MACHINE LEARNING APPROACHES FOR PROCESSING AND DECODING ATTENTION MODULATION OF SENSORY REPRESENTATIONS FROM EEG

By

Sari Saba-Sadiya

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Computer Science—Doctor of Philosophy
Psychology—Dual Major

2023

ABSTRACT

This thesis presents novel machine learning algorithms that achieve state-of-the-art performance on a variety of electroencephalography (EEG) tasks, including decoding, classification, and unsupervised / semi-supervised artifact detection and correction. These algorithms are then used within the scope of an EEG experiment that explores how attention to multiple items modulates sensory representations. Using a signal detection paradigm, we demonstrate that attending to multiple items impacts the sensitivity of our participants, causing a sharp increase in false-alarm rates and only slightly decreasing hit-rate. We conclude that our behavioral and EEG decoding results contradict simultaneous attention guidance by multiple items (the multiple item template hypothesis).

ACKNOWLEDGEMENTS

First and foremost, I am eternally grateful to my advisors, Dr Mohammad Ghassemi and Dr Taosheng Liu, for their tutelage, support, and dedication throughout my graduate school journey. Their vast knowledge and creative thinking have been a source of inspiration both inside and outside the lab.

I would also like to thank my committee members and other academic mentors: Dr Susan Ravizza, Dr Pang-Ning Tan, Dr Tuka Alhanai, and Dr Joyce Chai. All of which provided excellent feedback and our collaborations were instrumental to my growth as a researcher.

I acknowledge that many of my better ideas came from conversations with friends and colleagues. Especially Dr Burgoyne, Dr Reynolds, Dr Ming, Dr Masrour, Dr Gao, Dr Mills, Dr Sherry, and Dr Bergan. I would also like to thank current and previous lab mates Eric Chantland, Brendan Valentine, Reza Khan Mohammadi, Shaohua Yang, Allen Williams, Niloufar Eghbali, and Sanaz Hasanzadeh. They all made this journey much more joyful, and I learned how to be a better researcher and human from each of them.

Last but not least, I would like to thank my family members. Those back home - Ahmad, Sylvia, Yara, Abed-Alaziz, and Joud - for their support from the very beginning of my academic journey. And those that became my family during my studies - Emily and Barsha - for more than I can put in words.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Introduction

This dissertation focuses on the two research areas I pursued during my graduate studies at Michigan State University: cognitive attention and deep representation learning. My cognitive neuroscience research utilized electroencephalography (EEG) to explore how attention modulates sensory perceptions. EEG decoding is challenging for many practical engineering reasons, as the signal is a consequence of an ensemble of neural activations from various processes. The difficulties are further exacerbated in the case of attention modulation, which is expressed as a latent variable embedded within the already noisy EEG signal. Successful EEG research requires developing better machine learning techniques to eliminate sources of noise, and more generally, algorithms capable of successfully deriving insights about latent variables from noisy EEG signals.

1.1.1 Electroencephalography in Attention Research

Common sense and science studies alike attest to the importance of attention for task performance: drivers that attend to their phones are twice as likely to experience a car accident than those that attend to the roadway exclusively [168], and eliminating distractions during study can significantly improve academic performance [33]. Controlled behavioral experiments in lab settings have repeatedly demonstrated that attention can improve performance in search, detection, and memory retrieval tasks [173, 186]. However, as will be discussed in section 5.2, understanding the mechanisms responsible for the behavioral effects of attention requires direct access to the latent cognitive processes behind perception. Prior work has demonstrated that EEG signals reflect latent cognitive processes at a high degree of temporal fidelity, and are suited for the study of dynamic cognitive processes such as attention. Within the last two decades, EEG experiments have corroborated many theories proposed based on behavioral results; for instance, attending to an event causes us to process it faster (known as the law of prior entry) [178], and attending to a feature enhances its sensory representation [25].

Despite an ever-growing body of literature, much about cognitive attention remains unknown, and many open debates in attention research stand to benefit from the increasing ubiquity of EEG as a research modality. This dissertation utilized EEG to explore attention modulation of performance in multi-template tasks; if attention improves performance on single-target tasks, what happens when multiple items are attended to at once? In less colloquial terms: Attention to a feature value enhances its sensory representation, but how does attending to multiple feature values modulate sensory representations, if at all?

To successfully decode attention modulation in the EEG experiment multiple machine learning algorithms were developed. These tools enabled more effective data preprocessing, facilitating the study of the latent attention phenomenon with greater fidelity.

1.1.2 Electroencephalography in Machine Learning Research

Electroencephalography devices are unique in being cheap, portable, and non-invasive neuroimaging technology. Moreover, EEG has a variety of applications such as brain-computer interfaces [170, 41, 192], emotion recognition [94, 31], and medical diagnostics [56]. Considering the above, the exponential growth of "machine learning for EEG" literature in recent decades is unsurprising (see figure 1.1). However, successful application of machine learning to EEG tasks remains difficult, as it requires overcoming a number of challenges inherent to EEG data:

- ***Low Signal-to-Noise Ratio***: EEG data is extremely noisy, and any EEG classification or decoding task requires heavy preprocessing and artifact removal [146].
- ***Few Rows, Many Columns***: Traditionally, machine learning has focused on image or text data sets that contain hundreds of thousands of data points, each represented by relatively short vectors. In contrast, EEG data sets often contain only a few dozen subjects, each having only a few hundred trials; at the same time, EEG data is extremely dense, containing thousands of measurements per second of recording.
- ***Data Scarcity:*** One consequence of EEG's low signal-to-noise ratio is that data collected from subjects is often unusable (for instance [119] and [171] both discarded 10% of their subjects due to noise). Moreover, data collection, annotation, and maintenance all require

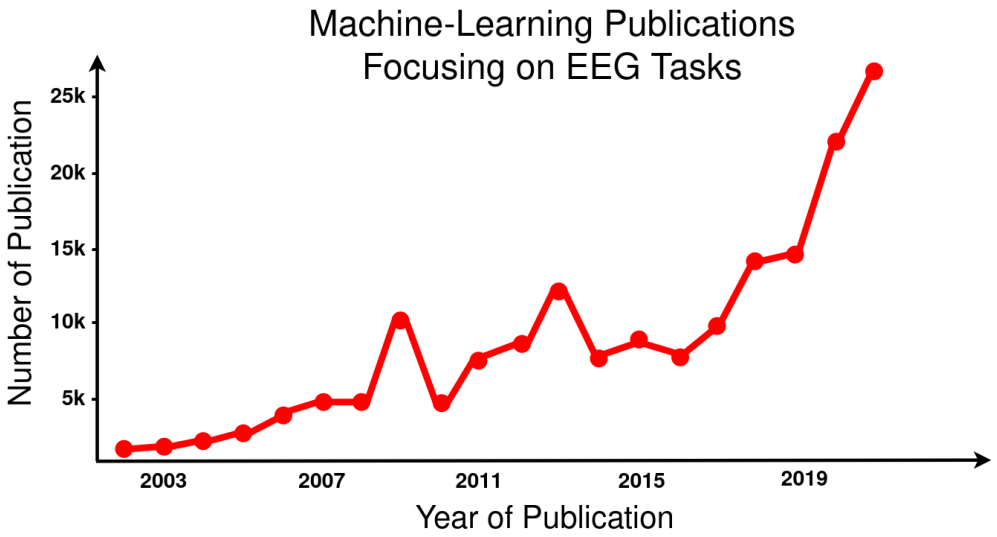

Figure 1.1 Number of publications containing involving machine learning and EEG by year of publication. The figure was produced using data from the dimensions.ai analytics tool.

.

considerable effort and data scarcity affects EEG research reliability. This is reflected by the low sample sizes in Table 2.2.

- ***Temporal and Spatial Covariability***: The EEG signal is global and continuous across space (all signal components appear in multiple electrodes) and time. This high level of covariance between dimensions is relatively unique.
- ***Inter-Subject Variability:*** Research shows consistent individual differences in EEG activity even when performing the same task, under the same circumstances. These differences are so distinct that it has enabled researchers to design EEG based user authentication methods [64]. However, this is an issue when designing any kind of application that seeks to generalize to new subjects; for instance, BCI applications.
- ***Inter-Task Variability:*** EEG classification models are often trained to decode a narrow set of labels. This inhibits the reusability of developed models. As we will discuss in Chapter 2, this weakness is inherent to *discriminative*, as opposed to *generative*, models.
- ***Interpretability:*** While high accuracy might be a priority when designing BCI, this is not the case when the underlying goal is the study of specific cognitive phenomena; for a decoding methodology to be widely adopted it must also be interpretable.

My machine learning work focused on leveraging state-of-the-art representation learning methods to address these weaknesses. Sections 3.1 and 3.2 present algorithms that utilize data properties to learn signal representation that enable unsupervised detection and correction of corrupted data. Section 3.3 presents a novel approach for integrating expert knowledge into neural networks based classification approaches via modular pre-trained sub-networks; that section demonstrates how this approach can mitigate scarcity and inter-task variability concerns while achieving state of the art results on a variety of EEG classification tasks. Finally, it is repeatedly demonstrated that our algorithms can generalize to out-of-training subject data after simple fine-tuning.

1.2 Contributions

This thesis has three main contributions:

- Development of novel unsupervised learning methods for artifcat detection and correction [145, 147, 146]. Facilitating easier preprocessing, and alleviating data scarcity concerns.
- Proposes a novel framework for integrating expert knowledge, and insights from the cognitive neuroscience literature, into neural networks via weight initialization [143].
- Directly evaluate differences in attention modulation of sensory representation across different attentional load conditions. And evaluate how attention load impacts performance using a signal detection theory framework.

1.3 Thesis Outline

Here we provide a brief outline of the upcoming chapters in this thesis. In Chapter 2, we provide a comprehensive review of current machine learning approaches for artifact detection and correction in EEG data, as well as popular EEG decoding algorithms. In Chapter 3, we present novel algorithms that achieve state-of-that-art performance on channel interpolation, unsupervised artifact detection and correction. Section 3.3 also discusses a novel framework for integrating expert knowledge into neural networks that achieves state-of-the-performance on multiple EEG classification tasks. All algorithms are based on representation learning in neural networks, and are designed to generalize for unseen data. Chapter 4 briefly describes a pilot study we conducted as a precursor to our main experiment. Finally, Chapter 5 focuses on cognitive neuroscience themes;

the chapter begins with an introduction to the current debate surrounding the capacity of attention (our ability to attend to multiple items simultaneously), followed by an in-depth literature review. Sections 5.3 to 5.6 present and discuss the main experiment and how our results fit within the previous literature. Section 5.4 also demonstrates how the algorithms presented in Chapter 3 can be useful in a cognitive neuroscience setting. Finally, the last chapter was reserved for general conclusions and reflections.

CHAPTER 2

EEG PREPROCESSING AND DECODING LITERATURE REVIEW

Parts from this chapter are adapted from a published manuscript titled "Artifact Detection and Correction in EEG data: A Review" that appeared in the 2021 proceedings of the 10th International IEEE/EMBS Conference on Neural Engineering (NER) [147]. Other sections were adapted from an unpublished manuscript that was reviewed by Dr Ghassemi and Dr Liu.

2.1 Artifact Correction and Rejection

Electroencephalography (EEG) is a non-invasive, inexpensive, and portable neuro-imaging technology, but the low signal-to-noise ratio of EEG limits its ease of adoption and use for the research and commercial communities alike. The low signal-to-noise ratio of EEG is due, in part, to a variety of artifacts including ocular artifacts from blinks and eye movements, and muscle artifacts from movements. While EEG data is affordable to collect, it is challenging to use in practice because artifacts correction is a necessary prerequisite for meaningful use.

To reduce the human labor associated with EEG experimentation (and the requisite data cleansing) researchers have developed several methods for automated artifact detection. Once an artifact has been detected, the corrupted segment may be discarded but discarding segments introduces discontinuities to the signal that may limit its applications. To circumvent discontinuities, artifact correction techniques may be utilized to "correct" the signal. Implementing effective strategies for artifact detection and correction requires careful review of approaches scattered across the scientific literature. In this review, we highlight the key research contributions in the EEG artifact detection and correction domain over the last 7 years, and identify promising directions for further research and development efforts.

Paper	Artifact	Type	Datasets	Method	Requirements	Performance
[154]	Blinks	D	4256 trials†	ICA	labeled ICA scalp-maps	> 0.80 AUC
[113]	Blinks Muscle	D	47752 trails† 1955 Blinks 4203 Muscle Mov.	Supervised learning algorithms	labeled trials	0.98 F1
[43]	Muscle	R	‡	Hand Crafted EEMD	Uses expert knowledge	.83 F1
[58]	All	R	‡	LDA, SVM, KNN, ICA	labeled trials	< 0.50 F1
[114]	All	D	‡	CNN classifier	labeled trials	0.92 F1
[162]	All	R	2 new datsets real and simulated†	MWF	labeled trials assumes stationarity	6.20 SNR
[2]	Blinks	D	†4 new datasets 2350 blinks	Hand Crafted	Assumes artifact frequency	> 0.94 F1
[57]	Blinks	R	†2000 trials 1000 Blinks	SVM Autoencoder	labeled trials	> 0.98 F1 .024 RMSE
[134]	Blinks, Muscle Heart, Channel	D	†6352 subjects	ICA CNN classifier	Labeled ICA components	0.80 F1 (multi-class)
[17]	Blinks	R	‡2 new dataset simulated and real	ICA with ASR	Labeled ICA components	Downstream tasks
[146]	All	R	†2 new datasets 4578, 4569 trails 628, 570 artifacts	Classical classifiers and Autoencoder	Assumes artifacts are uncommon	Downstream tasks, 0.54 F1
[133]	Blinks	R	EEGLAB dataset with simulated blinks	ICA, SVM and Autoencoder	Uncorrelated signal and noise	0.97 F1 0.04 NMSE
[194]	Blinks Muscle	C	simulated artifacts	Autoencoder	Simulates only specific artifacts	0.56 RRMSE

Table 2.1 Artifact rejection / correction papers being reviewed in chronological order. Type: (D)etection, (C)orrection or (R)emoval. See 2.1.2 for a breakdown of the different metrics. † marks a new dataset. ‡ Data characteristic were not reported by the authors.

2.1.1 Definition of Artifact

For the EEG community, an "artifact" refers to a diverse set of signal distortions that span spatial, frequency and temporal scales [100]. While different taxonomies of artifacts have been proposed [100], the exact distinction between signal and artifact is often dependant on the specific purposes of those collecting the data. For instance, muscle artifacts are unwanted in a motor-imagery Brain Computer Interface (BCI) application, but are useful for tasks such as sleep stage identification [54]. Given the variety of phenomena that could be classified as an artifact for any given EEG use-case, it is not surprising that artifact detection algorithms are narrowly-focused on correcting the intruding artifact in a specific context [154]. Al alternative approach argues that a distortion to an EEG segment is an artifact *if and only if* the distortion negatively impacts the performance of a downstream tasks [146].

2.1.2 Scope of Review

This review includes algorithms for artifact detection and correction using EEG data, *alone*. That is, we do not discuss algorithms that rely on external signals (e.g. electrooculography). Furthermore, we exclude research focused on electrode *'pops'* or other spatially localized artifacts as their unique characteristics enable ease of detection by simple unsupervised and self-supervised techniques [145]. Finally, for the sake of brevity, when a group of papers constitutes a sequence of incremental improvements, we select only the work which presents the accumulation of that line of research [27, 17]. Table 2.1 provides an overview of the literature surveyed in this review.

2.1.2.1 Removal vs. Correction

This review distinguishes between two approaches: artifact *removal* and artifact *correction*. For an algorithm to perform correction (rather than removal) it must have access to an artifact free version of the EEG waveform to be used as ground truth for correcting an artifact ridden version of that same waveform. Note that this necessitates that artifact correction algorithms are trained on datasets with simulated artifacts (for instance see the data-set proposed by [194]).

2.1.2.2 Metrics

The performance of artifact detection algorithms are often measured using manually annotated EEG signals. Common metrics to evaluate artifact detection methods include the F1 score, accuracy, sensitivity, specificity, Area Under the Receiver Operator Curve (AUC), and Cohen's Kappa (inter-rater reliability). For the purpose of comparing performance in this review, we standardized these metrics when possible. For instance, if an author did not report the F1 score, we attempted to derive it from the other metrics [57].

For artifact detection, we compare algorithms using several common performance metrics. We note that not all metrics are equally valid for evaluating EEG artifact detection algorithms. The F1 score and accuracy are appropriate for the assessment of tasks with balanced outcome class labels, which is not common in artifact annotation settings; a classifier graded on an unbalanced dataset may achieve a high accuracy but suffer from a high false negative rate.

Artifact correction algorithms are more challenging to assess compared to detection algorithms as (barring simulated data) the ground truth is unknown. When artifacts are simulated, and access to the artifact free waveform is available, metrics such as normalized mean square error (NMSE) and root mean square error (RMSE) are used [133, 194]. When the data is not simulated the same metrics are calculated using artifact free EEG data collected under similar circumstances (i.e. stimuli and task) [57]. The signal-to-noise ratio (SNR) between clean and noisy EEG post artifact removal is another popular metric [162]. Finally, some researchers use the improvement in downstream task performance as a measure of the reconstruction fidelity; for instance, artifact removal was demonstrated to improve stimuli decoding and visual-evoked potentials recognition [146, 17].

2.1.2.3 Datasets

Table 2.1 lists a summary of investigations conducted for the purpose of developing algorithms for artifact detection and correction. We note that investigators typically evaluate their approaches on data they have collected themselves, as opposed to a standard community benchmark dataset; this highlights a larger issue in the EEG research community around data sharing practices. When

data is shared, it is often to study a particular downstream task, so to facilitate this end, artifacts are often removed which renders the dataset irrelevant for the purpose of artifact detection research. For papers surveyed in this review, only a few made their datasets publicly available [114, 2, 134, 146].

2.1.3 Artifact Detection Methods

Various machine learning and statistical approaches have been applied to the domain of artifact detection. We elaborate on these methods below.

2.1.3.1 Hand Crafted Methods

The *BLINK* algorithm was tailor-made to detect the specific signal characteristics of artifacts caused by eye blinks. Like many hand crafted methods, this approach performs well for the specific task it was engineered to accomplish, but can not be easily extended, tuned, or adapted to detect other types of artifacts [2].

2.1.3.2 Signal Decomposition Methods

Blind source separation methods, most prominently Independent Component Analysis (ICA), treat EEG as a composite signal; ICA decomposes EEG signals into their constituent signal components from which an expert may identify and remove artifact components. While there are rules-of-thumb to distinguish artifact from signal components (for instance, higher power aggregates in frontal areas of scalp maps for blinks), expert annotation is still often required. One notable exception to this is the work of *Shamlo et al.*, who side-stepped the need for an expert annotator by collecting thousands of scalp maps of blink artifacts to contrast new EEG segments against [154].

2.1.3.3 Supervised Approaches

Supervised classification approaches including Support Vector Machines (SVM), Decision Trees, and K-nearest neighbors (KNN) have been applied to a variety of EEG artifact detection problems. Deep learning and Neural Network methods are a relatively recent development in the field of EEG artifact detection. Multiple recent efforts have applied Convolutional Neural Networks (CNN) to EEG by representing data as an $n \times t$ image of n channels and t samples. *Nejedly et*

al. used a CNN in conjunction with fully automated image processing procedures to automatically detect artifacts in intracerebral EEG data [114]. Transfer learning has also be used to improve the performance of network models previously trained on different datasets [114]. Ultimately, supervised classifiers have been shown to effectively discriminate artifact from signal segments [58, 57], but require annotated artifact data to do so, which is not commonly available for many EEG datasets.

2.1.3.4 Unsupervised Approaches

Sadiya et al. proposed a general-purpose artifact detection algorithm [146]; their method extracted 58 different EEG features that are commonly used in EEG research and prognostication, and made the assumption that the frequency of artifacts in the datasets was relatively low. While, this assumption may not always be true (for instance, seizure detection), it is usually valid. The authors benchmarked multiple unsupervised methods. For instance, an auto-encoder was trained to reconstruct EEG waveform segments. Assuming artifact are infrequent, the auto-encoder minimizes the reconstruction error for artifact free trials, hence high reconstruction error is taken as indicative of an outlier EEG segment likely to be an artifact. Their results showed artifact detection rates comparable to the inter-annotator agreement reported in the literature, but as expected, unsupervised algorithms are outperformed by methods tailor-made to detect a given artifact type (Table 2.1).

2.1.3.5 Hybrid Approaches

Hybrid methods that use deep learning classifiers in conjunction with other methods have shown great promise. *ICLabel* is a recently available artifact rejection plugin for EEGLab[1] that uses a CNN to label the components of the ICA decomposed waveform [134]. The classifier distinguishes between seven different artifact types with a binary accuracy (artifact vs signal) of 0.83. Like other ICA based algorithms, ICLabel is capable of online artifact rejection.

2.1.4 Artifact Removal and Correction Methods

Detecting and excluding artifact ridden trails allows researchers access to clean data. However, these trials could constitute a non-trivial portion of the collected data, and rejecting them may

[1] https://github.com/sccn/ICLabel

introduce discontinuities into data that is fundamentally temporal in nature. Recent research efforts have focused on approximating an artifact free version of the affected segment, instead of discarding it all together. It is important to note that all artifact removal methods discussed below are supervised, even when constituting a component of a larger unsupervised pipeline.

2.1.4.1 Signal Decomposition Methods

As previously stated, ICA decomposes EEG signals into their constituent components from which noise components may be identified. A natural extension of the detection algorithms discussed above is to reconstruct the EEG signal from all but the identified noise components. *Gilbert et al.* trained several classifiers (LDA, SVM, KNN) to distinguish between signal and noise independent components [58], and as previously mentioned, [134] trained a CNN classifier to distinguish between noise and signal components. Notably, these methods involve some global loss of information when the signal is reconstructed [133].

Another approach to blind source separation is Artifact Subspace Reconstruction (ASR) which learns statistical characteristics of the components resulting from Principal Component Analysis (PCA). While the performance of ASR and ICA based methods are comparable, the former is faster and less computationally demanding, and is therefore more suitable for online artifact correction [17].

Extended Empirical Mode Decomposition (EEMD) has also been applied to EEG artifact removal [43]. Empirical mode decomposition methods can be used as filters but are not strictly in the same category. EMDs decompose signals into a special class of generating functions that maximizes the signal-to-noise ratio of the reconstruction. While EMDs might appear reminiscent of ICA, the nature of the decomposition is different. ICA decomposes the data for all EEG channels simultaneously, while EMD and the other filtering methods decompose the signal at each channel separately.

2.1.4.2 Filter-based Methods

In signal processing, filters are basic sequence-to-sequence elements that suppress unwanted temporal phenomenon. The Multi-Channel Wiener Filter (MWF) has been used to great effect in

audio and speech processing; Wiener filters use labeled examples to estimate parameters of the signal and noise waveforms such that that noise waveform may be filtered out while the NMSE between a clean signal and its output is minimized. The amount of labeling required to use MWF is minimal and an EEGLab plugin is publicly available [162][2]. MWF assumes stationary of the EEG and noise profiles but to be fair, many simple classifiers make a similar assumption. With sufficient depth, neural encoder-decoder models can learn to correct multiple artifacts drawn from different distributions.

2.1.4.3 Supervised Approaches

Artifact removal with neural networks is a recent development that was been made possible with breakthroughs in sequence-to-sequence modeling tasks using encoder-decoder neural network architectures. Since the ground truth is not usually available, researchers use noisy trials as the input sequence to the encoder-decoder model and artifact free trials as the target sequence [57]. To facilitate work in artifact correction, *EEGdenoiseNet* was recently published as a bench-marked data set of simulated ocular and muscle artifacts [194]. The package provided by the authors allows for the simulation of various artifacts at various signal-to-noise ratios. The authors implemented fully-connect, convolutional, and recurrent neural networks to bench mark the data-set.

2.1.4.4 Unsupervised Approaches

As discussed, section 3.2 proposes an unsupervised approach for artifact detection. Assuming a low false positive rate, trials marked as artifact free are used to train a CNN to reconstruct EEG segments using surrounding samples. The trained network is then used to reconstruct artifact ridden segments. By training with artifact free trials, the method ensures that the reconstructed signal approximates an artifact free signal. While the artifact removal component itself was supervised the pipeline as a whole does not require any labeling (due to the artifact detection being unsupervised). Note that this same approach could be used with any other supervised artifact removal component such as [57, 162]. This approach remains highly limited by the low accuracy of unsupervised artifact detection (Table 2.1).

[2]https://github.com/exporl/mwf-artifact-removal

2.1.4.5 Hybrid Methods

Phadikar et al. suggested a hybrid model that uses SVMs to detect noise components in the ICA deconstructed signals and a denoising autoencoder to remove artifacts from the ICA components rather than the raw EEG [133]. By denoising the ICA components, instead of excluding them from the reconstruction all together, the reconstruction was found to be more accurate.

2.1.5 Conclusion

In this review, we provide a succinct overview of EEG artifact detection and correction methods, with a focus on the last 5 years of research. We reviewed many more papers than formally discussed in this chapter; indeed, there has been an increased interest in artifact detection and removal as EEG devices become more prevalent in multiple fields.

As evident from Table 2.1, the research community is in dire need for a standardized metric, database,and terminology surrounding the EEG artifact detection task, especially if the goal is to produce usable application that will generalize to multiple datasets, and heterogeneous tasks. The more recent entries in Table 2.1 imply a growing popularity of deep learning techniques comes at the expense of traditional approaches and expert knowledge. However, we note that recent papers successfully drew on the rich history and knowledge developed within the EEG preprocessing community to build hybrid approaches that synthesize deep learning, ICA frameworks [133], or features borrowed from EEG prognostication [146]. We believe that hybrid frameworks are an interesting future direction of work in this domain and uniquely situated to combine the strengths of multiple approaches that will advance the current state-of-the-art.

2.2 Decoding EEG Signals

EEG applications span a wide spectrum: from healthcare [36] and accessibility [88, 29] to entertainment [20] and user authentication [64].EEG decoding in particular, namely the decoding of internal cognitive representations has been an of extensive research. Researchers working in cognitive science use EEG decoding to investigate how stimuli is represented and stored in working memory [119, 7, 171, 66]. In contrast, many biomedical applications leverage advanced machine learning techniques to decode user intentions, such as motor-imagery [92, 29, 192], envisioned speech [88], and environmental control via brain-computer interfaces (BCI) [36]. Here we compare the most common EEG decoding approaches, highlighting the particular circumstances that led to the adoption of different approaches across disciplines, and the strengths and weaknesses of each of them. See Table 2.2 for a quick overview of the literature discussed in this section.

2.2.1 Challenges to EEG Decoding

Despite the plethora and variety of research, there remain a number of challenges that limit EEG decoding applications. While some of these challenges are universal, the impact of others is unevenly felt across disciplines. The following are the most relevant challenges for the sake of this review:

- ***Inter-Subject Variability:*** Research shows consistent individual differences in EEG. This has enabled researchers to design EEG based user authentication methods [64]. However, this is an issue when designing any kind of BCI application as pre-trained models might face difficulties when used by out-of-training subjects.
- ***Inter-Task Variability:*** Models are often trained to decode a narrow set of labels. As we will discuss in subsection 2.2.2, this weakness is inherent to *discriminative*, as opposed to *generative*, models.
- ***Data Scarcity:*** EEG data may suffer from an extremely low signal-to-noise ratio that frequently renders data collected from subjects unusable (for instance [119] and [171] both discarded 10% of their subjects due to noise). Moreover, data collection, annotation, and maintenance all require considerable effort and data scarcity effects EEG research reliability.

This is reflected by the low sample sizes in Table 2.2.

- ***Interpretability:*** While high accuracy might be a priority when designing BCI, this is not the case when the underlying goal is the study of specific cognitive phenomena; for a decoding methodology to be widely adopted it must also be interpretable.

Note that these challenges are interrelated; relieving *Inter-Task Variability* for instance can alleviate the *Data Scarcity* issues facing the researcher. Moreover, *Interpretability* might provide insights that can improve transfer learning methods used to combat *Inter-Subject* and *Inter-Task Variability*.

Paper	**Feature**	**Discipline**	**Subjects**	**Method**
[88]	SR	BCI	23	RF
[119]	Cat	P	20	ECOC-SVM
[7]	Ori Loc	P	16	ECOC-SVM
[171]	Loc	P	8	IEM
[66]	Clr Ori	P	30	LDA
[49]	Ori	P	16	IEM
[192]	MI	BCI	9	CNN
[170]	MI	BCI	5,5	DNN
[149]	MI	BCI	10,14	CNN
[94]	EM	AS	24	CNN
[41]	MI	BCI	109	CNN
[199]	MI	BCI	25,9	CNN
[31]	EM	AS	58	CNN
[190]	Ori	P	24	MHL
[184]	Clr	P	34	MHL

Table 2.2 A breakdown of the papers reviewed in this section. Feature: the signal being decoded; orientation (Ori), location (Loc), color (Clr), category (Cat, for instance faces, scenes, tools), Motor-Imagery (MI), Emotion Labeling (EM). Discipline: Perception (P), Brain Computer Interfaces (BCI), Affective Science (AS). Subjects: number of subjects. Algorithms: Mahalanobis distance (MHL), Support Vector Machine (SVM), Inverted Encoding Models (IEM), Linear Discriminant Analysis (LDA).

2.2.2 Review

Our review includes decoding studies using various features (motor intentions as well as item location, color, category) and decoding methodologies. As shown in Table 2.2, the literature is dominated by three decoding methods. First classic methods such as Support Vector Machines (SVM) and Linear Discriminant Analysis (LDA), and Mahalanobis distance (MHL) are still widely used. Second, Inverted Encoding Models (IEMs) have recently become popular among cognitive scientists. Finally, advanced machine learning techniques, and particularly neural networks, are now the state-of-the-art for most BCI based applications. In the following subsections we explore each of these methods and examine their strengths and weaknesses.

2.2.2.1 Classical Classification Algorithms

Support Vector Machines These classifiers were developed by Valdimir Vapnik and his colleagues in a series of papers during the mid 90s. This method quickly peaked in popularity, and by 2001 it has been applied to multiple EEG classification problems [109]. While the engineering community has come to favor neural network approaches for most classification problems, SVMs remain the predominant method in other non-engineering focused disciplines [7, 8]. This is not surprising; SVMs use a relatively low number of parameters, are simple to train without costly computational resources (i.e. GPUs), and can find globally optimal solutions for most problems with the correct kernel selection.

Linear Discriminant Analysis Another classification algorithm that is commonly used with EEG data is Linear Discriminant Analysis (LDA). Given two or more (normal) distributions, LDA finds the projection that maximizes the separation of the clusters. LDA is reminiscent of decomposition algorithms such as PCA. However, while principle component analysis (PCA) solves for a projection that captures the direction of maximum variation in the data set without requiring any labels (thereby projecting the data into a lower dimension in a way that allows for low reconstruction error), LDA maximizes separability between clusters. Mathematically, given two clusters $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$ with means μ_x, μ_y and covariance matrices Σ_x, Σ_y respectively. LDA finds a transformation W that 1) Maximizes the difference between the means of the transformed clusters

$(W\mu_x - W\mu_y)^2$ and 2) Minimizes both within class scatter values. Since $\sigma_x = E[XX^t] - \mu_x\mu_x^t$ after the transformation we get a new scatter matrix $W\sigma_x W^t$. The two terms are combined, and we get the objective function

$$argmax_W \frac{(W\mu_x - W\mu_y)^2}{W\Sigma_x W^t - W\Sigma_y W^t}$$

.

Empirical studies on multiple EEG data-sets have demonstrated that LDA consistently achieves decoding accuracy on par with other classification methods [61]. For this reason LDA was chosen as the default classifier in the ADAM toolbox [44]. Despite implementing a Mahalanobis distance classifier, for reasons that would be discussed in future sections, decoding of data collected during this dissertation was completed using the ADAM LDA classifier.

Multidimensional Scaling Visualization Due to the similarities between LDA and SVM many visualization techniques are applicable to both methods. One such technique is Multidimensional Scaling (MDS): a distance-preserving dimensionality reduction technique that can be used to visualize complex high dimensional data. MDS originated in psychometrics and is still a common tool in cognitive science. For instance, using the distances that SVM and LDA methods provide, it is possible to make deductions regarding the similarity of EEG measurements for different items. By making the reasonable assumption that, for a given subject, EEG pattern similarity correlates with neural representation similarity Hanjonides et al. demonstrated that the color representations follows the color circle in representational space (orange is between green and red) [66]. That is, they showed that the representations being decoded were sensory in nature, not merely the result of a verbal label (the decoded activation wasn't simply the result of subjects repeating the color names in their head). This is an extremely valuable insight for cognitive scientists as it speaks to the organization of neural representation in the brain. Note that such techniques can not be applied used with deep learning black box modules that do not preserve stimuli properties.

This type of visualization highlights the *interpretability* of results obtained using classical methods, which contributes to their persistent popularity amongst cognitive scientists.

Mahalanobis Distance Finally, many EEG decoding papers have used Mahalanobis distance for EEG labeling [190, 184]. The Mahalanobis distance is calculated between a data point and a distribution. Let X_k be a set of all data points labeled k in the training set, and Σ_k and μ_k be the covariance matrix and mean for X_k. Given a data-point from the testing set y the Mahalanobis distance of the point from the distribution of each label is:

$$MHL_k = \sqrt{(y - \mu_k)^t {\Sigma_k}^{-1} (y - \mu_k)}$$

The smaller the Mahalanobis distance from a distribution, the more likely it is that the data point belongs to the same label as the cluster used to generate the distribution. Therefore, for each data point the label that satisfies $argmin_k(MHL_k)$ is assigned as the class. The reason behind the popularity of Mahalanobis distance in EEG research stems from signals being global, hence producing highly correlated electrode readings. By multiplying by the inverse of the covariance matrix we essentially uncorrelate and z-score the data. This is becomes apparent when decomposing the covariance matrix to eigenvalues and vector $\Sigma = v\lambda v^t$ as can be seen in figure 2.1.

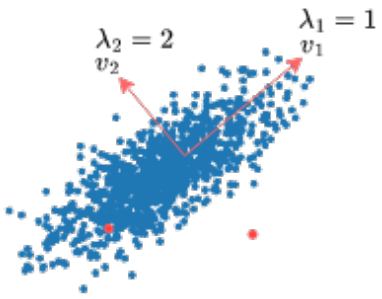

Figure 2.1 The two red data-points have the same euclidean distance from the center of the cluster. By scaling the distance in the direction of the eigenvalue v_k by $\frac{1}{\lambda_k}$ we correct for covariance.

2.2.2.2 Neural Networks

Neural Networks have facilitated massive advances in all area relating to signal processing; this "Deep learning tsunami" [103] did not spare the field of bioinformatics. With respect to the decoding problem, most deep learning applications are seen in BCI applications. This is not surprising as BCI engineers are primary interested in high algorithm performance; statistical

significance alone is typically insufficient. For instance, the decoding accuracy in the Neuroscience paper [119] peaked at 37.5% against a chance level of 33.3% but the BCI study [41] achieved a 79.25% accuracy for a comparable chance level.

Convolutional Neural Networks (CNN) in particular have proven to be extremely capable for EEG decoding. Schirrmeister *et al.* tested multiple CNN architectures on multiple data-sets against traditional decoding methods and found that CNN dramatically outperformed all baselines [149]. The authors suggest that the temporal nature of EEG signals might be especially suited for CNNs, as these neural network *'can capture the temporal hierarchies of local and global [temporal] modulations in the deeper architectures'*.

