A comparison of stochastic and recursive approaches for EEG signal denoising

How does a stochastic algorithm (Independent Component Analysis) compare to a recursive algorithm (Empirical Mode Decomposition) in removing artifacts in simulated EEG time series data?

I. Introduction

Electroencephalography (EEG) is a widely used neuroimaging technique that measures electrical brain activity (Biasiucci et al.). EEG data is essentially a time series - a sequence of measurements taken at regular intervals over time. EEG is measured using EEG electrodes, in which electric potentials generated by the brain's neural activity are measured in microvolts (μV)(Chorlian and Cohen).

EEG has proven to be invaluable in various fields, including clinical diagnosis and monitoring of various neurological conditions like epilepsy (Noachtar and Rémi), cognitive neuroscience for identifying patterns of brain activity elicited by stimuli (Sarma and Barma), and brain-computer interfaces to establish a direct communication pathway between the brain and an external device (Pfurtscheller and Neuper). However, EEG recordings are often affected by artifacts, which are unwanted signals that contaminate the acquired data and hinder accurate interpretation and analysis(Sheoran et al.). With this, denoising, the removal of artifacts or noise in data, is vital in the use and analysis of EEG.

.

The evaluation of denoising algorithms lacks standardized protocols and benchmark datasets. Different studies often use different evaluation metrics, noise models, and performance criteria, making it difficult to compare the efficacy of different algorithms directly. The variability in EEG data and the lack of consensus in evaluation methodologies hinder the ability to determine which algorithm is most efficient across various EEG denoising tasks. Therefore, developing effective artifact removal methods is essential to enhance the quality and reliability of EEG signals (Jiang et al.).

In recent years, several algorithms have been proposed for artifact removal in EEG data, among which Independent Component Analysis (ICA) and Empirical Mode Decomposition (EMD) have gained considerable attention - 34% of papers citing EEG denoising utilize ICA while 8% employ EMD(Jiang et al.) ICA is a stochastic algorithm meaning that the procedure incorporates randomness and thus the behavior and outcome of the algorithm may vary across different runs, even with the same input. ICA aims to separate statistically independent sources from mixed observations and has been widely applied in EEG artifact removal (Delorme and Makeig; Hyvärinen and Oja). ICA has been proven effective and flexible in denoising under the following premises: statistical independence of the source signals, instantaneous mixing of the sources(Jiang et al.), the dimensionality of the observed signal being greater than or equal to the dimensionality of the source signals(Schlögl et al.; Jung et al.), and the non-Gaussian nature of the sources, or alternatively, having only the sources be Gaussian(Jiang et al.).

On the other hand, EMD is an iterative algorithm employed to break down a signal into intrinsic mode functions (IMF), which capture distinct oscillatory components. It operates recursively, meaning it repeats the decomposition process multiple times to extract these components (Sweeney-Reed et al.). For application to a few channels, EMD is an ideal choice although it suffers from the drawback of IMFs no longer representing distinct oscillatory components of the original signal but instead being a mixture of multiple oscillations(Jiang et al.).

This manuscript aims to compare the performance of two denoising algorithms, Independent Component Analysis (ICA) and Empirical Mode Decomposition (EMD), in removing artifacts from simulated EEG time series data. The research is limited to ocular artifacts, which are unwanted electric potential during eye blinks.

The research assesses how well each algorithm can separate the desired EEG signal from ocular artifacts. Each algorithm will be evaluated based on computational time, memory usage, Mean Squared Error (MSE). The metrics chosen allow for a comprehensive understanding of the trade-offs and advantages of each algorithm in terms of denoising quality and use of computational resources. I will use simulated EEG data in this research for controlled experimentation. Simulated data allows for a systematic evaluation of the algorithms' performance, providing a known ground truth for artifact presence and characteristics. Simulated EEG data also offers the advantage of reproducibility, as the experiments can be replicated consistently, ensuring consistent experimental conditions for evaluating the algorithms' effectiveness. In summary, this research aims to compare the performance of ICA and EMD algorithms in removing artifacts from simulated EEG time series data.

