Enhancing EEG Data Quality: A Comprehensive Review of Outlier Detection and Cleaning Methods

Sofia Utoft*
utoft@bc.edu
Boston College
Chestnut Hill, MA, USA

Jingwen Dou*
jd5668@nyu.edu
New York University
New York, NY, USA

Jade Wu*
jw32946@pausd.us
Henry M. Gunn Senior High School
Palo Alto, CA, USA

ABSTRACT

Electroencephalography (EEG) is a crucial tool in neuroscience and clinical diagnostics, offering valuable insights into brain function. However, EEG data is often marred by noise and outliers, compromising data quality and analysis accuracy. This paper presents a comprehensive review of outlier detection and dataset cleaning techniques specifically for EEG data, with an additional application on the EEGEyeNet dataset. Our systematic review covers recent advancements in statistical, machine learning, signal processing, and visual inspection methods for noise reduction and outlier removal. We evaluate these methods based on their accuracy, robustness, computational efficiency, and applicability to EEG data. Our results highlight the strengths and limitations of current techniques and utilize the findings to propose potential improvements to EEGEyeNet data processing. This review aims to guide researchers in selecting effective outlier detection and cleaning strategies, ultimately enhancing the reliability of EEG data analysis.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

Review, EEG, eye-tracking, outliers, EEG, time series, spatiotemporal data, data preprocessing, data cleaning, artifact removal, gaze estimation, machine learning, deep learning.

ACM Reference Format:
Sofia Utoft, Jingwen Dou, and Jade Wu. 2024. Enhancing EEG Data Quality: A Comprehensive Review of Outlier Detection and Cleaning Methods. In *Proceedings of KDD Undergraduate Consortium (KDD-UC '24).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

*Utoft and Dou are undergraduate students and Jade is a high school student. They contributed equally to this research.

KDD-UC '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 978-1-4503-XXXX-X/23/08…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Electroencephalography (EEG) is a widely used method in neuroscience and clinical diagnostics for monitoring electrical activity in the brain. In this technique, electrodes are positioned on the scalp to capture and register electrical signals produced by neurons in the brain. Its non-invasive nature and high temporal resolution make it a valuable tool for studying brain function and diagnosing neurological disorders. EEG signals are typically classified by their frequency, amplitude, and shape. The most common classification is based on signal frequency in Hertz (Hz), identifying signals as either Alpha, Beta, Theta, Delta, or Gamma as shown in Table 1. These frequency bands provide valuable insights into different states of brain activity, such as relaxation, alertness, and deep sleep.

Name	Frequency Band (Hz)	Predominant Brain Activity
Delta	0.5-4	Sleeping
Theta	4-8	Dreaming, Meditation
Alpha	8-13	Relaxation
Beta	13-36	Alert, Active
Gamma	36-100	Problem Solving

Table 1: EEG Signal Frequency Bands. Modified from [31]

However, the quality of EEG data is often compromised by various forms of noise and outliers, which arise from environmental interference, physiological artifacts, and technical issues. These contaminants obscure true neural signals, leading to inaccurate analyses and interpretations.

To address these challenges, effective outlier detection and dataset-cleaning techniques are essential. The objective of this paper is to provide a comprehensive review of the methods used for cleaning EEG data, with a particular focus on the EEGEyeNet dataset. The EEGEyeNet dataset presents unique challenges due to its size and complexity, making it an ideal case study for evaluating the efficacy of various data-cleaning methods.

1.1 Research Questions

In addressing these objectives, our study aims to answer three key questions:

- Which specific analysis techniques are used for outlier detection and data cleaning in EEG datasets?
- What data preprocessing methods have proven effective in the analysis of EEG datasets?
- Which specific analysis techniques have been employed in the EEGEyeNet dataset for outlier detection & data cleaning?

By examining these questions, we hope to provide insights that will enhance the reliability of EEG data analysis and contribute to the development of more robust data-cleaning protocols.

Abbreviation	Definition
EEG	Electroencephalography
BCI	Brain-Computer Interfaces
ICA	Independent Component Analysis
CCA	Canonical Correlation Analysis
PCA	Principal Component Analysis
BSS	Blind Source Separation
EMD	Empirical Mode Decomposition
EEMD	Ensemble EMD
MEMD	Multivariate EMD
EOG	Electrooculography
EMG	Electromyography
DWT	Discrete Wavelet Transform
SWT	Stationary Wavelet Transform
SVM	Support Vector Machine
GAN	Generative Adversarial Network

Table 2: Abbreviation Table

2 RELATED WORK

In EEG signal processing, artifact removal is essential for accurate data analysis and interpretation. Several studies have explored various methods to address this challenge[6, 11, 14, 19, 25–29, 38, 39]. Notably, review papers by Sadiya et al. (2021), Mumtaz et al. (2021), Kotte et al. (2020), Huang et al. (2021), Rashmi et al. (2022), Ranjan et al. (2021), and Shad et al. (2020) provide comprehensive insights into the current techniques used in the field. This section summarizes four key review papers that outline the state-of-the-art methods in EEG artifact removal [5, 9, 31, 33].These papers highlight common artifacts in EEG recordings, evaluate various artifact removal strategies, and discuss the strengths and limitations of each approach. By synthesizing insights from these reviews, this section provides a comprehensive understanding of the current landscape and identifies potential directions for future research.

The review by Jiang, Bian, and Tian (2019) examines artifact removal techniques for EEG signals, emphasizing the challenges posed by artifacts such as ocular, muscle, cardiac, and extrinsic factors. Their work discusses a range of methods including regression, wavelet transform, blind source separation techniques like Principal Component Analysis (PCA) and Independent Component Analysis (ICA), as well as hybrid approaches that combine these methods for enhanced performance, highlighting the ongoing innovation in this field. They conclude that the choice of method depends on factors such as real-time processing needs, computational cost, and the number of recording channels. [9].

