accurate and reliable results. The structure of this paper as follows section 2 outlines the related work carried out in this field, section 3 carries out the proposed method, section 4 shows the results and discussion of the comparison between the methods to identify and deal with outliers, finally, in section 5 we put an acknowledgment about this study.

II. Related Work

A. Overview of Data Distribution in Preprocessing

Data preprocessing is a data mining process because data quality must be checked before applying data mining algorithms to understand the distribution of data in each attribute. If the data does not follow a normal distribution, such as there are outliers or the data is asymmetrical, it can identify problems in data quality and collection errors[9]. Identifying abnormal data are steps that can be taken to clean up the data before being executed into the model. Data distribution is important to understand the characteristics of the data before proceeding to the next stage, such as modeling or analysis[10].

Outliers can interfere with statistical analysis and cause bias in predictive models. It is important to detect detail outliers as possible to avoid data processing failures because anomalous events can cause minor to severe damage to the data distribution[11]. Identifying and treating outliers, a more accurate and representative data distribution can be achieved. In ensuring good data distribution, the preprocessing stage may involve steps such as normalization, detection and handling of outliers, class sampling, data transformation, and standardization or scaling. All of these steps aim to achieve a more representative data distribution, minimize bias, and improve the performance of machine learning algorithms[12].

B. Existing Approaches and Techniques

In data preprocessing, there are several approaches and techniques that can be used to deal with data distribution that is not ideal or does not meet certain assumptions as follows:

Boxplot

A statistical graph used to visually present the distribution of data. Boxplots provide information about quartiles, interquartile ranges, extreme values, and the presence of outlier values in the dataset[13]. A boxplot usually consists of a vertical line that divides the box into two equal parts, as well as horizontal lines around the box that describe the range of data beyond the quartiles. The boxplot highlights the values of the quartiles in the dataset, namely the first quartile (Q1), median (Q2), and third quartile (Q3)[14].

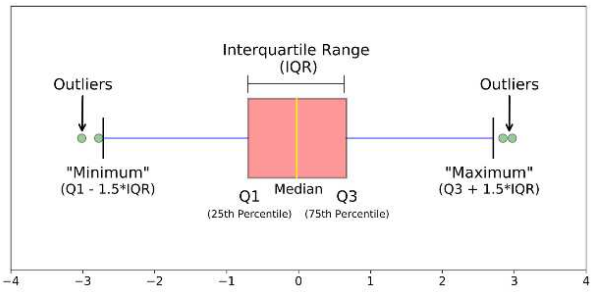

Fig. 1. Boxplot

Histogram Analyzing

Histogram Analyzing for graphically visualizing and analyzing the frequency of pixels in an image for each pixel or intensity value between 1 and 255. Histograms are especially useful for analyzing the distribution of graphical display. Attempts to infer the pixel frequency distribution of the encrypted image to apply any attack. However, the proposed algorithm generates a histogram of encrypted images with a constant level for each original image. Outliers can also be detected using a histogram, with most of the observations being on one side, and some of the observations appearing far from the main group[15]. By analyzing the histogram, researchers can easily identify the presence of anomalies or outlier values. If a bar in the histogram is very high or low compared to other bars, it indicates the presence of such rare values. These outliers can be important for various reasons, as they might signify errors in data collection, measurement, or data entry, or they could represent genuinely significant events or data points[16].

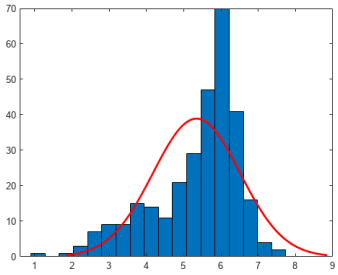

Fig. 2. Histogram Analyzing

Anderson

A statistical test used to test the suitability of a data sample with a certain probability distribution, such as the normal distribution. This test measures the extent to which the data follow the distribution being tested by comparing the empirical data distribution function with the theoretical distribution function being tested. This test provides an Anderson statistical value that can be used to evaluate the null hypothesis that the data comes from a certain distribution. Anderson's role is to help identify whether a data sample can be ascribed to the distribution being tested[11].

Shapiro Test

The Shapiro Test is used to test the normality assumption of the data sample. Many statistical methods and data analysis rely on assumptions of normality, such as parametric tests, regression analysis, and other hypothesis testing. By testing for normality, we can find out whether these assumptions are met or not. The Shapiro test has

limitations, especially when the sample size is large, small sample differences from the normal distribution can produce statistically significant results. Therefore, interpretation of the results of these tests should be made with caution and taking into account the data context and sample size[17].