II. Materials & Methods

2.1 Empirical Mode Decomposition

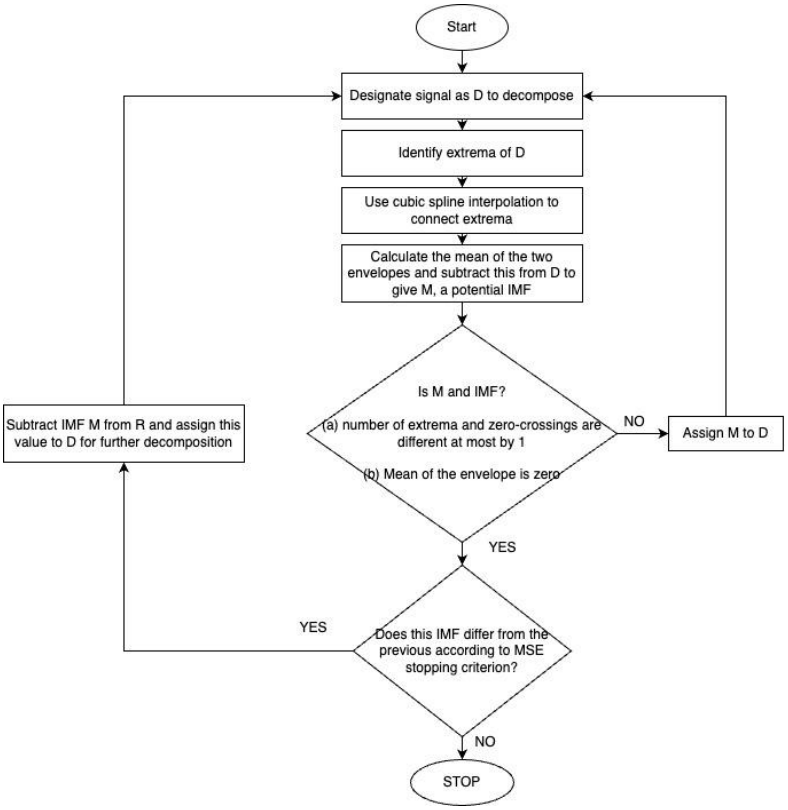

Figure 1. Flowchart of the Empirical Mode Decomposition Algorithm

Empirical Mode Decomposition (EMD) is a data analysis technique used for extracting intrinsic modes or underlying components from a given time series or signal. It was introduced by Norden Huang in the late 1990s as part of the Hilbert-Huang Transform. EMD is particularly effective for analyzing nonlinear and non-stationary signals, which are often encountered in real-world applications (Huang et al.)

At its core, EMD follows the notion that a signal is merely an aggregate of fast oscillations superimposed to slow oscillations. EMD is a data-driven, adaptive method that decomposes a signal into a set of oscillatory components called intrinsic mode functions (IMFs). These IMFs are assumed to have a well-behaved envelope and can capture the different scales or modes present in the data.

A zero crossing refers to a point in a signal where the signal changes its direction from positive to negative or vice versa. In a time series or waveform, zero crossings are significant because they indicate transitions in the signal's behavior. With this established, it’s important to note that zero crossings play a vital role in determining IMFs to be stored throughout the EMD process. Specifically, IMF candidates must satisfy two conditions: (1) in the whole data set, the number of extrema and the number of zero crossings must either equal or differ at most by one; and (2) at any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero.

As seen in Figure 1, EMD is a recursive algorithm, which means it operates in a repeated cycle of sifting and decomposition, where each cycle refines the IMFs by extracting one IMF at a time. Specifically in each iteration, EMD finds all the local maxima and minima in the current signal as these points define the envelope of the current IMF candidate. Together, the upper and lower envelopes define a smooth envelope that encapsulates the underlying oscillatory patterns of the signal. Then, it uses cubic spline interpolation to construct the upper and lower envelopes by connecting the identified extrema. The mean of the upper and lower envelope is calculated to establish a mean envelope. The mean envelope is then subtracted from the original signal to obtain the IMF candidate. If the extracted IMF candidate satisfies the two conditions previously mentioned, it is stored as an IMF and is subtracted from the current signal to calculate the residue or for the algorithm to further its iteration and obtain more IMFs. Otherwise, if the conditions are not met, the current IMF candidate is set as the new signal and the sifting process is repeated. If the residue satisfies a stopping criterion (ex. number of zero crossings, amplitude, percentage of signal captured), the decomposition process ends.

The recursive nature of EMD affects its performance, memory usage, and how it handles different types of data. In terms of computational complexity, EMD can be computationally expensive due to its recursive nature. As the algorithm iterates over the data multiple times, the number of iterations can be high, especially for complex signals. This can lead to slower processing times, making EMD less suitable for real-time applications or large datasets.

EMD doesn't require a fixed set of basis functions or predefined parameters. Instead, it adapts to the data's characteristics, allowing it to handle a wide range of data patterns (patterns in forms such as oscillations, trends, cycles, or transient behaviors). However, this adaptability comes at the cost of memory usage. The algorithm needs to store intermediate

results and residuals at each iteration, which can lead to significant memory consumption, particularly when dealing with long signals or multiple iterations (Huang et al.)

Although the algorithm mainly operates on arrays or vectors representing the signal at various stages of decomposition, EMD does not rely on specific data structures beyond the input signal and the extracted components.