Chen et al. (2019) focus specifically on methods for removing muscle artifacts from EEG recordings induced by involuntary muscle contractions. They discuss traditional filtering methods, linear regression, and various signal decomposition techniques like Empirical Mode Decomposition (EMD) and Singular Spectrum Analysis (SSA), emphasizing the importance of tailored approaches based on the type and intensity of movement artifacts encountered [5].

Roy et al. (2021) categorize features extracted from technical research papers to enhance the accuracy of EEG artifact removal for healthcare applications. Their taxonomy covers methodologies such as Noise-Assisted Ensemble EMD (EEMD), wavelet transform, and BSS techniques including ICA and Canonical Correlation Analysis (CCA), underscoring the effectiveness of combining these algorithms to improve EEG signal quality [31].

Saeidi et al. (2021) systematically review machine learning and deep learning models for decoding EEG signals, with a focus on tasks like mental workload assessment and motor imagery. They highlight the prevalence of Convolutional Neural Networks (CNNs) and Support Vector Machines (SVMs) for classification tasks, alongside preprocessing techniques including regression, BSS methods, and wavelet transform for artifact removal and feature extraction [33].

2.1 Review of Additional Papers

In addition to the key papers discussed above, we reviewed several other studies on EEG artifact removal and dataset cleaning. The findings and insights from these additional papers are summarized in the results section of this paper, providing an extensive overview of the current state of research in this field.

3 METHODS

In this section, we detail the systematic approach used to conduct our review of outlier detection and dataset cleaning methods in EEG research. First, we describe the keywords and search strategy employed to identify relevant literature, ensuring a thorough and targeted search across multiple databases. Next, we outline the selection criteria to filter the identified papers, focusing on relevance, recency, impact, and empirical evidence. Finally, we present the review framework, which categorizes and evaluates the selected papers based on accuracy, robustness, computational efficiency, and applicability to EEG data. This structured methodology ensures a thorough and objective review of current advancements in the field.

3.1 Keywords

We conducted a systematic review using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to identify relevant papers on outlier detection and dataset cleaning in EEG research. This approach ensures a rigorous and exhaustive search strategy, aligning with methodologies used in prior studies. The search encompassed multiple research databases, including Google Scholar, Paperwithcode, and arXiv.

To retrieve relevant literature, we employed the following keywords: *('EEG' AND ('Outlier Detection' OR 'Anomaly Detection' OR 'Data Cleaning' OR 'Noise Reduction' OR 'Artifact Removal'))*. This keyword strategy was designed to capture a wide range of studies focused on EEG data analysis, with a particular emphasis on those utilizing machine learning techniques.

Our search strategy aimed to identify papers most relevant in the context of EEG data analysis. The inclusion criteria focused on studies that discussed methods for outlier detection, anomaly detection, data cleaning, and noise reduction in EEG datasets. Exclusion criteria were applied to filter out studies that did not meet the scope of our review.

Figure 1 visually represents the search process, illustrating the number of papers identified at each step and the number of papers excluded based on predefined inclusion and exclusion criteria. This

systematic approach ensures that the selected papers are highly relevant to our research questions and objectives.

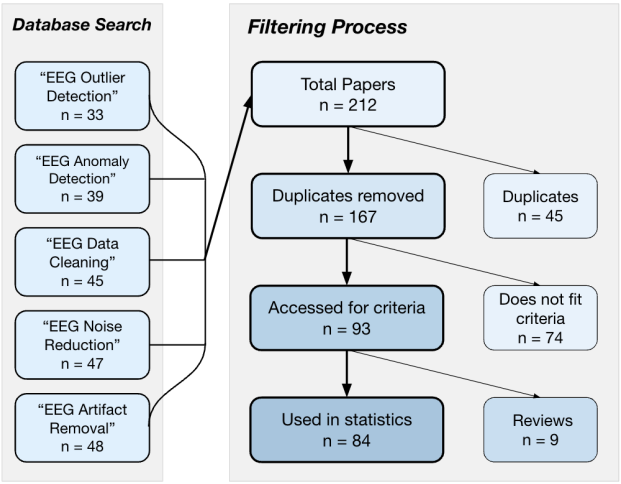

Figure 1: Flowchart of Paper Search and Selection Process

3.2 Paper Selection Criteria

To ensure the relevance and quality of the research papers included in our review, we applied the following selection criteria:

- **Relevance:** Papers must focus on outlier detection and dataset cleaning methods specifically in EEG data. Studies addressing related areas without a clear emphasis on EEG were excluded.
- **Recency:** Only studies published from 2020 and beyond were included, to capture the most recent advancements in the field.
- **Impact:** Preference was given to peer-reviewed articles from reputable journals and conferences, particularly those with significant citations, indicating their influence and recognition within the research community.
- **Empirical Evidence:** Papers must provide empirical evidence demonstrating the effectiveness of the discussed methods. Theoretical papers without empirical validation were excluded.

These criteria ensured that the selected papers were both current and highly relevant, contributing valuable insights to our review of outlier detection and dataset cleaning methods in EEG research.

3.3 Review Framework

The selected papers were analyzed and categorized based on the type of outlier detection and dataset cleaning methods employed. The framework for analysis included the following criteria:

- **Accuracy:** The ability of the method to correctly identify and remove outliers or noise from EEG data, ensuring the integrity of the remaining signal.
- **Robustness:** The method's performance across different EEG datasets and conditions, highlighting its generalizability and reliability.
- **Computational Efficiency:** The time and computational resources required to apply the method to large datasets, such as EEGEyeNet, evaluating its practicality for real-world applications.
- **Applicability:** The suitability of the method specifically for EEG data, considering the unique characteristics and challenges associated with EEG signals, such as low signal-to-noise ratio and non-stationarity.

This framework ensures a comprehensive evaluation of each method, considering multiple dimensions that are crucial for effective outlier detection and data cleaning in EEG research.