Transfer Learning A well known weakness of deep learning methods is the need for large amounts of data. In our case this is further exacerbated by the data scarcity issues discussed above. One potential way to combat data scarcity is solving the *Inter-class variability* problem by using pretrained models with transfer learning. For instance [192] used VGG-16, a CNN classifier trained for image classification [195], to improve performance on an EEG motor imagery data-set. Xu *et al.* freezed the first 11 VGG layers and allowed tuning the remaining five. The intuition being that the initial layers have *'extracted low-level universal features ... appropriate for general image processing tasks'*.

Transfer learning can also be used to elevate *Inter-Subject Variability* problems. Emotion detection EEG data is particularly noisy and decoding accuracy is often extremely low. Using a model pre-trained on a subset of subjects can improve the decoding for specific subjects with low accuracy [94]. This was demonstrated as possible even across different data-sets that were collected using different paradigms [31]. A different approach to inter-subject transfer learning is to train a model using a few trials from all available subjects before fine tuning the model for a specific subject. This approach was used by [41] to improve the performance of the models proposed by [149]. Finally, [199] have used CNN with hand crafted features to implement a *'training free'* classifier that can outperform traditional methods such as SVM and LDA for subjects it did not previously encounter.

Interpretability Machine Learning researchers have recently drawn a distinction between *Interpretability*, the ability to associate cause and effect, and the more general attribute of *Explainability* which relates to justifying this association. For example, in image classification, interpretability techniques, such as the Shapely values, reflect the contribution of each input pixel to the final decision model. However, they do not *explain* why the presence of a tail increases the probability of an image being classified as a dog [26]. In the specific context of EEG decoding, *interpretability* is sufficient, as explaining the chain of cause and effect often falls in the realm of cognitive science. Many examples of such "attribution" techniques can be found in the cognitive neuroscience literature [149, 170]. For instance, "Layer-Wise Relevance Propagation" was used in a motor-imagery classification from EEG task [170]. As could be expected, results indicated that activity in the contralateral sensorimotor areas was crucial for the accurate classification of the motor-imager action. The same technique was also used for Alzheimer's disease classification from fMRI data. Not only did the techniques attribute relevancy to areas known to be implicated in the progression of Alzheimer's disease, but the attribution also had high inter-patient variability, enabling the researchers to identify Alzheimer's disease "subtypes" [18].

2.2.2.3 Encoding Models

In contrast to BCI researchers, cognitive scientists prioritize understanding the underlying neural representation over decoder performance. This has led to the development of Encoding Models (EM)s that aim to predict brain responses r given the stimuli s. In other words, EMs model cognitive functions as conditional distributions $P(r|s)$. For instance by characterizing how every fMRI voxel (cluster of neurons) in the early visual areas responds to different spatial frequencies and orientations, researchers were able to identify images by comparing voxel responses with model predicted activations with over 80% precision [112]. Moreover, by inverting the encoding model and calculating $P(s|r)$ the authors were able to reconstruct images from brain activation. Later work expanded on this by constructing entire video segments from fMRI data [118][3]). Another development is the so-called Inverted Encoding Models (IEMs), a specific type of EMs that uses

[3] A demonstration of video reconstruction

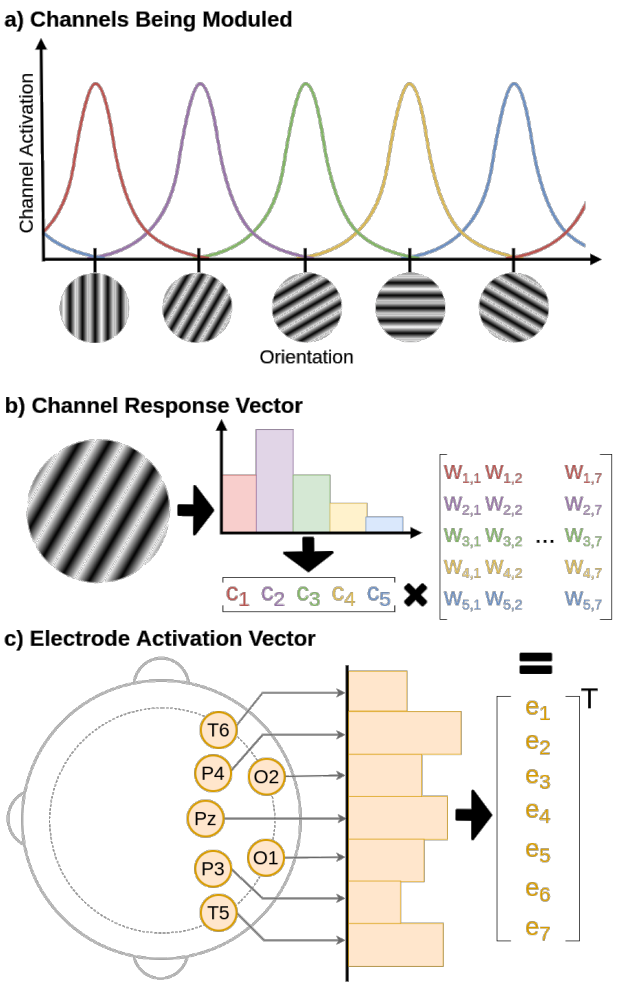

Figure 2.2 The *Inverted Encoding Model*. a) The channel tuning functions. Each channel is sensitive to a different orientation. b) predicted activation for a specific stimuli. c) the electrode values for the stimuli. We can calculate W, and use the inverse transformation to predict channel response from electrode activation, their-by concluding stimuli properties from raw EEG.

"channel responses" instead of stimuli as input. The IEM assumes that the activation measured by each of the m EEG electrodes reflects a weighted sum of n response channels. Each response channel is selective towards a specific stimuli value (Figure 2.2 a). Computationally, the problem is equivalent to solving a system of linear equations $c\dot{W} = e$ where c_n is the vector of the channels' activation for a given stimuli, e_m are the electrode measurements, and the weight matrix, $W_{m\times n}$, describes the contribution of each channel to each electrode. Solving $W_{m\times n}$ on the training data and then applying the inverse operation allows us to infer the channel responses, and thereby stimuli, from EEG data [171, 49]. See figure 2.2 for a visualization of IEM training.

Flexibility Encoding models can be used to reconstruct any stimuli, even if it was not encountered during training. Meaning that, unlike the methods previously discussed in this review, EMs are *generative* rather than *discriminative*. For instance while the EM in figure 2.2 is trained for orientations that are 30 deg apart, we can still predict the channel response for a 45 deg orientation. As noted by Brouwer et al. training an EM is equivalent to creating "a lookup table of channel outputs for an arbitrarily large number of different [stimuli]" [23].

Limitations EMs model the neural representations that underline human perception. Neural mechanisms underlying human perception can be investigated using the performance of EMs that incorporate them. For example, EMs that account for the semantic categories present in the stimuli accurately predict voxel responses at some brain regions but not others, indicating which regions represent such concepts [118]. However, as two recent papers have demonstrated, the reconstructed channel response functions of IEMs are highly contingent on model assumptions [98, 50]. For instance, the decoding would work just as well if instead of assuming that the channel functions are shaped like a normal distribution (Figure 2.2 a), the researcher uses bimodal (or arbitrary) distributions [50]. Proponents of IEMs responded by suggesting that as longs as “sensible” models that follow the current consensus in the research community are used, IEM methods are still useful for intuition regarding the inner working of the cognitive-neural systems [165]. This demonstrates how dependant EMs of the expert knowledge.

Another related limitation that is more specific to IEMs is that they require handcrafted channel responses, but it is not apparent how these can be extended to more complex stimuli categories. For instance [119] used an SVM to decode stimuli category (faces, natural scenes, and tools), it is not clear how one can model a set of channels for such high-dimensional stimuli, or even non-perceptual domains such as motor imagery or semantic categories.

2.2.3 Conclusion

In this review we examined the currently most widely used EEG decoding modalities and the context in which they are utilized. While a unified EEG decoding methodology could be beneficial, by carefully choosing the modalities most appropriate for their use cases researchers seem to be

able to side-step many of their inherent weaknesses. As discussed in subsection 2.2.2, each of the current popular methods has different strengths that counter-act a specific subset of the challenges discussed in 2.2.1. This can be summarized as follows:

- *Classical Methods* require relatively small amount of data and their results can provide insights into the cognitive processes underlying the EEG representations being decoded. These methods are suited to researchers concerned with *Interpretability* and *Data Scarcity* and are therefore most popular amongst non-engineering disciplines focusing on cognitive science research as opposed to building practical EEG applications.
- *Inverted Encoding Modules* A new method that gained popularity in cognitive science. This is a *generative* rather than *discriminative* EEG decoding method as it can decode stimulus values that were not encountered during training. Moreover, as discussed earlier it has been demonstrated that it can be used to decode stimuli with multiple features even if only a single value stimuli was available during training (and vice versa). This method is therefore uniquely flexible as far as *Inter-task Variability* is cornered.
- *Deep Learning Methods* Deep learning methods achieve the highest accuracy, and transfer learning is a promising approach that can alleviate *Inter-subject* and *Inter-task Variability* related issues. However, even with transfer learning these methods remain relatively data-intensive. Considering the above, these are the most popular methods for BCI applications.

Considering the cross-disciplinary challenges of working with EEG data it is more than likely that practices from different fields will eventually find their way into other disciplines. We hope that this review will serve to help the reader consider decoding approaches from new angles.

CHAPTER 3

DEEP LEARNING METHODS FOR PREPROCESSING EEG

3.1 EEG Channel Interpolation Using Deep Encoder-decoder Networks

This section was published as a manuscript titled "EEG Channel Interpolation Using Deep Encoder-decoder Networks" in the proceedings for the 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) [145].

3.1.1 Introduction

Electroencephalography (EEG) devices have become increasingly popular in recent years and are used in a wide range of applications. Naturally, the medical applications of EEG are centered on neurological diagnosis, but EEG has proven useful for other problems in healthcare domain [153, 138]. Moreover, the use of EEG devices extends far beyond the medical domain; novel applications of EEG may be found in wide a variety of fields including advertising [177], education[161], entertainment [20], and security [82].

A fundamental challenge of EEG data is the low signal to noise ratio. Different sources contribute to this noisiness but, in general, they can be categorized as either movement artifacts or electrode artifacts. The most common, and particularly persistent, electrode artifact is the electrode *"pop"* [180, 100]. These artifacts result from abrupt changes in impedance, usually due to a loose electrode or bad conductivity. Furthermore, these artifacts are difficult to avoid because, even if the greatest care is taken when applying electrodes, the most minor subject movement or change in perspiration can cause the electrode to "pop".

A common solution to EEG *"pops"* is to interpolate the missing segments using recordings from nearby electrodes [48]. In practice, this interpolation is most commonly performed using *eeglab*, which contains a tool for spherical interpolation [131, 45]. Within the last few years, alternative interpolation methods reporting improved performance have been proposed: Petrichella *et al.* proposed a euclidean inverse distance method [132] while Courellis *et al.* demonstrated an interpolation approach that (while also being based on an inverse distance calculation) used geodesic

lengths and electrode localization to extract more exact channel locations, thereby performing more accurate interpolation. [35]. Effective interpolation is a necessary preliminary step to any subsequent preprocessing or formal analysis of the EEG, including Independent Component Analysis (ICA). As noted by Ullsperger *et al*: "*activity from bad channels should be removed before ICA decomposition, as it can massively deteriorate otherwise good decomposition results.*" [175]

One shortcoming of these previous solutions is their dependence on knowledge of the precise locations of the electrodes (i.e. electrode localization / registration), which are not collected in most practical settings. Furthermore, the existing methodologies assume that the incidence and specific characteristics of the *"pops"* are similar across both subjects and tasks (i.e. "one-size-fits-all"). Furthermore, as far as the authors are aware, none of the studies surveyed for the purposes of this work provided publicly available software repositories to enable practical use, reproduction of their methodologies, or ease of extension.

To address the aforementioned challenges, we propose a novel electrode interpolation framework using representation learning. Our method autonomously identifies the spatio-temporal properties of EEG data measured at a set of electrodes, that predict the values of a given neighbor to those electrodes. Our model, which has been made publicly available (at `github.com/sari-saba-sadiya/EEG-Channel-Interpolation-Using-Deep-Encoder-Decoder-Networks`), can be used "out of the box" to more effectively interpolate EEG for any missing channel, at any time.

One important advantage of our model over existing approaches is its amenability to transfer learning: the ability to easily fine tune it using clean data from a novel subject or EEG experiment. This property of our model allows for interpolation that is tailor-made to the specific task and subject at hand, enabling the model to learn even idiosyncratic relations in new data.

To determine the usefulness of our method we evaluate the model on unseen tasks and subjects with and without further tuning. To summarize, our main contributions are:

- We propose and implement a new framework for EEG channel interpolation using encoder-decoder deep representation learning.

- We compare our method against contemporary algorithms for channel interpolation.
- We make our code publicly available for the benefit of the community, and demonstrate how it may be tuned to novel subjects and tasks using transfer learning.

3.1.2 Related Work

3.1.2.1 Current Interpolation Solutions

The Spherical interpolation method (as implemented by *eeglab*) was first proposed by Perrin *et al.* in their 1989 work [131] and remains the most widely-used interpolation solution [131, 45, 163, 132]. More recent work describes improved interpolation methods by, for instance, using ellipsoid geodesic lengths [35] to better estimate the distances between electrodes. A weighted signal reconstruction scheme that favors electrodes closer to the location of the missing channel is then used to achieve higher accuracy. This however requires that the electrode positions be digitally registered so that an ellipsis can be fitted to the shape of the subject's head. That ellipsis is then used when calculating the distances between electrodes. Unfortunately, most data tends not to provide the specific channel locations.

3.1.2.2 Artifact Detection Using Neural Networks

EEG is a signal with spatial structure that unfolds in time. Convolutional neural networks (CNNs) are an obvious candidate for the EEG interpolation problem because they naturally capture hierarchical spatio-temporal relationships.

A few recent papers have leveraged CNNs for EEG classification [55, 106]. Previous works have also used CNNs to annotate EEG wave-forms for the presence of artifacts: Nejedly *et. al.* framed artifact detection as a classification problem and used CNNs to robustly annotate eye blink segments [115]. However, while the use of deep learning approaches for artifact *detection* has shown promising results, there is relatively less work on the use of deep learning to *reconstruct* artifact-ridden data segments.

Finally, recently researchers working with EEG data started to use *generative* rather than *discriminative* neural networks. For instance [34] used generative auto-encoders (GANs) to specially

up-sample EEG data by interpolating non-existing channels. While not interchangeable, this is a similar task to channel interpolation. However, no previous work was validated on either subjects or tasks that did not appear during training. Moreover, to the best of the authors' knowledge no previous work tested how transfer learning could facilitate further fine tuning on specific data-sets, or even made the trained model available. We hope that our rigorously tested ready to use model will allow for wider access to state-of-the-art machine learning techniques.

3.1.3 Data

3.1.3.1 Data Collection

The data in this study was collected from the recently published *EEG During Mental Arithmetic Data Set* [201]. The data consists of 24 subjects performing two tasks: a *resting state task*, and a *mental arithmetic task*. EEG data was collected as subjects performed the tasks using the 10-20 international system with the linked ears serving as a reference electrode (see Figure 3.2, left). The sampling rate was $500Hz$. Each *resting state task* lasted 180 seconds while each *mental arithmetic task* lasted for 60 seconds; hence, the total number of samples was $90,000$ and $30,000$ for each *resting state* and *mental arithmetic* task respectively.

We segmented the data into $16ms$ (8 samples) epochs. Hence, each subject ended up with $11,250$ and $3,750$ epochs for the *resting state* and *mental arithmetic* tasks respectively.

3.1.3.2 Data Partitioning

In Figure 3.1, we illustrate how the data was partitioned for the training and evaluation of our method. Training data from the *resting state task* was used for model development while data from the *mental arithmetic task* was held out for intra-task evaluation. We further partitioned data from both tasks into training subjects (67% of subjects) and evaluation subjects (33%). This resulted in the following four partitions of the data, which we will refer to later when discussing our results:"*Seen Task, Seen Subjects*" (n=16 subjects),"*Seen Task, Unseen Subjects*" (n=8), "*Unseen Task, Seen Subjects*" (n=16), and "*Unseen Task, Unseen Subjects*" (n=8).

The "unseen" data sets contain some deviation from the data the model was trained on, and are

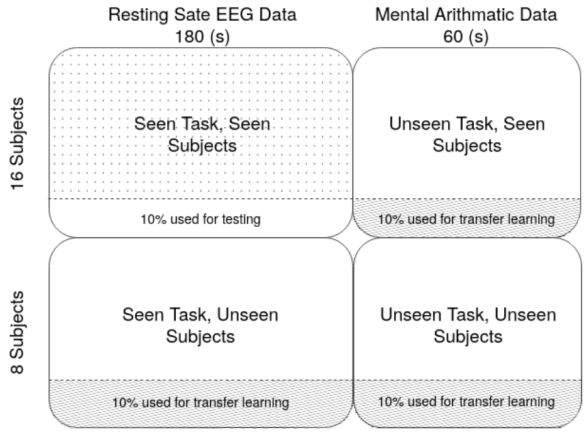

Figure 3.1 The four-way split of the data into partitions. The dotted area (top left) was used to train our main algorithm. The areas with the diagonal shading were used for transfer learning. The areas with white background were used to test our algorithm. Each shaded area was used separately to tune the model for testing on the remaining 90% of the data in its own partition.

thus a good test for the generalizability of the method. Because we are utilizing transfer learning 10% of the data for all unseen sets was held-out for tuning, to test the impact of this additional context on the model's performance.

In general, the fundamental difference between the two tasks is crucial to our evaluation. Researchers will often not have enough data to train neural networks "from scratch", not to mention training a neural network often requires exploration of a hyper-parameter space that may consume significant temporal (and financial) resources. This is especially true for deep learning frameworks that have strong tendencies to over fit and produce remarkable results for a specific data set while failing to generalize to other contexts. With this in mind, it is important that our models achieve good results on both unseen tasks and subjects to enable their continued development and utility within the greater research community. We therefore structured our data to assess its ability to generalize across tasks and subjects.

3.1.4 Methods

3.1.4.1 Pre-processing

To begin, all EEG data were Z-scored at the subject-level (i.e. converted to a zero mean, and unit variance representations). Deep networks require a large volume of training data; hence, we

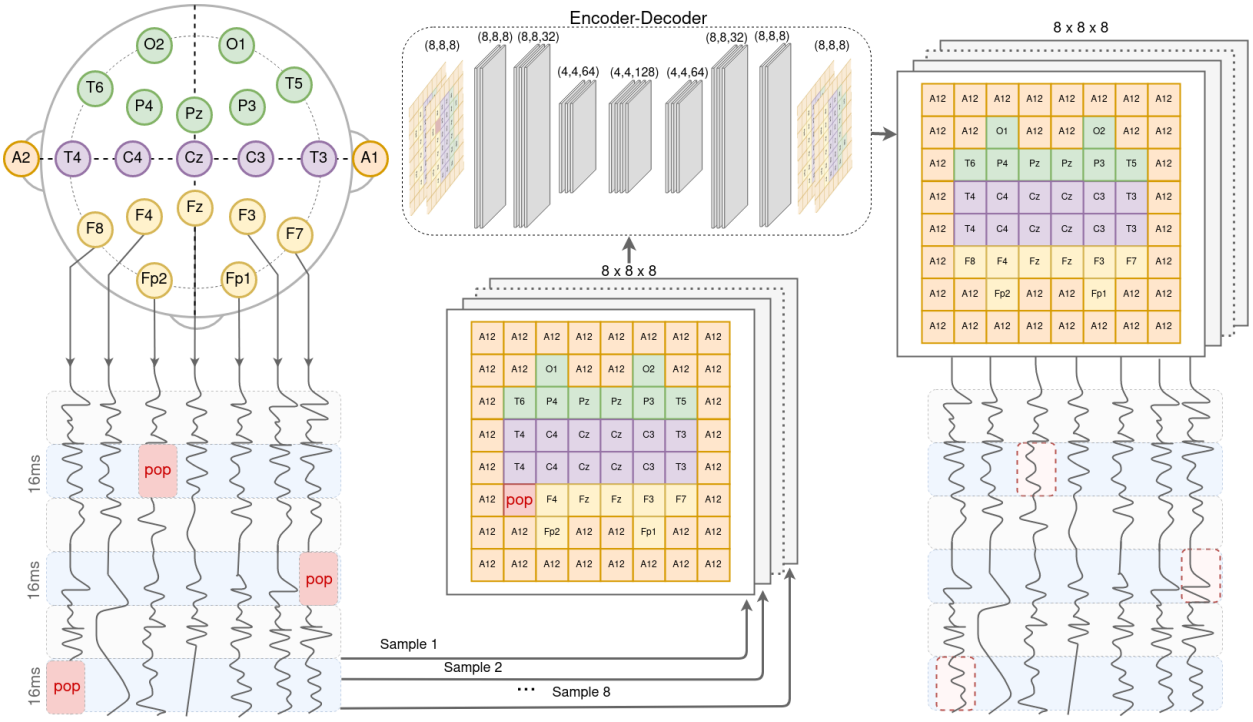

Figure 3.2 A diagram of our framework. The EEG data is first segmented the 500Hz data into 16ms epochs, each segment is then mapped to a 5x5 matrix that roughly reflects the spatial locations of the EEG electrodes (e.g. *F7* is located in position 1,1 of the matrix). The electrodes at the sagittal and median planes were duplicated and the tensor was padded with the linked ear channel data to create a $8 \times 8 \times 8$ tensor per epoch. This data serves as the input to an encoder-decoder model. Finally, the output is transformed back into a signal.

supplemented our training data by transforming each subject's EEG data into 10 distinct *pseudo-subjects*. Each pseudosubject's data was an elementwise addition of the Z-scored subject's EEG data and random draws from a Gaussian distribution with ($\mu = 0$, $\sigma = 0.05$). The pseudosubject's (already normalized) EEG data was then Z-scored again following the introduction of the noise. The utility and validity of this data augmentation approach for EEG research has been demonstrated in prior work [56]. Next, a simple transformation was applied to project each sample of EEG data from a spherical channel representation onto a quantized two dimensional surface represented by a 5×5 matrix (Figure 3.2, panel 2).

Finally, EEG data was epoched into 16ms segments, with no overlap across segments [1]. This resulted in a 5 channels ×5 channels ×8 samples tensor. The electrodes at the sagittal and median planes (the central electrodes) were then duplicated and the tensor was padded with the linked ear

[1] Average pop artifact duration exceeds 1 second, hence 16 ms is more than sufficient for reconstruction.

channel data to create a $8 \times 8 \times 8$ tensor. This manipulation of input size is common place in deep learning and is mainly the result of networks being optimized to work with input sizes that are powers of two [95]. This $8 \times 8 \times 8$ tensor formed a single sample of input data, from the perspective of the network when training.

To create training data, we iterative occluded each of the 19 non reference electrodes (All electrodes except $A1$ or $A2$, see Figure 3.2). Thus each $8 \times 8 \times 8$ tensor became the prediction target for 19 input tensors, each with one distinct occluded channel.

3.1.4.2 Proposed Approach

An Encoder-Decoder model for EEG Interpolation Inspired by research on image inpainting we deployed an encoder-decoder model for EEG interpolation. Image inpainting is a classical problem in computer vision: given a corrupted image the aim is to complete or “fill in" missing pixels. This is a similar problem to electrode interpolation.

Encoder-decoder models are the combinations of two networks that are trained simultaneously: the encoder first learns a lower dimensional embedding of the data, and the decoder attempts to recover the original data from the embedding. Encoder-decoder networks are a popular tool in image inpainting [129, 193, 181], so much so that this technique is now leveraged for image compression as selective removal of pixels might greatly enhance compression ratios [11].

We determined the optimal topological configuration of our encoder-decoder network via a random search of the network hyper-parameter space [15]. The tested topologies varied in the number of convolution layers, the existence of max-pooling, dropout, and batch normalization layers after each convolution layer, and whether the decoder was based on transposed convolution layers or simple up-sampling with convolution. More specifically, we trained 300 distinct architectural configurations of the encoder-decoder networks, and retained the configuration that best generalized within a held out subset of the training data itself. The best network was then used after training for our transfer learning evaluation. The code to run this search in the topological space, as well as the trained winning algorithm before and after the transfer learning tuning is available online. The optimal topology is shown in Figure 3.3, and discussed in Subsubsection 3.1.5.1.

Subject+Task Enhancement via Transfer Learning Transfer learning experiments were carried out by taking the model trained on the original data set and tuning it on a small subset (10%) of the testing data. Realistically, it is highly likely that a small sample of clean data will be available for a researcher to use when tuning our model.

To assess performance enhancements associated with transfer learning, we held out 10% of each data partition (See Figure 3.1) and tuned our network for 100 epochs. These numbers were intentionally small as to showcase how even minimal training that can be easily completed on non specialized hardware and using very little data can lead to significant improvements. By tuning the model for specific subjects and tasks, we assessed the flexibility and practical extensibility of our proposed approach.

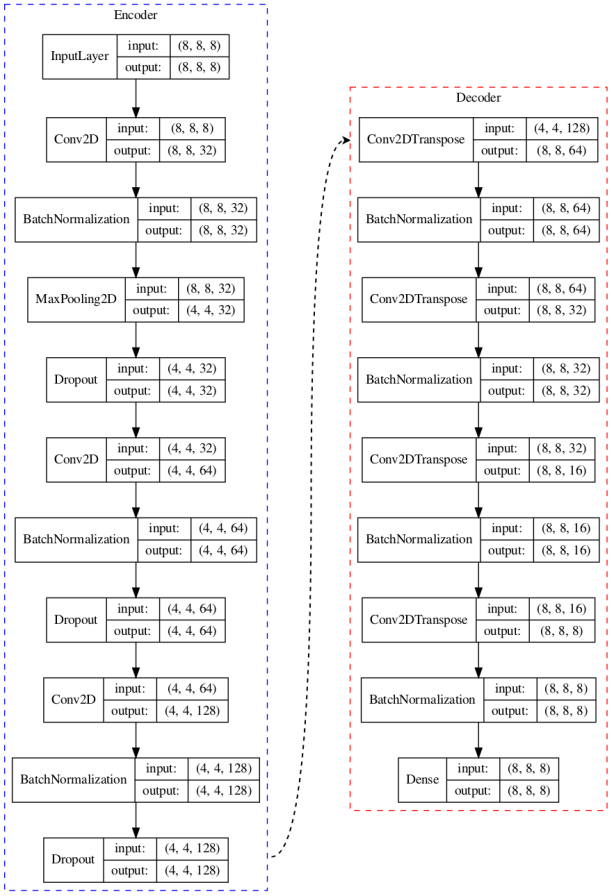

Figure 3.3 Our network, the dashed black arrow denotes the $4 \times 4 \times 128$ embedded data tensor. This embedded representation is passed form to the decoder which reconstructs the original input sans the occlusion.

3.1.4.3 Methodological Baselines

For our baseline we implemented the three methods described in the Related Work subsection [131, 132, 35].

The Euclidean Baseline Both [132, 35] suggest methods that employ an inverse distance metric where the interpolated channel $\hat{s_i}$ is calculated using the following equation:

$$\hat{s_i} = \frac{\sum_{j \neq i} w_{ij} s_j}{\sum_{j \neq i} w_{ij}}, w_{ij} = \frac{1}{d_{ij}^p}$$

where p is the power parameter. The variable d_{ij} represents the distance between electrodes i and j. The original channel j is represented by s_j. The power parameter is an integer (usually between 2 and 5) that is set using a small amount of data; while it is usually set to be the same value across a given data-set where interpolation is happening, we optimized the power parameter separately for each baseline and data-set to maximize the performance of the baselines.

The calculation of the distance, d_{ij}, is the main difference between the two baselines. The fist euclidean baseline (*EUD*) uses a simple euclidean distance formula. This distance calculation is done using the generic electrode positions in space that are always available for every cap.

The Geodesic Baseline The Geodesic Length baseline (*EGL*) is also based on the inverse distance equation. However, instead of using euclidean distances for d_{ij} the geodesic length is calculated. The geodesic distance is calculated using the *Vincenty* algorithm which was originally used in geodesy to calculate the distance between points on the surface of a spheroid. The method is iterative and not theoretically guaranteed to converge. Previous work have demonstrated that interpolation calculated using this method outperforms simple euclidean interpolation [35]. However, as previously discussed, this was tested by the original baseline works when specific electrode locations were available [35]. It should therefore be expected that results obtained for the *EGL* may be lower than those reported in previous studies.

The Spherical Splines Baseline Finally, we also followed the *eeglab* MATLAB implementation of spherical splines method (*SS*) [131]. According to this implementation at each point in time the value of of the interpolated channel $\hat{s_i}$ can be approximated using the equation:

$$\hat{s_i} = c_0 + \sum_{j \neq i} c_j g\left(cos(\theta_{i,j})\right)$$

Where $\theta_{i,j}$ is the angle between the electrode locations i and j. Instead of calculating the angle, given the positions of the electrodes in space $p_i = (x_i, y_i, z_i)$ it is possible to directly calculate the cosine value: $\cos(\theta_{i,j}) = \frac{x_i \cdot x_j}{\|x_i\| \cdot \|x_j\|}$. The function $g(x)$ is defined as the sum of the series:

$$g(x) = \frac{1}{4\pi} \sum_{n=1}^{\infty} \frac{2n+1}{n^m (n+1)^m} P_n(x)$$

Where P_n is the Legendre polynomial and following [131] we set $m = 4$. Additionally, following the *eeglab* implementation we limited the infinite sum to the first seven values. Finally, the coefficients $C = (c_1, c_2, \ldots, c_n)$ are set to be the solution for the system of equation $GC+Tc_0 = S$ with the constraint $T'C = 0$ where $G_{k,l} = g\left(cos(\theta_{k,l})\right)$, $S = (s_1, s_2, .., s_n)$ and T is a vector of ones. The interpolated channel is excluded from the calculations of G and S.

3.1.4.4 Model Evaluation Approach

The selected baseline approaches [35] used the averaged normalized mean square error (ANMSE) as the main evaluation measure. The normalization of the mean square error is used to prevent a specific channel's performance from skewing the results, in case of a bad reconstruction. Having Z-scored the data however all channels are guaranteed to have the same mean amplitude. Therefore we do not normalize our mean square error results. Hence our final evaluation is:

$$AMSE = \frac{1}{M} \sum_{j=1}^{M} \left(\frac{\sum_{i=1}^{N} (s_i - \hat{s}_i)^2}{N} \right)_j$$

Where N is the number of channels (19 in our specific case). M is the number of samples. Note that the expression for mean reconstruction error (the inner average) changes for every sample. s_i and $\hat{s}_i$ are as previously defined. This measure was used both for optimizing the power parameter for the different baselines (see subsection 3.1.4.3) and calculating the final results presented momentarily.

Interpolation Method	Seen Task, Seen Subjects (% Improvement)	Seen Task, Unseen Subjects	Unseen Tasks, Seen Subjects	Unseen Task, Unseen Subjects
SS Baseline	0.728	0.694	0.8238	0.779
EUD Baseline	0.5215	0.561	0.665	0.566
EGL Baseline	0.585	0.501	0.566	0.622
Our Encoder-Decoder Model	**0.446** (14.47%)	0.478	0.552	0.465
Our Encoder-Decoder Model +Transfer Learning	—	**0.392** (21.75%)	**0.439** (21%)	**0.446** (19.78%)

Table 3.1 Comparison between Encoder-decoder model and baselines using Averaged mean square error (AMSE); *lower is better*. Note that for transfer learning (last row) the training data set differed on each column. There was no transfer learning for the *Seen Task, Seen Subject* partition as this is the original data used to train the model. SS: spherical splines baseline EGL: geodesic length calculation; EUD: euclidean baseline. The best result is bolded and percentage of improvement over the most competitive baseline is given.

3.1.5 Results

3.1.5.1 Model Hyper-parameter Optimization

After exploring the topological space by testing different network architectures (using the *Seen task, Seen subjects* data, see Figure 3.1), the best performing network is visualized in Figure 3.3. This network consisted of a simple encoder with three convolution layers and one max pooling layer, as well as four transposed convolutions in the decoder. Additionally there was a dropout and a batch normalization layer after each convolution in the encoder. All results described in this subsection were achieved using this particular architecture.

3.1.5.2 Baseline Power-parameter Optimization

For our evaluation to be extra rigorous. we optimized the power parameter for each baseline data-partition configuration separately. In Figure 3.4, we illustrate the results of our power parameter optimization for the baseline methods. As seen in the Figure, the optimal power parameters were comparable with those reported in previous literature (between 2 and 5) [35]. All the results that are reported in this subsection were for the optimized baseline on the specific data set being discussed. The spherical splines baseline has no analogous parameter we can optimize.

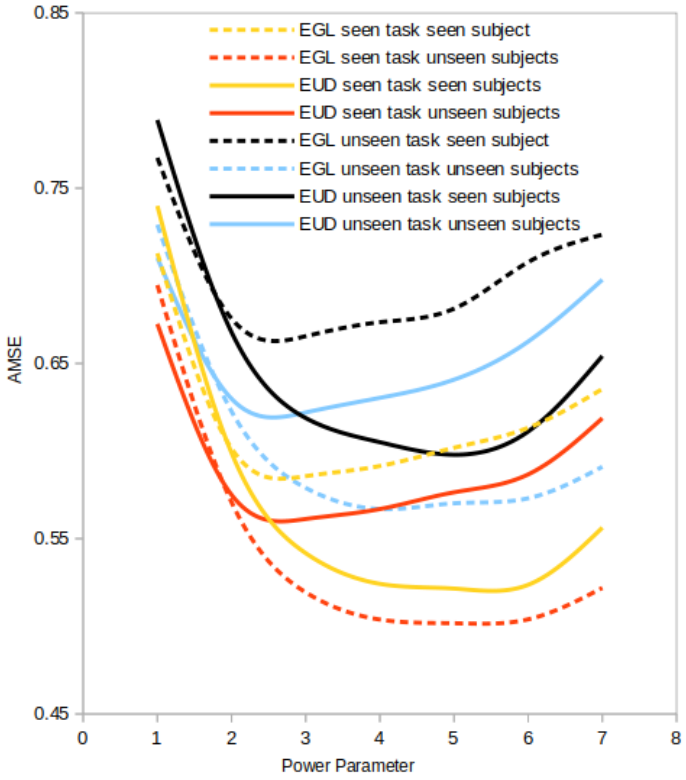

Figure 3.4 Power parameter optimization to maximize the performance of the baseline approaches. For each of the four data partitions and for the two, EUD and EGL (solid and dashed lines respectively), methods. EGL: Geodesic Length calculation; EUD: euclidean baseline; AMSE: Averaged mean square error.