2.2 Independent Component Analysis

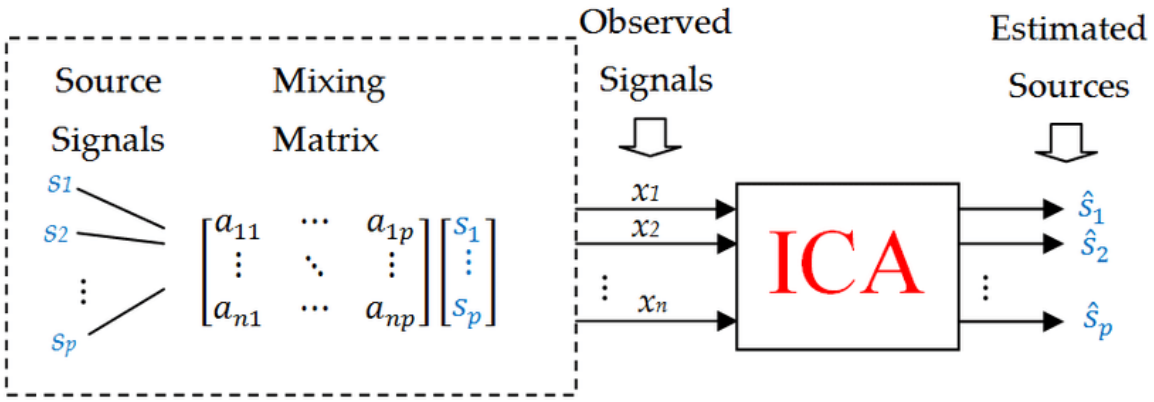

Figure 2. Block Diagram of the Independent Component Analysis Algorithm

In the context of algorithms, "stochastic" refers to a process or method that involves randomness or probabilistic elements. Independent Component Analysis (ICA) is a stochastic algorithm.

ICA is mainly considered a statistical technique used to separate a multivariate signal into additive, statistically independent components. It's particularly useful in scenarios where the observed signals are assumed to be mixtures of unknown source signals (such as artifact-contaminated EEG), and the goal is to recover those original sources (denoising).

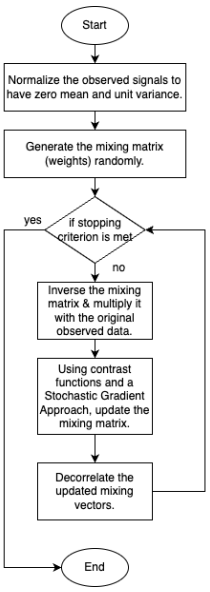

Figure 3. Flow Chart of the Independent Component Analysis Algorithm

With reference to Figure 2, the mixing matrix is a mathematical representation of how much each original brain activity from various brain regions contributes to the signals we actually measure on the scalp. The procedure typically begins with preprocessing the observed signals, ensuring they have zero mean. Subsequently, the main iteration loop is initiated. In each iteration, ICA estimates the sources by updating a mixing matrix, which captures the linear combinations responsible for the observed mixtures. The key concept is to find signals that are not correlated in any linear fashion, as these are the most informative and meaningful sources. To achieve this, ICA employs mathematical tools known as contrast functions. These functions quantify how different the distribution of a given signal is from a Gaussian (normal) distribution. Since most natural signals are not Gaussian, ICA seeks to maximize the non-Gaussianity of the projected data through these contrast functions. In other words, it looks for signals that deviate significantly from the patterns expected in Gaussian distributions. During the loop, the mixing matrix is iteratively adjusted to improve the estimation of independent sources. Stopping criteria such as convergence of the sources or a predefined number of iterations are used to halt the process. The resulting estimated sources are then further processed for analysis. This process unveils underlying components that were mixed in the observed signals, enabling a deeper understanding of the complex interactions within the data.

One component (source) is estimated and randomly updated at a time, while the other components are held fixed. This isolation allows the algorithm to concentrate its efforts on refining a single source without being influenced by other components. This lead to a faster convergence — the point at which an iterative process has reached a state where further updates might lead to negligible changes in the outcome.

Of course however, ICA has its challenges brought upon by its stochastic nature. In essence, due to the randomness introduced in every mixing matrix update, there can be variance in the estimates of the sources and mixing matrix.

2.3 Metrics

The algorithms will be compared using Normalized Mean-Squared Error (NMSE), time complexity, and space complexity on the same computer system.

Time complexity measures the computational efficiency of an algorithm in terms of the time required to execute the algorithm as a function of the input size. It provides insights into the algorithm's scalability and efficiency. Comparing the time complexity of ICA and EMD allows us to assess their computational efficiency in denoising EEG data. We can analyze how the execution time of the algorithms scales with increasing data size, which is valuable for selecting the most efficient algorithm for real-world applications.

Space complexity refers to the memory or storage requirements of an algorithm as a function of the input size. It helps assess the algorithm's efficiency and scalability, particularly in resource-constrained environments, and find a balance between computational performance and memory usage.