4 RESULTS

4.1 Data Cleaning Methods in EEG Data

Data cleaning is a critical aspect of preprocessing EEG data to ensure its reliability and integrity for subsequent analysis in EEGEyeNet. In this section, we present a comprehensive overview of various data cleaning methods employed in EEG preprocessing, aimed at removing artifacts and unwanted signals. We discuss statistical techniques, machine learning algorithms, signal processing methods, and visual inspection, showcasing their respective strengths in addressing different types of noise and artifacts present in EEG recordings. Table 3 illustrates the comparative popularity of these method types. By leveraging a combination of these methods, researchers can enhance the quality of EEG data, thereby enabling more accurate and insightful analyses of brain activity patterns. This section elucidates the significance of robust data-cleaning practices in EEG research, laying the foundation for rigorous and reliable findings in neuroscience studies utilizing EEG data.

Method	Paper Count
Statistical Methods	8
Machine Learning Methods	32
Signal Processing Techniques	40
Visual Inspection	4

Table 3: Most Common EEG Data Cleaning Methods

4.1.1 ***Statistical Methods***. In the realm of anomaly detection, statistical methods play a pivotal role in uncovering irregularities within datasets, often assigning an "anomaly score" to individual instances. Particularly, the Histogram-Based Outlier Detection (HBOS) method utilizes dynamic bin widths in histograms to reveal clusters and anomalies across various feature dimensions [32]. Despite its simplicity, HBOS demonstrates remarkable versatility across diverse data types. Similarly, the Local Outlier Factor (LOF) method evaluates outlier scores by assessing the local density of data points, providing a nuanced perspective on anomalies [32]. The Angle-Based Outlier Detector (ABOD) calculates anomaly scores by measuring the cosine similarity of data points with their neighbors and analyzing the variance of these scores [32].

Additionally, Modified Z-Scores have been employed, in addition to other methods, to remove eye-blink artifacts from EEG signals [45]. In the context of EEG artifact detection, these data-driven approaches have proven immensely valuable. Notably, global classifiers such as HBOS and certain variants of LOF exhibit superior

performance, suggesting that EEG artifacts are distinguishable by their global characteristics, which often represent sporadic and task-specific occurrences of uncorrelated noise.

Recent studies have expanded the repertoire of statistical methods by introducing the utilization of power spectrum features alongside null hypothesis testing, normal distribution approximation (NDA), and Poisson distribution approximation (PDA) [42]. Furthermore, novel methodologies for artifact detection and removal in single-channel EEG signals leverage statistical measures such as entropy, kurtosis, skewness, and periodic waveform index, followed by stationary wavelet transform for artifact removal [8].

4.1.2 ***Machine Learning Methods.*** Machine learning (ML) methods play a crucial role in enhancing the quality of Electroencephalogram (EEG) data by automatically removing artifacts and noise. Various ML techniques, including K-means clustering and Adaptive Noise Cancelling, have been employed to automatically remove eye-blink artifacts from EEG signals [45]. Moreover, artificial intelligence (AI) techniques such as recurrent neural networks (RNNs) and artificial neural networks (ANNs) have been effectively utilized for analyzing anomalies in EEG signals [22].

Various convolutional neural network (CNN) implementations have been explored for enhancing EEG signal quality through artifact removal. The 1D-ResCNN model employs a one-dimensional residual convolutional neural network (1D-ResCNN) in an end-to-end manner to directly map noisy EEG signals to clean ones, demonstrating significant improvements in signal-to-noise ratio (SNR) and root mean square error (RMSE) when compared to methods like ICA, FICA, RLS filter, wavelet transform, and DNN [36]. Another approach, the 1D-CNN with hybrid optimization, integrates an improved 1D-CNN with Spider Monkey-based Electric Fish Optimization (SM-EFO), effectively enhancing artifact removal by optimizing CNN parameters, thereby achieving cleaner waveforms and superior performance metrics [17].

Additionally, general CNN architectures incorporating ascending feature dimensions and downsampling techniques have been employed to effectively remove muscle artifacts from EEG data, utilizing multiple layers to capture diverse features and thereby improving the denoising process [41]. Self-supervised learning methods have been developed for anomaly detection in EEG signals, leveraging multi-class classifiers trained on self-labeled EEG data generated through scaling transformations. These methods have been compared with classic anomaly detection techniques such as Support Vector Machines (SVMs) and autoencoders [37].

Additionally, the one-class SVM detector (OCSVM) utilizes SVMs trained on the entire dataset to assign anomaly scores based on the distance from the class boundary; this technique is particularly effective for detecting infrequent outliers [32]. Generative adversarial networks (GANs) have been employed to denoise EEG time series data artifacts. These methods map noisy EEG signals to clean signals in a supervised learning approach [4].

Overall, machine learning techniques are instrumental in cleaning EEG data by removing artifacts and noise, thereby enhancing the reliability and accuracy of EEG-based applications.

4.1.3 ***Signal Processing Techniques.*** Signal processing methods are pivotal for enhancing Electroencephalogram (EEG) data quality by effectively removing artifacts and noise. Blind source separation (BSS), including techniques such as principal component analysis (PCA), independent component analysis (ICA), canonical correlation analysis (CCA), and morphological component analysis (MCA), addresses the challenge of isolating original source signals from mixed signals in EEG data, enhancing data quality by effectively separating artifacts from neural activity [30].

Independent component analysis (ICA) assumes that source signals are independent and non-Gaussian, requiring manual intervention to remove artifacts, particularly ocular artifacts. Canonical correlation analysis (CCA) leverages second-order statistics to automatically separate uncorrelated features, effectively identifying artifacts such as muscle artifacts based on minimal autocorrelation. Principal component analysis (PCA) constructs a mixing matrix using eigenvectors, prioritizing variance to maintain signal orthogonality and independence. Morphological component analysis (MCA) decomposes signals based on their morphological characteristics, offering targeted removal of artifacts with predefined shapes, notably addressing ocular and specific muscle artifacts. These methods exemplify advanced signal processing techniques aimed at enhancing EEG data fidelity by accurately isolating neural signals from unwanted noise and artifacts.