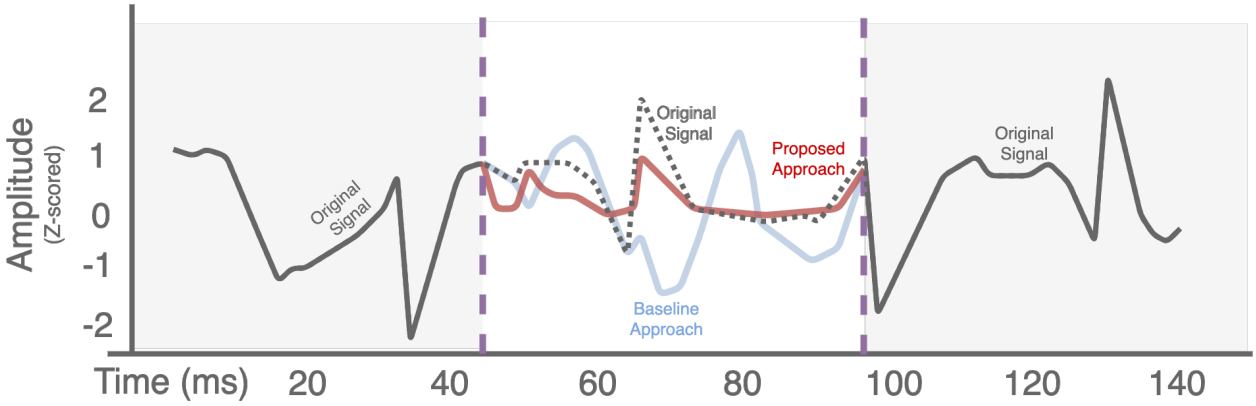

Figure 3.5 An exemplary 48ms reconstruction of the EEG data for Subject 0 *resting state task* for channel *P4*. The original channel data was removed and interpolated using best performing baseline (geodesic length calculation, in blue) and our method (in red).

3.1.5.3 Main Result

In Table 3.1, we compare the results of our proposed approach against the baselines for the EEG interpolation task on the test sets. The baseline methods are highly unstable, giving a high

variability in performance relative to our approach.

The Encoder-decoder model consistently outperformed the baselines by at least 10%. Moreover, by utilizing transfer learning the network was able to improve it's accuracy even with minimal additional data and training time.

Interestingly, in contrast to results reported in the literature, the *EGL* method did not clearly outperform the *EUD* baseline [35]. This might be due to our data not having the precise electrode locations in contrast to previous research (see Subsubsection 3.1.6.2 in the discussion).

Another interesting result is the pattern of improvements after transfer learning. As can be seen in the last row of Table 3.1, the biggest improvement was for *Unseen task, Seen subjects* data. This hint that there was more variability between EEG data from different tasks compared to data from different subjects. Additional testing will be needed to verify this hypothesis. However, this can be seen as a compelling argument in favor of flexible models that can be tuned for the specific data the researcher is working with.

Finally, we also extracted the delta (0.5-4Hz), theta (4-8Hz), alpha (8-12Hz), beta (12-30Hz), and gamma (30-100Hz) bands and tested the models performance for each band separately. Our method significantly improved over the baselines method in all bands. This is crucial as different bands have different functions (for instance, the theta band is especially responsive during observation and memorization tasks [177]). Hence for an interpolation method to be useful the reconstruction fidelity must be consistent across all frequency bands. In the interest of brevity we will not present the results for all these bands separately. The code to extract the sub-bands is also available online.

3.1.5.4 Performance on Exemplary Data

In Figure 3.5, we present an example of our method's interpolation on an exemplary portion of the data, compared against the baselines. As shown in the figure, the best baseline reconstruction contains voltage fluctuations that do not appear in the original signal, or the one reconstructed using our method. These fluctuations were quite common in baseline reconstructions. We speculate that our method might have learned to not only the optimal weights to use to approximate the occluded

channel, but also a more complicated relationship that enables our method to suppress potential artifacts that are localized to one electrode and therefore do not effect the original electrode that is being reconstructed. All things being equal, this is evidence that our framework was able to learn the nuanced relationships between electrode measurements that are not captured by baseline approaches.

3.1.6 Discussion

Our work used a deep encoder-decoder model to tackle the problem of EEG channel interpolation. While discriminative frameworks are able to only detect and label bad data segments, our results demonstrate that a generative approach can reconstruct the missing channel with high fidelity to the original signal. The success of our method suggests that deep learning can capture complex relationships between electrodes that are not sufficiently expressed by the relatively simple inverse distance calculations predominant in contemporary solutions.

3.1.6.1 On Self-supervised Learning

Data labeling is often a tenuous and resource consuming process. Unfortunately, training deep learning models often requires extensive data collection and labeling efforts. Therefore, deep learning researchers have recently began to focus on finding ways to mitigate the need for labeled data. As we showed in this study, one approach to mitigate this is to frame problems as self-supervised learning tasks. Specifically, our work is a special case of a popular self-supervised learning task: the prediction of occluded parts of data from visible ones. By using this framing we were able to circumvent a common hurdle faced by deep learning approaches.

3.1.6.2 On the Challenges of Electrode Localization

As discussed previously, prior research that compared different interpolation methods used electrode localization to extract exact channel locations for each specific subject. While generic and imprecise locations are always available, electrode localization methods attempt to alleviate the noisiness inherent to EEG by providing exact electrode locations. This localization can be done in many ways; one expensive option is to equip EEG caps with spatial sensors, or motion capture

sensors [35, 139]. Other methods that require less specialized hardware including a simple DSLR camera [32] and Kinect with an Neural Network [53]. However, despite these recent advances, electrode localization remains uncommon. For instance, no EEG data set in physionet[2] or gigadb[3] contain an EEG database with electrode localization. Therefore, and to ensure our method is applicable to the vast majority of databases, the data we used also did not include electrode localization [201]. A possible future work could incorporate location data into the deep learning framework.

3.1.6.3 On Baseline Approaches

It is worth noting that there are multiple other interpolation methods such as the nearest neighbors method, planar-spline technique [163]. We selected the baselines methods described in Subsubsection 3.1.2.1 as they were the most contemporary approaches on the topic. Furthermore, the performance improvement of our model are especially impressive considering that the *EUD* and *EGL* baselines were optimized to maximize their performance on each and every separate partition of the data.

The *SS* method requires a system of equations to be solved for each and every time point. This is not a trivial requirement as it necessities complex calculations. This demand renders the *SS* method ill-suited for any online interpolation, and by extension many BCI applications [20, 93, 138]. In contrast to the taxing nature of the training procedure, piping data foreword in neural network is computationally cheap. Therefore our approach could potentially satisfy a growing need for accurate interpolation from online data.

3.1.6.4 On Transfer Learning

Transfer learning involves training a model on a problem similar to the one being solved. This is especially useful when only scarce data is available for the problem being solved, hindering the training of the model. While transfer learning is possible for many machine learning algorithms such as Bayesian networks and Markov chains, this technique became essential to deep learning

[2] https://physionet.org/about/database/#ecg
[3] https://gigadb.org/search/new?keyword=eeg

especially due to its reliance on huge amounts of training data. Transfer learning is considered to be essential for the success and ubiquity of neural networks [117]. Our work for instance would be considerably less useful if it required every researcher to train the neural network from scratch, or if the results on data-sets that the model was not trained on were considerably worse.

3.1.7 Conclusion and Future Work

With the increasing prevalence of EEG devices, there is a need for methodologies that better address common EEG artifacts. In this work, we developed a deep encoder-decoder based method to interpolate EEG segments impacted by the most common EEG artifact: the electrode *"pop"*. We demonstrated that our method improved EEG reconstruction performance compared to existing approaches, and that our method generalized well to unseen tasks and subjects.

Future work will extend this method to tackle other kinds of electrode artifacts. Moreover, an end-to-end system that automatically detects artifacts and replaces the corrupted data with an interpolated reconstruction of the original might be of particular interest to the community.

3.2 Unsupervised EEG Artifact Detection and Correction

This section was published as a manuscript titled "Unsupervised EEG Artifact Detection and Correction" January 2021 in Frontiers in Digital Health [146].

3.2.1 Introduction

Electroencephalography (EEG) devices are pervasive tools used for clinical research, education, entertainment, and a variety of other domains [161]. However, most EEG *applications* remain limited by the low signal to noise ratio inherent to data collected by EEG devices. EEG noise sources include: movement artifacts, physiological artifacts (e.g. from perspiration), and instrument artifacts (resulting from the EEG device itself). While researchers have developed a number of methods to identify specific instance of these artifacts [176] in EEG data, most methods require manual labeling of exemplary artifact segments [4] or special hardware such as Electrooculography electrodes that are placed around the eyes, or large data-sets of templates such as independent component scalp maps [154].

Manual annotation of artifacts in EEG data is problematic because it is time-consuming and may even be untenable if the specific profiles of artifacts in the EEG data vary as a function of the task, the subject, or the experimental trial within a given task, for a given subject - as they so often do. These realities quickly scale the complexity of the artifact annotation problem, and make the use of a one-size-fits-all artifact detection method infeasible for many practical use cases.

Even if artifacts could be identified with perfect fidelity, their simple removal (e.g., by deletion of the corrupted segment) may introduce secondary analytic complications that confound the performance of downstream methods that leverage these data. For instance, methods that rely on the stationarity of EEG segments will be confounded by simple removal of the artifact segments. Even the simplest approaches, such as averaging many EEG trials before extracting features [30], may be less effective if artifact occurrence is correlated with the trail type or experimental condition, thereby increasing the likelihood of a type II error and the consequent reduction in experimental power.

[4] which may be used as "templates" by statistical or rule-based methods for the identification (and potential rejection) of noisy data epochs

An essential challenge of artifact detection in EEG processing is that the definition of "artifact" depends on the specific task at hand. That is, a given EEG segment is an artifact if and only if it impacts the performance of downstream methods by manifesting as uncorrelated noise in a feature space that is relevant to those methods. For instance, muscle movement signatures confound comma-prognostic classification but are useful features for sleep stage identification [54].

The task specific nature of artifacts makes their detection especially suitable for data-driven unsupervised approaches as the only requirement for the identification of artifacts using such methods is that the artifacts are *relatively* infrequent. That is, when mapping our data into feature spaces that are relevant to the specific EEG task, artifacts should stand out as rare anomalies. Indeed, many state-of-the-art approaches use unsupervised methods for the detection of specific artifact types, under specific circumstances. For instance, the *Blink* algorithm described by Agarwal et al. is a fully unsupervised EEG artifact detection algorithm [2] that is effective for the detection of eye-blinks. While existing methods provide excellent performance for specific artifact types, there is a need for additional progress toward generalized artifact detection approaches, that make no assumptions about the task-, subject- or circumstances.

It is also possible to go beyond artifact detection to *correct* the EEG trial by removing the artifact signal. EEG artifact removal is one instance of a more general class of noise reduction problems. The removal of noise from signal data has been a topic of scientific inquiry since Shannon's laid the foundation of information theory in the 1940s [155]; and over the years multiple signal processing approaches to this problem have found their way into EEG research. One such technique for artifact removal that is ubiquitous for EEG processing is Independent Component Analysis (ICA). This method and it's modern derivative remain popular among the research community for unsupervised artifact correction. However, ICA still requires EEG experts to review the decomposed signals and manually classify them as either signal or noise. Furthermore, while ICA is undeniably an invaluable tool for many EEG applications, it also has limitations that are particularly poignient when the number of channels is low; ICA can only extract as many independent components as there are channels, and will therefore be unable to isolate all independent noise components if the

total number of independent noise components and signal sources exceeds the number of EEG electrodes [39].

Artifact removal is an especially common practice for a particular artifact type: the electrode *"pop"*. These artifacts result from abrupt changes in impedance, often due to loose electrode placement or bad conductivity [180, 100]. Unlike muscle and movement artifacts, electrode pop is extremely localized, often effecting only one electrode channel. Channel interpolation is the process of replacing the signal of a corrupted channel with one that is interpolated from surrounding clean channels. Patrichella et al. demonstrated that knowing specific electrode locations (namely the exact electrode locations for each subject) and the distances between them can improve interpolation results [132, 35]. However this type of additional information is rarely available and often requires special dedicated hardware. Recently, Sadiya et al. proposed a deep learning convolutional auto-encoder based approach to learn task and subject specific interpolation [146]. By iteratively occluding channels in the input and using original data as the ground truth, the model learned how to interpolate channels in a self-supervised manner with no human annotation. Moreover, not only was the model able learn idiosyncratic information such as subject specific electrode location, beating state of the art models, it was also possible to use transfer learning to improve performance on previously unseen tasks and subjects.

In this paper, we extend the aforementioned state-of-the-art approaches in artifact detection and rejection by building an end-to-end pipeline that solves both the detection and rejection problems together, without making any assumptions concerning the task or artifact type.

Our artifact detection approach uses a collection of quantitative EEG features that are relevant for a wide variety of tasks including coma prognostics [172], diagnosing mental-illness [174], decoding mental representations [70], decoding attention deployment [200], and brain computer interface design [6]. Unsupervised outlier detection algorithms utilize these extracted features to identify artifacts in the EEG data. These unsupervised algorithms only require an estimate of the *frequency* of artifacts in the data, and can detect any artifact type, irrespective of the task. To guarantee that our results accurately represent the capabilities of these unsupervised outlier

detectors we carefully selected algorithms that are qualitatively different from each other (for instance relying on local vs global characteristics of the data distributions) and explored hundreds of different possible configurations. Sub-subsection 3.2.2.2 provides a comprehensive review of the feature extraction process. Sub-subsection 3.2.2.2 details our experimentation with different outlier detection algorithms.

Our artifact correction approach uses a deep encoder-decoder network to correct artifacts that are *not restricted to only one channel*. Specifically, we frame our learning objective as a modified *"frame-interpolation"* task. Frame interpolation is the filling in of missing frames in a video [79]. To the best of our knowledge this is the first work that takes this approach to EEG artifact correction. The proposed approach is also unique in that it does not require the maintenance of any large data-set of templates or annotated data similarly to other state-of-the-art artifact removal methods [2]. The model architecture, as well as the exact objective formulation are discussed in detail in subsubsection 3.2.2.3.

The data-sets used in this work are discussed in detail in subsubsection 3.2.2.1. And the results of the different experiments we conducted can be found in subsubsection 3.2.3. Finally we, discuss our findings, their broad implications, and the limitations of our approach in subsection 3.2.4.

3.2.2 Methods

In this paper we propose an end-to-end pre-processing pipeline for the automated identification, rejection and removal / correction of EEG artifacts using a combination of feature-based and deep-learning models which is intended for use as a general-purpose EEG pre-processing tool. To begin, we provide a brief overview of the data and methodological pipeline, calling out the specific subsubsections where the full details of each component of the pipeline is discussed.

In Figure 3.6 we provide a visualization of our proposed pre-processing pipeline; our method begins by performing unsupervised detection of epoched EEG segments in a 58 dimensional feature space (subsubsection 3.2.2.2). The trials that were not rejected in this initial stage are used to train a deep encoder-decoder network designed to correct artifacts segments (subsubsection 3.2.2.3).

While we demonstrate this method on a particular data set (described below), it is applicable

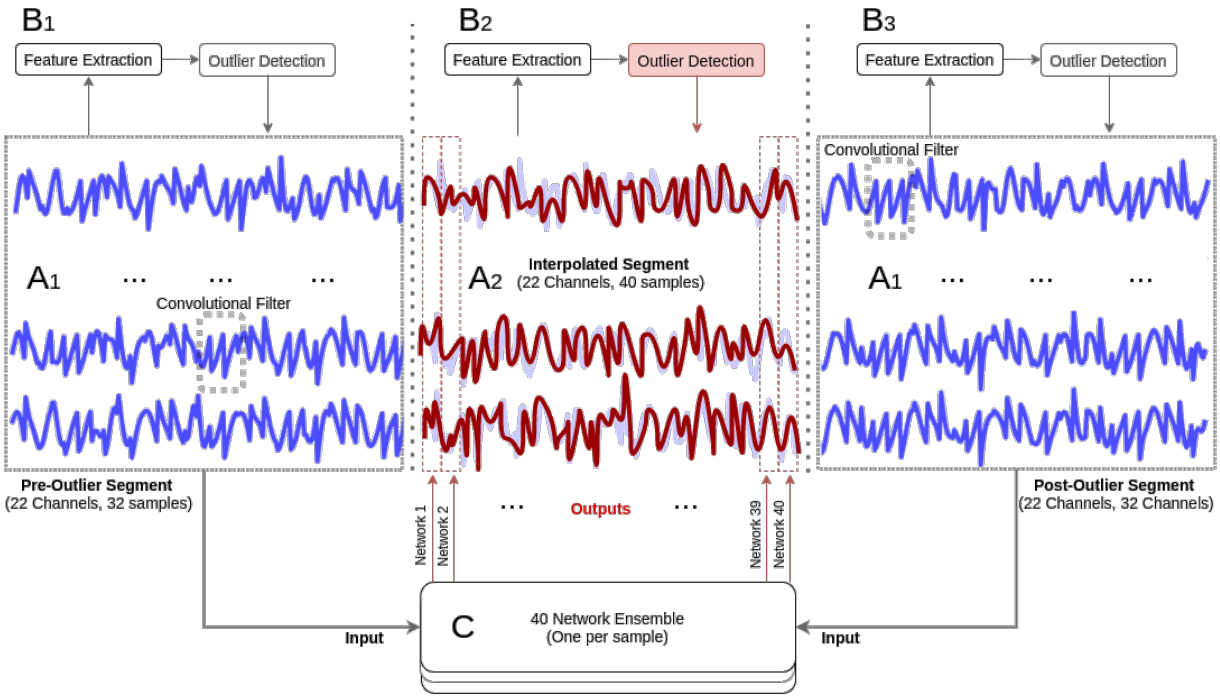

Figure 3.6 Our methodological approach. The EEG data is first segmented into epochs (see A_1, A_2, A_3). Next, 58 features are extracted and an ensemble of unsupervised outlier detection methods are used (see B_1, B_2, B_3) to identify EEG epochs that are artifact-ridden and require interpolation (see A_2 and B_2). The artifact-ridden epochs are then interpolated by an ensemble of deep encoder-decoder networks (see red line in C).

(with no modifications) for any EEG pre-processing work. The methods are presented in the order of their processing within our proposed pipeline.

3.2.2.1 Data-sets

Data acquisition Our aim is to demonstrate that unsupervised anomaly detection be successfully used to identify artifacts in EEG data, and that these artifacts can be corrected via representation learning methods (see subsection 3.2.2.3). To demonstrate the feasibility of our approach, it is necessary to not only have ground truth artifact annotations, but also the ground truth labels for all trials, including those that were annotated as artifacts. While the artifact annotations allow us to test the unsupervised outlier detection methods, the trial labels allow us to verify that corrected EEG data can indeed be used in conjunction with that regular data for downstream analytic tasks (e.g. training a classification model). Unfortunately available data sets usually do not contain rejected trials, and even when these annotations are available the original trial label is not included

[5]. Therefore, our work is validated on two data-sets, hereinafter referred to as the *orientation* and *color* data-sets, that were previously collected by Saidya et al. [144]. We briefly describe these datasets here; additional information about the data-sets are provided in the Supplementary materials.

Both experiments were passive viewing tasks. The orientation task stimulus consisted of 6 oriented gratings, the color task stimulus consisted of random dot fields in 6 different colors. The stimulus was generated using MGL, a library running in Matlab (Mathworks). The data was collected using a 32 electrode actiCHamp cap at $1000Hz$. For each task we collected data from 7 subjects (4 male) for a total of $\sim 10,000$ EEG Trials. All subjects reported normal or corrected to normal vision. The data was examined for noisy trials by expert annotators. Fully annotated and anonymized data-sets will be made available online. Participants gave informed consent and compensated at the rate of 15$ per hour. The experimental procedures were approved by the Michigan State University Institutional Review Board and adhered to the tenets of the Declaration of Helsinki.

3.2.2.2 Unsupervised artifact detection

To benchmark the different outlier detection methods we collected a list of common features used in EEG research in different domains and applied various unsupervised outlier detection algorithms. Our main objective was to thoroughly investigate the feasibility of unsupervised artifact rejection for EEG.

Feature extraction Building on the previous work of Ghassemi et al. [56], we reviewed the EEG literature and constructed a permissive list of several features that are commonly used for EEG classification tasks. In total we identified and extracted 58 features. The code that extracts these features was written to allow for parallelization of the calculations, and is accessible as a downloadable python 3.5 package[6]. See Table 3.2 for breakdown and references for all 58 features.

These features can be grouped into three categories that measure the complexity, continuity and connectivity of EEG activity. Before continuing to discuss our pipeline we will provide a high

[5]For instance BCI competitions data: http://bbci.de/competition/

[6]Code available at: https://github.com/sari-saba-sadiya/EEGExtract

Signal Descriptor	Ref.	Brief Description
Complexity Features		degree of randomness or irregularity
Shannon Entropy	[156]	additive measure of signal stochasticity
Tsalis Entropy (n=10)	[52]	non-additive measure of signal stochasticity
Information Quantity ($\delta, \alpha, \theta, \beta, \gamma$)	[158]	entropy of a wavelet decomposed signal
Cepstrum Coefficients (n=2)	[125]	rate of change in signal spectral band power
Lyapunov Exponent	[185]	separation between signals with similar trajectories
Fractal Embedding Dimension	[1]	how signal properties change with scale
Hjorth Mobility	[120]	mean signal frequency
Hjorth Complexity	[120]	rate of change in mean signal frequency
False Nearest Neighbor	[69]	signal continuity and smoothness
ARMA Coefficients (n=2)	[21]	autoregressive coefficient of signal at (t-1) and (t-2)
Continuity Features		clinically grounded signal characteristics
Median Frequency		the median spectral power
δ band Power		spectral power in the 0-3Hz range
θ band Power		spectral power in the 4-7Hz range
α band Power		spectral power in the 8-15Hz range
β band Power		spectral power in the 16-31Hz range
γ band Power		spectral power above 32Hz
Median Frequency		median spectral power
Standard Deviation	[142]	average difference between signal value and it's mean
α/δ Ratio	[172]	ratio of the power spectral density in α and δ bands
Regularity (burst-suppression)	[172]	measure of signal stationarity / spectral consistency
Voltage $< (5\mu, 10\mu, 20\mu)$		low signal amplitude
Diffuse Slowing	[167]	indicator of peak power spectral density less than 8Hz
Spikes	[167]	signal amplitude exceeds μ by 3σ for 70 ms or less
Delta Burst after spike	[167]	Increased δ after spike, relative to δ before spike
Sharp spike	[167]	spikes lasting less than 70 ms
Number of Bursts		number of amplitude bursts
Burst length μ and σ		statistical properties of bursts
Burst band powers ($\delta, \alpha, \theta, \beta, \gamma$)		spectral power of bursts
Number of Suppressions		segments with contiguous amplitude suppression
Suppression length μ and σ		statistical properties of suppressions
Connectivity Features		interactions between EEG electrode pairs
Coherence - δ	[172]	correlation in in 0-4 Hz power between signals
Mutual Information	[6]	measure of dependence
Granger causality - All	[16]	measure of causality
Phase Lag Index	[166]	association between the instantaneous phase of signals
Cross-correlation Magnitude	[81]	maximum correlation between two signals
Crosscorrelation - Lag	[81]	time-delay that maximizes correlation between signals

Table 3.2 The 58 EEG features fell into three EEG signal property domains: Complexity features (25 in total), Category features (27 in total), Connectivity features (6 in total).

level intuition behind the inclusion of each category. We encourage the interested reader to refer to the previous work of Ghassemi et al. for a more detailed discussion of of the specific features [56].

Complexity features (n = 25): These features measure the complexity of the EEG signal, from an information theoretic perspective, and are known to correlate with impaired cognitive functions and the presence of degenerative illnesses. Therefore our first set of features is a collection of information theoretic complexity measures. Of special interest are the first three features shown in Table 3.2 as they are particularly prominent in EEG research: *Shanon's entropy* has been associated with neurological outcomes in post-anoxic coma patients [172]; the entropy of the decomposed EEG wavelet signals (known as the *Subband Information Quantity*) have similarly been used in cardiac arrest studies [157, 78]. *Tsalis entropy* is a generalization of Shannon's entropy that does not make assumptions about the independence of data channels (as Shannon's entropy does) and has been shown to be particularly useful for the characterization of complexity in EEG data [52].

Continuity features (n = 27): These features capture the regularity and volatility of EEG activity. Bursts, spikes, and unusual changes in the mean and standard deviation in the frequency and power domains are examples of continuity features that are relevant for a variety of clinical tasks. See Hirsh et al. for an in-depth review of continuity and it's relevance to clinical care [71].

Connectivity features (n = 6)): These features reflect the statistical dependence of EEG signal activity *across* two or more channels. Functional connectivity networks are an established features of normal brain functioning. We draw on the rich literature on measuring connectivity from EEG signals [150] extracting features that have previously been used for designing brain computer interfaces [6] as well as in mental-illness, perception, and attention research (See [174], [70], and [200] respectively).

Outlier detection methods We explored a set of ten algorithms for unsupervised artifact detection; the explored algorithms were inspired by the work of Zhao et al. [196]. The algorithms can be divided into two general groups: statistical methods and representation learning methods; they are described in more detail in the *"Statistical Methods"* and *"Representation Learning Based Methods"* subsections below. The hyper-parameters of each method were determined by randomly exploring the hyper-parameter space and choosing the settings that yielded the best performance of the methods on the data according to our artifact annotations.

Statistical methods: Statistical methods identify anomalies based on statistical measures extracted from the data, thereby producing an "anomaly score" for each trial. The Histogram Based Outlier detection **(HBOS)** method uses histograms with dynamic bin widths to detect clusters and anomalies in different feature dimensions. Despite the simplicity of the approach it has been shown to work well on a variety of data types [60]. The Local Outlier Factor **(LOF)** method similarly calculates an "outlier score", however instead of global measures it relies on the local density of the data as it's main indicator [22]. Another popular local algorithm, the Angle-Based Outlier Detector **(ABOD)**, calculates the cosine similarity of data points with their neighbors and uses the variance of these scores to generate anomaly scores [85]. Finally, we also trained a One Class SVM Detector **(OCSVM)**, a classic algorithm for outlier detection [151]. In this algorithm a SVM is trained on the entire data-set and afterwords every instance is scored based on its distance from the class boundary; the intuition is that the infrequent outliers will contribute less to the decision boundary calculation and will be more likely to be on the margin of the learned boundary.

As previously mentioned, we selected these detectors to be different in the type of statistical measurements they use. Therefore, it makes sense to also train ensemble classifiers to further improve the outlier detection accuracy. Specifically we trained five hundred *Locally Selective Combination in Parallel* **(LSCP)** Outlier Ensembles [197] with different combinations of the algorithms mentioned above.

Representation learning based methods: Unlike statistical methods, representation learning based outlier detectors do not simply calculate statistical properties of featuarized data. The most basic classifier uses auto-encoder **(AUTO)** based deep learning architectures to learn a lower dimensional representation of the data that enables the best possible reconstruction of the original signal; the embedding would be optimized for the common regular data points thereby producing distinctly noisy reconstructions for the outlier trials [3]. This classifier can be viewed as a modern update of similar classic outlier detection methods that use methods such as PCA reconstruction instead of of a training a deep auto-encoder **(PCA)** [159]. A more sophisticated approach uses Variational Auto-Encoders **(VAE)**. This class of algorithms try to ensure that the learned embedding captures

the structure of the original data by penalizing the classifier if the embedding does not follow a standard normal distribution [4]. Finally we also examine a Generative Adversarial Active Learning **(GAAL)** outlier detector [99] which uses a generative adversarial networks to generate outliers. This method can be used to improve any of the statistical methods described in 3.2.2.2. We also use an extension of the original method to learn multiple generators (**MGAAL**).

3.2.2.3 Artifact correction

As previously mentioned, encoder-decoder based deep learning methods have proven useful for channel interpolation [146]. In this subsection we discuss an extension of this approach that utilizes the same framework for artifact correction. Namely, given an EEG data segment with an isolated artifact we remove the corrupted segment and use the data samples preceding and proceeding it to fill in the resulting void. This problem is equivalent to the *"frame-interpolation"* task of filling in missing frames in a video [79].

The model Input representation: The channel interpolation model proposed in [146] represented the EEG as a time series of 2D topologically organized arrays. This reflects the spatial nature of the EEG channel interpolation issue; the intrepolated values at different time points are treated as independent. To the best of the author's knowledge this is a standard assumption for EEG interpolation algorithms. For instance, Petrichella et al. and Courellis et al. calculate the interpolated values of the missing data at each time point separately [132, 35]. However, research on convolutional neural networks for EEG decoding and visualization have shown performance benefits from presenting the input as a column of electrodes unfolding in time, as this facilitates the learning of temporal modulations [149]. Since artifact correction is first and foremost a process of completing gaps across time we decided to depart from [146] and use a 2D array representation with the number of time steps as the width of the array.

Architecture: The best frame interpolation models involve calculating object trajectory and accounting for possible occlusion (e.g. if one object moves behind another). With these "flow computations" and a stack of the frames before and after the missing image a convolutional encoder-decoder can generate realistic intermediate images [79]. Unlike video, EEG data has only one spatial dimen-

sion (see subsubsection 3.2.2.3) and no analogues to local phenomena such as occlusion or object movement can occur as EEG modulations are often thought of as mostly global in nature [149]. Therefore we only concern ourselves with a stacked convolutional auto-encoder. This architecture is shared by previously discussed state-of-the-art algorithms for both frame interpolation and channel interpolation [149, 146].

The interpolation of each frame is done separately, thus to predict n frames it is necessary to train n networks. Technically this is equivalent to training one ensemble model, however by separating the networks we allow for easier parallelization of the training process. Specifically, given a series of EEG frames $x_1, x_2, \ldots, x_n$ where x_t is a vector of all the channel values at time t, and assuming that the series is missing all frames between time points t_b and t_e, our network learns to predict x_{t_q} from the two stacks, $x_{t_b-h}, x_{t_b-h+1}, \ldots, x_{t_b}$ and $x_{t_e}, x_{t_e+1}, \ldots, x_{t_e+h}$ where $t_q \in (t_b, t_e)$ and h is some small positive integer representing how many frames before and after the missing segment can be perceived. Every network is trained to predict the value at one specific value of q. Every network takes the same $2h$ frames (half preceding the missing segment and half following it) to calculate the value at a given frame.

3.2.2.4 Model validation approach

Artifact Detection Method: The performance of the artifact detection methods was assessed by inspecting the agreement between the artifact detection approach and the expert annotations from the two data sets (color and orientation). More specifically, the agreement was measured using the f-score and Cohen's Kappa (first and second values in each cell respectively). We compared the performance of our model against the expected performance of a classifier with knowledge of the exact number of artifacts; this random classifier is expected to have an f-score of 0.172 and a Kappa of 0.029. We ran the detection algorithms in two configurations, for each subject separately and for the entire aggregated data. We hypothesise that the performance will drop when using the aggregated configuration as each individual setup for an EEG recording is likely introduce unique artifacts (due to loose connections, or subject specific circumstances such as perspiration).

Artifact correction method: To optimize the parameters of the artifact correction model, we pro-

duced training data from trials that were marked as artifacts free by our unsupervised artifact detection method (subsection 3.2.2.2) and randomly removed a segment from the middle of the trial. The h samples proceeding the removed segment and h samples preceding it were used as input for the model while the removed segment was the ground truth (h was a hyper parameter optimized on the training set). For the purposes of validating the artifact correction model, all EEG data was re-sampled to $200Hz$. The reconstructed segments were $200ms$ each.

End-to-end assessment approach: We ran a number of tests to examine if the trials reconstructed by our artifact correction method could be used to enhance the performance of downstream EEG tasks. More specifically, we trained two SVM models to predict the label of the trial from the color data-set: one SVM was trained using the *raw data*, and the other was trained using the raw data *after artifact correction*. Both models were validated using 5 fold cross validation, and the performance of the models on the test set (μ and σ) was reported.

We also evaluated the impact of our artifact correction method on downstream EEG tasks when applied to *clean trials, exclusively*; this evaluation allowed us to test for inadvertent degeneration in signal quality of clean segments when processed by our method. More specifically, we applied our artifact correction method to 20% of *clean* trials and used the resulting data to trained an additional SVM model.

3.2.3 Results

This subsection presents the results of the two main components in our pipeline, the artifact detection method and the artifact correction method on the data described in 3.2.2.1.

3.2.3.1 Artifact detection results

In Table 3.3, we compare the *average* performance of the outlier detection methods described in subsection 3.2.2.2 when applied to each subject *separately*. Therefore, each value is the mean of the algorithm's performance across subjects. As previously mentioned, the expected performance of a baseline random classifier with knowledge of the exact number of artifacts is an f-score of 0.172 and a Kappa of 0.029. Hence, all models other than the *ABOD* classifier performed significantly better than the baseline (one tailed t-test with a $p = 0.05$ significance level).

Statistical Methods	**HBOS**	**LOF**	**ABOD**	**OCSVN**	**LSCP**
Orientation	0.564 0.473	0.218 0.065	0.11 0.06	0.41 0.29	0.577 0.489
Color	0.5 0.4	0.241 0.091	0.1 −0.08	0.36 0.23	0.51 0.411
Representation Learning	**AUTO**	**PCA**	**VAE**	**GAAL**	**MGAAL**
Orientation	0.53 0.44	0.527 0.426	0.477 0.368	0.429 0.311	0.428 0.309
Color	0.51 0.42	0.477 0.367	0.478 0.368	0.241 0.086	0.389 0.263

Table 3.3 Comparison of the different unsupervised outlier detection methods when applied to each subject separately. We calculated the mean f-score and Cohen's Kappa (first and second row in every cell) across all subject. HBOS: Histogram based outlier detection, LOF: Local outlier factor Method, ABOD: Angle-based outlier detector, OCSVM: One class support vector machine, LSCP: Locally selective combination of parallel outlier Ensembles, AUTO: Auto-encode based method, VAE: Variational auto-encoder based method. GAAL: Generative Adversarial Active Learning, MGAAL: Multi-object Generative Adversarial Active Learning.

Unsurprisingly, the best outlier detector was an *LSCP* ensemble classifier that performed 16.86x better than the baseline method, and 1.03x better than the next best approach; the best performing configuration of the classifier consisted of two *HBOS* classifiers and one *OCSVM*. While it is difficult to interpret ensemble classifiers it is worth noting that the two histogram based classifiers diverged quite substantially; one using a high number of histogram bins and a rigid outlier scoring policy ($tol = 0.1$) while the other using a smaller number of bins and more relaxed policy ($tol = 0.5$). A simple auto-encoder was the best representation learning algorithm, closely followed by the PCA algorithm. We speculate that the auto-encoder could have possibly had better performance if more data was available for each subject. See our supplementary material for a breakdown of trial and artifact numbers for each subject.

In Table 3.4, we compare the performance of the outlier detection methods described in subsection 3.2.2.2 when applied to the subjects *aggregated* data; that is, subject were not considered separately as they were in the results from Table 3.3. When compared to the results shown in

Table 3.3, the performance decreased for most models. This is not surprising as the fundamental assumption of unsupervised methods is that the data is homogeneous with the exceptions of the outliers. Here again the LSCP method performed the best of the tested approaches. A comparison of the results in Tables 3.4 and 3.3 provide motivation for the development of subject-specific anomaly detection approaches. Moreover, the comparison also highlights that the unsupervised algorithms and the features we extracted can successfully capture both common EEG artifacts and subject specific idiosyncrasies.

Statistical Methods	**HBOS**	**LOF**	**ABOD**	**OCSVN**	**LSCP**
Orientation	0.502 0.4	0.246 0.095	0.07 −0.11	0.362 0.234	0.537 0.441
Color	0.476 0.35	0.305 0.15	0.09 −0.108	0.377 0.238	0.463 0.332
Representation Learning	**AUTO**	**PCA**	**VAE**	**GAAL**	**MGAAL**
Orientation	0.488 0.338	0.448 0.338	0.447 0.336	0.383 0.246	0.393 0.258
Color	0.414 0.283	0.437 0.312	0.436 0.31	0.185 0.022	0.393 0.258

Table 3.4 The performance of the models trained on data aggregated from all the subjects. The f-score and Cohen's Kappa are presented in the first and second row in every cell.