The performance of ICA and EMD will also be evaluated by computing the following Normalized Mean-Squared Error (NMSE):

Lastly, the performance of ICA and EMD will also be evaluated by computing the following Normalized Mean-Squared Error (NMSE):

$$NMSE = \sum_{n=1}^{N} \left(\frac{\Sigma_{l=1}^{L} \Sigma_{k=1}^{K} (x_n^{(e)}[k] - \hat{x}_n^{(e,l)}[k])^2}{L \Sigma_{k=1}^{K} x_n^{(e)}[k]^2} \right) \quad (1)$$

where $\{x_n^{(e)}[k]\}$ is the n-th sample or signal, $\{\hat{x}_n^{(e,l)}[k]\}$ is the reconstructed surface EEG after denoising from the l-th run, L is the number of Monte Carlo runs (essentially repetitions of the experiment to account for randomness or variability), K is the data length, and N is the number of electrodes (in this case, 32).

2.4 Simuated EEG Dataset

To compare the effectiveness of stochastic and recursive algorithms in removing artifacts from EEG time series data, synthetic contaminated data is used. This approach offers controlled experimentation with varying Signal-to-Noise ratios. A known clean data ground truth enables objective algorithm evaluation, measuring Signal-to-Noise Ratio improvement. Synthetic data simplifies comparison without ethical or recruitment complexities.

This study employs EEGDenoiseNet, a benchmark EEG dataset for training and testing Deep Learning denoising models. EEGDenoiseNet comprises quantitative time series data from public repositories, provided in platform-independent .npy format. It contains clean EEG segments, ocular, and muscular artifact segments, enabling the synthesis of diverse artifact-contaminated EEG segments. The dataset contains 4514 clean EEG segments, 3400 ocular artifact segments and 5598 muscular artifact segments,

allowing users to synthesize up to 15,347,600 unique ocular-artifact-contaminated and up to 25,269,372 unique myogenic-artifact-contaminated EEG segments with the ground-truth clean EEG. The dataset underwent preprocessing and standardization, resulting in 2-second epochs per sample.

Equation (2) allows the creation of contaminated signals by linearly combining pure EEG segments with EOG or EMG artifact segments.

$$y = x + \lambda \cdot n, \tag{2}$$

In this context, y represents the combined one-dimensional signal of EEG and artifacts, while x represents the clean EEG signal considered as the reference or ground truth. The term n denotes ocular or myogenic artifacts. The hyperparameter λ is used to regulate the signal-to-noise ratio (SNR) in the contaminated EEG signal y. To adjust the SNR of the contaminated segment, the parameter λ can be modified following equation (3).

$$SNR = 10log\frac{RMS(x)}{RMS(\lambda \times n)}, \tag{3}$$

in which the root mean squared (RMS) value is defined as equation (4):

$$RMS(g) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} g_i^2} \tag{4}$$

Where N denotes the number of temporal samples in the segment g and g_i denotes the i^{th} sample of a segment g. Notably, lower λ represents higher SNR, as less EOG or EMG artifacts are added in the EEG signal. In return, lower SNR means higher noise level.

Finally, Figure 4 below shows a sample of a synthetically generated EOG-artifact-contaminated EEG signal.