Additionally, other signal processing techniques such as multiple wiener filtering (MWF) and wavelet-enhanced independent component analysis (wICA) have been applied to artifacts identified by ICLabel using the RELAX preprocessing pipeline [3]. A data-driven approach based on the Koopman operator has also been proposed to analyze EEG data dynamics [24]. SEEG data cleaning methods, including common average reference and Laplacian reference, aim to enhance brain-computer interface (BCI) decoding performance [15]. Innovative approaches propose robust distortion measures, including weighted signal-to-noise ratio (WSNR) and weighted correlation coefficient (WCC), aimed at accurately quantifying band-wise distortion introduced during EEG signal denoising [34].

Moreover, techniques like CCA, Empirical Mode Decomposition (EMD), and median filters are combined to eliminate EEG artifacts effectively [35]. Algorithms such as artifact subspace reconstruction (ASR) further contribute to removing artifacts from EEG data [13]. Integration of discrete wavelet transform (DWT) with meta-heuristically optimized thresholding facilitates efficient artifact removal [21]. These diverse techniques collectively ensure the integrity and reliability of EEG data for accurate analysis.

4.1.4 ***Visual Inspection and Manual Correction.*** In addition to automated methods, visual inspection and manual correction are critical for data cleaning. Trained experts meticulously review EEG recordings to identify artifacts, anomalies, and inconsistencies that automated algorithms may overlook. This process enhances the quality of EEG data, complementing automated methods and leading to more accurate and reliable analyses of brain activity patterns.

Specifically, researchers have presented case studies where event-based EEG recordings were scrutinized, and artifacts were precisely identified and removed through visual inspection and manual correction [16], ensuring the accuracy of EEG data analysis.

These methodologies collectively enhance the reliability and accuracy of EEG-based applications, facilitating advanced research

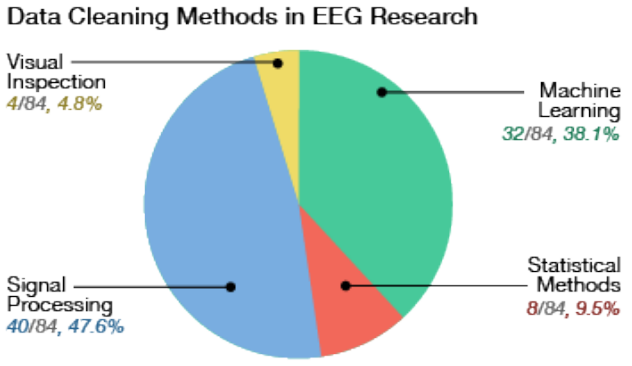

Figure 2: Distribution of EEG Data Cleaning Methods: the Proportions of Statistical Methods, Machine Learning Techniques, Signal Processing Approaches, and Visual Inspection

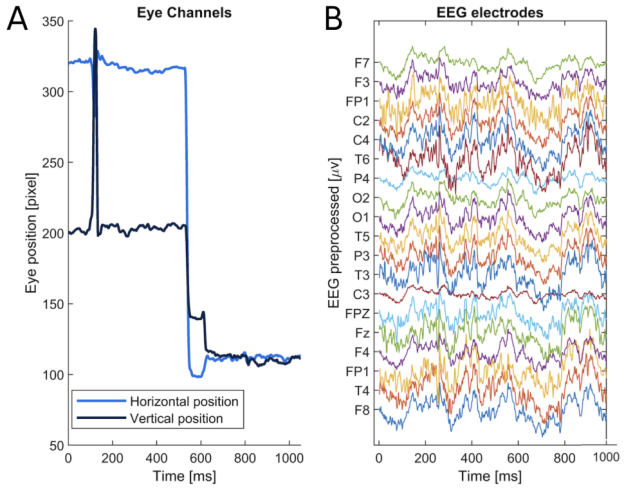

Figure 3: Representation of EEG and Gaze Data: Each EEG data sample is 500 × 128, indicating 500 time points across 128 EEG channels. Panel A depicts gaze data along the XY-axes for a one-second sample, while Panel B shows a subset of preprocessed EEG data using electrodes conforming to the 10–20 system. Adapted from [10].

and clinical practices in neuroscience. As shown in Figure 2, signal processing and machine learning methods are predominant, underscoring their critical role in data-cleaning processes.

4.2 Current Methods in the EEGEyeNet Dataset

4.2.1 ***Background.*** The EEGEyeNet dataset [10] provides a comprehensive collection of EEG and eye-tracking data from 356 participants (190 males and 166 females, aged 18 to 80). The dataset includes three benchmark tasks: Left-right, Angle/Amplitude, and Absolute position. EEG data were recorded using the EEG Geodesic Hydrocel system with 128 channels at 500 Hz, ensuring electrode impedance below 40 kOhm. Eye position was recorded with the ET EyeLink 1000 Plus infrared video system, also at 500 Hz. Figure 3 illustrates a sample of this data. Participants kept their heads stable with a chin rest, positioned 68 cm from a 24-inch monitor. During the large grid task, participants fixated on a series of 25 dots displayed in a pseudo-randomized order for 1.5 to 1.8 seconds each, with 27 dots per block and a total of 810 stimuli per participant.

4.2.2 ***Current Preprocessing Methods.*** The original EEGEyeNet dataset underwent minimal and maximal preprocessing using the toolbox from [20]. Maximal preprocessing involved using Independent Component Analysis (ICA) and a pre-trained classifier, to remove data with a high probability (>0.8) of reflecting external activity [23]. Minimal preprocessing used a 40 Hz high-pass filter and a 0.5 Hz low-pass filter to preserve signal integrity and reduce information loss. When utilizing models on the data, Kastrati et. al. used the minimally preprocessed data as it produced better performance.

Furthermore, performing thorough data cleaning on the minimally preprocessed data for the absolute position task within the large grid paradigm involves excluding all data points with eye positions falling beyond the 800 x 600 pixel screen dimensions. This process leads to the exclusion of 15 samples out of the total 21,464 samples.