C. Gaps and Limitations in the Literature

Some common research gaps in the approach to the outlier problem during the automatic processing of measurement data series received from technical devices are considered. Strategies for detecting outliers in time series of noisy data containing unknown trends[18]. We will examine the importance of identifying and treating outliers and the impact of these observations on the results of statistical analyzes and machine learning models using the N2O Emissions Dataset, Fertilizer Prediction and Crop Yield Prediction Dataset[13].

This study first proposes a validation process for each feature data set. We will validate each process starting with raw data, then check for outliers using boxplots, delete data containing small outliers, check the distribution of data using several algorithmic techniques including Histogram Analysis, Percentage of Immuter Data, Anderson and Shapiro Test. We will generate the best algorithm to check the distribution of the data, then we need to remove the outliers to output the normal data distribution. Each process described above will be validated individually. The approach taken will depend on the specific context and purpose of the analysis, and it is important to carefully consider the impact of outliers on the results. A systematic literature review with a focus on data preprocessing outliers is presented in a paper that concentrates on papers explaining systematic literature reviews (SLR)[14].

III. METHODOLOGY

A. Systematic Literature Review (SLR) Process

In this study SLR as a methodology for studying current research work related to preprocessing data on data preprocessing outliers. Systematic literature reviews provide a means for the evaluation of research related to a particular topic area. The goal of conducting an SLR is to systematically collect and evaluate all relevant published studies with a predetermined RQ to inform the research community[19].

In this case the research took some published literature from popular database journals namely IEEE Xplore, Springer Link, MDPI, Science Direct and ACM from 2015 to 2023. The aim of the Research Questions was to maintain the focus of the literature review. This condition facilitates the process of finding the required data. In this case, the research employed a data collection approach that involved gathering published literature from popular database journals, namely IEEE Xplore, Springer Link, MDPI, Science Direct, and ACM, spanning the period from 2015 to 2023. It allowed the researchers to efficiently sift through the vast amount of available literature, enabling them to identify and extract the most pertinent information that aligned with the research questions and objectives. As a result, the data collected were more precise and relevant, enhancing the overall quality and rigor of the research findings [20]. The questions posed by this study in Table I are as follows:

TABLE I. RESEARCH QUESTION (RQ)

No	Research Question (RQ)	
	Research Question	**Motivation**
1	Which international journals often publish research on handling data preprocessing on outliers?	Identify which international journals often publish research on handling data preprocessing on outliers.
2	In what year was the trend of research on data preprocessing on outliers?	Identify research trends about data preprocessing on outliers.
3	What problems arise in research on handling preprocessing data on outliers?	Identify problems that often arise in research about handling preprocessing data on outliers.

To determine the research questions in Table 1, the process involved several key steps. First, an extensive review of the existing literature on handling preprocessing data on outliers was conducted from 2015 to 2023. This literature review helped identify the current gaps and areas that required further investigation. The researchers brainstormed and discussed potential research questions that would address the identified gaps and contribute valuable insights to the field. These questions were formulated to be specific, measurable, and relevant to the research topic, ensuring that they would lead to meaningful findings.

The search process is in accordance with the Systematic Literature Review (SLR) stages above which consists of several processes, including selecting a digital library and setting keywords. Before starting the search, it is necessary to determine or select the appropriate database to find relevant journals. During this process, we collected a total of 50 international journals and conference papers which we kept for the inspection stage[21]. The following is a digital library in this study: Science Direct, SpringerLink, IEEE Explorer, ACM Digital Library, MDPI.