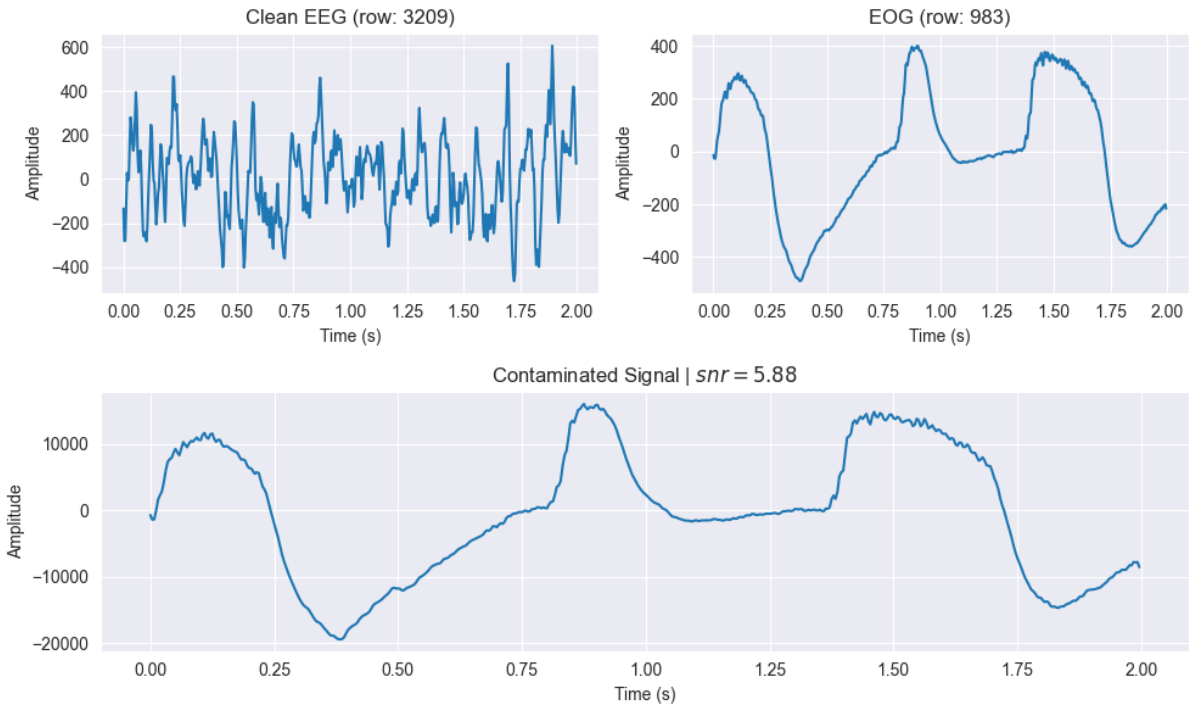

Figure 4. Plot of a sample of a contaminated signal

2.5 Experimental Setup

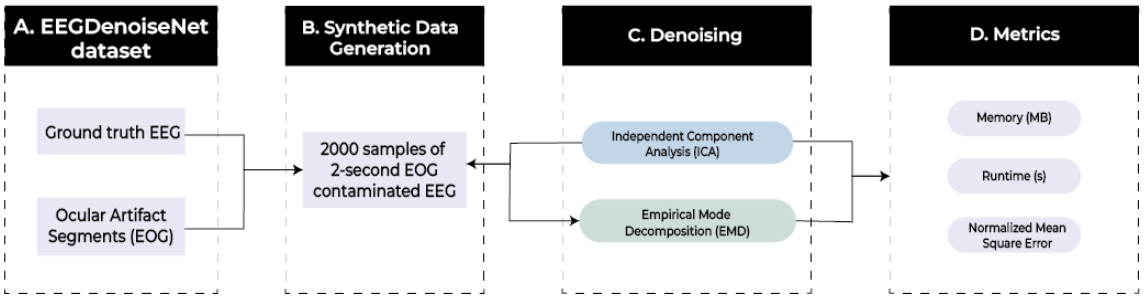

Figure 5. Methodology

It is important to mention that previous research indicates that the Signal-to-Noise Ratio (SNR) of EEG contaminated by ocular artifacts typically falls within the range of -7 to 2 dB (Wang et al.). With this, 5 sets of 2000 samples of 2-second EOG contaminated EEG will be created for each SNR in the given ocular SNR range.

The contaminated signal dataset is contained in a Python list, a versatile data stricture that can hold elements of different types. Each sample in the list is represented as an array of 512 float values, which represents the EEG measurements for 2 seconds at a sampling rate of 257Hz.

It is important to calculate the memory taken up by the dataset as it may impact the memory requirements of the denoising algorithm during its execution. To calculate the memory usage of the dataset, we need to consider the size of each data type involved. Python float values are represented as 64-bit double-precision values. A single Python float value takes up 8 bytes (64 bits). Each sample contains 512 Python gloat values, so the memory for one sample is 4096 bytes. As there are 2000 samples in the list, the total memory for the contaminated signal dataset alone is 8,192,000 bytes or 8.192 megabytes(MB).

Finally, it is important to control the environment during the conduct of the experiment. Thus, the experiment will be conducted on the same system. The relevant details of the system can be found in the table below. The experiment will be conducted through a Python Notebook on PyCharm, an integrated development environment used for programming in Python.

Component	**Detail**
CPU	Model: MacBook Air M1 2020 Chip 8-core CPU with 4 performance cores and 4 efficiency cores 7-core GPU 16-core Neural Engine
Memory	8GB
Storage	256GB SSD

III. Results & Discussion

SNR	MEMORY (MB)		RUNTIME (s)		MSE	
	EMD	ICA	EMD	ICA	EMD	ICA
-7	43.237 ± 1.879	24.568 ± 6.487	478.046 ± 2.616	8.132 ± 3.73	0.0385 ± 0.00229	0.0193 ± 0.00114
-6	43.261 ± 1.737	22.258 ± 4.958	477.235 ± 1.007	6.088 ± 1.401	0.0363 ± 0.00163	0.0182 ± 0.00813
-5	44.235 ± 2.142	26.859 ± 5.82	474.707 ± 3.58	7.133 ± 1.978	0.0348 ± 0.00179	0.0174 ± 0.00897
-4	43.705 ± 1.831	21.593 ± 6.752	475.496 ± 2.31	5.789 ± 1.319	0.0313 ± 0.00150	0.0158 ± 0.00751
-3	43.732 ± 2.149	23.079 ± 5.07	472.818 ± 1.341	7.029 ± 1.596	0.0269 ± 0.00282	0.0134 ± 0.00141
-2	44.058 ± 1.252	19.085 ± 3.485	463.839 ± 11.885	3.923 ± 2.032	0.0237 ± 0.00840	0.0119 ± 0.00424
-1	42.976 ± 2.177	26.883 ± 4.293	454.4289 ± 13.735	4.335 ± 1.052	0.0182 ± 0.0067	0.00927 ± 0.00436
0	45.578 ± 0.185	22.748 ± 4.143	470.478 ± 4.702	6.882 ± 2.053	0.0139 ± 0.00624	0.0694 ± 0.00312
1	44.052 ± 1.489	21.01 ± 5.556	463.201 ± 10.665	5.069 ± 0.872	0.0984 ± 0.00329	0.0492 ± 0.00164
2	45.625 ± 0.236	22.116 ± 0.175	466.585 ± 2.324	5.503 ± 1.734	0.0677 ± 0.00338	0.0338 ± 0.00169