5 DISCUSSION

5.1 Comparative Analysis of EEG Cleaning Methods

The effectiveness of various methods for cleaning EEG data can vary based on the nature of the artifacts, the specific research objectives, and the characteristics of the EEG recordings. Statistical methods, including HBOS, LOF, and ABOD, show promise in EEG anomaly detection [32]. While they offer versatility, their effectiveness may depend on data characteristics. ML and DL techniques, such as RNNs and GANs, efficiently remove EEG artifacts [4, 22, 37, 45]. RNNs and ANNs excel in anomaly analysis, while GANs focus on denoising EEG time series. Signal processing methods like MWF, wICA, and ASR enhance EEG data quality [3, 13, 15, 21, 24, 35]. Integration of DWT with thresholding improves artifact removal, and innovative approaches further enhance data quality. Visual inspection and manual correction complement automated methods by identifying nuances not captured by algorithms. While each method category offers distinct advantages, conducting a comparative analysis is crucial for selecting the most appropriate approach for specific EEG data-cleaning tasks. Statistical methods provide valuable insights into outlier detection but may lack sophistication for complex artifact removal. ML techniques offer robustness and scalability but may require extensive training data and computational resources. Signal processing approaches offer versatility and efficiency but may necessitate domain-specific knowledge for effective implementation. Ultimately, the choice of method should align with the study's requirements, considering factors like data complexity, artifact characteristics, and computational resources. Researchers can leverage insights from our review to navigate the

various outlier detection and data-cleaning techniques, advancing EEG research and analysis.

5.2 Assessment of Present Techniques in EEGEyeNet

The current preprocessing methods in EEGEyeNet, including high-pass and low-pass filtering and ICA with a pre-trained classifier, address artifacts effectively; however, their effectiveness depends on rigorous evaluation. Meticulous data cleaning post-preprocessing, tailored to task-specific criteria, ensures data quality.

5.3 Enhancing EEGEyeNet Dataset Quality: Potential Methods

To enhance the EEGEyeNet dataset, various advanced methods can be employed. ML techniques like CNNs and RNNs provide automated feature extraction and temporal dependency analysis, crucial for identifying and correcting EEG signal artifacts. GANs can generate clean EEG signals from artifact-contaminated data. Hybrid approaches, such as combining ICA with regression or integrating Wavelet Transform with BSS, offer improved artifact removal for the multi-channel EEGEyeNet dataset. Signal processing techniques, including EMD variants and Adaptive Filtering like Kalman Filtering, efficiently handle real-time artifacts. Additionally, statistical and data-driven methods like HBOS and ABOD, along with robust measures like WSNR and WCC, can effectively identify and mitigate artifacts, enhancing the dataset's accuracy and reliability.

5.4 Insights and Recommendations

Future research should compare EEG data-cleaning methods using standardized benchmarks to identify the most effective techniques for various artifacts. Cross-domain techniques may inspire novel outlier detection methods in EEG datasets [1, 2, 43, 44]. Hybrid methods combining statistical, machine learning, and signal processing approaches could enhance artifact removal efficiency. Real-time processing methods are crucial for scalability in handling large datasets and providing immediate feedback. Personalized techniques that adapt to individual EEG patterns could improve efficacy and generalizability. We recommend the following five key papers on EEG data-cleaning methodologies for further reading:

(1) **A Deep Convolutional Neural Network Model for Automated Identification of Abnormal EEG Signals [40]** proposes a novel deep one-dimensional CNN model for the automatic recognition of normal and abnormal EEG signals. The model features an end-to-end structure that classifies EEG signals without requiring feature extraction, highlighting the effectiveness of deep learning approaches. This paper is valuable for its demonstration of modern deep learning techniques in EEG signal classification, offering a robust solution for automated analysis.

(2) **Probability Mapping Based Artifact Detection and Removal from Single-Channel EEG Signals for Brain-Computer Interface Applications [8]** presents an innovative method for detecting and removing artifacts in single-channel EEG signals. It utilizes statistical measures such as entropy, kurtosis, skewness, and periodic waveform index to probabilistically map artifacts. Subsequently, the stationary wavelet transform is employed for artifact removal, offering an effective solution for enhancing EEG signal quality in brain-computer interface applications. This paper is valuable because it addresses the challenges of single-channel EEG signal processing, providing a robust technique that combines statistical and wavelet-based methods to improve artifact detection and removal accuracy.

(3) **Identifying Key Factors for Improving ICA-based Decomposition of EEG Data in Mobile and Stationary Experiments [12]** evaluates the impact of movement, the number of channels, and high-pass filter cutoff during preprocessing on the effectiveness of ICA decomposition. It aims to optimize preprocessing for ICA decomposition, considering the specific requirements of mobile and stationary experiments. This research is important because it addresses the variability in EEG data collection conditions, helping to standardize preprocessing methods.

(4) **TMS Combined with EEG: Recommendations and Open Issues for Data Collection and Analysis [7]** examines the integration of transcranial magnetic stimulation (TMS) with EEG for studying cortical reactivity and connectivity. It provides guidelines for standardization, artifact correction, and data analysis to improve reproducibility and promote standard practices in TMS-EEG studies. The value of this paper lies in its comprehensive guidelines, which help to standardize TMS-EEG studies and improve data quality.

(5) **Review of Challenges Associated with the EEG Artifact Removal Methods [18]** addresses algorithm-specific and general challenges in EEG artifact removal. It provides recommendations for overcoming these challenges, reviews Matlab and Python toolboxes for EEG preprocessing, and offers an overview of artifact types and removal methods. This paper is valuable for its comprehensive examination of the challenges in EEG artifact removal and its practical recommendations, which serve as guidelines for selecting appropriate tools and methods for EEG artifact correction.