3.2.3.2 Artifact correction results

Network optimization: Our first step was to optimize the network hyper-parameter configurations. This included testing different sizes of both the layers and convolution filter, as well as exploring different hyper-parameters such as optimization algorithms, dropout rates, and activation functions. To train the network we followed the method discussed in subsection 3.2.2.2: we randomly extracted 104 samples from the data, the first and last 32 samples were stacked and used as the input to the model, the sample at position i from the remaining 40 samples was used as the ground truth. Essentially we are training a network to predict the values after removing 40 samples ($200ms$) using the 32 samples that before and after the removed segment. The best performing network

(lowest loss) was different for different ts. The optimal topology for reconstructing sample 20 is available in the supplementary material as a reference of the type of convolutional U-net architecture used.

End-to-end assessment In Table 3.5 we compare the classification accuracy of a 5-fold SVM model trained to perform a downstream classification of trial type using down sampled EEG data with three different configurations of the data: (1) the raw EEG data, (2) the data after correction of artifact segments and (3) the data following "correction" of a random 40 samples of 20% of the non-artifact segments. Note that while simple this type of analysis is used in actual EEG research [30].

The performance remained comparable after using the artifact correction on trials that did not contain any artifacts. This is a strong indication that the model is indeed able to learn how to reconstruct the original EEG signal. When using the corrected trials with EEG artifacts the classification accuracy improved by 10% overall and over 20% for trials that were marked as containing artifacts. These results successfully demonstrate that our unsupervised end-to-end artifact correction pipeline improves down-stream analysis.

	Original EEG	EEG with Random Correction	EEG with Artifact Correction
All trials	0.3	0.31	0.33
Rejected trials	0.23	0.23	0.29

Table 3.5 Mean accuracies of simple SVM classifiers. A simple t-test confirmed that all accuracies were significantly above chance level (1/6 for 6 different colors) at a $p = 0.05$ level. Original EEG: The Original EEG data. EEG with Random Correction: The EEG data after random artifact free trials were "corrected". EEG with artifact correction: The data after we applied the EEG artifact correction on the trials that were marked as artifact ridden.

3.2.4 Discussion

Significance of our results: In this paper we presented an end-to-end pipeline that is capable of unsupervised artifact detection and correction. Our results demonstrate that data driven approaches for unsupervised outlier detection can be an extremely useful when applied to the problem of EEG

artifact detection. Interestingly the classifiers with the best performance (HBOS, OCSVM, and the best performing LSCP) are global classifiers, this might indicate that EEG artifacts are better discriminated by global characteristics. This supports our previous observation that artifacts are task specific and infrequent occurrences of uncorrelated noise. It is worth noting that, as demonstrated in Table 3.4, the classifiers we trained were able to learn subject specific idiosyncrasies.

While the accuracy and agreement between the annotators and the detectors was far from perfect, the Cohen Kappa of the best performing algorithm was comparable to the inter-rater agreement levels of expert annotators reported in the literature; for instance when asked to annotate "periodic discharges" (a specific type of artifact) and "electrographic seizure" annotators had a cohen's Kappa of 0.38 and 0.58 respectively [67]. Our results indicate that unsupervised outlier detection is a feasible approach for generalized EEG artifact detection.

The data-sets: We validated our framework on two novel data-sets. To test the impact of artifact correction algorithms on downstream analysis it is necessary to have ground truth artifact annotation as well as knowledge of the labels of all trials, including those that are artifact ridden. Unfortunately public data sets often exclude trials that contain artifact. Even in the rare occasions in which these trials are made available the labels are often replaced with a special identifier for rejected trials [7]. We hope our data-sets inspire other researchers to adopt more thorough data publishing practices as data-availability is perhaps the primary limiting factor in artifact correction research.

The strength of unsupervised end-to-end methods: The accuracy of simple classifiers improved modestly after artifact removal. It is possible that replacing our deep learning based artifact removal components with an ICA artifact removal algorithm [80] could yield better results. However, two important distinctions should be made: First, the proposed method does side step many weaknesses inherit to ICA [39] (such as the number of independent components being limiting by the number of channels, which is particularly problematic for lightweight commercial EEG setups). Secondly, while the independent component deconstruction itself is data driven and unsupervised, the ICA method still requires visual inspection and analysis of the decomposed signal by human experts. In

[7] For an example of standard EEG publishing practices see the BCI Competition data-sets

contrast, our method can be put into effect without any human intervention, making it is suitable for online EEG applications or as a no-cost first step before more thorough analysis. In general, supervised methods unquestionably out-perform unsupervised ones and we fully acknowledge that the pipeline proposed in this work is no different. It is therefore useful to consider unsupervised methods not as replacements of currently existing algorithms, but as complimentary additions to the toolbox of the EEG researcher. With this in mind we intentionally designed our end-to-end pipelines to be highly modular; An experienced researcher can easily substitute our last component with an ICA artifact removal algorithm, and in contrast, researchers that have access to artifact annotations (for instance by virtue of employing specialized hardware during data acquisition) will be able use their method in conjunction with ours or side step the first processes completely and apply only the artifact correction component before carrying on with the analysis process.

Limitations: We did not formally evaluate the reconstruction performance of the model because (1) there is not an authoritative literature baseline and (2) insofar as the reconstruction enhances the ability of the downstream classification model to perform their intended classification tasks, the reconstruction is valid and valuable. There are a few limitations that we hope to address in future work. First and foremost, this artifact detection method can only be used if the frequency of the artifacts is low enough for them to be considered outliers. While this is indeed the case for the vast majority of EEG use cases, tasks such as seizure detection often involve long periods of unusually low signal to noise ratio. Additionally, the performance of our artifact correction network would likely benefit from introducing more complex component into the architecture. For instance, introducing temporal dependencies via and LSTM component would guarantee that the corrected frame at time t influences the frame at time $t + 1$. Finally, our method is in dire need of being validated on additional tasks and data-sets.

Despite the challenges described above, we believe that our work demonstrates the feasibility of a EEG pre-processing pipeline which if adopted could facilitate and expedite the often tenuous process of artifact annotation and removal, and could therefore be extremely beneficial for the general EEG research community.

3.2.5 Conclusion and future Work

The applications of EEG are numerous and diverse, and while this impacts the particularities of what components are classified as part of the signal versus artifacts, data homogeneity is a common concern in this area of research. Building on this data science perspective, in this work we appropriated state-of-the-art data driven methods to construct an end-to-end unsupervised pipeline for general artifact detection and correction. We introduced two new data-sets and demonstrated that the inter-rater reliability of our artifact detection component against expert annotators is comparable to reported inter-human levels. Furthermore, we demonstrated how applying the complete pipeline on a data-set can improve the performance of a common downstream analysis. The pipeline makes use of a wide range of handcrafted clinically relevant features, and we believe the released python package will be of use to many in EEG research community.

3.3 Feature Imitating Networks

This section was published as a manuscript titled "Feature Imitating Networks" in the proceedings for the 2022 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) [143].

3.3.1 Introduction

The successful application of deep learning to new problem domains has three conditions: (1) access to large data sets, (2) access to sufficient computing resources for hyper-parameter optimization and (3) modest expectations about model interpretability. Deep learning models require large data-sets to learn representations that generalize on future unseen data. Additionally, extensive exploration of the model topological space is often necessary to identify a network architecture with sufficient representational power for a given task. Lastly, despite ample recent work on interpretability of deep learning models, the community remains without normative standard for how deep networks should be interpreted; this is problematic for many problem domains (e.g. healthcare) where the importance of interpretability may supersede performance [141]. [19].

Contributions In this section we introduce Feature-Imitating-Networks (FINs): a FIN is a neural networks with weights that are initialized to approximate one or more closed-form statistical features. In this section, we will demonstrate how this property of FINs improves their interpretability while also reducing data and hyper-parameter tuning requirements compared to other networks with similar or greater representational power. More specifically, we demonstrate how, when combined with a careful application of transfer-learning, and by taking into account expert knowledge, FINs can be used to quickly build and deploy robust and better performing models using less training epochs. Our validation of FINs focused on tasks involving biomedical signals; the data-sets in this domain are often smaller, and therefore stand to benefit the most from the introduction of our framework.

Section Organization The remainder of the Section is organized as follows: First we review relevant literature regarding transfer learning. The *Related Work* subsection is followed by the

Methodology subsection where we discuss how to build and design different FINs. The *Experiments* subsection contains three experiments - including a brief discussion of the data and results for each. Finally, the *Discussion* subsection examines all the results in aggregate and discusses how our framework might be expanded.

3.3.2 Related Work

Transfer learning is the application of a pre-trained model to tasks it was not originally intended to perform [117]. Transfer learning enabled researchers to make significant progress on various tasks in Machine Vision[73], Speech [89], and Natural Language Processing [75].

Most applications of transfer learning are *within-domain*; these involve fine-tuning a pre-trained model for new tasks. For instance, AlexNet [87], VGG [160], and ResNet [68] are computer vision models trained to classify the ImageNet data-set. The features learned by these models (in later layers), and their more fundamental image components (in earlier layers) can be re-purposed to solve other tasks using only a fraction of the training data required by original models.

Models that are trained on large heterogeneous data-sets are good candidates for "transfer". But for biomedical signal processing problems, there isn't a sufficiently large data-set to train such a model. Indeed, the largest publicly available biomedical signal archives contain only a few thousand subjects, which is too small by most data standards in other domains [56]. Consequently, transfer learning for biomedical signals is often performed *across-domains*. For instance, computer vision models such as VGG have been adapted to emotion recognition from speech [89], motor-imagery classification [192] and mental task classification [124], but these *cross-domain* transfers are not as effective as those performed *within-domain*. Finally, interpretability is greatly hindered when models are trained on broad data-sets with objective functions that differ from the final application [19].

The performance of transfer learning is proportional to the proximity of the domains across which knowledge is being transferred. Feature imitating networks were designed to address this limitation of current transfer learning paradigms; they provide the power and flexibility of transfer learning without the "Big data" and heavy computational requirements.

3.3.3 Methodology

A FIN is a neural network with weights that are initialized to approximate one or more closed-form statistical features. In this section, we train FINs that approximate five commonly used features in biomedical signal processing: Shannon's Entropy, kurtosis, skewness, fundamental frequency, Mel-frequency cepstral coefficients (mfcc), and regularity [146]. We evaluate the utility of the FINs on three biomedical signal processing experiments, which we describe in subsection 4, below. The pre-trained FINs, and code to reproduce all experimental results are available online [8].

Network Construction For each feature, we used a simple gradient descent optimizer with mean square error (MSE) loss to train a simple dense network to approximate that feature on synthetic signals. The topological space explored for all the FINs was between 2 to 10 layers with the number of parameters in the $3 - 15$ million range. All best performing FINs used simple $relu$ and $tanh$ activation functions. See Figure 3.7 for density functions for the errors of the FINs reconstruction.

Input The input data for the FINs consisted of synthetic signals generated randomly (zero mean, unit variance) and converted to the time-frequency domain using the wavelett transform [89, 192].

Outcome The outcome data for the FINs consisted of closed-form feature values calculated on the synthetic signals using *SciPy* and *EEGExtract* packages [146].

Transfer When applying the models to new classification tasks, (i.e. subsection 3.3.4) the very last layer was discarded in favor of a randomly initialized softmax layer with dimensions suitable to the task.

Baseline model The baseline in each experiment was the best performing neural network with similar (or greater) representational power to the corresponding FIN, trained using the same training data and schedule, but with weights that were randomly initialized; In total one hundred different topologies were explored for each baseline. The baseline with the best average performance on the validation data was retained for comparison against the FINs.

Training All models (both FIN and Baseline) were trained using early stopping and a simple gradient descent optimizer. Non topological hyper-parameters such as learning rate and momentum

[8] https://github.com/sari-saba-sadiya/Feature-Imitating-Networks

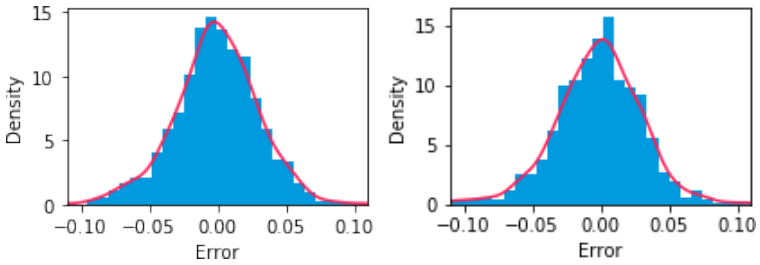

Figure 3.7 Density plots for the errors for the entropy (left) and regularity (right) FIN reconstructions. The feature values were scaled and normalized, making the biggest possible error 1, as can be observed the FINs faithfully recreate the closed form equations.

had minor effects in comparison and therefore will be omitted from future discussions. Training was conducted using *CUDA* on a *Tesla K80* GPU with 25GB of RAM.

3.3.4 Experiments

To evaluate our framework we ran three experiments on three different biomedical data-sets and tasks. The first experiment was an Electrocardiography (ECG) classification task; our goal was only to demonstrate that our FINs framework can successfully improve performance on small low-accuracy data-sets. The second experiment was an Electroencephalography (EEG) artifact detection task; our goal was to demonstrate the modular nature of FINs, and the potential benefits of using FIN ensembles. The third experiment was a drowsiness detection task using EEG; our goal was to compare both the performance and speed of FINs against state-of-the-art transfer-learning techniques under conditions of varying data scarcity.

For all three experiments, the data was regularized and transformed to the time-frequency domain as discussed in the Methods subsection. The data was partitioned into training, validation, and testing sets. In the first two experiments this was achieved by randomly partitioning the data 15% – 85% for testing and training respectively, before repeating the same split for the training data to extract a validation subset. This was repeated for a 50-fold cross-validation. In the third experiment, where subject data is balanced, the partitioning was achieved by iteratively leaving two of the twelve subjects out for validation and testing. To compare training time we used similar instances of nodes with *Tesla K80* GPUs and 25GB of RAM, all training times are reported in seconds.

Model	Baseline	SVM	kNN	Fine-tuned FIN
Mean (± std)	.443 (±0.174)	.5233 (±0.016)	.525 (±0.018)	.543 (±.0245)

Table 3.6 Mean and standard deviation of the accuracy for the experiment I classification task. As demonstrated the FIN based network out performs both randomly initialized neural networks and classical statistical approaches.

3.3.4.1 Experiment I

In this experiment, we explored the potential of FINs for the detection of artifact ridden ECG signals [116, 59].

Data and Prepossessing We used data made available by The *Brno University of Technology ECG Quality Database* [116]. ECG segments of variable lengths from 18 subjects were classified by experts into three categories according to signal quality. After standardizing the lengths we ended up with 2544 trials. Preprocessed data will be made available.

Models We hypothesized that a FIN trained to imitate Kurtosis might be useful in the context of this task [198]. The Kurtosis FIN was adapted for our classification task by replacing the very last layer with a softmax classification layer. In addition to the baseline neural network, we also compare against several non deep learning classification algorithms.

Results As can be seen in Table 3.6, The Kurtosis FIN consistently outperformed the baseline models. Moreover, the standard deviation in the performance of the FINs was an order of magnitude smaller than the deep network baseline models. A Levene's test indeed indicates a statistically significant ($p < .05$) difference in variance between the performance of the two methods throughout the iterations. This highlights the fact that our framework helps with the robustness of the models.

3.3.4.2 Experiment II

In this experiment, we investigate how different FINs can be used in conjunction to build complex networks suited for EEG artifact detection. Moreover, we demonstrate how theoretical knowledge regarding the features and their relevancy to the task is helpful when using the FINs Framework.

FIN	Regularity	Fundamental Frequency	Entropy+ Regularity	Kurtosis+ Regularity	Baseline
Mean (± std)	.6527 (±.1066)	.6825 (±.0591)	.6991 (±.0662)	.7134 (±.0807)	.7142 (±.0587)
FIN	MFCC	Entropy	Kurtosis	Entropy+Kurtosis +Regularity	
Mean (± std)	.7167 (±.01411)	.7195 (±.03783)	.0.7214 (±.02397)	.0.724 (±.0451)	

Table 3.7 Mean and std of the accuracy for experiment 2. Corrected one-tailed t-tests demonstrated that models imitating features known to be useful for EEG artifact detection (last three columns) significantly out-performed models imitating ill-suited features (first two columns).

Data and Prepossessing The data used in this experiment is from an EEG artifact detection data-set [146]. The data contains EEG segments from a $1kHz$ recording made using 32 electrodes during a passive viewing task. Each segment is a second long and was labeled as artifact ridden or clean by expert annotators. We re-sampled the data at $500Hz$ and converted the EEG setup to the international $10-20$ system that contains only 19 electrodes.

Models We evaluated individual FINs and FIN ensembles trained to imitate Kurtosis, Shannon's Entropy, Regularity, Fundamental Frequency, MFCCs, and ensemble combinations thereof. We expect some of these FINs to outperform others based on the task-relevance of the feature being imitated. For instance, the fundamental frequency of the signal, defined as the lowest periodic frequency of the waveform should be irrelevant, while the Kurtosis is highly relevant to the task [37, 77]. Similarly, we expect 'Complexity Features' such as the cepstrum coefficients and Shannon's entropy to outperform clinically grounded 'Continuity Features' such as the fundamental frequency or EEG regularity (burst suppression) [146]. Following these theoretical considerations we hypothesize that the MFCC, Entropy, and Kurtosis FINS, as well as a combination of these FINS will outperform the Regularity FINs. As we have multiple hypotheses, appropriate Bonferroni correction for multiple comparisons was used. To have enough statistical power after the correction we increased sample size by repeating each experiment 50 times.

Each FIN was applied on each electrode signal in parallel, the outputs were then concatenated and passed forward to a binary softmax classification layer. We compared the FINs against the best

performing baseline dense neural network.

Results As summarized in Table 3.7, our experiment demonstrates how (when appropriately selected) FIN ensembles may be used in combination to further enhance task performance. We note here that deliberate consideration when combining FINs can improve task performance, while keeping the size of the ensemble small.

A corrected one tailed t-test showed that after correction the Kurtosis, Entropy, and Ensamble Network (last three columns in the table) performed significantly better than the Fundamental Frequency or Regularity FINs.

3.3.4.3 Experiment III

In this experiment, we compare the performance of FINs against state-of-the-art approaches for a fatigue and drowsiness detection from EEG task on a recently published data-set.

Data and Prepossessing We used data made available by [108]. This section identified multiple subsets of electrodes as especially predictive. We tested separately on every subset. The data was then partitioned for six fold intra-subject cross validation. In other words, at each iteration ten out of twelve available subjects were used for training, one was used for validation, and the testing accuracy was calculated on the remaining subject. Finally, each cross-validation step was repeated 5 times. If only a fraction of the training data was being used different trials were picked at each of these iterations.

Models Prior work indicates that entropy is a useful feature for the prediction of drowsiness and fatigue from EEG data [108]. Thus, we compared a FIN trained to imitate Shannon's entropy against the baseline models (described in the methods), as well as a fine tuned VGG network pre-tarined on the ImageNet data-set [160]. The VGG model was similar to the 19 layer convolutional neural network introduced in [160] sans the last 3 dense layers and with the addition of a final softmax classification layer. This is the standard way in which the VGG model is used in biomedical tasks [124, 192].

Additionally, to test the performance of our pre-trained FIN when only very limited data is available we ran the same models but with varying fractions of the data being made available

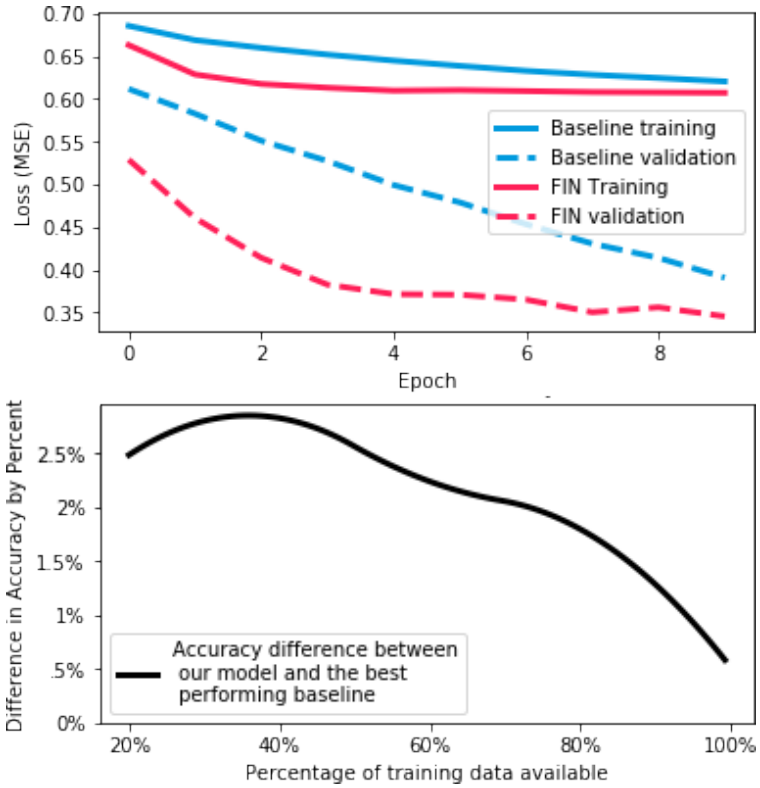

Figure 3.8 Experiment 3 results. (top): Training and Validation loss for the baseline and pre-tuned entropy FIN. (Bottom): Difference in accuracy between our pre-trained FIN and the best performing baseline as a function of the percentage of data available.

during training.

Results The pre-tuned Shannon entropy FIN outperformed the baseline and reinitialized FIN in each of the four subsets and at over 83% of the iterations of the cross validation. Additionally, as can be seen in Figure 3.8 (top), the pre-tuned FIN had lower loss at every epoch compared to the baseline. It is important to note that the FIN also beat the performance of classical classifiers with different entropy measures that was reported in the literature [108].

The performance improvements were even greater under limited data availability conditions As small training data sets are known to increase performance noise, we also repeated this process 10 times and reported the average accuracy; the difference in the average accuracy between our method and the baseline for each data percentage is plotted in figure 3.8 (bottom).

The VGG transfer learning network under-performed both the FINs and other baseline models and was particularly sensitive to small data sizes. When only a fraction of the data was available,

Mean (± std) Training Seconds	Baseline	VGG	FIN
20% of data	.746(±0.019) 37.2(±8.9)	0.615(±0.122) 21911(±520)	**.771**(±**0.016**) 50.8(±5.7)
40% of data	.895(±0.04) 37.2(±0.2)	.644(±0.18) 10881(±1441)	**.922**(±**0.013**) 106.0(±23.3)
60% of data	.939(±.07) 189.8(±79.9)	.69(±.197) 18434(±1058)	**.962**(±.**006**) 322.4(±201.2)
80% of data	.983(±0.0203) 574.8(±132.3)	.802(±.103) 17141(±1838)	**.996**(±.**0013**) 441.6(±92.2)
100% of data	.993(±0.022) 645.7(±187.1)	.846(±.088) 19267.3(±2014)	**.998**(±**0.001**) 552.6(±129.7)

Table 3.8 Experiment 3 results. The models were trained using varying subsets of the data. We report the mean and standard deviation for both accuracy and training duration on a node with a *Tesla K80* GPU and 25GB of RAM running *CUDA*.

VGG performed at close to chance.

3.3.5 Discussion

The feature imitating networks framework proposed in this section is an innovative way to use transfer learning. Traditional transfer learning requires large, slow to train, black-box, networks such as *VGG* and *AlexNet* tuned on hundreds of thousands of labeled data. In contrast, FINs require no human labeling, are small and fast to train, and can be combined to create ensemble FIN networks in accordance with insights from the literature surrounding the task being performed. Therefore, our network facilitates the integration of domain specific knowledge into modern data driven machine learning practices. This is also beneficial for alleviating some *interpretability* concerns; while the final FIN sub-models likely do not reproduce the exact feature they were trained to imitate after the end to end tuning, the fact that FINs perform better than a network with the same architecture that with weights initialized at random suggest that the tuned FINs are most likely computing a slightly modified version of the original statistical feature. Moreover, the fact FINs that are ill-suited for a task continue to underperform even after fine-tuning (see Experiment II) also suggests that our

the measures computed by the FINs do not drastically change after the fine-tuning. Beyond these considerations there are several practical benefits to using our framework:

- **Robustness:** Our experiments indicated that FINs are more robust than other networks and techniques with similar representational power. This is evident in statistically significant differences in variations in accuracy when performing leave subject out and cross validation. Deep learning in general is sensitive to weight initialization randomness and data idiosyncrasies. Transfer learning of weights tuned to calculate task relevant features seems to guarantee we start at a 'neighborhood' of a good solution. Moreover, FINs expedite the hyper-parameter optimization step which remains resource-consuming despite recent research [191, 107].
- **Performance:** Data scarcity still plagues many domains. In the case of biomedical research data collection is especially costly and can prohibit researchers from applying deep learning to their tasks altogether [147]. Our experiments indicate that FINs are useful especially when only limited data is available. The intuition behind this is straightforward; pre-trained weights already extract useful task-relevant information, resulting in a better performance and lower loss when from the very first epochs of the training procedure, as can be seen in Figure 3.8.
- **Flexibility:** By tuning on task-specific data sets our framework also out performs methods that pass the calculated features as input to the classifiers. This is not surprising as our FINs are allowed to tune the extracted features to better suit the task (for instance by focusing on specific parts of the signal). Additionally, the modular nature of the FINs lends itself to easily building and testing ensembles networks.
- **Speed:** *VGG* and *AlexNet* are powerful networks that have been successfully applied in various domains. However, these architectures are extremely large. The *VGG* based network consists of at least 17 layers and contains over 20 million parameters. In contrast, FINs are simple shallow networks consisting of up to 4 layers and a quarter of that numbers of parameters. The shallowness of the models in particular guarantees that even when using an

ensemble of FINs gradient descent calculations, and therefore training and inference times, remain simple and fast. This can be observed in the results of the third experiment presented in Table 3.8, training the VGG network was in some cases over 60 times slower than the FINs network training despite lower performance.

3.3.5.1 Future Directions

Designing the dense implementation of the FINs can be streamlined by considering the closed form equation of the signal and training layers to imitate each operator separately. For instance, the mathematical expression for Shannon's entropy requires discretization of the signal to create a histogram before averaging each bin. Partitioning the operations allows us to reuse pre-trained operation-specific layers to quickly construct FINs that are then fine tuned to mimic specific features.

3.3.6 Conclusion

In recent years, some have critiqued the current state of the machine learning community. These critiques often focus on disregard of traditional techniques in favor of data driven approaches [103] and the different ways deep learning have struggled to live up to it's promise, especially when it comes to real world applications [141]. In this section, we presented *Feature-Imitating-Networks*, a variation over traditional transfer learning that uses networks trained to imitate simple closed form statistical features, that we believe elevates these concerns. We demonstrated that our framework is superior in both the speed and accuracy to deep and transfer learning techniques with similar (or greater) representational ability. Especially when only very limited data is available. The experiments were conducted on a variety of tasks and domains. Future work will extend this initial exploratory work. An extensive library of Feature Imitating Network bench-marked on many data-sets and other signal processing domains might be of particular use and interest to the research community.

3.4 Conclusion and Future Work

This chapter presented multiple novel methods for unsupervised artifact detection and correction [145, 146], as well as a framework for integrating expert knowledge into deep learning models via weight initialization [143]. All presented methods were validated using unseen subject (and when possible unseen task) EEG data. The models were designed to streamline EEG preprocessing and decoding, and can help mitigate some of the challenges facing the research community; unsupervised artifact detection eliminates the need for a manual data examination by an expert, and artifact correction reduces the amount of EEG trials rejected. Finally, the modular nature of our feature imitating networks framework enables researchers to rapidly test different intuitions when decoding EEG data. While the methods achieve state-of-the-art performance on EEG tasks, each could have many potential uses in various other fields of research which share some of the difficulties inherent to working with EEG signals. For instance, the feature imitating network framework has already been utilized in natural language processing, computer vision, and even predicting athletic performance [83]. The models developed in these works are being added to a library of pre-tuned models which we hope will be useful for other researchers interested in building on our framework.

CHAPTER 4

PILOT: DECODING COLOR FROM PASSIVE VIEWING

Before focusing on attentional capacity, a small pilot study was first conducted. The goal was to determine the feasibility of color decoding from EEG signals. The experiment design was a simple passive viewing paradigm that used stimuli similar to what will later be used in the main experiment. This pilot study also laid important foundations for the work that follows by allowing the identification of electrodes and frequency bands that optimize decoding performance. Most importantly, results of this pilot study addressed concerns regarding the classifier decoding capturing verbal label, rather than feature information. Since the completion of this pilot study, a number of papers that decoded color from EEG have been published [66], confirming its main conclusions.

4.1 Pilot Experiment

The pilot was conducted with the eventual main experiment in mind. Therefore, this pilot experiment utilizes dot-field stimuli reminiscent of those used in [96].

4.1.1 Methodology

4.1.1.1 Participants

8 participants were recruited from the Michigan State University student body. The protocol was approved by the MSU institutional review board and written informed consent was obtained from every subject.

4.1.1.2 Stimuli and Apparatus

The experiment was programmed in MATLAB and using the MGL extension [51]. This experiment was a passive viewing task involving six different stimuli. Participants were directed to focus on a fixation cross in the middle of the screen. Each trial consisted of $1000ms$ stimulus presentations, followed by a $1000 - 1500ms$ inter-trial interval. The stimulus consisted of random dot fields (240 dots each) in six different colors. The dots were drawn in an annulus (inner radius=1.5°, outer radius=6°). In total there were 648 trials per experiment per subject (108 trials

per condition). We collected additional trials for some subjects but kept the balance between the different trial types. See figure 4.1 for an example of the stimuli.

Isoluminance Procedure To eliminate potential confounds, each subject adjusted the brightness for every hue separately to achieve isoluminance between all colors. The isoluminance procedure employed heterochromatic flicker photometry [90]. Observers viewed grey and chromatic square tiles (size: $1.8°x1.8°$, luminance: 6.3 cd/m^2) arranged in a checkerboard pattern and constrained within an annulus (inner radius=1.5°, outer radius=6°) centered around a white fixation cross. The gray and chromatic tiles flickered at $8Hz$ in a counterphase fashion. Subjects were instructed to adjust the brightness of the chromatic tiles to minimize the flicker, resulting in isoluminance between the color and the constant gray. We ran three separate blocks for each of the six colors. The three "isoluminant values" obtained were averaged to get the final luminance value for each color. These value were calculated for each participant and used during the main task in this pilot study.

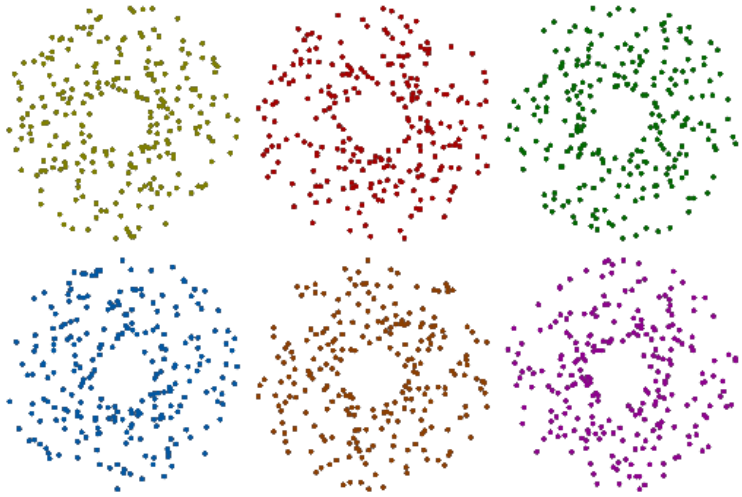

Figure 4.1 Examples of pilot experiment stimuli. The background color was changed for visibility (RGB value 240, 240, 240).

4.1.1.3 Data Acquisition and Preprocessing

Continuous EEG activity was recorded using the actiCHamp system with BrainVision recorder software. The participants were fitted with a 64-channel actiCap with active electrodes. The screen refresh rate was set to $120Hz$ and data sampling was at $1000Hz$. Additionally, electrooculogram

(EOG) activity was recorded from from horizontal and vertical electrode pairs, and used to detect and reject horizontal eye movements, eyeblinks, and vertical eye movements. Electrode impedance was maintained $< 50K\Omega$. The data from the inter-trial interval was discarded. We used EEGLAB and ERPLAB to process the data. First, we resampled the data to 500Hz, removed the AC line noise (using the cleanline plugin), applied a bandpass filter between 1 and 100Hz, and used ICA decomposition to separate and remove components originating from blinks and other artifacts. Finally, waveforms were manually examined by experimenters and noisy epochs were rejected.

4.1.2 EEG Data Analysis

We used the ADAM decoding toolbox default LDA algorithm with a 10 cross-fold process and 2000 iterations of cluster-based significance testing [44]. The data was not down-sampled or filtered beyond what was previously described. Different subsets of electrodes were examined. For detailed description of the LDA decoding algorithm see Section 2.2.2.1. The decoding and significance testing will also be discussed more thoroughly in the Methodology section of the main experiment. The results plotted in this chapter were achieved using the following electrode subset: Pz, P3, P7, O1, Oz, O2, P4, P8, TP10, T8, P1, P5, PO7, PO3, POz, PO4, PO8, P6, P2, CP4, CP2, CP1, CP3.

4.2 Pilot Results

We were successfully able to decode color from EEG data. Performance peaked when limiting decoding to sub-alpha frequencies ($< 10Hz$). Follow up decoding after time-frequency decomposition revealed significant decoding cluster only in this sub-alpha band.