B. Inclusion and Exclusion Criteria

The inclusion and exclusion criteria section is used to select the main research from the results of the articles based on the criteria which will later be reviewed by the researcher. Exclusion and inclusion criteria ensured that only relevant studies were included in the analysis of the preprocessing outliers data[22]. As this review focuses on addressing data preprocessing on outliers, only papers published in English from 2015 to 2023 were included in this study. This study aims to collect, analyze, and synthesize articles in a systematic literature review from 2015 to 2023. The inclusion and exclusion criteria section can be seen in Table II as follows[23]:

TABLE II. INCLUSION AND EXCLUSION CRITERIA

No	Inclusion and Exclusion Criteria	
	Inclusion criteria	***Exclusion Criteria***
1	This research is written in English	This research the full text is not available
2	This research was published starting in 2015−2023	Duplicate research
3	Articles related to data preprocessing outliers	This research the full text is not available
4	Articles are fully accessible	Unranked conference articles

C. Data Extraction and Analysis Methods

Data Extraction

We focus on specific information in each article related to data pre-processing, cleaning, preparation, purification, and sanitization related to existing unbalanced data sets. We also consider articles that address key data related issues such as detection of outliers and other data preprocessing. Data collected from each study included journal or conference resources, scope of research and topic areas, details of author, institution and country of origin and Research summary, including RQ and responses to each question[24].

Data Analysis

In data analysis we report how the information extracted from the study was analyzed. In this way, the information obtained can help not only in handling our RQ, but in guiding the machine learning community in decision making with an emphasis on data preprocessing We systematically apply statistical techniques to ensure clear presentation of information[25].

IV. Result and Discussions

4.1 Research design

In this study, the data used is the N2O Emissions Dataset, Fertilizer Prediction and Crop Yield Prediction Dataset which is one that offers information about anomaly readings. Data distribution uses several algorithmic techniques including Histogram Analysis, Percentage of Immuter Data, Anderson and Shapiro Test. In this study several stages will be carried out as depicted in Fig. 3.

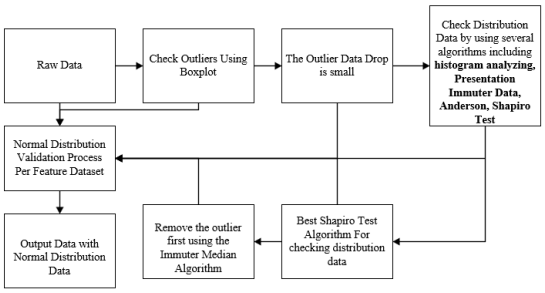

Fig. 3. Research Steps

The steps for preprocessing data outliers in Fig. 3 above, first is the validation process for each feature data set. We will validate each process starting with raw data, then check for outliers using boxplots, delete data containing small outliers, check the distribution of data using several algorithmic techniques including Histogram Analysis, Percentage of Immuter Data, Anderson and Shapiro Test. We will generate the best algorithm to check the distribution of the data, then we need to remove the outliers to output the normal data distribution. Each process described above will be validated individually.

The characteristic of the N2O Emission dataset we used are DAF_SD, NO3, PP2, WFPS25cm, DAF_TD feature with normal and not normal data, the characteristic data will be identified by Histogram test, Shapiro test and KS (Kolmogorov-Smirnov) – test. The methods for handling outliers are Interquartile Range (IQR) method, Z – Score method and Winsorizing method. Relevant studies we used for summaries are data preprocessing for handling outlier topics, the research years were taken from 2015 to 2023.

In this experiment we used the average data from each method used to detect the percentage of outliers from the data before being removed. Next, we removed the outlier data according to the value closest to the average result of each method compared, where the results showed that the Interquartile Range (IQR) had a result that was not far from the average result of each method, indicating that the highest error rate was 5.426%. Based on the results, the farthest from the average result was from feature PP2, where the outlier detection in feature PP2 using Z-Score was still better. The next step is to perform data interpolation for the 6 features that were detected as not normally distributed to balance them with other features, where after removing the outliers, the data had an imbalanced amount compared to the other features that did not have their outliers removed.

4.2 Data Collection

In this paper, we use outlier detection in public datasets including the N2O Emissions Dataset[26], Fertilizer Prediction[27], and Crop Yield Prediction Dataset[28] with data sources from the Kaggle repository collection.

4.3 Reference Systematic Literature Review

The list of main studies is shown in Table III which consists of 4 attributes (Journal Name, Publication, Year and Problem) and 50 main studies were obtained from 2015 to 2023 which are sorted by year of publication.