By building on the insights provided in these papers and the recommendations outlined above, researchers can advance the field of EEG data cleaning, leading to more accurate and reliable EEG analysis and applications.

6 CONCLUSION

Our paper reviewed outlier detection and data cleaning methods in EEG research, identifying and analyzing statistical techniques, machine learning algorithms, signal processing methods, and visual inspection. Each method has unique strengths and limitations, emphasizing the need to select appropriate techniques based on specific research needs and artifact characteristics. Combining these methods can achieve robust data cleaning, ensuring a reliable and accurate EEG foundation for analyses. We suggest that future research could focus on hybrid approaches, real-time optimization, and validation across diverse datasets to enhance EEG data quality and support more accurate neuroscience studies.

REFERENCES

[1] Sizhe An, Ganapati Bhat, Suat Gumussoy, and Umit Ogras. 2023. Transfer learning for human activity recognition using representational analysis of neural networks. *ACM Transactions on Computing for Healthcare* 4, 1 (2023), 1–21.

[2] Sizhe An, Yigit Tuncel, Toygun Basaklar, and Umit Y Ogras. 2023. A survey of embedded machine learning for smart and sustainable healthcare applications. In *Embedded Machine Learning for Cyber-Physical, IoT, and Edge Computing: Use Cases and Emerging Challenges*. Springer, 127–150.

[3] N. W. Bailey and et al. 2022. Introducing RELAX (the Reduction of Electroencephalographic Artifacts): A fully automated pre-processing pipeline for cleaning EEG data-Part 1: Algorithm and Application to Oscillations. *BioRxiv* (2022). 2022-03.

[4] Eoin Brophy et al. 2022. Denoising EEG signals for real-world BCI applications using GANs. *Frontiers in Neuroergonomics* 2 (2022), 805573.

[5] Xun Chen, Xueyuan Xu, Aiping Liu, Soojin Lee, Xiang Chen, Xu Zhang, Martin J McKeown, and Z Jane Wang. 2019. Removal of muscle artifacts from the EEG: A review and recommendations. *IEEE Sensors Journal* 19, 14 (2019), 5353–5368.

[6] Guangyao Dou, Zheng Zhou, and Xiaodong Qu. 2022. Time Majority Voting, a PC-Based EEG Classifier for Non-expert Users. In *HCI International 2022-Late Breaking Papers. Multimodality in Advanced Interaction Environments: 24th International Conference on Human-Computer Interaction, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings*. Springer, 415–428.

[7] Julio C. Hernandez-Pavon et al. 2023. TMS combined with EEG: Recommendations and open issues for data collection and analysis. *Brain Stimulation* 16, 2 (2023), 567–593. https://doi.org/10.1016/j.brs.2023.01.002

[8] Md Kafiul Islam, Parviz Ghorbanzadeh, and Amir Rastegarnia. 2021. Probability mapping based artifact detection and removal from single-channel EEG signals for brain–computer interface applications. *Journal of Neuroscience Methods* 360 (2021), 109249.

[9] Xiao Jiang, Gui-Bin Bian, and Zean Tian. 2019. Removal of artifacts from EEG signals: a review. *Sensors* 19, 5 (2019), 987.

[10] Ard Kastrati, Martyna Beata Płomecka, Damián Pascual, Lukas Wolf, Victor Gillioz, Roger Wattenhofer, and Nicolas Langer. 2021. EEGEyeNet: a simultaneous electroencephalography and eye-tracking dataset and benchmark for eye movement prediction. *arXiv preprint arXiv:2111.05100* (2021).

[11] Matthew L Key, Tural Mehtiyev, and Xiaodong Qu. 2024. Advancing EEG-based gaze prediction using depthwise separable convolution and enhanced preprocessing. In *International Conference on Human-Computer Interaction*. Springer, 3–17.

[12] Marius Klug and Klaus Gramann. 2021. Identifying key factors for improving ICA-based decomposition of EEG data in mobile and stationary experiments. *European Journal of Neuroscience* 54, 12 (2021), 8406–8420. https://doi.org/10.1111/ejn.14992

[13] Velu Prabhakar Kumaravel et al. 2021. Efficient artifact removal from low-density wearable EEG using artifacts subspace reconstruction. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE.

[14] Weigeng Li, Neng Zhou, and Xiaodong Qu. 2024. Enhancing Eye-Tracking Performance Through Multi-task Learning Transformer. In *International Conference on Human-Computer Interaction*. Springer, 31–46.

[15] Shengjie Liu et al. 2021. Investigating data cleaning methods to improve performance of brain–computer interfaces based on stereo-electroencephalography. *Frontiers in Neuroscience* 15 (2021), 725384.

[16] Danyal Mahmood, Humaira Nisar, and Yap Vooi Voon. 2021. Removal of physiological artifacts from electroencephalogram signals: a review and case study. In *2021 IEEE 9th Conference on Systems, Process and Control (ICSPC 2021)*. IEEE.

[17] Battula Tirumala Krishna Mariyadasu Mathe, M. Padmaja. 2021. Intelligent approach for artifacts removal from EEG signal using heuristic-based convolutional neural network. *Biomedical Signal Processing and Control* 70 (2021), 102935. https://doi.org/10.1016/j.bspc.2021.102935

[18] Wajid Mumtaz, Suleman Rasheed, and Alina Irfan. 2021. Review of challenges associated with the EEG artifact removal methods. *Biomedical Signal Processing and Control* 68 (2021), 102741. https://doi.org/10.1016/j.bspc.2021.102741

[19] Nathan Koome Murungi, Michael Vinh Pham, Xufeng Caesar Dai, and Xiaodong Qu. 2023. Empowering Computer Science Students in Electroencephalography (EEG) Analysis: A Review of Machine Learning Algorithms for EEG Datasets. *SIGKDD* (2023).

[20] Andreas Pedroni, Amirreza Bahreini, and Nicolas Langer. 2019. Automagic: Standardized preprocessing of big EEG data. *NeuroImage* 200 (2019), 460–473.