4.3 Discussion and Conclusion

Having identified the setup necessary to achieve robust decoding of color, our main experiment will use the same setup to explore how attentional load impacts this decoding. For the sake of the following chapters, it is important to emphasize of two conclusions of this this pilot experiment. First, feature information is available mostly in the sub-alpha frequency band (below $10Hz$). While this was demonstrated for other features such as orientation and motion direction [7, 8], to the best of

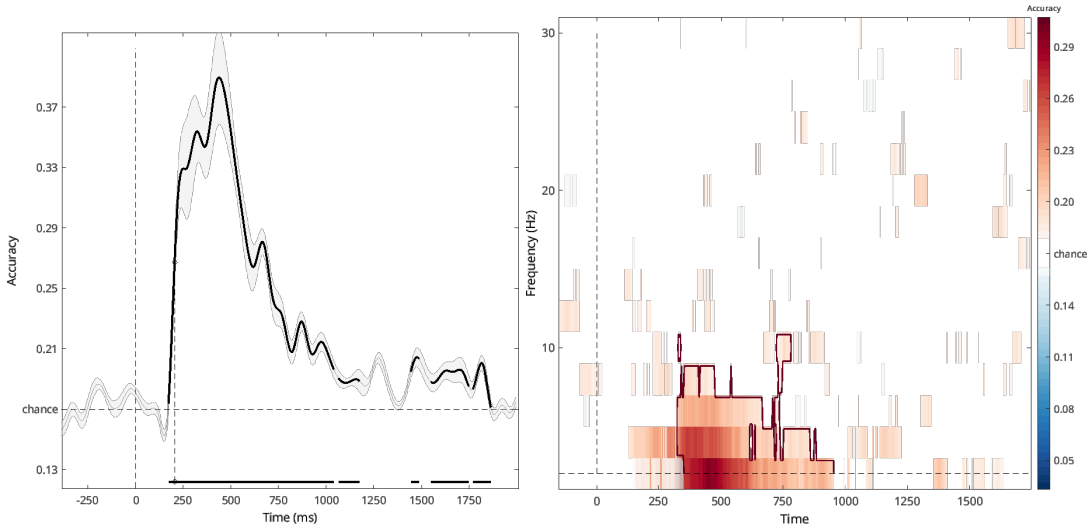

Figure 4.2 Right: Decoding performance using sub-alpha frequencies and visual-cortex electrodes. Horizontal black lines indicate significant decoding clusters. Left: Decoding after time-frequency decomposition, significant decoding clusters are highlighted in red.

our knowledge, this is the first experiment that demonstrates this to be the case also for color. Note that previous EEG studies that decoded color demonstrate that the color decoding reflects sensory qualities (such as a circular order with obvious color categories) and not, for instance, verbal labeling [66]. With all the above in mind we assume that decoding color from sub-alpha frequency band will reflect feature information, therefore we expect differences in attantional modulation of sensory representations to manifest in the classifier decoding accuracy. Finally, having tested different combinations, we conclude that the color feature information is carried mostly by the occipital and parietal electrodes. This is not surprising considering similar "visual-cortex" electrode subsets have been used in other EEG papers that decoded feature values [7, 8, 128]. Following these conclusion from the pilot study the main experiment will use a similar electrode subset and a $10Hz$ low pass filter will be applied before the feature decoding.

CHAPTER 5

THE NUMBER OF ATTENTIONAL TEMPLATES MODULATES SENSORY REPRESENTATIONS

Attention to a specific feature value enhances its sensory representation [25]. This everyday phenomenon is essential for the completion of search tasks and has been thoroughly researched [186, 173]. However, modulation of sensory representation when attending to *multiple* feature values remains the topic of much debate. Traditionally there have been two opposing views; Many researchers argue that there is a hard limit on attentional capacity (only one "attentional template" can be maintained at a time). On the other hand, experiments have demonstrated that multiple working memory items can influence behavior simultaneously. Recent theories reconcile these two views by separating the search task into different processes with different capacity limits [128] or arguing that task characteristics might be influencing attentional capacity limits [135]. Either way, the exact mechanism responsible for the difference in performance under varying "attentional load" conditions remains a mystery. Inspired by recent studies, we conducted Behavioral and EEG experiments to investigate the capacity of early attentional processes in tasks that demand active guidance by attentional templates (rather than the suppression of irrelevant distractors). Our study utilized a detection, rather than search, paradigm, which enabled us to explore the specifics of how attentional load impacts sensory representations using both behavioral modulations and their neural correlates. Results indicate that maintaining multiple attentional templates increases false-alarm rates while only slightly diminishing hit rates, and are overall incompatible with many versions of multiple-item templates theories. Generally, both the signal detection theory and EEG decoding analyses indicate that sensitivity deteriorated when maintaining multiple templates. Finally, analysis of behavioral performance and EEG decoding in target-present trials conforms with some current theories of attentional load [128, 136]. However, considering cue-condition effects are more pronounced in target-absent trials, further evaluation of current attentional load theory is necessary.

5.1 Introduction

A friend asks for help locating a misplaced notebook. One natural response could be "what is the notebook's color?". When color alone is not enough to go by, one might ask for additional characteristics (shape, size, etc.) and conjure up a *mental image* to *attend to* and guide the search. This is a real-life example of the single-item search paradigms that have been studied extensively in attention research. Attention to a target feature value enhances its sensory representation, enabling efficient completion of search tasks [148, 25]. This phenomenon of sensory representation modulation has driven the development of many influential theories of attention (such as Treisman's Feature Integration Theory [173] and Wolfe's guided search model [186]). The general consensus is that after being cued to attend a specific feature value an "attentional template" forms in working memory. This "attentional template" then enhances the saliency of targets matching the represented feature value [186, 38]. However, despite the phenomenon of guidance by a single attentional template being a cornerstone of modern attention research, the debate surrounding the modulation of attention by multiple working memory representations remains unresolved. Moreover, while the capacity of working memory has been studied for decades, research focusing on the capacity for attentional templates - the number of working memory items that can guide attention simultaneously - is mainly limited to the previous decade.

Broadly speaking, there have been two conflicting theories. Proponents of the (SIT) Single-Item-Template hypothesis [123, 74, 110] argue that only one attentional template can exist at a time. On the other hand, a growing body of literature seems to favor an opposing (MIT) Multiple-Item-Template hypothesis [72, 28, 9]. Behavioral, eye-tracking, and neuroimaging studies found evidence in favor of both theories. Moreover, despite utilizing very similar paradigms, researchers repeatedly arrived at conflicting results that withstood scrutiny via large-scale replications [47]. Recently there have been attempts to reconcile this conflicting body of literature. In a recent EEG study, the original authors of the SIT hypothesis suggest that participants can in fact maintain multiple attentional templates simultaneously, but only one can be effectively deployed at a time [128]. The conclusions of this study remain limited by the fact the authors decoded target location

rather than attended feature value, thus possibly missing modulations in early attention processes. Other attempts to reconcile the literature focused on the fact that, studies that support the SIT hypothesis often require active guidance by the working memory representations, while studies that support the MIT hypothesis focus on attention capture by distractors instead [84, 47]. Very recently, researchers have explored this possibility by slightly modifying a search task to either require distractor suppression or active guidance [135]. The authors concluded that, while multiple working memory representations can be behind distractor costs, only a single "active control set" that improves search performance can exist at a time.

Finally, diminishing performance during guidance by two representations (as opposed to one) was often interpreted as evidence in support of the SIT [128]. However, the exact mechanism behind this diminished performance remains a mystery. Does attending to multiple, rather than a single, feature values not enhance the sensory representation to the same degree? Or are participants more prone to errors when the target feature value is ambiguous? Previous attention capacity studies focused on search - rather than detection - paradigms. And while search paradigms are very common in general attention research, they are ill-suited for shining a light on the exact nature of the performance differences observed between single vs multiple cue conditions (see section 5.2.2). In one exception, researchers investigated modulations in the psychometric function under different attentional template load conditions [96]. The authors demonstrated that the target signal in multiple template trials must be stronger than in single template trials to achieve the same level of performance (as measured by hit-rate minus false-alarm). However, the exact way in which the number of attentional templates modulate sensory representations is yet to be thoroughly investigated.

Considering the recent studies discussed above, it becomes imperative to refine and reformulate the original question. First, following [135], we differentiate between distractor interference and active guidance by working memory items. To directly investigate how the number of templates *actively* guiding attention impacts sensory modulation we modified experiment 2 from [96]. Additionally, to assess the cueing effect modulation directly we decode the cued feature value. We

hope that by decoding the stimulus feature matching the working memory content - as opposed to stimulus location [128] - we will be able to capture early attention effects. The authors of [128] argued for different capacity limits during the template maintenance and deployment processes. They observed delayed and suppressed decoding in two-cue-one-target compared to single cue trials, an effect they attributed to a bottleneck in the template maintenance stage. However, there is no guarantee that deployment when multiple templates are available is indeed comparable to single attentional template deployment. Therefore, the conclusion that their decoding results reflect limitations in template maintenance capacity is contestable. In simpler terms, the differences in location decoding observed by the authors of [128] could reflect attentional load effects on later processes and not necessarily a template maintenance bottleneck. In contrast, our paradigm enables us to investigate early differences in perception via modulations in representations of the attended feature value (instead of a subsidiary target location). Moreover, the detection paradigm we employ enables us to examine how number of templates impacts the decision making process. Finally, very few studies had conditions with more than two attentional templates (one notable example is [84]). Attentional guidance might diminish with the number of templates (following a $\frac{1}{num-cues}$ decay rate as observed in [96]), or there might be a sudden drop when the number of templates exceeds two. To explore this further, the behavioral session of our study includes one, two, three, and no-cue conditions.

5.2 Literature Review

5.2.1 Behavioral Studies

As previously mentioned, many paradigms in attentional capacity research are cued search tasks. More specifically, these paradigms are often some version of a nested Memory-Search task (see figure 5.1). In this paradigm, subjects are asked to remember a feature value for a memory task while first performing a simple cued search task (for which the remembered feature value is irrelevant). Under some conditions, the search array contains items with feature values matching the remembered item (memory cue in figure 5.1). In addition to these "matching distractors" trials, "mismatched distractors" and "no distractor" trials are used to measure baseline performance. Longer response time for trials containing "matching distractors" (in comparison to "mismatching distractors") indicates that the remembered feature value interferes with the search task. Since the search cue is also guiding attention, this would indicate that multiple items can indeed drive attention simultaneously. Trials in which the subject answered the search task or the memory task incorrectly are usually excluded from the analysis, as this could indicate failure to memorize the cue or find the search target. One variation of Memory-Search tasks are nested Memory-NoCue-Search paradigms. Instead of the search being guided by a cued feature value, the search target is a consistent item across the experiment. The paradigm in figure 1 can be modified to a Memory-NoCue-Search paradigm by removing the "search cue" and asking subjects to always find the triangle in the array. While the difference between the two paradigms might seem minute, this difference has become central to recent debates that we will discuss in future sections [135].

Downing and Dodds were among the first to use a nested Memory-Search paradigm [42]. Trials started with two cues, one for a search task that immediately followed and another for a later memory task. The researchers found that even when the distractors in the search task were identical to the memory cue there was no increase in response time, indicating that the remembered item did not bias attention. This conclusion was in direct conflict with other contemporary studies [164, 122]. Both of these studies used a Memory-NoCue-Search paradigm. For instance, in [122], participants remembered a color while performing a search task in which the target value

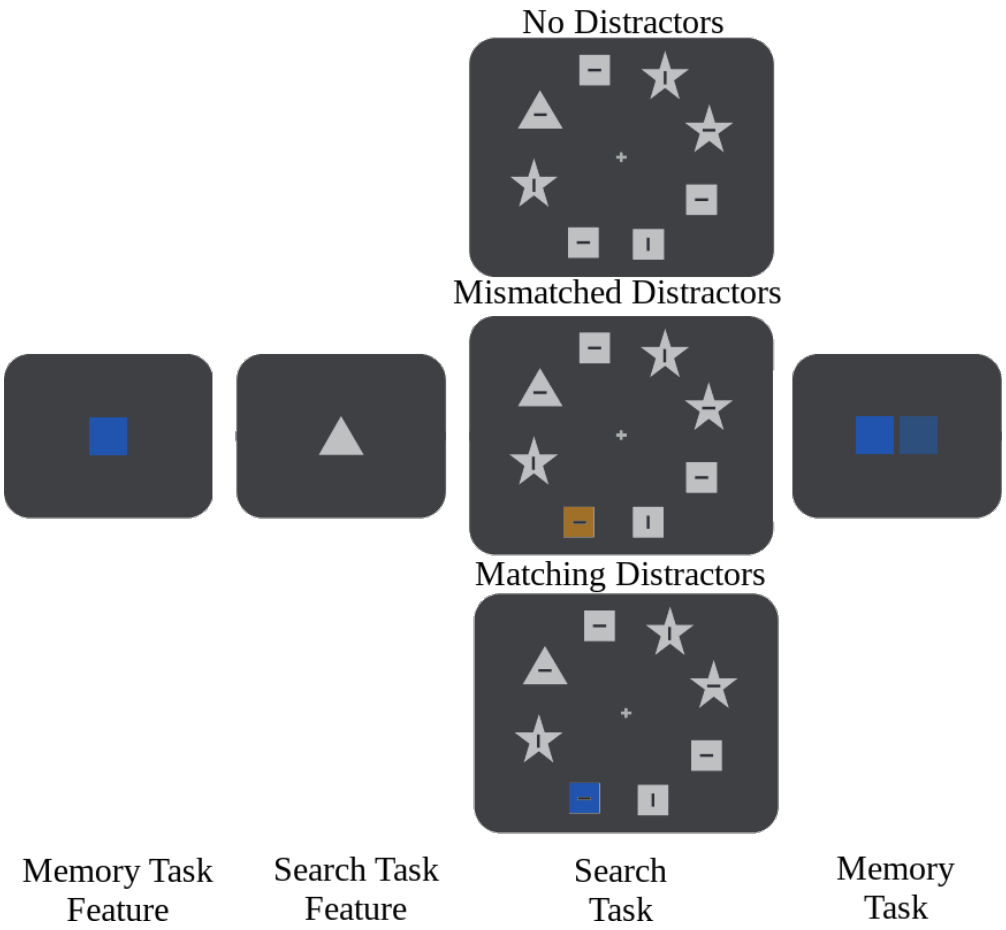

Figure 5.1 Example of a Nested Memory-Search task, stimulus order from left to right. The subjects are directed to remember two feature values (blue and triangle). One of the features is relevant to an immediate search task (the shape) and another for a later memory task (the color). The search array task is to report the orientation of the line in the target item. The Memory is to report the color matching the memory sample.

was constant throughout the experiment. The search task had different distractor conditions (see figure 5.1). Analysis showed that "matching distractors" captured attention significantly more than "mismatched distractors". Additionally, when the memory test was conducted before the search task the difference between matching and mismatching distractors disappeared. Together these results indicate that only task-relevant working memory items bias attention. And secondly, the behavior modulation is not simply due to some sensory after-effect, but is driven by a working memory item that is actively maintained (relevant to a future task).

The discrepancy between these results was replicated by other studies in this time period. Olivers later reviewed these three papers as well as two others and discussed design similarities

and differences [121]. Olivers argued that other than search task difficulty the only difference in experimental design that correlates with finding distractor-interference from working memory items was a lack of "varied mapping". In other words, experiments with a constant search target (Memory-NoCue-Search paradigm) found distractor interference, while experiments with a search cue (Memory-Search paradigm) did not.

Olivers did not elaborate further in [121] on the mechanisms that might be behind the importance of a block-consistent search target. However, this review is an important precedent to the attentional capacity debate. The conclusion suggested that maintaining multiple items (as required by the Memory-Search paradigm) impacts working memory modulation of attention. The first paper to explicitly frame the discussion in terms of "limits of attentional capacity" was another review from Olivers' lab [123]. The authors of this work coined multiple definitions that became fundamental to the attentional capacity debate. Most importantly, they established the main components of what is now known as the single-item template hypothesis (SIT)

- Working Memory Representations guide attention via "templates". This statement is widely accepted. Moreover, the concept of "attentional template" is essential to many theories of visual search that preceded the attentional capacity debate (consider Wolfe's guided search [186] or the biased competition theory [12]).
- When there is only one item in the working memory, such as in [122], this item becomes "active" and will become an attentional template that influences even irrelevant tasks. If the item is not relevant to future tasks it will no longer be active and will not influence attention (again, see [122]).
- When there are two items in working memory, such as in [42] or other papers discussed in [121], only the representation that is relevant to the immediate task is made active. And the irrelevant representation exists in a different passive working memory state that does not impact the deployment of attention or influences behavior.

Beyond simply suggesting that some 'accessory' representations in visual working memory do not drive attention, this review commits to the notion of a "hard limit" on the capacity of attention.

Moreover, the authors explicitly argued that only a single memory representation can become a search template, and this representation will then "block" attentional guidance by all other memory representations.

The same authors later went on to make an addition to this SIT hypothesis. Using a Memory-NoCue-Search paradigm they demonstrated that when the memory task involves multiple items no “attention capture” is observed [110]. Crucially, attentional capture was observed in a condition with a single memory item. In other words, results indicated that in conditions with more than one memory cue, all relevant to the memory task, there was no reaction time difference between “matched” and “mismatched” distractors (even when distractors matched all the memory cues). This result does not simply follow from the SIT hypothesis: If only one item out of three was active, one would expect attentional capture during one-third of the trials. The effect sizes would only be reduced but not disappear completely. The complete lack of evidence of attentional capture prompted the researchers to conclude that when multiple items are held in working memory they automatically compete with each other, preventing any item from becoming an attentional template and eliminating any memory contents driven attentional capture. In other words, multiple representations compete to be the active trace in a mutually detrimental fashion.

Despite the success of the SIT hypothesis in making sense of previous conflicting results, subsequent research quickly challenged the core notions of this theory. The authors of [72] showed, using a very similar paradigm to the one used in [110], that even when subjects had to remember multiple colors for a subsequent memory test, there was evidence for attentional capture by matching color distractors during the search task. Moreover, not only was attentional capture observed for either color in two-cue trials, it also appeared to be stronger than attentional capture during one-cue trials. Specifically, there seemed to be a compound effect when distractors matching both cues were present. Despite using a nearly identical paradigm, the results here are in direct contrast to the results of [110].

Furthermore, taking the original SIT hypothesis (before the addition of this competition component) at face value, one would expect reduced memory capture. Since [72] found the exact opposite,

a compounded effect, even the original more conservative version of the SIT hypothesis does not accommodate this result. The discrepancy surprised the research community and large-scale replications of both papers followed [47]. However, both studies were successfully replicated. The authors of the replication hypothesized that this outcome might be due to a difference in paradigm: [110] had only a single distractor in the search array and the delay period between the cue and the search task was 1250ms, in contrast, the paradigm in [72] contained multiple distractors and the delay period lasted only 700ms. The number of distractors is an unlikely culprit, as [72] found an effect even when only a single distractor matched the memory cue. However, the delay period duration could be relevant as there is evidence that the strength of memory representations diminishes over time [40]. However, based on the growing body of literature, delay period duration also fails to provide an adequate explanation. Two similar papers [28], and recently [84] showed that in Memory-NoCue-Search tasks (virtually identical to the one used in [110]) when two and even three items are maintained in the visual working memory all representations bias search results. This further supports the idea that the number of distractors in an array is not the culprit behind the conflicting results reported by [47]. Moreover, in the three experiments reported [28] the delay period varied between 300 and 2000 ms, but the same behavioral patterns emerged: response times were longer for memory-cue matching distractors. These results demonstrated that subjects can maintain multiple working memory representations simultaneously and that all of these items, in-fact, guide attention to some extent. Finally, even previous experiments such as [122] that reported attention capture by WM items had delay periods as long as 3,000 ms. It is unlikely that storing two templates instead of one would degrade the effect to the point that it becomes undetectable within 1250 ms.

Other studies completed by Beck and Hollingworth over the years arrived at similar conclusions [13, 9]. In their most recent paper [9] the authors conducted a series of experiments in which the feature dimensions of the memory and search cues were (unlike for instance [42]) different (color and shape, see figure 1 for an example). Response time increase for matching distractors indicated robust attention capture during the search task by the irrelevant feature dimension memory cue.

These results have prompted the authors to suggest the Multiple-Item-Template (MIT) hypothesis which states that the limit of attentional capacity is flexible and at the very least two attentional templates can be active simultaneously.

Finally, instead of focusing on single-item search, the authors of [86] used a foraging task. Subjects had to tap on target-colors circles on an iPad while avoiding distractor-colors circles. This foraging paradigm is designed to mimic animal foraging settings. The SIT hypothesis would predict longer reaction times due to *template switching cost* when the number of target colors increases from 1 to 2, but no such increase should occur for further increase. The pattern of results in [86] shows an almost linear reaction time increase when the number of target-colors increased, despite the number of total targets remaining the same and regardless of the number of distractors. The authors of [86] argue that this indicates that some evidence taken to support the SIT hypothesis might in fact just be the result of increased cognitive load. It is important to note that one can argue that an increase in reaction times between 2 and 3, as well as 3 and 4 target types, does in fact support the SIT hypothesis, as long as the subjects focus on each target type separately before moving to the next. However, this behavior of focusing on different target types sequentially was not observed in previous foraging studies conducted by the authors.

Finally, a very recent study by the authors of the MIT hypothesis used a paradigm where participants would benefit from multiple template usage [10]. Unlike previously, here performance increase (rather than decrease due to distractor interference) would support the MIT hypothesis. The task was a search paradigm with two search cues (color and shape). Targets could match either one or both cues. Response time was faster when the target matched both cues (the target had both the cued shape and color) as opposed to only one (shape or color). Moreover, the response time distribution in two-cue-match-target trials violated the "race model inequality" which indicates "coactive" guidance by both cue's attentional templates as opposed to a parallel but separate or sequential activation of two mechanisms [111]. Of course, proponents of the SIT hypothesis might argue that this pattern of results would be possible if participants were always simply using a single item. If participants were attending to the feature value that does not match the target half of the

time their response times would increase. Since any cue always matches the target when it has both the cued shape and color, this condition will always have faster response times. To reiterate, the "race model inequality" uses the sum of the probabilities observed in the single-cue-match-target conditions to calculate an upper bound on the response-time probabilities that should be observed in the two-cue-match-target condition if no coactive mechanisms are involved. However, if half the single-cue-match-target trials are "corrupted" by a switching cost the calculated upper bound might not be accurate.

At this point, one might be under the impression that an overwhelming number of studies draw conclusions that are in stark contrast to the core components of the SIT hypothesis. However, there are a few things to consider; As the authors of [84] point out, in the paradigms used in these studies no cue is 'actively' guiding visual attention during the search. This raises the questions regarding whether one can indeed maintain multiple attentional templates, or if in the absence of such a template all contents of working memory (which Olivers et al referred to as accessory memory items) are able to drive attention. Interestingly, [84] seem to believe the latter 'capacity to actively maintain irrelevant items during visual search is higher than our capacity to actively maintain relevant items for a visual search.'. In other words, while, as noted by [47], there are still contradictory results even for very similar paradigms, there might be a way for a relaxed or modified version of the SIT hypothesis to explain these results as well. This idea was expanded in a few studies we will discuss in the following section.

5.2.2 Challenges for Nested Search-Memory Paradigms

As the reader has no doubt gathered, nested Memory-Search paradigms (with and without a search cue) seem to dominate the attention capacity research. However, these paradigms have several clear limitations. Firstly, many influential theories of attention, as well as recent "attentional capacity theories", speculate the existence of analogous fast parallel stage followed by a separate slow stage in which potential target locations are sequentially evaluated [173, 186, 128]. As previously discussed, attentional load could affect both stages, making it difficult to attribute target-location decoding or behavioral differences specifically to template maintenance. In other words,

search-paradigms do not provide direct evidence of early differences in sensory modulation under different attentional capacity conditions. The issue is further exacerbated when participants are required to complete an additional task (such as reporting the orientation of a bar or letter [122, 121, 28, 9, 84], also see Figure 5.1), as previous research demonstrates that working memory load can impact the processing speed even in unrelated tasks [65]. Secondly, while it is well known that attentional templates improve performance by enhancing the sensory representation of the attended feature value [25], the exact impact of load on how attention modulates sensory representations is an important component that remains missing when focusing only on search-based paradigms. Analysis of search based paradigms focuses on reaction time and accuracy. In contrast, detection tasks enables, for instance, detailed analysis of the decision making process via signal detection theory models. One interesting example of such modeling can be observed in [74]. Participants in this study were cued to look for one or more items in a subsequent rapid serial presentation (an RSVP paradigm). The authors modeled the ROC curves expected under the SIT and MIT hypotheses, and estimate the number of attentional templates for each subject. Behavioral results were most consistent with having only a single attentional template. Furthermore, these results persisted when participants attended to objects instead of colors, as well as a mix of these two cue types, supporting the theory that working memory representations might be object-based [46]. One possible explanation for this surprising result might be found in individual differences. As noted by the authors, almost half the subjects had faster reaction time in the cue (in comparison to the non-cue) condition despite the cue being irrelevant to the search task [46]. This results can not be explained by either the single or multiple item template hypotheses, and might point out that attentional capacity modulation might simply have a small effect size that could be masked in nested search-memory paradigms by noise or individual differences.

Another relevant signal detection experiment can be found in Liu et al 2017 [96]. The authors attempted to test if multiple attentional templates are maintained when their guidance directly benefits the task, as opposed to negatively impacting the behavior via distractor interference. The task was to indicate if one of the six colors in a cloud of colored dots was more coherent (had more

dots) than the others. In some conditions, cues indicated which colors might be coherent. For instance, when green and red were cued the subjects knew that if there will be a coherent color it would be either one of those two. There were both one and two cue conditions. If multiple templates can indeed be maintained then the increase in performance should be similar for both conditions. However, analysis indicated that only one of the two cues guided attention. This study builds on the results of a similar earlier work by the authors that showed decreased performance when attending to two motion directions relative to only one [97]. However, since the lower accuracy could be due to splitting resources between two templates instead of attending only to one, [97] did not fully commit to the interpretations supporting the SIT hypothesis. Crucially, [96] found that the two-cue condition was equivalent to a condition with a 50% valid single cue. Suggesting that participants are able to successfully attend only to a single cue. Therefore the authors argue that their results (and possibly also [97]) support a strong limit on attentional capacity.

Finally, recent research uncovered another disadvantage particular to the Memory-NoCue-Search paradigms common in the attentional capacity literature. Throughout the literature two types of results were interpreted as supporting the MIT hypothesis: interference from multiple representations, and performance enhancement when multiple attentional templates are maintained. However, it is possible that this conflates two separate phenomena. Indeed, experiments in which guidance by working memory items would enhance performance (relative to a no cue baseline) seem more supportive of harder attentional capacity limits [96, 74]. In comparison, experiments that test for distractor interference [72, 28, 84]. This pattern of results was systematically tested in a recent study [135]. The authors ran experiments using both the classic Memory-NoCue-Search paradigm and a slight variation that require participants to "actively search" for items matching working memory content. Their analysis demonstrates a significant distractor effect in the first experiments, even for distractors that do not match WM content (not to mention multiple WM representations). However, analysis of results from experiments with the varied paradigm indicate that participants were able to adopt a WM "active state" for only a single item. The authors conclude that while (possibly multiple) working memory representations can bias distractor costs, only when

participants are actively using a WM representation an "active control set" is created to improve search performance for a *single* target stimulus.

Considering the above, the next few sections will focus on eye-tracking and neuro-imaging papers, as nested memory-search paradigms are significantly less common in these modalities. Finally, we will discuss the challenges inherent to the nested search-memory task in the context of the detection paradigm used in our experiment.

5.2.3 Eye-Tracking Studies

Shortly after proposing the SIT hypothesis, the authors of [123] presented eye-tracking results supporting their theory. A straightforward conclusion of SIT is that when the search criteria isn't defined by a single feature-value (for instance, targets are in either one of two different colors), participants begin first employ one attentional template and eventually switch to a second template. Following this logic, the SIT hypothesis expects some evidence of *switching cost*. To explore this matter, [40] introduced a novel paradigm: a search array with two targets (at the left and right sides of the visual fields) was used. The color of the two targets was either the same (one-color condition) or different (two-color condition). In the two-color condition, targets at each particular side had a consistent color (for instance, the left target is always red). Additionally, distractors could be the color of the target on the opposite side of the visual field (for instance, if the left target is always red and the right target green, then distractors at the left could be green). Analysis of subject eye movements showed that saccades to the second target were slower and less accurate in two-color trials. Moreover, initial saccades often landed on other target color distractors, despite the color never being task-relevant in that half visual field (throughout the entire experiment). Time course analysis revealed that the proportion of saccades ending on the distractors trended down after 250-300 ms. The authors concluded that a template-switching cost does exist and is responsible for this pattern of results. Specifically, they conclude that fully switching attentional templates requires at least 250ms, and until then attention capture by distractors matching the feature value of the initial template could occur.

It is interesting to note that eye-tracking studies have previously demonstrated that multiple

items in long-term memory can guide search simultaneously [169]. The authors of the MIT hypothesis modified the paradigm used by this earlier study [14]. A search array containing 32 circles in 4 different colors was presented, two target colors were cued before the search array began and the participants had to fixate on all target color circles as fast as they could. The subjects were instructed to search for targets either sequentially (moving on to targets matching the second cue only after finding all targets matching the first) or simultaneously. Researchers coded segments of sequential fixations on targets of the same colors. The switching cost was analyzed by comparing the length of the fixation between these segments to the fixations at the beginning/end of these segments. Analysis revealed that there was a significant switching cost when subjects were instructed to search for targets sequentially, but no such cost was observed in the simultaneous search condition. There was no significant difference in the mean fixation time between the sequential and simultaneous search conditions, indicating that higher rates of switching did not incur any additional costs. Hence, it is unlikely that only a single template, which is being constantly switched, is maintained. The authors concluded that it is likely that there is no single item bottle-neck between working memory and attention, and that failure of previous studies to observe multiple-item guidance is simply due to the participants approaching search tasks serially instead of looking for multiple target cues simultaneously.

Recent technical developments in the field of eye-tracking enable eye-movement measurements beyond duration and location. New "gaze-contingent" studies enable researchers to tailor the experiment trial by trial to test hypothesized behaviors. These experiments usually require participants to fixate on one of the multiple items in a search display, and subsequent trials are generated in real-time based on previous participant behavior. An interesting application of this technology is another more recent study conducted by Beck et al [13]. In this paradigm, two target colors were consistent throughout the block. Each trial began with two circles, at least one of which was in a cued color, the participant had to fixate on a cued color circle to continue. Every subsequent trial belonged to one of three types: 1) forces the participants to pick the same cue color multiple times by only presenting one cue color among different non-cue color targets. 2) forces the participant

to switch by not presenting the cue color fixated on in the previous trial. And 3) present both cued colors and let the participant choose. Analysis of participant behavior showed that, when possible, participants switched the color from the previous trial an average of 46% of the time. Bahle et al argued that if only a single attentional template was active at a time the template matching the color used in the preceding trial would drive subsequent selection even if target-types matching the other template are available. This would result in very few switching between colors. The authors argue that the high switch rate implies that attentional templates matching both cues are simultaneously driving attention.

In response, the authors of SIT conducted a similar gaze-contingent experiment [126]. Virtually the only difference being that this experiment contained multiple distractors instead of only two circles at each trial, as before there were three possibilities when generating each new trial (only a match to cue color A, only a match to cue color B, or targets matching both colors). The authors again found a high 37%switching rate when both cues were presented. However, further analysis showed longer pre-eye-movement fixation periods in trials that forced participants to switch relative to the last fixation cue color. The authors argue that this reflects a higher "switching cost" when subjects were forced to change the state of the working memory items. According to the MIT hypothesis there should not be any bias to any of the two cues, contradicting these results. The authors argue that the lack of a switching cost when both targets are available, and the high switching rate, are the consequences of participants spontaneously switching targets between trials.

While the contradictory results above might seem confusing, there is room to critique both of these papers. The authors of [126] correctly point out that template switching can occur between trials and a high switching rate does not unequivocally support the MIT view. In the other hand, the switching cost at the heart of their argument in favor of the SIT hypothesis can be driven by selection history rather than bias towards the most recently active template. Recently selected items tend to be favored in subsequent trials, even in non-search experiments and when targets are fixed across blocks [188]. This selection history effect was not controlled for by any of the aforementioned experiments, and seemingly, there is no way to differentiate between this effect and

attentional template guidance in the aforementioned paradigms.

5.2.4 Neuroimaging Studies

In their original paper presenting the SIT hypothesis [123] the authors postulate different neuronal mechanisms that can account for their theory. Mainly, they focused on previous rhesus single-cell recordings studies by Warden and Miller [183, 182]. Results indicated that item working memory representations change when a newer item is introduced. Olivers et al argue that this change to an "orthogonal representation" might reflect an "active" item undergoing a transition into a "passive" working memory state. Subsequent research found further support for multiple working memory states. For instance, Lewis-Peacock et al [91] found that the identity of a single task-relevant item could be successfully decoded from fMRI activity. Other items, however, were initially decodable but became "deactivated" after a newer item that is relevant to an immediate task was introduced. Crucially, these same items become reactivated and decodable once they became task-relevant again. Most recently, Olivers' lab published an EEG study that demonstrated that it is possible to decode the status (active vs passive) of the content of the working memory [179]. The authors showed subjects a cue followed by two search tasks. The cue was relevant for the first task in half the trials and the second half in the other, and participants knew the type of the trial in advance. This paradigm encouraged the participants to switch the status of the working memory contents (cued feature value) from passive during the irrelevant search task to active (and vice versa). Analysis showed that the status of the working memory content can be decoded from a burst of power in the delta band (2-4 Hz), and a longer non-lateralized alpha (8-14 Hz) power. The delta decoding was brief and the authors concluded that it reflects the transition processes involved in changing the status of WM contents. However, this experiment is not without its issues. Participants knew which of the two tasks is upcoming, making it possible that the decoded working memory status is in fact related to processes involved in task-specific preparations. To put it simply, the tasks were to search for a cued memory item or a duplicate color. Since the subjects knew which task was coming up, decoding the delay period before the presentation of the search array might be "corrupted" by task-specific preparations unrelated to the status of working memory contents.

More specifically, the irrelevant search task always involved looking for a duplicate color, instead of a specific target, and did not involve any attentional template. Therefore, the delta band activity could therefore possibly relate to the existence of any attentional template (even if its from long term memory), or even a preparation specific to duplicate-search, rather than the status of the working memory contents.

Notwithstanding, while these papers make a strong case for the existence of different working memory states, as pointed out by many researchers [86, 47], **this by itself does not necessarily imply that only a single attentional template is active at a time**[1].

To measure the attentional capacity more directly several EEG studies have been conducted. These studies often borrowed from previous EEG research. Two specific EEG components that were repeatedly used to explore attentional capacity are the *Contralateral Delay Activity (CDA)* and *N2pc* EEG components. The CDA component is understood to reflect the number of items maintained in the visual working memory [101]. While the N2pc component is a transient contralateral signal that tracks the spatial deployment of attention in the visual field. N2pc is particularly useful for measuring attention capture by distractors (as long as the distractor and target appear on opposite sides of the visual field). Attention capture by distractors produces reliable N2pc components that do not appear in distractor-free trials. Both the CDA and N2pc are examples of an *Event-Related Potential* (ERP) obtained from EEG recordings. Both ERP components are measured by subtracting ipsilateral activity from contralateral activity with respect to an electrode (usually PO8 or PO7) and the cue location. For instance, N2pc contralateral activity is recorded from either 1) *right* electrodes (such as PO8) on trials with distractors on the left side of the visual field. or 2) *left* electrodes on trials with distractors that appear on the right. One crucial difference between the two components is that, while the N2pc ERP peaks around 200ms after stimulus onset, the CDA activity is sustained during delay periods in working memory tasks.

One example of the use of ERPs to investigate attention was a study by Gurbert et al. CDA

[1]The authors of [86, 47] also argued that a single item attentional bottleneck would probably require a centralized visual working memory specific neural mechanism, which is unlikely considering the apparent distributed nature of working memory. But discussion regarding the nature of WM representations is somewhat beyond the our scope.

components in trials with a changing cue were significantly larger than trials in fixed cue trials [62]. This was expected as processing fixed cues can be delegated to non-working memory processes such as long-term memory [24]. More interestingly, CDA components were larger in two-cue (in comparison to one-cue) trials, and larger still in three-cue trials. The authors argue that this indicates that all cues are represented simultaneously in working memory. However, N2pc components were attenuated in the multiple cue conditions, becoming smaller and delayed, indicating impairments in the deployment of spatial attention, and that multiple cue representations were less effective in guiding search. Further analysis revealed that there was a significant correlation between behavioral measures such as response time and N2pc, but not CDA, component characteristics. The authors conclude that this indicates that decreased performance in multiple cue trials (as measured by longer RTs) is driven by worse attentional guidance and deployment and not the cognitive load of having to maintain multiple templates. The authors of [62] concluded that these findings support the SIT hypothesis.