TABLE III. Reference Systematic Literature Review

No	*Journal Name*	*Publication*	*Year*	*Problems*
1	Outlier detection using AI: a survey	AI Assurance	2023	Outlier Detection
2	Performance evaluation of outlier detection techniques in production timeseries: A systematic review and meta-analysis	Elsevier	2022	Outlier Detection
3	Long-term variability in N2O emissions and emission factors for corn and soybeans induced by weather and management at a cold climate site	Elsevier	2022	Low Accuracy
4	[13][12][14][11][19][18] [8] [16][15] [8][6] [5] [29] [25] [23] [24][22] [21]			

4.4 Which international journals often publish research on handling data preprocessing on outliers?

From the study selection process, 50 journals were obtained related to data processing outliers in review analysis to answer research questions that had been made previously. Then from the selected journals, international journals that contribute in the field of data processing outliers are then identified. Fig. 4 shows a journal that publishes the topic of data processing outliers.

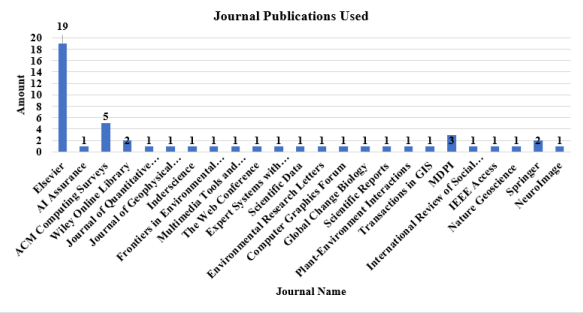

Fig. 4. Journal Publication Used in Literature Review

Based on Fig. 4 it is known that there are 2 highest journal publications that often publish journals with the topic of data processing outliers, namely Elsevier and ACM Computing Surveys. The two journal publications have a total of 19 and 5 respectively out of 50 journal publications that address the topic of data processing outliers in the identified SLR.

4.5 In what year was the trend of research on data preprocessing on outliers?

An overview of studies related to data processing outliers from year to year is shown in Fig. 5. In this review, the research years taken were from 2015 to the latest, namely 2023.

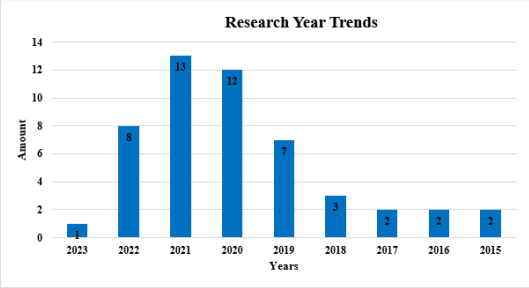

Fig. 5. Research Year Trends

Based on Fig. 5 it is known that there are 3 highest published years of journals with the topic of data processing outliers, namely the years from 2020 - 2022. The three published years of these journals totaled 13 from 2021, 12 from 2020 and 8 from 2022 based on 50 journal publications which raises the topic of data processing outliers on identified SLR.

4.6 What problems arise in research on handling preprocessing data on outliers?

The problems that arise in research on data processing outliers are very diverse. Of the many problems, this review broadly categorizes the problems that arise in research on data processing outliers into 3 categories, namely Outlier Detection, Data Processing and Low Accuracy.

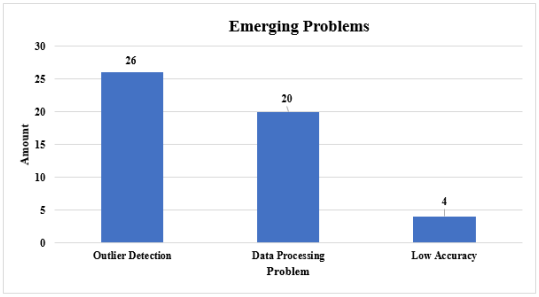

Fig. 6. *Problems that Arise in Data Preprocessing on Outliers*

Fig. 6 shows that the problem of Outliers Detection is the most common problem that arises in research on data processing outliers. Handling outliers in a data set includes identifying and correcting errors in the data, excluding outliers from analysis or changing the data to make it more normal. Then the second most problem is Data Processing.

In data preprocessing an important step in machine learning studies because preprocessing outliers in proper data processing can allow researchers to identify and correct errors in a data set, exclude outliers from analysis, and change data to make it more normal. Consequently, there are a limited number of systematic literature review studies on data processing outliers. In this study, the authors analyze the existing literature to identify the main issues related to data quality and handling and to provide a set of techniques used to overcome these problems when performing outlier detection processing. Implemented a systematic literature that involved searching for articles published from 2015 to 2023 for review. The quality of the existing studies used the assessment criteria of 50 relevant studies identified as having been conducted following systematic literature guidelines.