Table 1. Memory, runtime, and MSE performance of EMD and ICA on different SNRs

The experiment involved subjecting a 2000-sample EEG synthentically generated dataset contaminated with ocular artifacts to five iterative trials, each corresponding to distinct SNR values. The collective outcome of these trials was utilized by calculating the mean and standard deviation of memory consumption, computational runtime, and Mean Squared Error (MSE). These values are then used to finally compare EMD and ICA empirically.

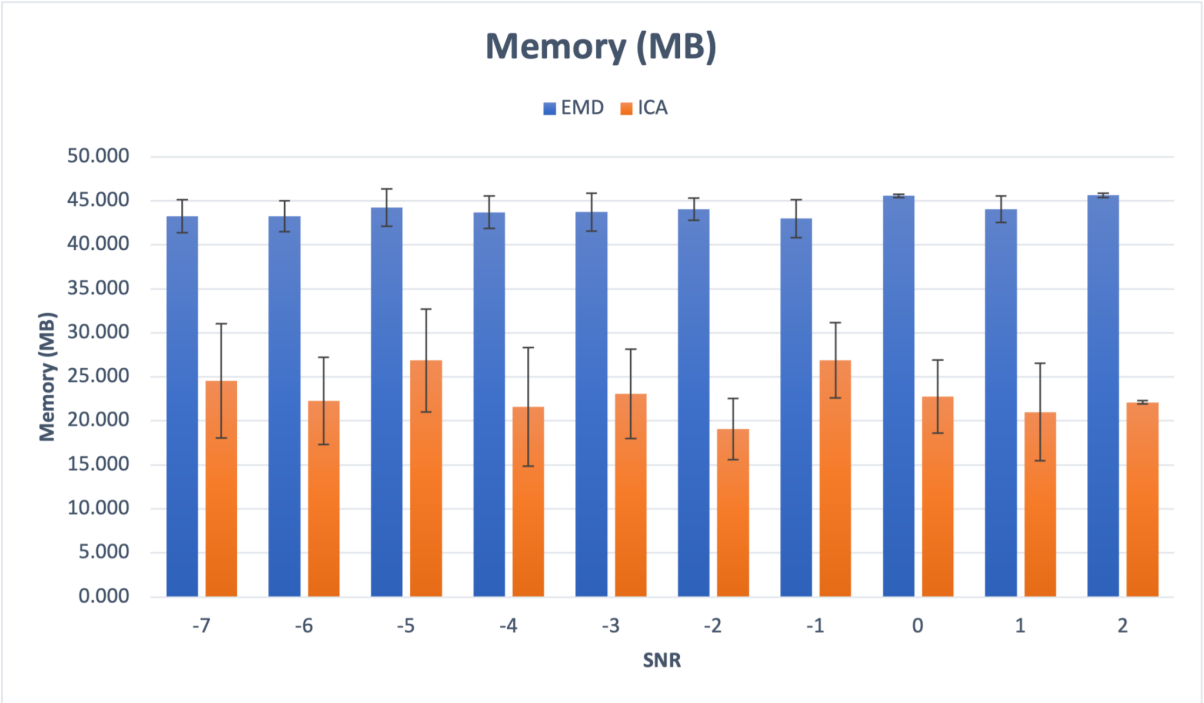

Figure 6. Memory performance in Megabytes of EMD and ICA in different SNRs

As previously mentioned in section 2.5, the dataset used for each trial of testing costs about 8.192MB.

As seen in Table 1, the average memory consumption of the EMD algorithm for a 2000-sample dataset range approximately from 42.976 MB to 45.625 MB. Based on Figure 6, across different SNR levels, EMD demonstrates a somewhat consistent memory usage pattern. This implies that EMD's memory requirements are somewhat resilient to changes in SNR condition, (at least for the given SNR range of -7 to 2). On the other hand, ICA's memory consumption, with values spanning from about 19.085 MB to 26.883 MB, exhibits variations across SNR levels. This suggests that ICA's memory usage is more sensitive to different levels of signal contamination and noise. To add, this more fluctuating memory consumption pattern could be attributed to ICA's stochastic nature. The statistical independence-based extraction process in ICA introduces variability in memory usage as it adapts to different source separation challenges posed by varying SNR levels.