Perhaps inspired by the notion of separating between deployment and maintenance, Olivers' lab conducted their own neuroimaging study [128]. The authors decoded the location of the target from EEG activity during a guided search paradigm with one-cue one-target, two-cue one-target, and two-cue two-target conditions. The authors found that classification accuracy (as well as behavioral measures) was similar for trials from the first two conditions (though slightly worse for the second condition) but significantly worst in two-cue two-target trials. They argue that this demonstrates that deploying two templates as opposed to preparing them is the true bottleneck. This conclusion would explain the difference between the CDA and N2pc ERP modulation in multiple cue conditions observed by Gurbert et al [62]. This research offers a possible resolution of the SIT vs MIT debate by postulating that multiple item templates can co-exist in a mutually suppressive manner, and when a stimulus that matches one of the two templates is presented the matching template strengthens at the cost of the unmatching one. However, when stimuli match both attentional templates (cues) there is a strong mutual suppression '*resulting in a substantially weakened and delayed selection*'. In other words, multiple templates can be maintained but not

selected simultaneously. This theory has similarities to popular models of attention. For instance, Treisman's Feature Integration Theory [173] and Wolfe's original guided search model [186] both have a quick parallel stage followed by a slow serial process with a strong bottleneck. However, these theories are all based on single-item search paradigms. Moreover, in more recent iterations of the guided search model, Wolfe proposed two completely separate attentional template mechanisms [187]. This conceptualization holds that, while guiding templates from working memory items are used to select objects based on features, items in an "activate long-term memory state" might capture attention nonetheless. While Oliver's new theory can account for some of the previously mentioned results, it fails to do so fully. For instance, this theory would predict that in [96], which had conditions only similar to the first two in this work, no strong difference between conditions would have been found. Moreover, as Ort et al decoded only the target location during the search task [128], there is no evidence that the two templates were actually maintained in an active state before the stimuli presentation. In fact, the small difference between the first two conditions could be accounted for by the original SIT hypothesis as simply a template becoming active in working memory (switching cost).

Recently an EEG study attempted to decode attentional templates directly[184]. Participants were cued to suppress a specific color during an upcoming search array. Maintaining a negative attentional template is beneficial for search performance. The authors had fixed and varied cue blocks. Decoding the to-be-suppressed color was sustained in the fixed cue condition and decoding strength correlated with search performance. However, in the varied cue condition, decoding happened only at the onset of the delay period and was negatively correlated with performance. The authors argued that this indicates that negative templates do not form under the varied-cue condition. Another possible interpretation is that attentional templates during delay periods are simply difficult to decode. This is a clear hurdle for anyone hoping to verify the simultaneous maintenance of multiple attentional templates [128], as decoding seems difficult even in single-template trials.

One other recent EEG experiment found similar evidence of multiple coactive attentional

templates [63]. In this work, the authors alternated the search task cue in an ABAB fashion across trials (red or green). Before presenting the search array seven task-irrelevant arrays with distractors matching the cue colors appeared. Knowing that one of the cues is irrelevant to the upcoming task, the researchers expect evidence of attention suppression. Specifically, if only the task-relevant color template is active then distractors matching the other color should not capture attention. However, analysis showed that distractors for both colors captured attention (evoked a significant N2pc component) in all but the last array out of the seven. The N2pc activity evoked by irrelevant color distractors was heavily suppressed in the array immediately preceding the search task. The authors argue that this indicates that two attentional templates were maintained until shortly before the search task began, at which point top-level processes suppressed the task-irrelevant template.

Other relevant studies attempted to explore working memory representations more directly using steady-state-evoked potentials (SSVEPs). When flickering a stimulus at a specific frequency, activity in early-visual area neurons representing the stimuli matches the SSVEP frequency. Moreover, attending to stimuli has been shown to increase the amplitude of the respective SSVEP oscillations. By having subjects attend to objects in two colors oscillating in different frequencies researchers have been able to confirm that, at the very least, attending to two colors increased the corresponding SSVEP for both colors simultaneously [5, 105]. While this seemingly supports a multiple attentional template theory, as noted in [127], a direct comparison between SSVEP modulation during one-cue and two-cue conditions is needed before any strong conclusions are taken.

5.2.5 Conclusion

As the reader might have gathered, despite attentional capacity research being mostly limited to the last decade, there is already a substantial body of literature. Before delving into the details of our experiment, it is worth highlighting some key take-ways that can help contextualize subsequent chapters. First and foremost, as evident from this literature review, and experimentally established in [135], there seem to be a fundamental difference between distractor interference and guidance by working memory templates. The overwhelming majority of experiments supporting the MIT hypothesis did not require participants actively search for multiple cues, instead focusing

on interference by distractors matching working memory contents. Secondly, as discussed, search tasks provide only indirect evidence of load during template maintenance, as confounds such as working memory load effects on processing speed [65] and attentional deployment related bottlenecks [128] could impact results. Previous neuro-imaging research theorized that modulation by attentional templates occurs shortly before and early on during stimulus onset [128, 63]. With this in mind, we will try to decode attentional modulations in early sensory representations. To summarize, given the literature we decided to use a detection paradigm that has no distractors and encourages active guidance by attentional templates. Moreover, we decode sensory modulations during stimulus presentation, which are a direct effect of attention, enabling direct observation of attentional load effects no matter how early or transitional. Finally, to understand how exactly the attentional load impacts the decision making process we use a detection paradigm and separately evaluated modulations in hit rate, false alarms rate, and signal detection theory measures.

5.3 Main Experiment

5.3.1 Motivation

This chapter presents our main experiment. Motivated by readings from the literature, our goal was decoding attentional modulation of sensory representations directly, and analyzing how - if at all - attentional load impacts these modulations. To our knowledge, all other neuro-imaging work on attentional capacity focused on other, less direct, measures of attention [128, 63]. Following previous neuro-imaging work we expect differences in these modulations to be most pronounced at stimulus onset [128, 63]. These differences are not confounded by working memory load effects on processing speed [65], unlike search-task accuracy and response time measurements [122, 121, 28, 84, 9]. Moreover, these differences can only be explained by template maintenance capacity limits rather than other bottlenecks in the later stages of perception [128]. In simple terms, modulation of sensory representations by attention are immediate and begin at stimulus perception. And early differences in these modulation can only be explained by attentional load. Another conclusion from our literature review was that guidance by attentional templates and interference by distractors matching working memory contents are two separate phenomena [135]. With the above in mind, we elected to use a detection paradigm with no distractors during stimulus duration. Using this paradigm has multiple additional benefits in bridging a few gaps that became apparent during the literature review. Firstly, in previous EEG experiments that used nested memory-search paradigms the target was always present, and analysis focused on correct trials only [128]. This left a gap in the literature; for instance, the attentional load theory proposed in [128] does not necessarily predict any difference between different attentional load conditions when the target is absent. However, previous results are consistent with an increase in false alarms [97, 96]. Since target-absent trials are impossible in the ubiquitous nested memory-search paradigm these consequences of increased attentional load that is yet to be thoroughly explored. Secondly, behavioral results of the nested memory-search paradigms can only be quantified in terms of accuracy and response time. Only few researchers employed a detection paradigm [74, 97, 96], and to the best of our knowledge only one - behavioral - experiment focused on signal detection modeling [74]. Diversity of

analytical methods and paradigms is essential for a comprehensive understanding of the underlying mechanisms responsible for the effects of different attentional load conditions. Our behavioral results could therefore be of particular interest, as they provide a different perspective from the vast majority of attentional capacity work published in the last decade, especially if neural correlates of behavioral differences are observed in target-absent trials.

5.3.2 The Experiment

We employed a modified version of experiment 2 from [96]. Our two-alternative-forced-choice experiment presented participants with an array of dots in different colors. One of the colors was over-represented in half of the trials, while in the other half the dots were equally distributed across five colors (for a total of six colors per trial). Participants were required to distinguish between trials with an over-represented color (target-present) vs trials with an equal distribution of colors (target-absent). The experiment had no-cue, one-cue, and two-cue conditions. The over-represented color in all one-cue and two-cue target-present trials was always one of the cued colors. The [96] experiment modified the color coherence (amount of over-representation) between trials. Performance at different coherence levels was then used to fit a psychometric function. The analysis used the parameters of the best fit function to study the impact of WM on attention guidance. EEG decoding is sensitive to such variability stimuli. Therefore, we separately calculated thresholds for each color per subject to be used in the actual experiment. The target was present in 75% of the trials, this was done to ensure we had enough EEG data to decode and compare the signal in one and two-cue target present trials, which are the most analogous to the types of trials analyzed in previous works [128]. The locations of the cues were randomized but where always 180° and 120° degrees apart in the two and three-cue trials respectively. The six colors were randomly selected from a pool of seven (red, green, blue, yellow, purple, orange, and cyan) in each trial.

Our experiment is divided into behavioral and EEG sessions. The behavioral session has no-cue, one-cue, two-cue, and three-cue conditions. Following [96] we expect the Hit-FA rate to drop when less information is available to the participants and more uncertainty is present. In other words, we expect the Hit-FA to be the highest in the one cue condition and consistently drop in the two, three,

and no cue. This can be explained by a drop in the hit rate, an increase in the false-alarm rate, or a combination of the two. With this in mind, we test the hypothesis that the Hit-FA monotonically decreases across conditions when less information is available (this order being one-cue, two-cue, three-cue, and no-cue). Additionally, we hypothesize the hit rate will similarly decrease, while the false-alarm rate would follow an opposite trend by monotonically increasing. Significant differences between memory load conditions would eliminate any possibility of a strong multiple item template hypothesis. Moreover, a complete failure in three-cue trials could indicate that even if two attentional templates can guide attention simultaneously, there is still a "hard limit". Furthermore, signal detection theory analysis can be used to explore how attentional load impacts the decision making process. Differences in sensory modulation are likely to manifest differences in the discriminability index (the d-prime) rather than the criterion for instance. Decreased performance when less information is available would correlate to decreased discriminability, hence we hypothesize that, similarly to the Hit-FA, the d-prime will also monotonically decreases across conditions.

EEG sessions focused on the one-cue and two-cue conditions. Our paradigm is designed so that trials of the same type are identical during the stimulus period, regardless of block condition. In other words, a one-cue and a two-cue target-present green trials will be identical after cue offset. Therefore, differences in EEG decoding during stimulus presentation must be the direct result of differing attentional loads. Similarly to previous research, we also expect to find a difference in EEG decoding early on in the stimulus presentation period [128, 63]. More specifically, we are directly decoding the sensory representations modulated by attention, therefore we should be able to observe any differences between the attentional load conditions no matter how transient or early in the perception process. We hypothesize that attention in the single template condition would results in bigger sensory modulations resulting in better decoding compared to the two-cue condition. This difference in modulation can be considered the neural correlate of a decrease in the discriminability index between the one and two cue conditions. We also decode false-alarm trials to confirm that false-alarms are driven by attention-capture of one of the attended cued color, as opposed to general performance decrease due to higher attentional load. Generally, any degradation

of attentional modulation between different memory load conditions could be interpreted as a cost for maintaining multiple templates.

Finally, we also completed an exploratory analysis of generalizations of different classifiers to preparatory period decoding.

5.3.3 Methodology

5.3.3.1 Participants

We surveyed the literature and examined the number of participants used in similar EEG decoding studies. The sample size of the most relevant EEG decoding papers ranged from 16 to 34 [8, 128, 119, 184]. Following [128] we collected data until we had 24 participants in total. In addition to the 24 subjects who's results were included in this study, 2 were rejected due to noisy EEG, 3 were rejected due to low performance during the behavioral session (accuracy was at threshold performance in all conditions), and 1 was rejected after his thresholding session failed to converge. The participants were recruited from the Michigan State University student body. The protocol was approved by the MSU institutional review board and written informed consent was obtained from every subject.

5.3.4 Stimuli and Apparatus

The experiment was programmed in MATLAB and using the MGL extension [51]. The stimulus consisted of random dot fields. For each trial, five out of the seven colors were selected and 240 dots were drawn in an annulus (inner radius = 1.5° , outer radius = 6°) and centered on a central fixation disc (white; size: 0.1°; luminance: 14.8 cd/m^2). A random color out of the five was selected to be the target-color. During target-present trials, the target-color was over-represented, and the remaining dots were divided equally between the four remaining colors. In target-absent trials, all five colors were equally represented. To eliminate potential confounds, each subject adjusted the brightness for every hue separately to achieve isoluminance between all colors. This was done because differences in luminance between colors could cause the overall brightness during stimulus presentation to differ - especially in target present trials with disproportionately represented colors

- thereby inflating the EEG decoding accuracy. The procedure for obtaining isoluminance between all colors was identical to the one used for the pilot experiment (see Section 4.1.1.2). During cue trials, the stimulus was preceded by a cue (size: 0.5°). One of the cues contained the target color. When more than one cue was present the other cue colors were randomly drawn from the remaining two colors that are not present in the stimulus dot field at that particular trial. The location of the cue was on a circle around the fixation (radius = 1.5°). The angle of the first cue was randomly drawn from (0°, 10°, 20°, .., 360°), in two-cue trials, the second cue was 180° away from the target-color cue. Finally, in three-cue trials the cues were 120° away from each other (see figure 5.2).

5.3.4.1 Procedure

Participants first performed a simple isoluminance task to prevent potential brightness confounds. The actual task trials began with a cue that lasted $500ms$, followed by an $800ms$ preparatory period and a $100ms$ stimulus segment. Finally, participants were required to make a target-absent or target-present trial judgment by pressing either 1 or 2 on a keypad with their right index or middle finger (see figure 5.2). Short auditory feedback was given after every incorrect answer. Before the behavioral and EEG sessions, we ran a separate thresholding task. All thresholding trials were similar to the no-cue condition. The target color coherence was manipulated to find the coherence (relative over-representation) that produces an intermediate level of performance (82%) for every subject. This thresholding was done separately for each color and was achieved by utilizing the Best PEST adaptive method with a Weibull psychometric function [130, 137]. In a classical paper, Quick proved that, if a few reasonable assumptions (such as Gaussian noise) hold, any psychometric function can be approximated using a Weibull probability distribution [140]. Given previous trials, the Best PEST adaptive function uses maximum likelihood estimation to select the parameter values that are most likely to induce the desired performance levels.

During the behavioral session, subjects ran 2 blocks of 84 trials in every condition. The order of the blocks was randomized across subjects.

The EEG session procedure differed from what was described above in multiple key aspects. First, the subjects completed 5 blocks of one-cue and two-cue conditions only (in an ABAB fashion

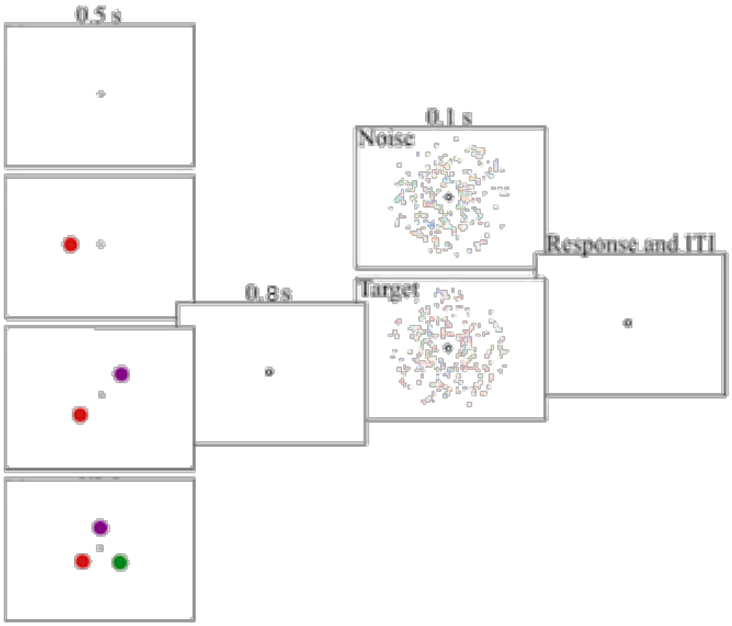

Figure 5.2 Example of behavioral session no-cue, one-cue, two-cue, and three-cue trials. The background color was changed for visibility (RGB value 240, 240, 240).

with the order balanced across subjects). Additionally, to prevent alpha frequencies phase issues the delay segment duration varied between 800ms and 1100ms. Finally, as we are mostly interested in decoding target-present trials, the ratio of target-present trials was 75% of the total number of trials. Since behavioral sessions were used to exclude subjects with abnormal performance, behavioral and EEG blocks shared the same task characteristics (such as proportion of target present trials).

5.3.4.2 Data Acquisition and Preprocessing

Continuous EEG activity was recorded using the actiCHamp system with BrainVision recorder software. The participants were fitted with a 64-channel actiCap with active electrodes. The screen refresh rate was set to 120Hz and data sampling was at 1000Hz. Additionally, electrooculogram (EOG) activity was recorded from from horizontal and vertical electrode pairs, and used to detect and reject horizontal eye movements, eyeblinks, and vertical eye movements. Electrode impedances were maintained $< 50K\Omega$. The data from the inter-trial interval was discarded. We used EEGLAB and ERPLAB to process the data. First, we resampled the data to 500Hz, removed the AC line noise (using the cleanline plugin), applied a bandpass filter between 1 and 100Hz, and used ICA decomposition to separate and remove components originating from blinks and other artifacts.

Finally, waveforms were manually examined by experimenters and noisy epochs were rejected. A simple peak-to-peak blink detection algorithm (available in ERPLAB) was also used to detect blink using the EOG channels. The threshold differed for every participant, and potential blinks were marked before an examiner manually checked the data and marked any additional noisy epochs for rejection. On average, we rejected 120 of the 840 recorded trials. This is comparable to the rejection rates reported in other EEG studies [7, 8].

5.3.5 EEG Data Analysis

We used the ADAM decoding toolbox [44], with a few modifications such as adding Gaussian smoothing to the raw data, and smoothing classifier accuracies. First, we resample the data at a rate of $100Hz$. Additionally, following the conclusion of our pilot experiment, we expect the color feature information to be mostly contained in the sub-alpha frequency band, and visual-cortex electrodes. Therefore we employ a $10Hz$ lowpass filter and use the subset of mostly parietal and occipital electrodes Pz, P3, P7, O1, Oz, O2, P4, P8, TP10, TP9, T7, T8, P1, P2, P5, PO7, PO3, POz, PO4, PO8, P6, P2, CP4, CP2, CP1, CP3, C1, C2, C3, and C4. The pilot decoding results for this subset were robust, and similar electrodes were used in the literature [7, 8, 128].

Following previous studies, we Gaussian smoothed (window size, $20ms$) the data along the time dimension, and smoothed the classifier outputs using a 40 ms moving average [7, 8, 184]. To verify that our pipeline does not inflate accuracy we simulated and tried to decode random noise (see Appendix B). After decoding trials from the one-cue and two-cue conditions we subtracted the two decoding results and used a similar cluster-based permutation testing to identify significant results. We used the default ADAM backwards decoding classifier with a 10 cross-fold permutation testing. For an in depth discussion of LDA decoders such as the ADAM backward decoding model see the previous EEG decoding section in literature review 2.2.2.1. At each permutation, 90% of the data (balanced across the seven colors) was used to train an LDA classifier and accuracy was computed on the withheld 10% of the data. However, an exception to this procedure is when using different training and testing sets (Figure 5.10), as in these cases no splitting of the data is necessary and there is only a single iteration. Given that we always decode seven classes representing the seven

different colors, the chance level of the classifier was $\frac{1}{7} = 14.28\%$. Finally, for evaluating statistical reliability we used the default ADAM monte-carlo sampling with 2000 iterations of cluster-based significance testing [128]. Since the classifier accuracy is being compared against chance, the t-tests used to generate the significance values before the cluster based permutation testing are always one-tailed [44].

To investigate temporal generalizations we employed a similar pipeline. First a classifier was trained using samples at a specific time point, then decoding accuracy is calculated on the testing data at every time point (instead of just the same time point). Analogous significance cluster analysis is performed on the resulting two dimensional (training time x testing time) accuracy matrix. Generalization can also be calculated across completely different EEG segments. For instance, by training classifiers for each stimulus duration time point, and classifying preparatory period data. Here however no cross-fold validation is needed, as the number of trials available for training and testing are independent.

Finally, we also implemented a Mahalanobis distance classifier. The implementation was directly integrated into ADAM. While the classification accuracy was comparable to the LDA performance, the Mahalanobis classifier failed to reproduce some transient effects that were present in the default LDA decoding. We believe that this is due to the Mahalanobis classifier requiring additional trial averaging and binning [184], see Appendix C for Mahalanobis classification results. Both the LDA and Mahalanobis classifier algorithms were described previously in section 2.2.2.1. All results in the following sections were achieved using the LDA classifier.

5.4 Unsupervised Artifact Rejection With Deep Encoder-Decoders

More often than not, electroencephalography preprocessing sections in cognitive neuroscience papers allude to manual trial-rejection by visual inspection [7, 8, 128]. This tasking process constitutes a bottleneck, especially at early research stages when making decision regarding the design and procedure of the experiment. To elevate this bottleneck one of the unsupervised artifact detection algorithms presented in section 3.2 was used. Code for converting EEGLAB files to and from formats that can be processed by our feature extraction and artifact detection packages is available online[2]. As can be observed in figure 5.3, using unsupervised outlier detection improved the decoding performance without requiring hours of commitment in the early stages of the experiment.

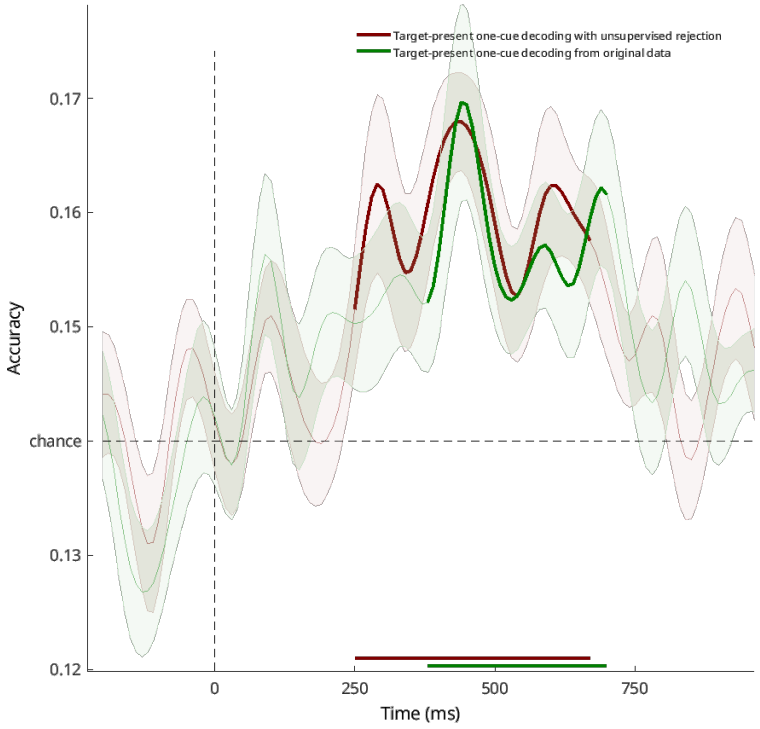

Figure 5.3 decoding performance on target-present correct trials after and before using unsupervised outlier detection (red and green lines respectively). Horizontal lines indicate a significant decoding cluster.

[2] github.com/sari-saba-sadiya/EEGLAB-Artifact-Detection

5.5 Results

5.5.1 Behavioral Results

Following [96] we hypothesized that the Hit-FA rate would drop when less information is available to the participants (in other words we expected a downward trend across the one, two, three, and no cue conditions). Moreover, we hypothesized that the hit rate would decrease following the same trend, and the false-alarm rate would do the opposite and increase.

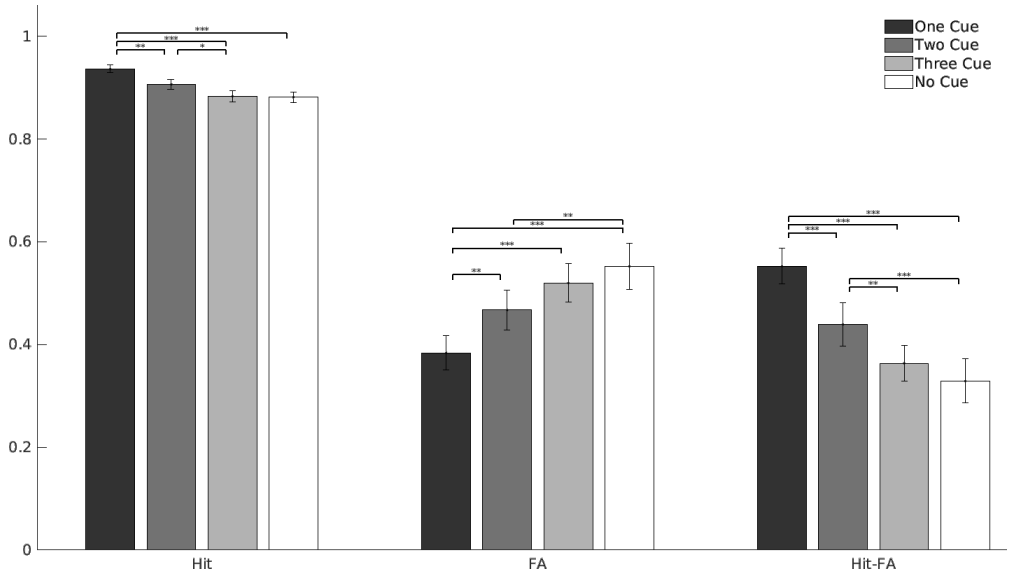

Figure 5.4 Results of our behavioral session. ⋆) p<.05 ⋆⋆) p<0.01 ⋆ ⋆ ⋆) p<0.005.

A repeated measures ANOVA revealed a significant main-effect of cue condition on Hit-FA ($F(3, 69) = 28.812$, $MSE = 0.008167$, $p < .005$, $\eta_G^2 = 0.556$). We followed up the ANOVA with a series of paired two-tailed t-tests (see figure 5.4). Results showed significant differences between one-cue trials and two, three, and no-cue trials (one vs two $t(23) = 4.652$ $p < 0.005$, one vs three $t(23) = 7.95$ $p < 0.005$, one vs no cue $t(23) = 8.24$ $p < 0.005$) as well as two-cue and three, and no-cue trials (two vs three $t(23) = 2.88$ $p < 0.01$, two vs no-cue $t(23) = 4.22$ $p < 0.005$). A repeated measures ANOVA revealed a significant main-effect of cue condition on hit rate ($F(3, 69) = 9.609$, $MSE = 0.001646$, $p < .000$, $\eta_G^2 = 0.294$). We followed up the ANOVA with a series of paired two-tailed t-tests. Results showed significant differences between one-cue trials and two, three, and no-cue trials (one vs two $t(23) = 2.80$ $p < 0.01$, one vs three

$t(23) = 4.86$ $p < 0.005$, one vs no cue $t(23) = 4.73$ $p < 0.005$), as well as between the two and three-cue conditions ($t(23) = 2.12$ $p < 0.05$). Finally, a repeated measures ANOVA revealed a significant main effect of cue condition on false-alarm rate ($F(3, 69) = 14.015$, $MSE = 0.009272$, $p < .001$, $\eta^2_G = 0.378$). We followed up the ANOVA with a series of paired two-tailed t-tests. Results showed significant differences between one-cue trials and two, three, and no-cue trials (one vs two $t(23) = 3.3025$ $p < 0.01$, one vs three $t(23) = 4.94$ $p < 0.005$, one vs no-cue $t(23) = 6.38$ $p < 0.005$) as well as a significant difference between the two-cue and no-cue conditions ($t(23) = 2.84$ $p < 0.01$).

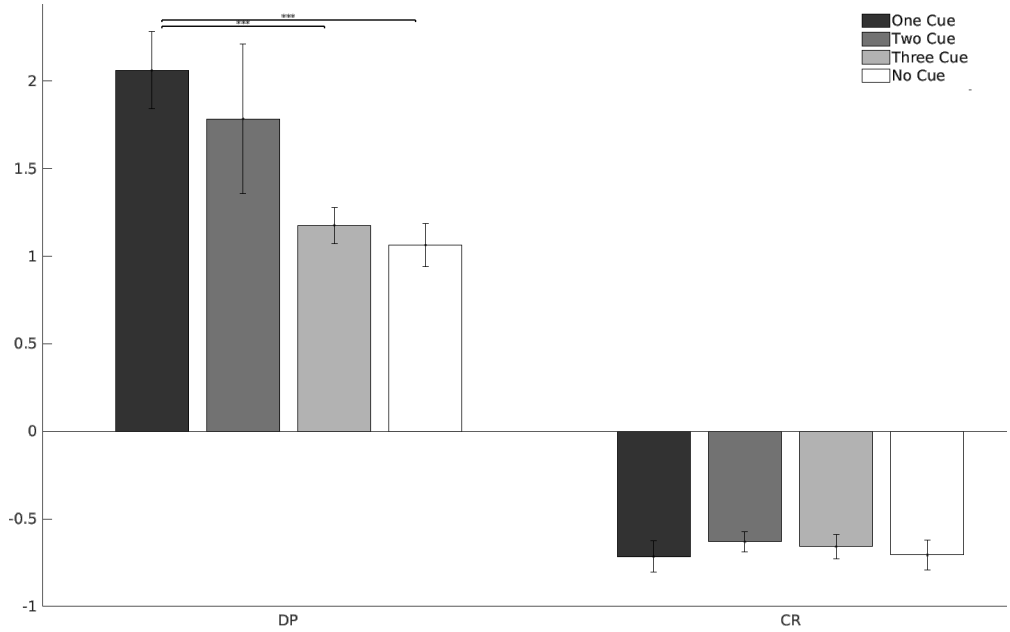

Figure 5.5 Discriminability index and Criterion results of our behavioral session. ⋆) p<.05 ⋆⋆) p<0.01 ⋆⋆⋆) p<0.005.

Further signal detection theory analysis was also conducted by calculating the d-prime and criterion for each condition. A repeated measures ANOVA revealed a significant main-effect of cue condition on d-prime ($F(3, 69) = 5.175$, $MSE = 1.064882$, $p < .01$, $\eta^2_G = 0.10106$). Follow up two-tailed t-tests showed significant differences between the discriminability indexes (d-prime) in the one and three as well as no-cue conditions (one vs three-cue $t(23) = 4.67$ $p < 0.005$, one vs no-cue $t(23) = 4.98$ $p < 0.005$). No significant main effect of cue condition on criteria was found ($F(3, 69) = 0.454$, $p = .715$).

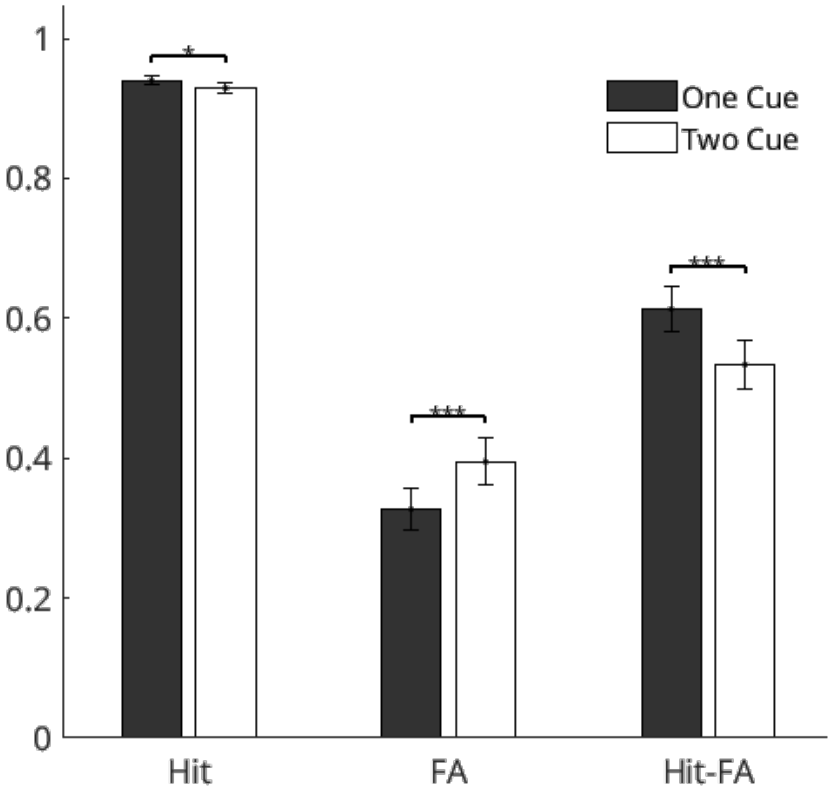

Figure 5.6 Results of our EEG session. ⋆) p<.05 ⋆⋆) p<0.01 ⋆⋆⋆) p<0.005.

Additionally, we found a main effect of condition on response time ($F(3, 69) = 3.851$, $MSE = 0.0118$, $p < .05$, $eta^2_G = 0.143$). Response-time generally followed the same trend as other behavioral measures (The mean response time was 0.6038, 0.6399, 0.7019, and 0.6803 seconds for one, two, three, and no-cue trials respectively). The effect of cue condition on response time for correct trials was also significant ($F(3, 69) = 3.9273$, $MSE = 0.00937$, $p < .05$, $eta^2_G = 0.145$) and followed the same pattern overall. In general, the response time, response time on correct only trials, and response time on target-present correct only trials (not plotted) followed an inverse trend to the accuracy, hit rate, hit-FA, and the discriminability index (See Figure 5.8 and Figure 5.5). Therefore there is no evidence of any speed accuracy trade-off in this experiment.

We ran the same analysis of performance during EEG trials (see Figure 5.6). A repeated measures ANOVA revealed a significant main effect of cue condition on Hit-FA rate ($F(1, 23) = 26.883$, $MSE = 0.0028$, $p < .001$, $\eta^2_G = 0.538$). Similarly, repeated measure ANOVA also showed

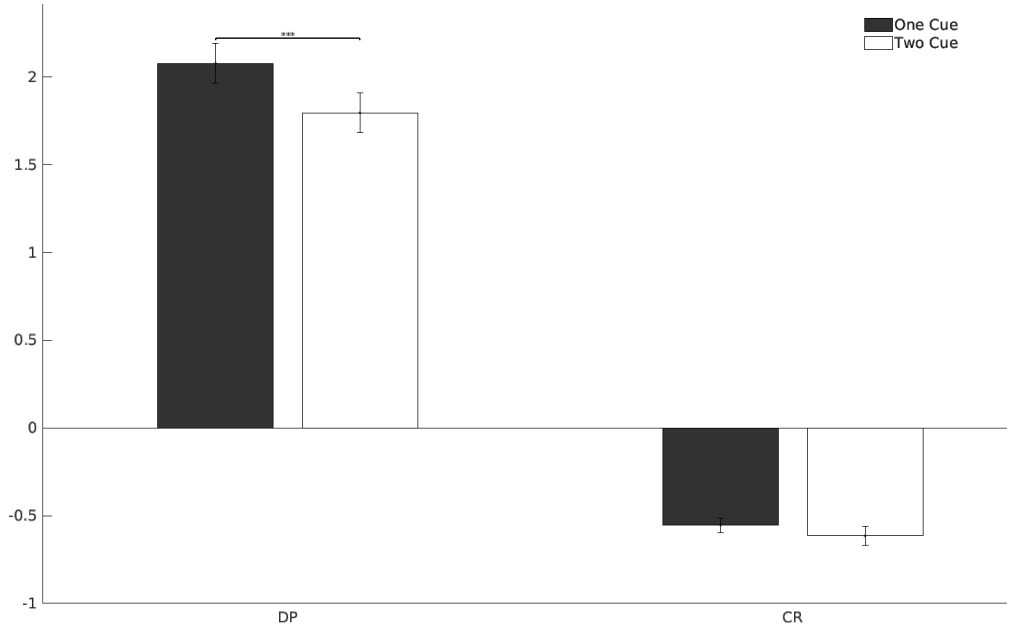

Figure 5.7 Discriminability index and Criterion results of our eeg session. ⋆) p<.05 ⋆⋆) p<0.01 ⋆ ⋆ ⋆) p<0.005.

a main-effect of cue condition on hit rate ($F(1, 23) = 4.293$, $MSE = 0.0003$, $p < .001$, $\eta^2_G = 0.157$) and false alarm ($F(1, 23) = 19.674$, $MSE = 0.0028$, $p < .001$, $\eta^2_G = 0.461$).