[21] Souvik Phadikar, Nidul Sinha, and Rajdeep Ghosh. 2020. Automatic eyeblink artifact removal from EEG signal using wavelet transform with heuristically optimized threshold. *IEEE Journal of Biomedical and Health Informatics* 25, 2 (2020), 475–484.

[22] Dejan Pilcevic et al. 2023. Performance evaluation of metaheuristics-tuned recurrent neural networks for electroencephalography anomaly detection. *Frontiers in Physiology* 14 (2023), 1267011.

[23] Luca Pion-Tonachini, Ken Kreutz-Delgado, and Scott Makeig. 2019. ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage* 198 (2019), 181–197.

[24] Shaodi Qian and Chun-An Chou. 2021. A Koopman-operator-theoretical approach for anomaly recognition and detection of multi-variate EEG system. *Biomedical Signal Processing and Control* 69 (2021), 102911.

[25] Yansheng Qiu, Ziyuan Zhao, Hongdou Yao, Delin Chen, and Zheng Wang. 2023. Modal-aware visual prompting for incomplete multi-modal brain tumor segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 3228–3239.

[26] Xiaodong Qu. 2022. *Time Continuity Voting for Electroencephalography (EEG) Classification*. Ph. D. Dissertation. Brandeis University.

[27] Xiaodong Qu, Peiyan Liu, Zhaonan Li, and Timothy Hickey. 2020. Multi-class time continuity voting for EEG classification. In *Brain Function Assessment in Learning: Second International Conference, BFAL 2020, Heraklion, Crete, Greece, October 9–11, 2020, Proceedings 2*. Springer, 24–33.

[28] Xiaodong Qu, Saran Liukasemsarn, Jingxuan Tu, Amy Higgins, Timothy J Hickey, and Mei-Hua Hall. 2020. Identifying clinically and functionally distinct groups among healthy controls and first episode psychosis patients by clustering on EEG patterns. *Frontiers in psychiatry* (2020), 938.

[29] Xiaodong Qu, Qingtian Mei, Peiyan Liu, and Timothy Hickey. 2020. Using EEG to distinguish between writing and typing for the same cognitive task. In *Brain Function Assessment in Learning: Second International Conference, BFAL 2020, Heraklion, Crete, Greece, October 9–11, 2020, Proceedings 2*. Springer, 66–74.

[30] CR Rashmi and CP Shantala. 2022. EEG artifacts detection and removal techniques for brain computer interface applications: a systematic review. *International Journal of Advanced Technology and Engineering Exploration* 9, 88 (2022), 354–383.

[31] Vandana Roy, Prashant Kumar Shukla, Amit Kumar Gupta, Vikas Goel, Piyush Kumar Shukla, and Shailja Shukla. 2021. Taxonomy on EEG artifacts removal methods, issues, and healthcare applications. *Journal of Organizational and End User Computing (JOEUC)* 33, 1 (2021), 19–46.

[32] Sari Saba-Sadiya, Eric Chantland, Tuka Alhanai, Taosheng Liu, and Mohammad M Ghassemi. 2021. Unsupervised EEG artifact detection and correction. *Frontiers in Digital Health* 2 (2021), 608920.

[33] Maham Saeidi, Waldemar Karwowski, Farzad V Farahani, Krzysztof Fiok, Redha Taiar, PA Hancock, and Awad Al-Juaid. 2021. Neural decoding of EEG signals with machine learning: a systematic review. *Brain Sciences* 11, 11 (2021), 1525.

[34] Manali Saini, Udit Satija, and Madhur Deo Upadhayay. 2020. Wavelet based waveform distortion measures for assessment of denoised EEG quality with reference to noise-free EEG signal. *IEEE Signal Processing Letters* 27 (2020), 1260–1264.

[35] Piyush Kumar Shukla et al. 2023. An advanced EEG motion artifacts eradication algorithm. *Comput. J.* 66, 2 (2023), 429–440.

[36] Weitong Sun, Yuping Su, Xia Wu, and Xiaojun Wu. 2020. A novel end-to-end 1D-ResCNN model to remove artifact from EEG signals. *Neurocomputing* 404 (2020), 108–121.

[37] Junjie Xu, Yaojia Zheng, Yifan Mao, Ruixuan Wang, and Wei-Shi Zheng. 2020. Anomaly detection on electroencephalography with self-supervised learning. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 363–368.

[38] Yiming Yao, Peisheng Qian, Ziyuan Zhao, and Zeng Zeng. 2022. Residual channel attention network for brain glioma segmentation. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2132–2135.

[39] Long Yi and Xiaodong Qu. 2022. Attention-Based CNN Capturing EEG Recording's Average Voltage and Local Change. In *International Conference on Human-Computer Interaction*. Springer, 448–459.

[40] Ö. Yıldırım, U.B. Baloglu, and U.R. Acharya. 2020. A deep convolutional neural network model for automated identification of abnormal EEG signals. *Neural Comput & Applic* 32 (2020), 15857–15868. https://doi.org/10.1007/s00521-018-3889-z

[41] Haoming Zhang, Chen Wei, Mingqi Zhao, Quanying Liu, and Haiyan Wu. 2021. A Novel Convolutional Neural Network Model to Remove Muscle Artifacts from EEG. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2021), 1265–1269. https://doi.org/10.1109/ICASSP39728.2021.9414228

[42] Hao Lan Zhang et al. 2020. EEG self-adjusting data analysis based on optimized sampling for robot control. *Electronics* 9, 6 (2020), 925. https://doi.org/10.3390/electronics9060925

[43] Zhengming Zhang, Renran Tian, and Zhengming Ding. 2023. TrEP: Transformer-based Evidential Prediction for Pedestrian Intention with Uncertainty. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37.

[44] Zhengming Zhang, Renran Tian, Vincent G Duffy, and Lingxi Li. 2024. The comfort of the soft-safety driver alerts: Measurements and evaluation. *International Journal of Human–Computer Interaction* 40, 4 (2024), 904–914.