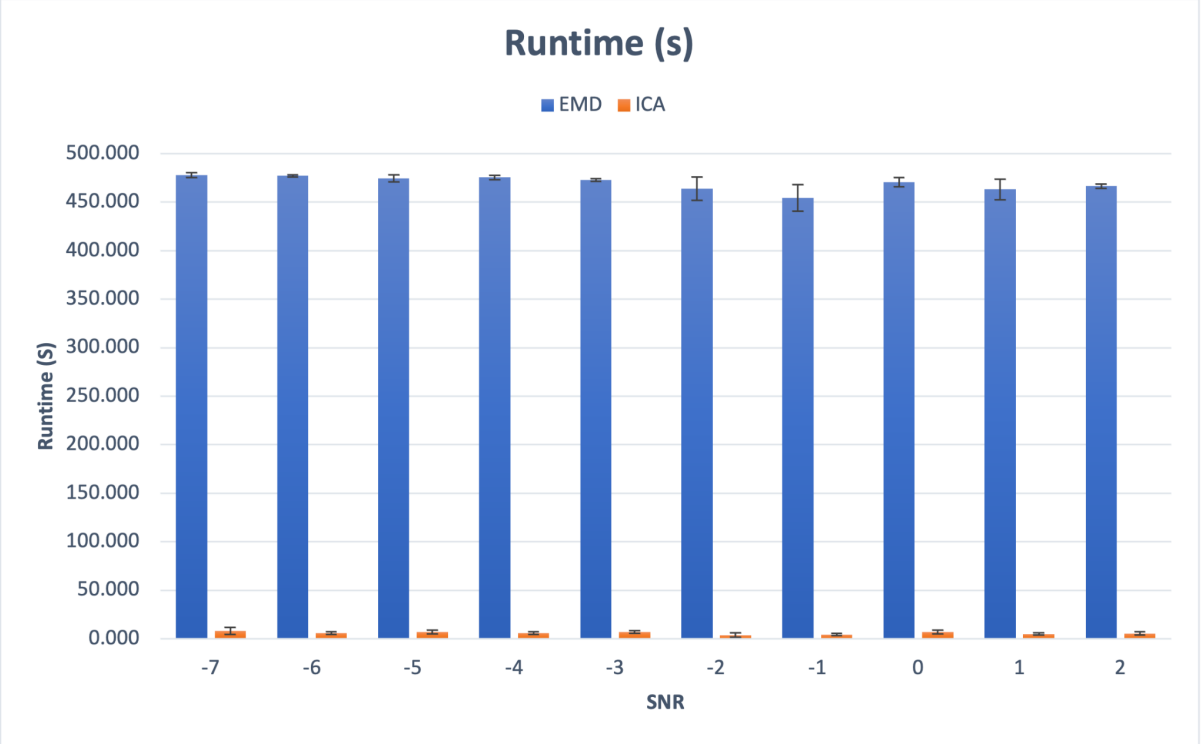

Figure 7. Runtime performance in seconds of EMD and ICA in different SNRs

For the range of SNRs used, EMD took about 454 to 478 seconds to run while ICA took 3.923 to 8.132 seconds. It's also important to note that EMD is about 59 to 118 more times slower than ICA.

Similar to EMD's memory usage previously discussed, its runtime performance can be attributed to its inherently complex and iterative algorithmic nature. EMD involves a recursive decomposition process that iteratively extracts intrinsic mode functions from the

input signal. Each iteration introduces computational overhead, contributing to longer processing times.

For ICA, in each iteration, only a subset of the data is used, reducing the algorithm's memory requirements compared to batch methods like EMD. Another factor that makes ICA memory efficient is that each cycle, after each mixing vector update, decorrelation is performed to maintain orthogonality (or a lack of correlation) between mixing vectors. This step helps preserve memory efficiency by preventing the need to store redundant information.