Signal detection theory analysis of the EEG session behavioral data showed a significant main effect of cue condition on the discriminability index ($F(1, 23) = 25.793$, $MSE = 0.0369$, $p < .001$, $\eta^2_G = 0.0606$), but no significant main effect was found for the criterion ($F(1, 23) = 3.264$, $p = .083$).

Additionally, we found a main effect of condition on response time ($F(1, 23) = 3.851$, $MSE = 0.0114$, $p < .05$, $\eta^2_G = 0.222$). The mean response time was 0.5417 seconds for one-cue trials and 0.6210 seconds for two-cue trials (Figure 5.8). There was no main effect of condition on response time when analyzing correct only trials ($F(1, 23) = 3.876$, $MSE = 0.023$, $p = .061$), however the mean for one-cue trials was still lower than for two-cue trials (0.4678 and 0.557 respectively). Following the pattern observed in the behavioral session, conditions with higher accuracy and discriminability index also had lower mean response time, thus we conclude that there was no speed accuracy trade-off.

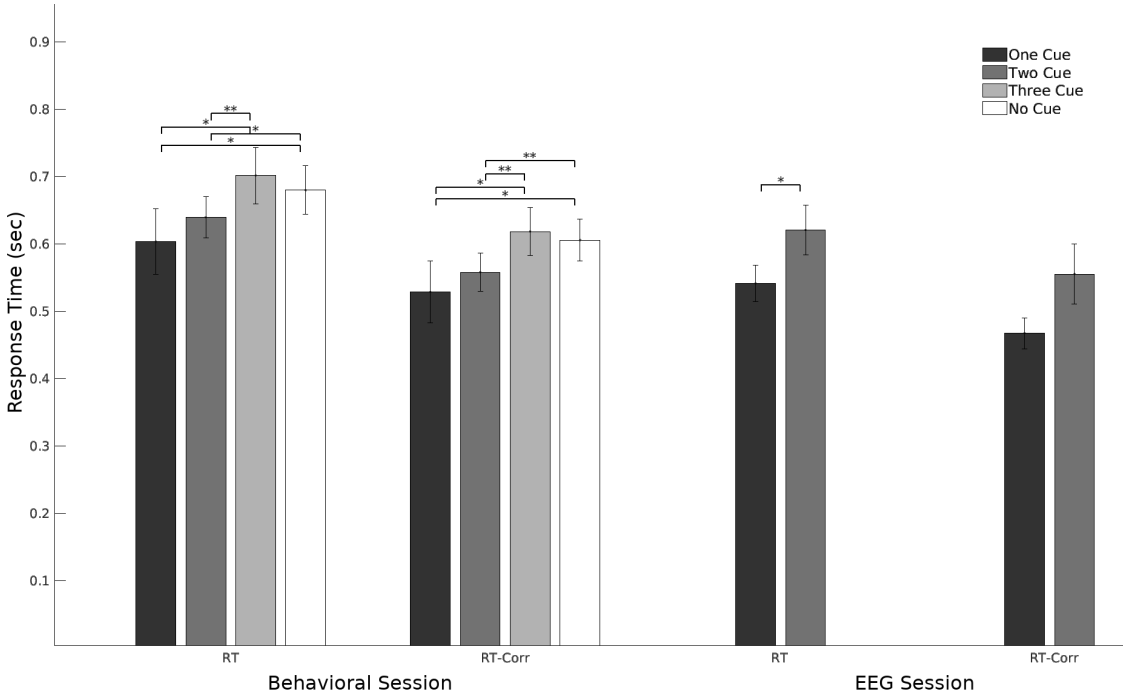

Figure 5.8 Response time (in seconds) for all trials and correct only trials, for the different conditions in the behavioral and EEG sessions. ⋆) p<.05 ⋆⋆) p<0.01 ⋆ ⋆ ⋆) p<0.005.

5.5.2 Bayesian Analysis of Behavioral Results

To further investigate if there is any evidence of behavior differing across three-cue and no-cue trials we followed up the previous analysis with Bayesian modeling using JASP [76]. Specifically, we tested the hypotheses that the hit rate, false-alarm, and d-prime values for the no-cue and three-cue trials were the same.

The default JASP prior of 0.707 on the Cauchy scale with 95% credibility interval was used for all Bayesian paired t-test. For follow up robustness analysis see Appendix D.

Bayesian paired t-test was used to explore if hit rate in three and no-cue trials differed. Analysis showed a Bayesian factor of $BF_{01} = 4.628$ (0.024 error %) in support of the null hypothesis. The median effect size was -0.022 and the confidence interval was $[-0.398, 0.353]$. Bayesian paired t-test was also used to explore if false-alarm rates in three and no-cue trials differed. Analysis showed a Bayesian factor of $BF_{01} = 2.879$ (0.026 error %) in support of the null hypothesis. The median effect size was 0.187 and the confidence interval was $[-0.191, 0.574]$. Finally, a Bayesian paired t-test was used to explore if the d-prime in three and no-cue trials differed. Analysis showed

a Bayesian factor of $BF_{01} = 2.122$ (0.026 error %) in support of the null hypothesis. The median effect size was -0.242 and the confidence interval was $[-0.634, 0.14]$.

The Bayesian factors suggest weak to moderate evidence in favor of the null hypotheses. Indicating no significant differences in behavior across three and no-cue trials.

5.5.3 EEG Results

We first compared target-present trials across the one-cue and two-cue conditions. Generally, the target feature value could be decoded from the very beginning of the one-cue trials, while decoding in the two-cue condition was possible from 150 ms after stimulus offset (Figure 5.9).

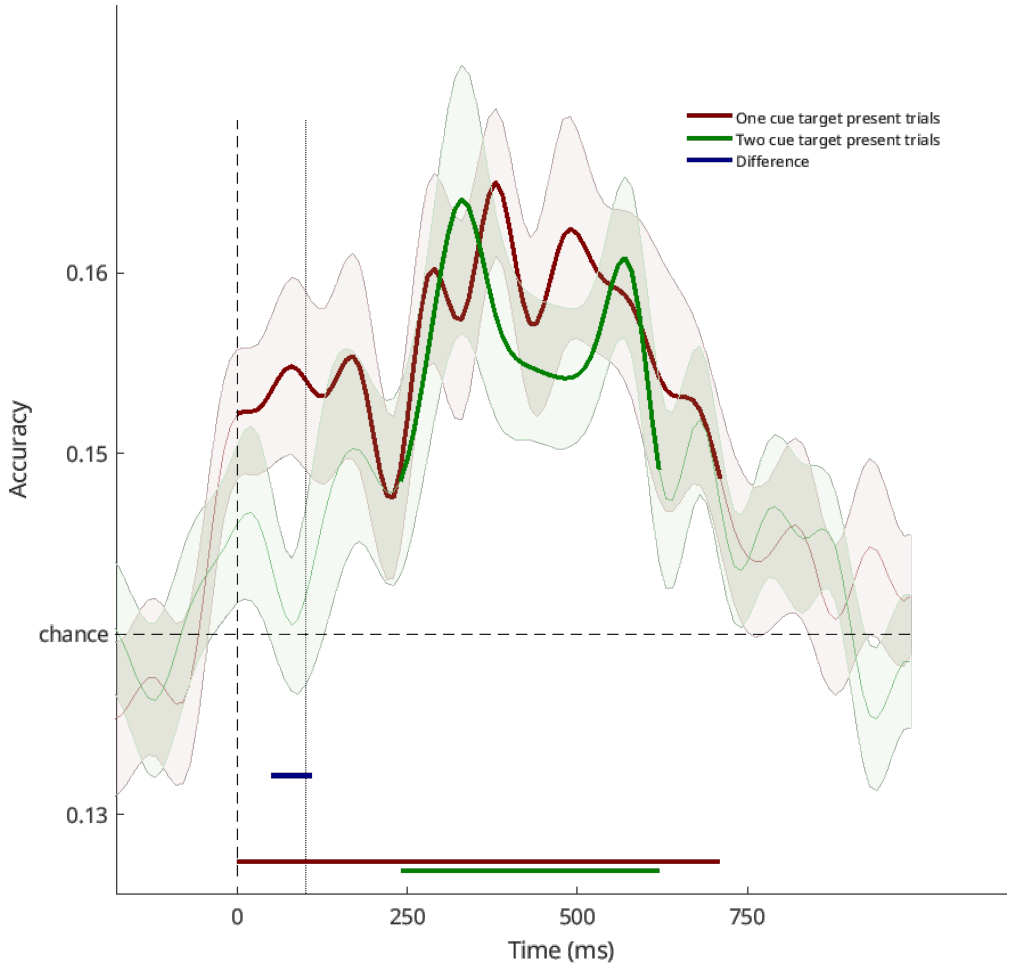

Figure 5.9 Decoding target present EEG trials. Bold red and green horizontal lines indicate significant decoding in the one cue and two trials respectively. Blue horizontal lines indicate clusters of significant differences between conditions. The dashed and solid lines indicate stimulus onset and offset respectively.

Additionally, to investigate the impact of attentional load on "false alarm" trials we decoded the cued color that was present (but not over-represented) in target-absent trials. Training the classifier using two-cue trials was generally unsuccessful and yielded no significant decoding. However, classifiers trained on target-present one-cue trials generalized to two-cue trials. The present (but not over-represented) color is likely to be the cause of the false alarm in two-cue target-absent trials (as the other color is completely absent from the array). Indeed, we were able to decode the present color in false alarm but not correct reject trials. Moreover, there was a significant difference in decoding between the two. Considering the two trials are identical stimulus-wise, this indicates the decoding in false-alarm trials is indeed driven by attention capture by the cued color. Surprisingly, there was no significant decoding in one-cue target-absent false alarm trials (see Appendix A). However, this is likely due to the low number of one-cue false-alarm trials rather than any attentional load modulation of perception. Finally, while decoding was not significant for all target-absent trials in either condition, it is worth noting that decoding accuracy was higher for two-cue, as opposed to on-cue, target-absent trials (Appendix A). The difference was not statistically significant, but the decoding results seem to correspond to the behavioral performance trends observed in the EEG and behavioral sessions (namely, false-alarms rate was higher in the two cue condition).

5.5.4 Generalizations For Stimulus and Preparatory Period Decoding

Classifiers trained on one-cue trials generalized well to two-cue trial stimulus period classification (figure 5.11). Preparatory period decoding was significant only in one-cue trials shortly after cue offset. Generalization across the preparatory period yielded no significant clusters for either one-cue or two-cue, trials (Figure 5.12). However, classifiers trained on the one-cue target present correct trials stimulus period successfully generalized to the one-cue preparatory period (Figure 5.13).

Mental representations are expected to be most robust during stimulus presentation in one-cue trials (when a coherent target color is present and after the sensory perception was enhanced by attention). The generalization of stimulus period one-cue classifiers to both one-cue preparatory

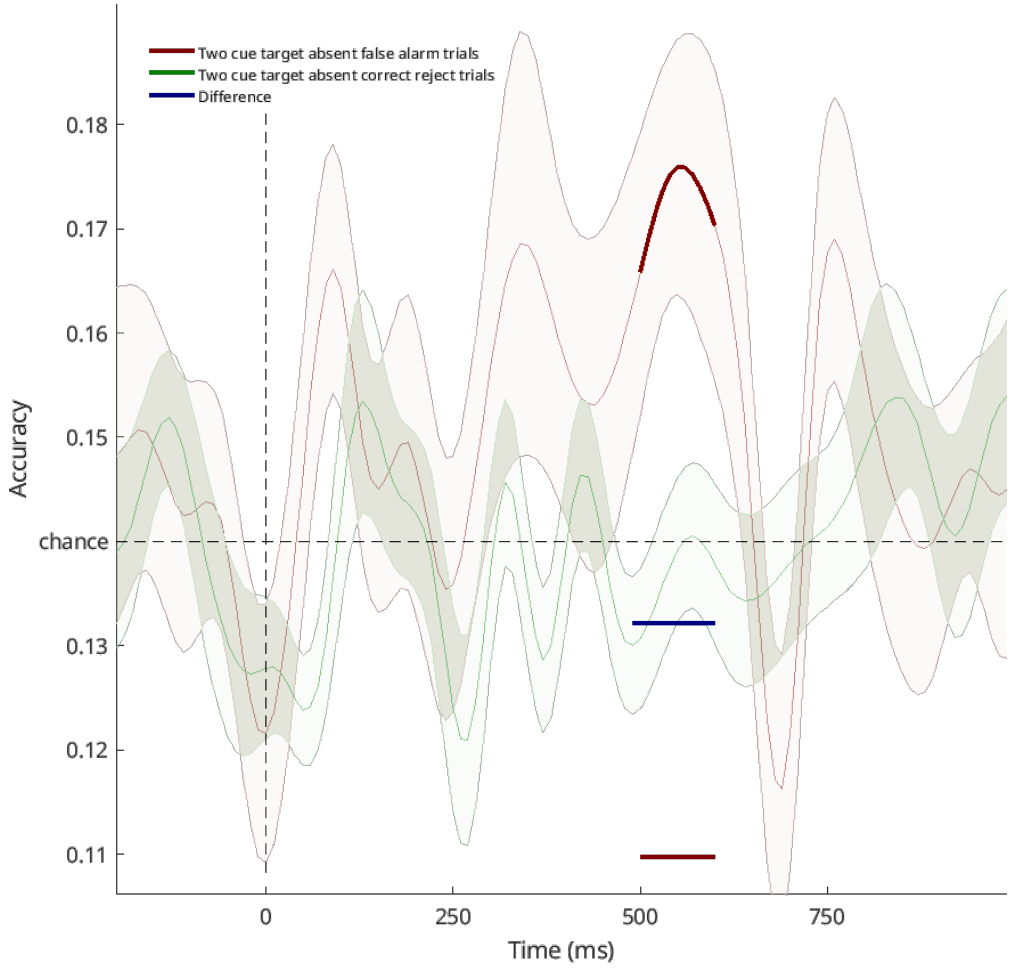

Figure 5.10 Decoding absent present EEG trials. Bold red and horizontal lines indicate significant decoding in the two-cue target absent false-alarm trials. Classifiers were trained using on target-present one-cue trials. Blue horizontal lines indicate clusters of significant differences between false-alarm and correct reject trials. The dashed and solid lines indicate stimulus onset and offset respectively.

period, and two-cue stimulus period, EEG could indicate a similarity in the representations being used. This indicates that when the target is coherent, representations in two-cue target present correct trials become similar to those of one-cue target present correct trials. Moreover, when only a single attentional template is maintained, representations in the preparatory period are similar to representations after perceiving the attended feature value.

Following the same logic, the lack of generalization for two-cue trials during the preparatory period could be indicative that a qualitatively different representation is being used when two attentional templates are being maintained simultaneously. There are a number of possible explanations

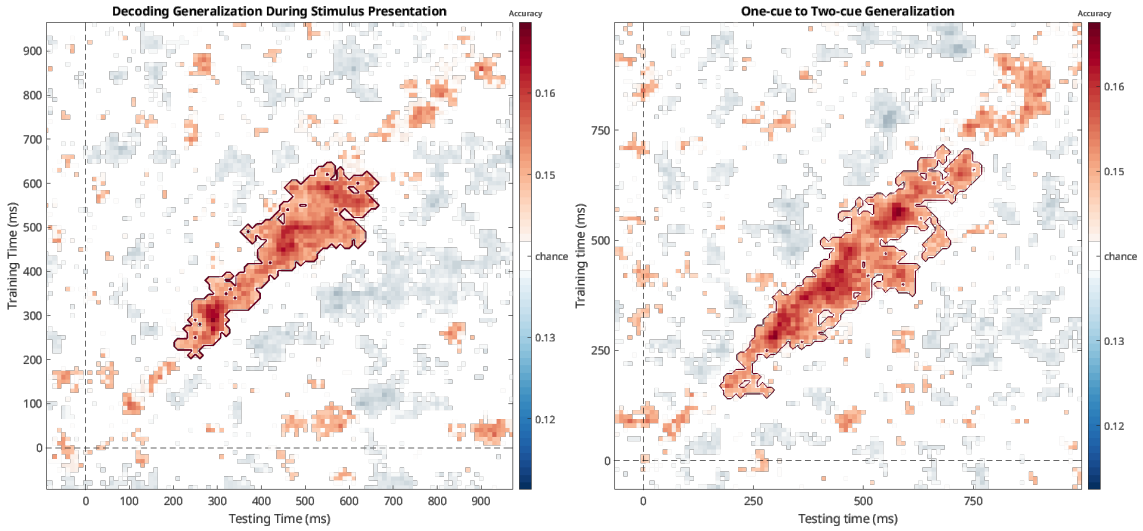

Figure 5.11 Decoding generalization for one-cue and two-cue target-present correct trials (left and right respectively). The classifiers were trained using one-cue target-present correct trials. The dashed lines indicate stimulus onset.

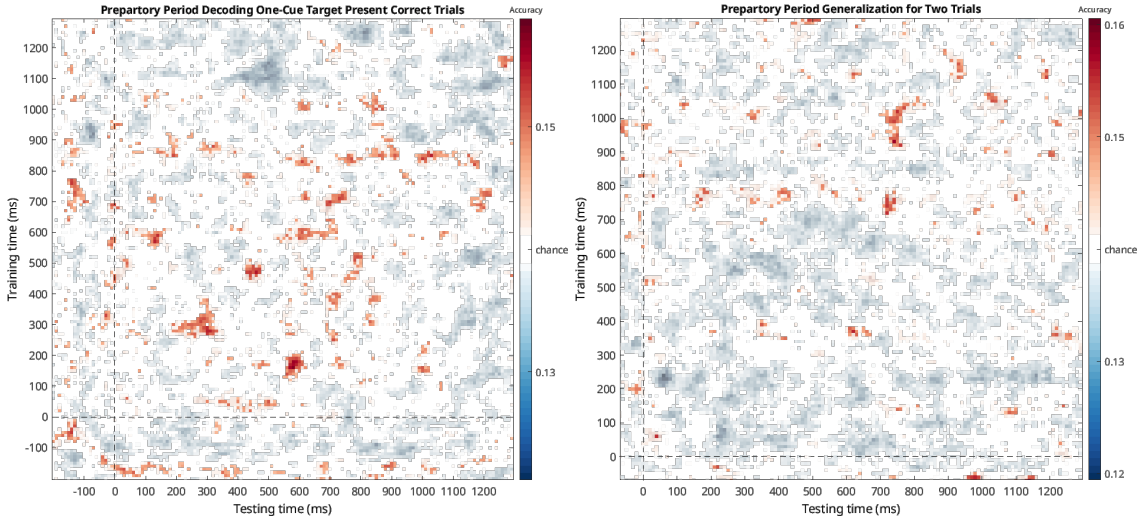

Figure 5.12 Decoding generalization for one-cue and two-cue target-present correct trials during the preparatory period (left and right respectively). The classifiers were trained using stimulus period EEG for one-cue and two-cue target present-correct trials (left and right respectively). The y-axis and x-axis dashed line indicate cue onset.

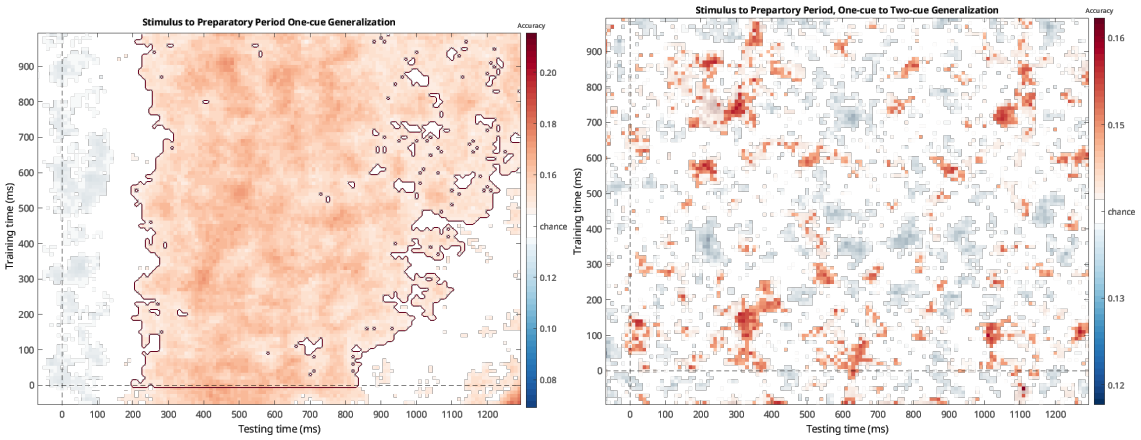

Figure 5.13 Decoding generalization for one-cue and two-cue target-present correct trials during the preparatory period (left and right respectively). The classifiers were trained using one-cue target-present trials after the stimulus onset. The y-axis dashed lines indicate stimulus onset, and the x-axis dashed line indicates cue onset.

for the above. One possibility is that, as suggested by [128], the two attentional templates are mutually suppressive until feedback from a sensory stimulus causes the corresponding template to win the competition. Another possibility is that only one of the items is being maintained in an active state. Researchers have demonstrated that passive and active working memory states are encoded using orthogonal representations where only the active template can be decoded [183, 182, 91]. If this is indeed the case, we expect decoding to not be possible during the preparatory period in half of the trials, limiting our statistical power. Finally, empirical limitations of machine learning complicate the interpretation of these results even further. Classifiers perform better with under-representative training noise (low training noise and high test noise) in comparison to over-representative training noise (high training noise and low test noise) [104]. Representations are expected to be most robust during stimulus presentation in one-cue trials (when a coherent target-color is present and the sensory perception is being enhanced by attention). Therefore, training stimulus presentation in one-cue trials is an example of under-representative training noise. Moreover, stimulus presentation could involve a variety of representations (visual, semantic labels, . . .). This is less likely during the preparatory period. Hence, preparatory period data might simply lack aspects of the representation needed to decode stimulus presentation data.

5.6 Discussion and Conclusion

EEG decoding analysis of the target feature values during and after stimulus presentation showed significant difference between one-cue and two-cue trials. This indicates the existence of a cost for maintaining multiple attentional templates. Having designed our paradigm so that one and two-cue trials differ only during cue presentation, it is safe to assume that decoding differences during stimulus presentation can be attributed to differences in attentional modulation of sensory representations. Specifically, under single attentional template conditions modulation of sensory representations occurs earlier. These neural differences are reflected behaviorally in the d-prime statistic which decreases when multiple templates are maintained. Therefore, we conclude that the enhanced attentional modulation in the single template condition (demonstrated by the EEG analysis) enables easier discrimination between noise and signal trials (as evidenced by differences in d-prime).

Moreover, there was a clear downward trend in hit minus false-alarm with decreased certainty regarding target color. Analysis revealed that maintaining multiple templates decreases performance mainly by increasing false-alarm rates. EEG analysis of two-cue target-absent trials demonstrated that we can decode the non-coherent signal in false-alarm trials but not correct-reject ones. This indicates that false-alarms occur due to participants mistaking the non-coherent signal for a target, rather than due to increased task difficulty or other confounds of increased attentional or working memory load. This further supports our conclusion that behavioral differences between the one and two-cue conditions originate from a decrease in sensitivity when multiple attentional templates are maintained. These differences in EEG decoding of target-absent trials also correspond to the results of the signal detection theory analysis. The d-prime significantly decreased between the one and two-cue conditions, indicating that it becomes increasingly harder to discriminate between signal and noise trials when maintaining multiple attentional templates, resulting in a higher false-alarm rate.

There was no significant main effect of attention load on the criterion, suggesting that the differences are not due to variation in response strategy across conditions. This held true in both

the behavioral and EEG session, indicating that the observed attentional load effects are likely due to low level mechanisms responsible for a change in sensitivity rather than any top-down change in response strategy. Additionally, response time and accuracy inversely correlated, hence there is no evidence to suggest speed any accuracy trade-off occurred.

Generally, three-cue trials did not differ significantly from no-cue trials. Moreover, follow up Bayesian analysis found weak evidence in favor of the null hypothesis (that behaviorally these trials are identical). Therefore, analysis indicates that attentional guidance might be limited to two items at most, and that maintaining three or more templates results in a complete failure of attentional guidance.

Considering the EEG results for the target-present and target-absent trials, as well as the overall pattern of the behavioral results, we conclude that maintaining multiple attentional templates comes at a significant cost. Moreover, we demonstrate that while the cost exists in target-present trials, impaired attentional modulation due to maintaining multiple templates also significantly increases false-alarms due to decreased ability to discriminate between signal and noise trials. This result is especially interesting as previous research utilized paradigms that make signal detection analysis impossible [123, 179, 9] and limited EEG decoding analysis to correct target-present trials only [128]. Overall, we conclude that attention modulation of sensory representations is significantly weaker when maintaining multiple templates, resulting in impaired signal perception in target-present trials, and more false alarms in target-absent trials due to difficulties discriminating between signal and noise. Following this conclusion, we reject many versions of the multiple item template (MIT) hypothesis, especially if they argue for a lack of cost (or even an additive effect) of maintaining multiple templates [14, 10, 28, 84]. However, significant performance differences between two-cue and three/no-cue conditions suggest that participants are still able to make a limited use of multiple cues. Previous literature discussed three different scenarios that could potentially account for our results.

First, participants could be maintaining only a single template while storing the other cues in working (or long term) memory. Switching occurs when a participant fails to detect a signal

matching the attentional template. Theoretically, both our results, and the data presented in [128], could be explained by template-switching. Many theories of attention postulate the existence of a quick, parallel pre-attentive perception stage and a slower sequential focused attention stage [173, 186]. Operating under such framework, the difference in decoding between our one-cue and two-cue trials (as well as the one-cue-one-target and two-cue-one-target conditions in [128]) could be attributed to template switching after quick pre-attentive processing revealed no stimulus matching the initial attentional template. The authors of [128] do acknowledge that sequential processing is indeed a possibility, but ultimately reject this interpretation after finding no trial-based correlation between classification confidence scores and target position, or any pattern of target location switching in individual subject data. We are unable to run an equivalent analysis using our data, as we did not systematically manipulate the spatial locations of neither cue nor target. Moreover, while many switching cost estimates reported in the literature are around 250ms [40, 189], more recent estimates go as low as 50ms [126]. The significant difference in decoding between one and two-cue trials in both our experiment and [128] lasted only 50 ms. Therefore, the literature is inconclusive as far as the possibility of attributing the decoding differences we observed to switching cost. Generally, while the sequential template-switching interpretation remains a possibility, considering previous research we believe it is a less likely alternative.

Another possibility is that multiple templates can exist simultaneously but at a cost. According to this view, while attentional capacity is limited it can be flexibly divided to accommodate multiple items at the expense of representation quality. Similar conceptualization of working memory have been suggested [102]. Moreover, the authors of [128] proposed an attentional load theory with simultaneous maintenance of mutually suppressive templates, which could be interpreted as a variation on the flexible attentional capacity theme.

Finally, it is also possible that the templates are rhythmically oscillating. One recent study demonstrated oscillatory patterns in behavioral performance when attending to two cues [136]. While interesting, analysis in [136] focused only on hit rate, ignoring the effects of attentional load on false-alarm rates which, according to our results, are more substantial. Unfortunately, the

literature mostly consists of paradigms with stimuli presentation until response, making post-hoc analysis for oscillatory behavior difficult.

Overall, it is important to note that both [128, 136] do not fully explain our results. Both paradigms consist of only target-present trials, and it remains unclear how either theory could be extended to account for the increased frequency and decoding accuracy of false-alarm trials in the multiple templates condition. Further experiments that center target-absent trials and deliberately manipulate cue presentation onset (similarly to [136]) are required to thoroughly test rhythmic template fluctuations. Similarly, a version of our experiment with multiple targets (similarly to [128]) could have interesting implications.

To summarize, we used a signal detection paradigm that requires active guidance by attentional templates. Our results seem sufficient to reject most versions of the multiple-item-template hypothesis. Several substantial differences in both behavior and EEG decoding indicate that multiple templates are not able to guide attention as well as a single template. One interesting result of our experiment was that maintenance of multiple templates increased the likelihood of false-alarms while only slightly decreasing hit rates. This could have practical implications, as many tasks (such as aviation security or radiology screening) prioritize low false-negative rates, even at the expense of a slightly inflated false-positive rates [152]. While not conclusive, some of our results are consistent with recent theories of attentional load such as the competition model suggested by [128] or the fluctuating templates suggested by [136]. However, further research is required into attentional load effects during target-absent trials.

CHAPTER 6

GENERAL CONCLUSIONS AND REFLECTIONS

6.1 What Was Accomplished

This thesis contains several relatively distinct components. Before concluding this document, it is worth highlighting a few of its more interesting elements:

- The Feature Imitating Network (FIN) framework enables the integration of expert knowledge into deep learning models. This middle ground between rigid hand-crafted feature engineering and unpredictable data-hungry deep learning models has already sparked some interest. So far, this framework has been utilized in various domains such as natural language processing, computer vision, and predicting athletic performance [83].
- EEG decoding results demonstrate the direct impact of attention load on modulation of sensory representation. Observing this latent variable provides direct evidence of the cost of divided attention. Moreover, behavior in target absent trials (and neural correlates of this behavior) reveals a manifestation of this cost that was not explored in previous literature.
- The EEG feature extraction pipeline, originally a sub-component of a larger project, achieved modest popularity among EEG researchers and have become a collaboratively maintained standalone library[1].

6.2 Future Directions

There are multiple limitations to the current Feature Imitating Networks framework. First and foremost, FINs do not currently accept variable length inputs. Remedying this is not as simple as may initially seem, as feature (for instance, entropy) calculation often requires access to the full input vector. Furthermore, a thorough examination of how the task specific tuning changes the embedding in each FIN sub-model is required to support our intuition that the network is adapting the hand-crafted features to task requirements, rather than for instance learning a completely new embedding.

The presented cognitive neuroscience research also warrants follow up experiments. Behavioral

[1] https://github.com/sari-saba-sadiya/EEGExtract

differences between three-cue and no-cue conditions were inconclusive. While it is likely that there is some attentional guidance in three-cue trials, future experiments with larger sample size are required to properly validate this intuition. Additionally, analysis indicated that attention load has significant effects on behavior in target-absent trials. Further experiments focusing on target absent-trials are needed to identify the changes to the decision making process responsible for the behavioral differences. For instance, it is unclear if the changes are driven by top-down changes in task parameters (such as a reduced decision threshold), or bottom-up changes in how the sensory evidence is processed.

6.3 Reflections on Deep Learning in EEG

The real-world applications of machine learning for EEG data are numerous. However, whether the application is a medical-diagnostics tool [56] or a thought controlled prosthetic [192], a number of conditions need to be met for the developed algorithms to have any real-world impact. Perhaps most importantly, performance needs to be consistent even when testing on data from *unseen* (out of training sample) subjects, and despite some variability in data acquisition circumstances (for instance, unseen EEG task). Preparing the literature review, I was dismayed to discover that the vast majority of researchers do not report any out-of-sample testing (one notable exception being [56]). As can be observed in Section 3.1, the difference in performance between "seen task seen subject" and out of sample data can significant. Generally, out of sample testing was reported for the algorithms proposed in throughout work. Hopefully, this practice will become more common as the field of EEG focused machine learning matures, and the demand for algorithms that perform well in the real world increases.

6.4 Reflections on Accessibility

Code and data for reproducing the experiments presented in this thesis will be made available. All developed machine learning algorithms are already available on several github repositories. The interest there repositories generated is a testament to the much discussed importance of accessibility in science. However, an unexpected hidden benefit I had the pleasure of experiencing is the friendship and collegiality that can blossom from an email inquiring about a run time error.

BIBLIOGRAPHY

[1] A. Accardo et al. "Use of the fractal dimension for the analysis of electroencephalographic time series". In: *Biological cybernetics* 77.5 (1997), pp. 339–350.

[2] M. Agarwal and R. Sivakumar. "Blink: A Fully Automated Unsupervised Algorithm for Eye-Blink Detection in EEG Signals". In: *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. 2019, pp. 1113–1121.

[3] Charu C. Aggarwal. "Outlier analysis". In: *Data Mining: The Textbook*. Springer Publishing Company, Incorporated, 2015, pp. 75–79.

[4] Jinwon An and Sungzoon Cho. *Variational Autoencoder based Anomaly Detection using Reconstruction Probability*. 2015.

[5] S. Andersen, S. Hillyard, and M Muller. "Global facilitation of attended features is obligatory and restricts divided attention". en. In: *J. Neurosci.* 33.46 (Nov. 2013), pp. 18200–18207.

[6] Kai Keng Ang et al. "Mutual information-based selection of optimal spatial–temporal patterns for single-trial EEG-based BCIs". In: *Pattern Recognition* 45.6 (2012), pp. 2137–2144.

[7] Giyeul Bae and Steven Luck. "Dissociable Decoding of Spatial Attention and Working Memory from EEG Oscillations and Sustained Potentials". In: *The Journal of Neuroscience* 38 (Nov. 2017), pp. 2860–17.

[8] Gi-Yeul Bae and Steven J. Luck. "Decoding motion direction using the topography of sustained ERPs and alpha oscillations". In: *NeuroImage* 184 (2019), pp. 242–255.

[9] Brett Bahle, Valerie M Beck, and Andrew Hollingworth. "The architecture of interaction between visual working memory and visual attention". en. In: *J. Exp. Psychol. Hum. Percept. Perform.* 44.7 (July 2018), pp. 992–1011.

[10] Brett Bahle et al. "The architecture of working memory: Features from multiple remembered objects produce parallel, coactive guidance of attention in visual search". en. In: *J. Exp. Psychol. Gen.* 149.5 (May 2020), pp. 967–983.

[11] Mohammad Haris Baig, Vladlen Koltun, and Lorenzo Torresani. "Learning to Inpaint for Image Compression". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 1246–1255.

[12] Diane M Beck and Sabine Kastner. "Top-down and bottom-up mechanisms in biasing competition in the human brain". en. In: *Vision Res.* 49.10 (June 2009), pp. 1154–1165.

[13] Valerie M Beck and Andrew Hollingworth. "Competition in saccade target selection reveals attentional guidance by simultaneously active working memory representations". en. In: *J. Exp. Psychol. Hum. Percept. Perform.* 43.2 (Feb. 2017), pp. 225–230.

[14] Valerie M Beck, Andrew Hollingworth, and Steven J Luck. "Simultaneous control of attention by multiple working memory representations". en. In: *Psychol. Sci.* 23.8 (Aug. 2012), pp. 887–898.