[45] Salim Çınar. 2021. Design of an automatic hybrid system for removal of eye-blink artifacts from EEG recordings. *Biomedical Signal Processing and Control* 67 (2021), 102543.

APPENDIX

- Yıldırım, Özal, Ulas Baran Baloglu, and U. Rajendra Acharya. "A deep convolutional neural network model for automated identification of abnormal EEG signals." *Neural Computing and Applications* 32, no. 20 (2020): 15857-15868. Available at: https://link.springer.com/article/10.1007/s00521-018-3889-z
- Sheela, Priyalakshmi, and Subha D. Puthankattil. "A hybrid method for artifact removal of visual evoked EEG." *Journal of Neuroscience Methods* 336 (2020): 108638. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0165027020300601
- Chen, Qiang, Yingying Li, and Xiaohui Yuan. "A hybrid method for muscle artifact removal from EEG signals." *Journal of Neuroscience Methods* 353 (2021): 109104. Available at: https://www.sciencedirect.com/science/article/abs/pii/S016502702100039X
- Qian, Shaodi, and Chun-An Chou. "A Koopman-operator-theoretical approach for anomaly recognition and detection of multi-variate EEG system." *Biomedical Signal Processing and Control* 69 (2021): 102911. Available at: https://www.sciencedirect.com/science/article/abs/pii/S1746809421005085
- Stalin, Shalini, et al. "A machine learning-based big EEG data artifact detection and wavelet-based removal: an empirical approach." *Mathematical Problems in Engineering* 2021 (2021): 1-11. Available at: https://www.hindawi.com/journals/mpe/2021/2942808/
- Cataldo, Andrea, et al. "A method for optimizing the artifact subspace reconstruction performance in low-density EEG." *IEEE Sensors Journal* 22, no. 21 (2022): 21257-21265. Available at: https://ieeexplore.ieee.org/abstract/document/9905486
- Yadav, Anchal, and Mahipal Singh Choudhry. "A new approach for ocular artifact removal from EEG signal using EEMD and SCICA." *Cogent Engineering* 7, no. 1 (2020): 1835146. Available at: https://www.tandfonline.com/doi/full/10.1080/23311916.2020.1835146
- Zhang, Haoming, et al. "A novel convolutional neural network model to remove muscle artifacts from EEG." *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021. Available at: https://ieeexplore.ieee.org/abstract/document/9414228
- Sun, Weitong, et al. "A novel end-to-end 1D-ResCNN model to remove artifact from EEG signals." *Neurocomputing* 404 (2020): 108-121. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0925231220305944
- Dora, Matteo, and David Holcman. "Adaptive single-channel EEG artifact removal with applications to clinical monitoring." *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 30 (2022): 286-295. Available at: https://ieeexplore.ieee.org/abstract/document/9694664
- Satpathy, Rudra Bhanu, and G. P. Ramesh. "Advance approach for effective EEG artefacts removal." *Recent Trends and Advances in Artificial Intelligence and Internet of Things* (2020): 267-278. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1111/psyp.13580
- Cline, Christopher C., et al. "Advanced artifact removal for automated TMS-EEG data processing." *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2021. Available at: https://ieeexplore.ieee.org/abstract/document/9441147
- Shukla, Piyush Kumar, et al. "An advanced EEG motion artifacts eradication algorithm." *The Computer Journal* 66, no. 2 (2023): 429-440. Available at: https://academic.oup.com/comjnl/article-abstract/66/2/429/6469155
- Geng, Xiaozhong, et al. "An improved feature extraction algorithms of EEG signals based on motor imagery brain-computer interface." *Alexandria Engineering Journal* 61, no. 6 (2022): 4807-4820. Available at: https://www.sciencedirect.com/science/article/pii/S1110016821007055
- Chen, Guangyuan, et al. "Anomaly detection in EEG signals: a case study on similarity measure." *Computational Intelligence and Neuroscience* 2020 (2020). Available at: https://www.hindawi.com/journals/cin/2020/6925107/
- Tahura, Sharaban, et al. "Anomaly detection in electroencephalography signal using deep learning model." *Proceedings of International Conference on Trends in Computational and Cognitive Engineering: Proceedings of TCCE 2020*. Springer Singapore, 2021. Available at: https://link.springer.com/chapter/10.1007/978-981-33-4673-4_18
- Xu, Junjie, et al. "Anomaly detection on electroencephalography with self-supervised learning." *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2020. Available at: https://ieeexplore.ieee.org/abstract/document/9313163
- Jindal, Komal, Rahul Upadhyay, and Hari Shankar Singh. "Application of hybrid GLCT-PICA de-noising method in automated EEG artifact removal." *Biomedical Signal Processing and Control* 60 (2020): 101977. Available at: https://www.sciencedirect.com/science/article/abs/pii/S1746809420301336
- Bullock, Madeleine, Graeme D. Jackson, and David F. Abbott. "Artifact reduction in simultaneous EEG-fMRI: a systematic review of methods and contemporary usage." *Frontiers in Neurology* 12 (2021): 622719. Available at: https://www.frontiersin.org/journals/neurology/articles/10.3389/fneur.2021.622719/full
- Jamil, Zainab, Afshan Jamil, and Muhammad Majid. "Artifact removal from EEG signals recorded in non-restricted environment." *Biocybernetics and Biomedical Engineering* 41, no. 2 (2021): 503-515. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0208521621000401
- Judith, A. Mary, S. Baghavathi Priya, and Rakesh Kumar Mahendran. "Artifact removal from EEG signals using regenerative multi-dimensional singular value decomposition and independent component analysis." *Biomedical Signal Processing and Control* 74 (2022): 103452. Available at: https://www.sciencedirect.com/science/article/abs/pii/S1746809421010491

  The remaining citations for the appendix (the 84 papers used for statistics) can be accessed via our GitHub repository: https://github.com/JadeW7/EEGViT-TCNet-pruned