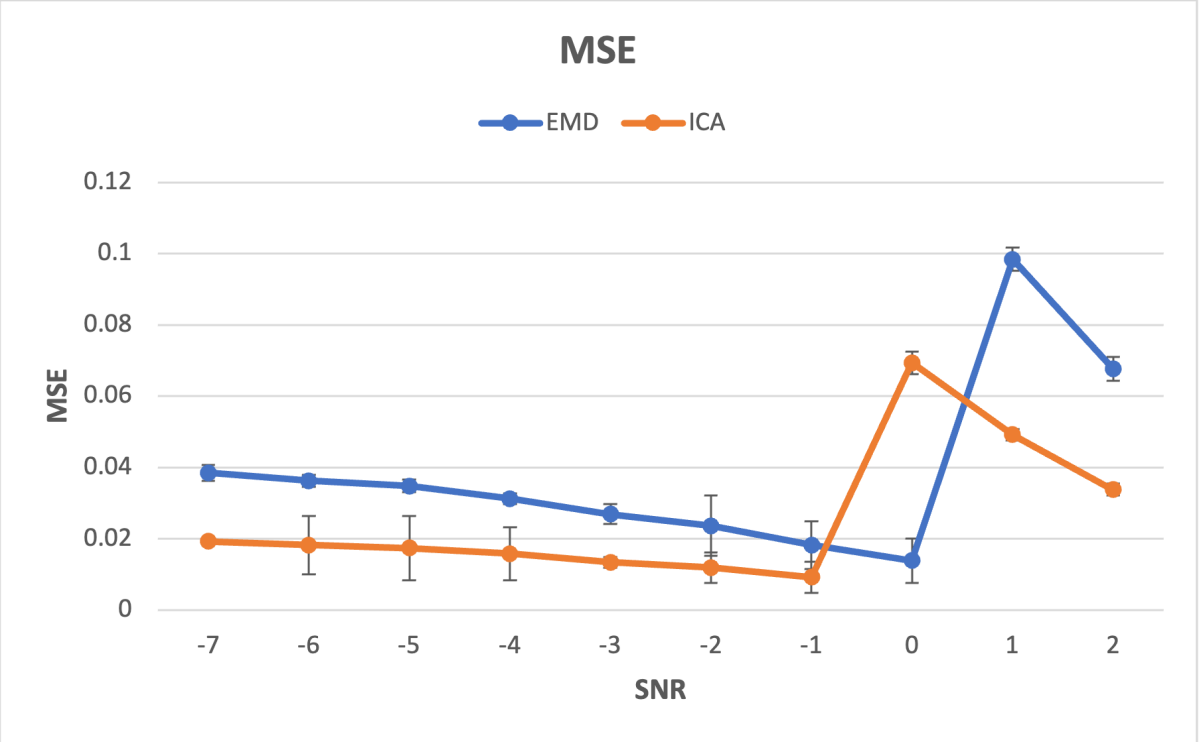

Figure 8. MSE performance of EMD and ICA in different SNRs

For EMD, the MSE values decrease as SNR levels improve, indicating that EMD's denoising performance becomes more effective in lower noise conditions.ICA's MSE values also show a decreasing trend with increasing SNR levels. Similar to EMD, lower noise conditions lead to better denoising performance.

For ICA, a notable observation is the significant increase in MSE at SNR 0, followed by a subsequent decrease at SNR 1 and 2. This pattern suggests that, under very low SNR conditions, ICA's denoising performance might be adversely affected, leading to higher errors. However, as SNR improves slightly in the positive direction, ICA's denoising efficacy recovers, resulting in reduced MSE values.

EMD exhibits a similar trend, but with some differences. At SNR 0, EMD also experiences a substantial increase in MSE, indicating that its denoising performance is

challenged in extremely noisy conditions. Interestingly, at SNR 1, EMD's MSE increases further before decreasing again at SNR 2. This observation suggests a non-linear relationship between EMD's performance and SNR levels.

The observed behavior at SNR 0 underscores the limitations of both methods in handling extremely noisy data. Both EMD and ICA seem to struggle when noise dominates the signal. The subsequent recovery in denoising efficacy at slightly positive SNR levels could be attributed to the methods' adaptive nature. As the SNR improves, both EMD and ICA are better able to differentiate between signal and noise components, resulting in reduced MSE values.

IV. Conclusion

In this study, a comprehensive investigation was conducted to analyze and compare the performance of two distinct denoising algorithms: Independent Component Analysis (ICA) and Empirical Mode Decomposition (EMD). The primary focus was on evaluating their efficiency in mitigating noise in the context of EEG artifact removal and how each algorithm's nature, recursive and stochastic, relate.

The memory consumption results presented trade-offs between memory requirements and algorithmic characteristics of EMD and ICA. ICA's memory consumption variability suggest a potential flexibility in resource allocation but also shows the need for sufficient memory resources in case of fluctuations. Thus for resource-constrained environments, EMD's consistent memory utilization may be advantageous as it provides a more predictable memory footprint.

Considering runtime, the observed significant disparity between EMD and ICA serves as a crucial indicator of the trade-offs inherent in computational efficiency. ICA's faster processing rate may be favored in cases demanding prompt analysis, even with the possibility of minor fluctuations in denoising efficacy due to its stochastic essence.

Upon considering MSE values, ICA performs better than EMD for SNR levels below -1, exhibiting lower MSE values while EMD maintains its lower MSE values for SNR levels above -1. Thus, the choice between Empirical Mode Decomposition (EMD) and Independent Component Analysis (ICA) depends on the specific SNR range.

While synthetic data offered the advantage of controlled conditions, there are inherent differences between synthetic and real EEG signals. Complex physiological and environmental factors shape the characteristics of real EEG data, and therefore, the translation of the research's findings to real-world scenarios is a vital consideration. I proposed that an avenue for further exploration would be the validation of the findings through the application of EMD and ICA to real EEG data contaminated by ocular artifacts.