ACKNOWLEDGMENT

The research described in this paper is supported by dissertation lecturers Pulung Nurtantio Andono, Aris Marjuni and Heru Agus Santoso.

REFERENCES

[1] D. Eilertz, M. Mitterer, and J. M. Buescher, "automRm: an r package for fully automatic LC-QQQ-MS data preprocessing powered by machine learning," *Anal. Chem.*, 2022, doi: 10.1021/acs.analchem.1c05224.

[2] G. Li *et al.*, "Outlier data mining method considering the output distribution characteristics for photovoltaic arrays and its application," *Energy Reports*, vol. 6, pp. 2345–2357, 2020, doi: 10.1016/j.egyr.2020.08.034.

[3] S. Sachan, F. Almaghrabi, J. B. Yang, and D. L. Xu, "Evidential reasoning for preprocessing uncertain categorical data for trustworthy decisions: An application on healthcare and finance," *Expert Syst. with ...*, 2021.

[4] C. V Castro and D. R. Maidment, "GIS preprocessing for rapid initialization of HEC-HMS hydrological basin models using web-based data services," *Environ. Model. &Software*, 2020.

[5] M. P. J. Oreska, K. J. McGlathery, L. R. Aoki, A. C. Berger, P. Berg, and L. Mullins, "The greenhouse gas

offset potential from seagrass restoration," *Sci. Rep.*, vol. 10, no. 1, pp. 1–15, 2020, doi: 10.1038/s41598-020-64094-1.

[6] Y. Zhang and Q. Yu, "Does agroecosystem model improvement increase simulation accuracy for agricultural N2O emissions?," *Agric. For. Meteorol.*, vol. 297, no. June 2020, p. 108281, 2021, doi: 10.1016/j.agrformet.2020.108281.

[7] E. Ibrahim, M. A. Shouman, H. Torkey, and A. El-Sayed, "Correction to: Handling missing and outliers values by enhanced algorithms for an accurate diabetic classification system," *Multimed. Tools Appl.*, vol. 80, no. 13, p. 20149, 2021, doi: 10.1007/s11042-021-10843-x.

[8] A. Blázquez-García, A. Conde, U. Mori, and J. A. Lozano, "A Review on Outlier/Anomaly Detection in Time Series Data," *ACM Comput. Surv.*, vol. 54, no. 3, 2021, doi: 10.1145/3444690.

[9] M. N. K. Sikder and F. A. Batarseh, "Outlier detection using AI: a survey," *AI Assur.*, pp. 231–291, 2023, doi: 10.1016/b978-0-32-391919-7.00020-2.

[10] S. J. Khagendra Raj Baral and C. W.-R. Guelph, Shannon E Brown, "Long-term variability in N2O emissions and emission factors for corn and soybeans induced by weather and management at a cold climate site," *Elsevier*, 2022.

[11] M. K. EJ Jamshidi, Y Yusup, JS Kayode, "Detecting outliers in a univariate time series dataset using unsupervised combined statistical methods: A case study on surface water temperature," *Elsevier*, 2022.

[12] D. P. Purbawa *et al.*, "Adaptive filter for detection outlier data on electronic nose signal," *Sens. Bio-Sensing Res.*, vol. 36, no. April, p. 100492, 2022, doi: 10.1016/j.sbsr.2022.100492.

[13] E. R. Stuchiner and J. C. von Fischer, "Characterizing the Importance of Denitrification for N2O Production in Soils Using Natural Abundance and Isotopic Labeling Techniques," *J. Geophys. Res. Biogeosciences*, vol. 127, no. 5, pp. 1–21, 2022, doi: 10.1029/2021JG006555.

[14] E. V. Karlovets, S. A. Tashkun, S. Kassi, and A. Campargue, "An improved analysis of the N2O absorption spectrum in the 1.18 μm window," *J. Quant. Spectrosc. Radiat. Transf.*, vol. 278, no. November, pp. 1–21, 2022, doi: 10.1016/j.jqsrt.2021.108003.

[15] W. Gruber, R. Niederdorfer, J. Ringwald, E. Morgenroth, H. Bürgmann, and A. Joss, "Linking seasonal N2O emissions and nitrification failures to microbial dynamics in a SBR wastewater treatment plant," *Water Res. X*, vol. 11, p. 100098, 2021, doi: 10.1016/j.wroa.2021.100098.

[16] S. Askari, "Fuzzy C-Means clustering algorithm for data with unequal cluster sizes and contaminated with noise and outliers: Review and development," *Expert Syst. Appl.*, vol. 165, p. 113856, 2021, doi: 10.1016/j.eswa.2020.113856.

[17] S. N. C. Mohammadi, Hamzeh Ali, "Performance evaluation of outlier detection techniques in production timeseries: A systematic review and meta-analysis," *Elsevier*, 2022.

[18] M. A. Bleken, T. F. Rittl, S. Karki, and S. Nadeem, "Data of biomass and N in grass and clover roots, stubbles, and herbage and associated N2O and CO2 emissions, inclusive soil air composition, following autumn ploughing – A field study," *Data Br.*, vol. 43, p. 108352, 2022, doi: 10.1016/j.dib.2022.108352.

[19] P. A. R. Kelly S. Aho, Jennifer H. Fair, Jake D. Hosen, Ethan D. Kyzivat, Laura A. Logozzo, Lisa C. Weber, Byungman Yoon, Jay P. Zarnetske, Peter A. RaymondKelly S. Aho, Jennifer H. Fair, Jake D. Hosen, Ethan D. Kyzivat, Laura A. Logozzo, Lisa C. Weber, Byungman Y, "An intense precipitation event causes a temperate forested drainage network to shift from N2O source to sink," *Wiley Online Libr.*, 2022.

[20] A. Ayadi, O. Ghorbel, A. M. Obeid, and M. Abid, "Outlier detection approaches for wireless sensor networks: A survey," *Comput. Networks*, vol. 129, pp. 319–333, 2017, doi: 10.1016/j.comnet.2017.10.007.

[21] T. Kieu, B. Yang, and C. S. Jensen, "Outlier detection for multidimensional time series using deep neural networks," *Proc. - IEEE Int. Conf. Mob. Data Manag.*, vol. 2018-June, pp. 125–134, 2018, doi: 10.1109/MDM.2018.00029.

[22] T. Kieu, B. Yang, C. Guo, and C. S. Jensen, "Outlier detection for time series with recurrent autoencoder ensembles," *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2019-Augus, pp. 2725–2732, 2019, doi: 10.24963/ijcai.2019/378.

[23] Y. Liu *et al.*, "Generative Adversarial Active Learning for Unsupervised Outlier Detection," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 8, pp. 1517–1528, 2020, doi: 10.1109/TKDE.2019.2905606.

[24] Z. Cheng, C. Zou, and J. Dong, "Outlier detection using isolation forest and local outlier," *Proc. 2019 Res. Adapt. Converg. Syst. RACS 2019*, pp. 161–168, 2019, doi: 10.1145/3338840.3355641.

[25] O. Alghushairy, R. Alsini, T. Soule, and X. Ma, "A review of local outlier factor algorithms for outlier detection in big data streams," *Big Data Cogn. Comput.*, vol. 5, no. 1, pp. 1–24, 2021, doi: 10.3390/bdcc5010001.

[26] "N2O Emissions Dataset", [Online]. Available: https://www.kaggle.com/datasets/saurabhshahane/predictions-of-agricultural-nitrous-oxide-emission

[27] "Fertilizer Prediction", [Online]. Available: https://www.kaggle.com/datasets/gdabhishek/fertilizer-prediction

[28] "Crop Yield Prediction Dataset", [Online]. Available: https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset

[29] Y. Almardeny, N. Boujnah, and F. Cleary, "A Novel Outlier Detection Method for Multivariate Data," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 9, pp. 4052–4062, 2022, doi: 10.1109/TKDE.2020.3036524.

9% Overall Similarity

Top sources found in the following databases:

- 9% Internet database
- Crossref database
- 0% Publications database
- Crossref Posted Content database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	ijaemr.com Internet	2%
2	mdpi.com Internet	1%
3	jurnal.iaii.or.id Internet	1%
4	catalog.libraries.psu.edu Internet	<1%
5	scholarworks.sjsu.edu Internet	<1%
6	atlantis-press.com Internet	<1%
7	osti.gov Internet	<1%
8	researchgate.net Internet	<1%
9	diva-portal.org Internet	<1%

10	orca.cardiff.ac.uk Internet	<1%
11	ieeexplore.ieee.org Internet	<1%
12	arxiv.org Internet	<1%
13	dspace.jaist.ac.jp Internet	<1%
14	frontiersin.org Internet	<1%