[15] James Bergstra and Yoshua Bengio. "Random search for hyper-parameter optimization". In: *Journal of machine learning research* 13.Feb (2012), pp. 281–305.

[16] Katarzyna J Blinowska, Rafał Kuś, and Maciej Kamiński. "Granger causality and information flow in multivariate processes". In: *Physical Review E* 70.5 (2004), p. 050902.

[17] Sarah Blum et al. "A Riemannian Modification of Artifact Subspace Reconstruction for EEG Artifact Handling". In: *Frontiers in Human Neuroscience* 13 (2019). ISSN: 1662-5161.

[18] Moritz Böhle et al. "Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification". In: *Frontiers in Aging Neuroscience* 11 (2019), p. 194.

[19] Rishi Bommasani et al. "On the opportunities and risks of foundation models". In: *arXiv preprint arXiv:2108.07258* (2021).

[20] L. Bonnet, F. Lotte, and A. Lécuyer. "Two Brains, One Game: Design and Evaluation of a Multiuser BCI Video Game Based on Motor Imagery". In: *IEEE Transactions on Computational Intelligence and AI in Games* 5.2 (2013), pp. 185–198.

[21] George EP Box et al. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[22] Markus M. Breunig et al. "LOF: Identifying Density-Based Local Outliers". In: *SIGMOD Rec.* 29.2 (May 2000), pp. 93–104.

[23] Gijs Joost Brouwer and David J. Heeger. "Decoding and Reconstructing Color from Responses in Human Visual Cortex". In: *Journal of Neuroscience* 29.44 (2009), pp. 13992–14003.

[24] Nancy B. Carlisle et al. "Attentional Templates in Visual Working Memory". In: *Journal of Neuroscience* 31.25 (2011), pp. 9315–9322.

[25] Marisa Carrasco. "Visual attention: the past 25 years". en. In: *Vision Res.* 51.13 (July 2011), pp. 1484–1525.

[26] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. "Machine Learning Interpretability: A Survey on Methods and Metrics". In: *Electronics* 8.8 (2019).

[27] C. -Y. Chang et al. "Evaluation of Artifact Subspace Reconstruction for Automatic Artifact Components Removal in Multi-Channel EEG Recordings". In: *IEEE Transactions on Biomedical Engineering* 67.4 (2020), pp. 1114–1121.

[28] Yanan Chen and Feng Du. "Two visual working memory representations simultaneously control attention". en. In: *Sci. Rep.* 7.1 (July 2017), p. 6107.

[29] Yaqi Chu et al. "A Decoding Scheme for Incomplete Motor Imagery EEG With Deep Belief Network". In: *Frontiers in Neuroscience* 12 (2018).

[30] Radoslaw Martin Cichy, Fernando Mario Ramirez, and Dimitrios Pantazis. "Can visual information encoded in cortical columns be decoded from magnetoencephalography data in humans?" In: *NeuroImage* 121 (2015), pp. 193–204.

[31] Yücel Çimtay and E. Ekmekcioglu. "Investigating the Use of Pretrained Convolutional Neural Network on Cross-Subject and Cross-Dataset EEG Emotion Recognition". In: *Sensors* 20 (Apr. 2020).

[32] Tommy Clausner, Sarang S. Dalal, and Maité Crespo-García. "Photogrammetry-Based Head Digitization for Rapid and Accurate Localization of EEG Electrodes and MEG Fiducial Markers Using a Single Digital SLR Camera". In: *Frontiers in Neuroscience* 11 (2017), p. 264.

[33] Harris Cooper, James J. Lindsay, and Barbara Nye. "Homework in the Home: How Student, Family, and Parenting-Style Differences Relate to the Homework Process". In: *Contemporary Educational Psychology* 25.4 (2000), pp. 464–487.

[34] I. A. Corley and Y. Huang. "Deep EEG super-resolution: Upsampling EEG spatial resolution with Generative Adversarial Networks". In: *2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*. 2018, pp. 100–103.

[35] H. S. Courellis et al. "EEG channel interpolation using ellipsoid geodesic length". In: *2016 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. Oct. 2016, pp. 540–543.

[36] Janis J Daly and Jonathan R Wolpaw. "Brain–computer interfaces in neurological rehabilitation". In: *The Lancet Neurology* 7.11 (2008), pp. 1032–1043.

[37] Arnaud Delorme, Scott Makeig, and Terrence Sejnowski. "Automatic Artifact Rejection for EEG Data Using High-Order Statistics and Independant Component Analysis". In: *Proceedings of the 3rd International Independant Component Analysis and Blind Source Decomposition Conference*. Jan. 2001.

[38] Robert Desimone and John Duncan. “Neural Mechanisms of Selective Visual Attention”. In: *Annual Review of Neuroscience* 18.1 (1995), pp. 193–222.

[39] Djuwari Djuwari, Dinesh Kumar, and Marimuthu Palaniswami. “Limitations of ICA for Artefact Removal”. In: *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference* 5 (Feb. 2005), pp. 4685–8.

[40] Isabel Dombrowe, Mieke Donk, and Christian N L Olivers. “The costs of switching attentional sets”. en. In: *Atten. Percept. Psychophys.* 73.8 (Nov. 2011), pp. 2481–2488.

[41] Hauke Dose et al. “An end-to-end deep learning approach to MI-EEG signal classification for BCIs”. In: *Expert Systems with Applications* 114 (2018), pp. 532–542.

[42] Paul Downing and Chris Dodds. “Competition in visual working memory for control of search”. In: *Vis. cogn.* 11.6 (Aug. 2004), pp. 689–703.

[43] K. K. Dutta, K. Venugopal, and S. A. Swamy. “Removal of muscle artifacts from EEG based on ensemble empirical mode decomposition and classification of seizure using machine learning techniques”. In: *2017 International Conference on Inventive Computing and Informatics (ICICI)*. 2017, pp. 861–866.

[44] Johannes J. Fahrenfort et al. “From ERPs to MVPA Using the Amsterdam Decoding and Modeling Toolbox (ADAM)”. in: *Frontiers in Neuroscience* 12 (2018).

[45] Thomas C. Ferree. “Spherical Splines and Average Referencing in Scalp Electroencephalography”. In: *Brain Topography* 19.1 (Dec. 2006), pp. 43–52.

[46] Rebecca M. Foerster and Werner X. Schneider. “Involuntary top-down control by search-irrelevant features: Visual working memory biases attention in an object-based manner”. In: *Cognition* 172 (2018), pp. 37–45.

[47] Marcella Fratescu, Dirk Van Moorselaar, and Sebastiaan Mathôt. “Can you have multiple attentional templates? Large-scale replications of Van Moorselaar, Theeuwes, and Olivers (2014) and Hollingworth and Beck (2016)”. en. In: *Atten. Percept. Psychophys.* 82.3 (June 2020), p. 1536.

[48] Laurel J. Gabard-Durnam et al. “The Harvard Automated Processing Pipeline for Electroencephalography (HAPPE): Standardized Processing Software for Developmental and High-Artifact Data”. In: *Frontiers in Neuroscience* 12 (2018), p. 97.

[49] J. O. Garcia, R. Srinivasan, and J. Serences. “Near-Real-Time Feature-Selective Modulations in Human Cortex”. In: *Current Biology* 23 (2013), pp. 515–522.

[50] Justin L. Gardner and Taosheng Liu. “Inverted Encoding Models Reconstruct an Arbitrary Model Response, Not the Stimulus”. In: *eNeuro* 6.2 (2019).

[51] Justin L. Gardner et al. *MGL: Visual psychophysics stimuli and experimental design package*. Version 2.0. June 2018.

[52] RG Geocadin et al. “Early Electrophysiological and Histologic Changes After Global Cerebral Ischemia In Rats”. In: *Movement Disorders* 15.S1 (2000), pp. 14–21.

[53] Nils Gessert et al. “Towards Deep Learning-Based EEG Electrode Detection Using Automatically Generated Labels”. In: *Computer Vision and Pattern Recognition* (2019).

[54] M. M. Ghassemi et al. “You Snooze, You Win: the PhysioNet/Computing in Cardiology Challenge 2018”. In: *2018 Computing in Cardiology Conference (CinC)*. vol. 45. 2018, pp. 1–4.

[55] MM. Ghassemi et al. “Quantitative EEG Trends Predict Recovery in Hypoxic-Ischemic Encephalopathy”. In: *Critical Care Medicine* 47.10 (2019), pp. 1416–1423.

[56] Mohammad M. Ghassemi. “Life After Death: Techniques for the Prognostication of Coma Outcomes after Cardiac Arrest”. PhD thesis. Massachusetts Institute of Technology, 2018.

[57] R. Ghosh, N. Sinha, and S. K. Biswas. “Automated eye blink artefact removal from EEG using support vector machine and autoencoder”. In: *IET Signal Processing* 13.2 (2019), pp. 141–148.

[58] R. C. M. P. Gilberet et al. “Automated artifact rejection using ICA and image processing algorithms”. In: *2017 International Conference on Signal Processing and Communication (ICSPC)*. 2017, pp. 354–358.

[59] A. Goldberger et al. *PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals.* Circulation [Online]. 2000.

[60] Markus Goldstein and Andreas Dengel. *Histogram-based Outlier Score (HBOS): A fast Unsupervised Anomaly Detection Algorithm*. 2012.

[61] Tijl Grootswagers, Susan G Wardle, and Thomas A Carlson. “Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data”. en. In: *J. Cogn. Neurosci.* 29.4 (Apr. 2017), pp. 677–697.

[62] Anna Grubert, Nancy B Carlisle, and Martin Eimer. “The control of single-color and multiple-color visual search by attentional templates in working memory and in long-term memory”. en. In: *J. Cogn. Neurosci.* 28.12 (Dec. 2016), pp. 1947–1963.

[63] Anna Grubert and Martin Eimer. "Preparatory Template Activation during Search for Alternating Targets". In: *Journal of Cognitive Neuroscience* 32.8 (Aug. 2020), pp. 1525–1535.

[64] Qiong Gui et al. "A Survey on Brain Biometrics". In: *ACM Computing Surveys (CSUR)* 51.6 (2019).

[65] Britt Hadar et al. "Working Memory Load Affects Processing Time in Spoken Word Recognition: Evidence from Eye-Movements". In: *Frontiers in Neuroscience* 10 (2016).

[66] Jasper E. Hajonides et al. "Decoding visual colour from scalp electroencephalography measurements". In: *NeuroImage* 237 (2021), p. 118030. ISSN: 1053-8119.

[67] J.J. Halford et al. "Inter-rater agreement on identification of electrographic seizures and periodic discharges in ICU EEG recordings". In: *Clinical Neurophysiology* 126.9 (2015), pp. 1661–1669.

[68] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778.

[69] Rainer Hegger and Holger Kantz. "Improved false nearest neighbor method to detect determinism in time series data". In: *Physical Review E* 60.4 (1999), p. 4970.

[70] JF Hipp, AK Engel, and M Siegel. "Oscillatory Synchronization In Large-Scale Cortical Networks Predicts Perception". In: *Neuron* 69.2 (2011), pp. 387–396.

[71] L. J. Hirsch et al. "American Clinical Neurophysiology Society's Standardized Critical Care EEG Terminology: 2012 version". In: *Journal of Clinical Neurophysiology* (2013), pp. 1–27.

[72] Andrew Hollingworth and Valerie M Beck. "Memory-based attention capture when multiple items are maintained in visual working memory". In: *J. Exp. Psychol. Hum. Percept. Perform.* 42.7 (July 2016), pp. 911–917.

[73] Marcia Hon and Naimul Mefraz Khan. "Towards Alzheimer's disease classification through transfer learning". In: *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2017, pp. 1166–1169.

[74] Roos Houtkamp and Pieter R Roelfsema. "Matching of visual input to only one item at any one time". en. In: *Psychol. Res.* 73.3 (May 2009), pp. 317–326.

[75] Jeremy Howard and Sebastian Ruder. "Universal Language Model Fine-tuning for Text Classification". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 328–339.

[76] JASP Team. *JASP (Version 0.16.4)[Computer software]*. 2022.

[77] Soroush Javidi et al. "Kurtosis based blind source extraction of complex noncircular signals with application in EEG artifact removal in real-time". In: *Frontiers in Neuroscience* 5 (2011), p. 105.

[78] X Jia et al. "Early Electrophysiologic Markers Predict Functional Outcome Associated With Temperature Manipulation After Cardiac Arrest In Rats". In: *Critical Care Medicine* 36.6 (2008), p. 1909.

[79] Huaizu Jiang et al. "Super SloMo: High Quality Estimation of Multiple Intermediate Frames for Video Interpolation". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (June 2018).

[80] Tzyy-Ping Jung et al. "Removing electroencephalographic artifacts by blind source separation". In: *Psychophysiology* 37 (Mar. 2000), pp. 163–178.

[81] Steven M Kay. *Fundamentals of statistical signal processing*. Prentice Hall PTR, 1993.

[82] W. Khalifa et al. "A survey of EEG based user authentication schemes". In: *2012 8th International Conference on Informatics and Systems (INFOS)*. 2012.

[83] Reza Khanmohammadi et al. *MambaNet: A Hybrid Neural Network for Predicting the NBA Playoffs*. 2022.

[84] Michael King and Brooke Macnamara. "Three visual working memory representations simultaneously control attention". In: *Scientific Reports* 10 (June 2020).

[85] HansPeter Kriegel, Matthias Schubert, and Arthur Zimek. "Angle-Based Outlier Detection in High-Dimensional Data". In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2008, pp. 444–452.

[86] Tómas Kristjánsson and Árni Kristjánsson. "Foraging through multiple target categories reveals the flexibility of visual working memory". In: *Acta Psychol. (Amst.)* 183 (Feb. 2018), pp. 108–115.

[87] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Commun. ACM* 60.6 (May 2017), pp. 84–90.

[88] Pradeep Kumar et al. "Envisioned speech recognition using EEG sensors". In: *Personal and Ubiquitous Computing* (Sept. 2017), pp. 1–15.

[89] Margaret Lech et al. "Real-Time Speech Emotion Recognition Using a Pre-trained Image Classification Network: Effects of Bandwidth Reduction and Companding". In: *Frontiers in Computer Science* 2 (2020), p. 14.

[90] B B Lee, P R Martin, and A Valberg. "The physiological basis of heterochromatic flicker photometry demonstrated in the ganglion cells of the macaque retina". In: *The Journal of Physiology* 404.1 (Oct. 1988), pp. 323–347.

[91] Jarrod A Lewis-Peacock et al. "Neural evidence for a distinction between short-term memory and the focus of attention". en. In: *J. Cogn. Neurosci.* 24.1 (Jan. 2012), pp. 61–79.

[92] Yurong Li et al. "EEG-based intention recognition with deep recurrent-convolution neural network: Performance and channel selection by Grad-CAM". in: *Neurocomputing* 415 (2020), pp. 225–233.

[93] C. Lin, S. Tsai, and L. Ko. "EEG-Based Learning System for Online Motion Sickness Level Estimation in a Dynamic Vehicle Environment". In: *IEEE Transactions on Neural Networks and Learning Systems* 24.10 (2013), pp. 1689–1700.

[94] Yuan-Pin Lin and Tzyy-Ping Jung. "Improving EEG-Based Emotion Classification Using Conditional Transfer Learning". In: *Frontiers in Human Neuroscience* 11 (2017), p. 334.

[95] Zhouhan Lin et al. "Neural Networks with Few Multiplications". In: *CoRR* (Oct. 2015).

[96] T Liu and M Jigo. "Limits in feature-based attention to multiple colors". In: *Attention, perception, and psychophysics* 79 (2017), pp. 2327–2337.

[97] Taosheng Liu, Mark W Becker, and Michael Jigo. "Limited featured-based attention to multiple features". en. In: *Vision Res.* 85 (June 2013), pp. 36–44.

[98] Taosheng Liu, Dylan Cable, and Justin L. Gardner. "Inverted Encoding Models of Human Population Response Conflate Noise and Neural Tuning Width". In: *Journal of Neuroscience* 38.2 (2018), pp. 398–408.

[99] Yezheng Liu et al. "Generative Adversarial Active Learning for Unsupervised Outlier Detection". In: *Proceedings of the IEEE Transactions on Knowledge and Data Engineering*. 2019, pp. 1–1.

[100] Erik K. St. Louis et al. *Electroencephalography (EEG): An Introductory Text and Atlas of Normal and Abnormal Findings in Adults, Children, and Infants*. American Epilepsy Society, 2016.

[101] Roy Luria et al. "The contralateral delay activity as a neural measure of visual working memory". In: *Neurosci. Biobehav. Rev.* 62 (Mar. 2016), pp. 100–108.

[102] Wei Ma, Masud Husain, and Paul Bays. "Changing concepts of working memory". In: *Nature neuroscience* 17 (Mar. 2014), pp. 347–56.

[103] Christopher D. Manning. "Computational Linguistics and Deep Learning". In: *Computational Linguistics* 41.4 (2015), pp. 701–707.

[104] Michael Mannino, Yanjuan Yang, and Young Ryu. "Classification algorithm sensitivity to training data with non representative attribute noise". en. In: *Decis. Support Syst.* 46.3 (Feb. 2009), pp. 743–751.

[105] Jasna Martinovic et al. "Neural mechanisms of divided feature-selective attention to colour". en. In: *Neuroimage* 181 (Nov. 2018), pp. 670–682.

[106] Ben McCartney et al. "A zero-shot learning approach to the development of brain-computer interfaces for image retrieval". In: *PLOS ONE* 14.9 (Sept. 2019), pp. 1–21.

[107] Risto Miikkulainen et al. "Chapter 15 - Evolving Deep Neural Networks". In: *Artificial Intelligence in the Age of Neural Networks and Brain Computing*. Ed. by Robert Kozma et al. Academic Press, 2019, pp. 293–312.

[108] Jianliang Min, Ping Wang, and Jianfeng Hu. "Driver fatigue detection through multiple entropy fusion analysis in an EEG-based system". In: *PLOS ONE* 12 (Dec. 2017), e0188756.

[109] Fumikazu Miwakeichi et al. "A comparison of non-linear non-parametric models for epilepsy data". In: *Computers in Biology and Medicine* 31.1 (2001), pp. 41–57.

[110] Dirk van Moorselaar, Jan Theeuwes, and Christian N L Olivers. "In competition for the attentional template: can multiple items within visual working memory guide attention?" en. In: *J. Exp. Psychol. Hum. Percept. Perform.* 40.4 (Aug. 2014), pp. 1450–1464.

[111] J T Mordkoff and S Yantis. "An interactive race model of divided attention". en. In: *J. Exp. Psychol. Hum. Percept. Perform.* 17.2 (May 1991), pp. 520–538.

[112] Thomas Naselaris et al. "Bayesian Reconstruction of Natural Images from Human Brain Activity". In: *Neuron* 63.6 (2009), pp. 902–915.

[113] E. Nedelcu et al. "Artifact detection in EEG using machine learning". In: *2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*. 2017, pp. 77–83.

[114] Petr Nejedly et al. "Intracerebral EEG Artifact Identification Using Convolutional Neural Networks". In: *Neuroinformatics* 17 (Aug. 2018).

[115] Petr Nejedly et al. "Intracerebral EEG Artifact Identification Using Convolutional Neural Networks". In: *Neuroinformatics* 17 (Aug. 2018).

[116] Andrea Nemcova et al. *Brno University of Technology ECG Quality Database (BUT QDB)*. PhysioNet. 2021.

[117] Andrew Ng. *Nuts and bolts of building AI applications using deep learning*. NIPS Tutorial. 2016.

[118] Shinji Nishimoto et al. "Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies". In: *Current Biology* 21.19 (2011), pp. 1641–1646.

[119] Sean Noah et al. "Neural Mechanisms of Attentional Control for Objects: Decoding EEG Alpha When Anticipating Faces, Scenes,and Tools". In: *Journal of Neuroscience* 40.25 (2020), pp. 4913–4924.

[120] Seung-Hyeon Oh, Yu-Ri Lee, and Hyoung-Nam Kim. "A novel EEG feature extraction method using Hjorth parameter". In: *International Journal of Electronics and Electrical Engineering* 2.2 (2014), pp. 106–110.

[121] Christian N L Olivers. "What drives memory-driven attentional capture? The effects of memory type, display type, and search type". en. In: *J. Exp. Psychol. Hum. Percept. Perform.* 35.5 (Oct. 2009), pp. 1275–1291.

[122] Christian N L Olivers, Frank Meijer, and Jan Theeuwes. "Feature-based memory-driven attentional capture: visual working memory content affects visual attention". en. In: *J. Exp. Psychol. Hum. Percept. Perform.* 32.5 (Oct. 2006), pp. 1243–1265.

[123] Christian N L Olivers et al. "Different states in visual working memory: when it guides attention and when it does not". en. In: *Trends Cogn. Sci.* 15.7 (July 2011), pp. 327–334.

[124] Sławomir Opałka et al. "Multi-Channel Convolutional Neural Networks Architecture Feeding for Effective EEG Mental Tasks Classification". In: *Sensors* 18.10 (2018).

[125] Alan V Oppenheim and Ronald W Schafer. "From frequency to quefrency: A history of the cepstrum". In: *IEEE signal processing Magazine* 21.5 (2004), pp. 95–106.

[126] Eduard Ort, Johannes J Fahrenfort, and Christian N L Olivers. "Lack of free choice reveals the cost of having to search for more than one object". en. In: *Psychol. Sci.* 28.8 (Aug. 2017), pp. 1137–1147.

[127] Eduard Ort and Christian N L Olivers. "The capacity of multiple-target search". en. In: *Vis. cogn.* 28.5-8 (Sept. 2020), pp. 330–355.

[128] Eduard Ort et al. "Humans can efficiently look for but not select multiple visual objects". en. In: *Elife* 8 (Aug. 2019).

[129] Deepak Pathak et al. “Context Encoders: Feature Learning by Inpainting”. In: *CVPR*. 2016.

[130] Alex Pentland. “Maximum likelihood estimation: The best PEST”. in: *Perception & Psychophysics* 28 (1980), pp. 377–379.

[131] F. Perrin et al. “Spherical splines for scalp potential and current density mapping”. In: *Electroencephalography and Clinical Neurophysiology* 72.2 (1989), pp. 184–187.

[132] S. Petrichella et al. “Channel interpolation in TMS-EEG: A quantitative study towards an accurate topographical representation”. In: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Aug. 2016, pp. 989–992.

[133] S. Phadikar, N. Sinha, and R. Ghosh. “Automatic EEG eyeblink artefact identification and removal technique using independent component analysis in combination with support vector machines and denoising autoencoder”. In: *IET Signal Processing* 14.6 (2020), pp. 396–405.

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

[135] Lindsay Plater et al. “Revisiting the role of visual working memory in attentional control settings”. en. In: *Vis. cogn.* 30.5 (May 2022), pp. 318–338.

[136] U Pomper and Ansorge U. “Theta-Rhythmic Oscillation of Working Memory Performance”. In: *Psychological Science* 32.11 (2021), pp. 1801–1810.

[137] Nicolaas Prins and Frederick A. A. Kingdom. “Applying the Model-Comparison Approach to Test Specific Research Hypotheses in Psychophysical Research Using the Palamedes Toolbox”. In: *Frontiers in Psychology* 9 (2018).

[138] Xing Qian et al. “Brain-computer-interface-based intervention re-normalizes brain functional network topology in children with attention deficit/hyperactivity disorder”. In: *Translational Psychiatry* 8 (Aug. 2018), p. 149.

[139] Pedro Reis and Matthias Lochmann. “Using a Motion Capture System for Spatial Localization of EEG Electrodes.” In: *Frontiers in Neuroscience* 9 (Apr. 2015).

[140] Quick Rf. “A vector-magnitude model of contrast detection.” In: *Kybernetika* 16 (1974), pp. 65–67.

[141] Michael Roberts et al. “Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans”. In: *Nature Machine Intelligence* 3 (Mar. 2021).

[142] Sheldon M Ross. *Introductory statistics*. Academic Press, 2017.

[143] Sari Saba-Sadiya, Tuka Alhanai, and Mohammad M Ghassemi. "Feature Imitating Networks". In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, pp. 4128–4132.

[144] Sari Saba-Sadiya, Eric Chantland, and Taosheng Liu. *Decoing EEG from Passive Viewing*. github.com/sari-saba-sadiya/DEPV. 2020.

[145] Sari Saba-Sadiya et al. "EEG Channel Interpolation Using Deep Encoder-decoder Networks". In: *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2020.

[146] Sari Saba-Sadiya et al. "Unsupervised EEG Artifact Detection and Correction". In: *Frontiers in Digital Health* 2 (2020), p. 57.

[147] Sari Sadiya, Tuka Alhanai, and Mohammad Ghassemi. "Artifact Detection and Correction in EEG data: A Review". In: *Proceedings of the 10th International IEEE/EMBS Conference on Neural Engineering (NER)*. May 2021, pp. 495–498.

[148] Melissa Saenz, Giedrius T Buracas, and Geoffrey M Boynton. "Global effects of feature-based attention in human visual cortex". en. In: *Nat. Neurosci.* 5.7 (July 2002), pp. 631–632.

[149] Robin Tibor Schirrmeister et al. "Deep learning with convolutional neural networks for EEG decoding and visualization". In: *Human Brain Mapping* 38.11 (Aug. 2017), pp. 5391–5420.

[150] J Schoffelen and J Gross. "Source Connectivity Analysis With MEG and EEG". in: *Human Brain Mapping* 30.6 (2009), pp. 1857–1865.

[151] B. Schölkopf et al. "Estimating the Support of a High-Dimensional Distribution". In: *Neural Computation* 13.7 (2001), pp. 1443–1471.

[152] Jeremy Schwark et al. "False feedback increases detection of low-prevalence targets in visual search". en. In: *Atten. Percept. Psychophys.* 74.8 (Nov. 2012), pp. 1583–1589.

[153] V. S. Selvam and S. Shenbagadevi. "Brain tumor detection using scalp eeg with modified Wavelet-ICA and multi layer feed forward neural network". In: *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2011, pp. 6104–6109.

[154] Nima Bigdely Shamlo et al. "EyeCatch: Data-mining over half a million EEG independent components to construct a fully-automated eye-component detector". In: *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*

(2013), pp. 5845–5848.

[155] C. E. Shannon. “Communication in the Presence of Noise”. In: *Proceedings of the IRE* 37.1 (1949), pp. 10–21.

[156] Claude E Shannon and Warren Weaver. *The mathematical theory of communication*. University of Illinois press, 1998.

[157] H Shin et al. “Quantitative EEG And Effect Of Hypothermia On Brain Recovery After Cardiac Arrest”. In: *IEEE Transactions on Biomedical Engineering* 53.6 (2006), pp. 1016–1023.

[158] Hyun-Chool Shin et al. “A subband-based information measure of EEG during brain injury and recovery after cardiac arrest”. In: *IEEE Transactions on Biomedical Engineering* 55.8 (2008), pp. 1985–1990.

[159] M.L. Shyu et al. *A Novel Anomaly Detection Scheme Based on Principal Component Classifier*. AD-a465 712. miami univ coral gables fl Department of electrical and computer engineering, 2003.

[160] Karen Simonyan and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *International Conference on Learning Representations*. 2015.

[161] Vorasith Siripornpanich et al. “Enhancing Brain Maturation Through a Mindfulness-Based Education in Elementary School Children: a Quantitative EEG Study”. In: *Mindfulness* 9 (Mar. 2018).

[162] B. Somers, T. Francart, and A. Bertrand. “A generic EEG artifact removal algorithm based on the multi-channel Wiener filter.” In: *Journal of neural engineering* 15 3 (2018), p. 036007.

[163] Anthony C.K. Soong et al. “Systematic comparisons of interpolation techniques in topographic brain mapping”. In: *Electroencephalography and Clinical Neurophysiology* 87.4 (1993), pp. 185–195.

[164] David Soto et al. “Early, involuntary top-down guidance of attention from working memory”. en. In: *J. Exp. Psychol. Hum. Percept. Perform.* 31.2 (Apr. 2005), pp. 248–261.

[165] Thomas C. Sprague, Geoffrey M. Boynton, and John T. Serences. “Inverted encoding models estimate sensible channel responses for sensible models”. In: *bioRxiv* (2019).

[166] CJ Stam, G Nolte, and A Daffertshofer. “Phase Lag Index: Assessment of Functional Connectivity From Multi Channel EEG and MEG With Diminished Bias From Common

Sources". In: *Human Brain Mapping* 28.11 (2007), pp. 1178–1193.

[167] John M Stern. *Atlas of EEG patterns*. Lippincott Williams & Wilkins, 2005.

[168] David Strayer, Frank Drews, and William Johnston. "Cell Phone-Induced Failures of Visual Attention During Simulated Driving". In: *Journal of experimental psychology. Applied* 9 (Apr. 2003), pp. 23–32.

[169] Michael J Stroud et al. "Using the dual-target cost to explore the nature of search target representations". en. In: *J. Exp. Psychol. Hum. Percept. Perform.* 38.1 (Feb. 2012), pp. 113–122.

[170] I. Sturm et al. "Interpretable deep neural networks for single-trial EEG classification". In: *Journal of Neuroscience Methods* 274 (2016), pp. 141–145.

[171] David W. Sutterer et al. "Item-specific delay activity demonstrates concurrent storage of multiple active neural representations in working memory". In: *PLOS Biology* 17 (Apr. 2019), pp. 1–25.

[172] MC Tjepkema-Cloostermans et al. "A Cerebral Recovery Index (CRI) For Early Prognosis In Patients After Cardiac Arrest". In: *Critical Care* 17 (2013), R252.

[173] Anne M. Treisman and Garry Gelade. "A feature-integration theory of attention". In: *Cognitive Psychology* 12.1 (1980), pp. 97–136.

[174] PJ Uhlhaas and W Singer. "Abnormal Neural Oscillations and Synchrony in Schizophrenia". In: *Nature Reviews Neuroscience* 11.2 (2010), pp. 100–113.

[175] Markus Ullsperger and Stefan Debener. *Simultaneous EEG and fMRI: recording, analysis, and application*. Oxford University Press, 2010, p. 132.

[176] Jose Antonio Urigüen and Begoña Garcia-Zapirain. "EEG artifact removal—state-of-the-art and guidelines". In: *Journal of Neural Engineering* 12.3 (Apr. 2015), p. 031001.

[177] G. Vecchiato et al. "Enhance of theta EEG spectral activity related to the memorization of commercial advertisings in Chinese and Italian subjects". In: *2011 4th International Conference on Biomedical Engineering and Informatics (BMEI)*. vol. 3. 2011, pp. 1491–1494.

[178] J Vibell et al. "Temporal order is coded temporally in the brain: early event-related potential latency shifts underlying prior entry in a cross-modal temporal order judgment task". en. In: *Journal of Cognitive Neuroscience.* 19.1 (Jan. 2007), pp. 109–120.

[179] De Vries, I E J Van Driel, and J Olivers. "Decoding the status of working memory representations in preparation of visual selection". In: *NeuroImage* 191 (2019), pp. 549–

559.

[180] Thaddeus S. Walczak and Sudhansu Chokroverty. “Electroencephalography, Electromyography, and Electro-Oculography. General Principles and Basic Technology”. In: *Sleep Disorders Medicine*. Elsevier Inc., Dec. 2009, pp. 157–181.

[181] Yi Wang et al. “Wide-Context Semantic Image Extrapolation”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 1399–1408.

[182] Melissa R Warden and Earl K Miller. “Task-dependent changes in short-term memory in the prefrontal cortex”. en. In: *J. Neurosci.* 30.47 (Nov. 2010), pp. 15801–15810.

[183] Melissa R Warden and Earl K Miller. “The representation of multiple objects in prefrontal neuronal delay activity”. en. In: *Cereb. Cortex* 17 Suppl 1.suppl 1 (Sept. 2007), pp. i41–50.

[184] Wen Wen et al. “Tracking Neural Markers of Template Formation and Implementation in Attentional Inhibition under Different Distractor Consistency”. In: *Journal of Neuroscience* 42.24 (2022), pp. 4927–4936.

[185] Alan Wolf et al. “Determining Lyapunov exponents from a time series”. In: *Physica D: Nonlinear Phenomena* 16.3 (1985), pp. 285–317.

[186] Jeremy Wolfe. “Guided Search 2.0 A revised model of visual search”. In: *Psychonomic Bulletin and Review* 1 (June 1994), pp. 202–238.

[187] Jeremy M Wolfe. “Guided Search 6.0: An updated model of visual search”. en. In: *Psychon. Bull. Rev.* 28.4 (Aug. 2021), pp. 1060–1092.

[188] Jeremy M Wolfe. “Visual attention: The multiple ways in which history shapes selection”. en. In: *Curr. Biol.* 29.5 (Mar. 2019), R155–R156.

[189] Jeremy M. Wolfe et al. “How fast can you change your mind? The speed of top-down guidance in visual search”. In: *Vision Research* 44.12 (2004). Visual Attention, pp. 1411–1426.

[190] Michael Wolff et al. “Revealing hidden states in visual working memory using electroencephalography”. In: *Frontiers in Systems Neuroscience* 9 (2015).

[191] Jia Wu et al. “Hyperparameter optimization for machine learning models based on Bayesian optimization”. In: *Journal of Electronic Science and Technology* 17.1 (2019), pp. 26–40.

[192] G. Xu et al. “A Deep Transfer Convolutional Neural Network Framework for EEG Signal Classification”. In: *IEEE Access* 7 (2019), pp. 112767–112776.

[193] Yang Yang et al. “LaFIn: Generative Landmark Guided Face Inpainting”. In: *arXiv preprint*. 2019.

[194] Haoming Zhang et al. *EEGdenoiseNet: A benchmark dataset for deep learning solutions of EEG denoising*. 2020.

[195] X. Zhang et al. “Accelerating Very Deep Convolutional Networks for Classification and Detection”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.10 (2016), pp. 1943–1955.

[196] Yue Zhao, Zain Nasrullah, and Zheng Li. “PyOD: A Python Toolbox for Scalable Outlier Detection”. In: *Journal of Machine Learning Research* 20.96 (2019), pp. 1–7.

[197] Yue Zhao et al. “LSCP: Locally Selective Combination in Parallel Outlier Ensembles”. In: *SDM*. 2019.

[198] Zhidong Zhao and Yefei Zhang. “SQI Quality Evaluation Mechanism of Single-Lead ECG Signal Based on Simple Heuristic Fusion and Fuzzy Comprehensive Evaluation”. In: *Frontiers in Physiology* 9 (2018), p. 727.

[199] Xuyang Zhu et al. “Separated channel convolutional neural network to realize the training free motor imagery BCI systems”. In: *Biomedical Signal Processing and Control* 49 (2019), pp. 396–403.

[200] Elana Zion Golumbic et al. “Visual Input Enhances Selective Speech Envelope Tracking in Auditory Cortex at a “Cocktail Party””. In: *Journal of Neuroscience* 33.4 (2013), pp. 1417–1426.

[201] Igor Zyma et al. “Electroencephalograms during Mental Arithmetic Task Performance”. In: *Data* 4 (Jan. 2019).

APPENDIX A

DECODING TARGET-ABSENT AND FALSE ALARM TRIALS

There was no significant decoding for either one or two-cue target-absent trials. Difference was similarly not significant. However decoding in two-cue target absent trials was higher than for one-cue target absent trials. This correlated to the behavioral results observed. Namely the higher false-alarm rate in the two-cue condition.

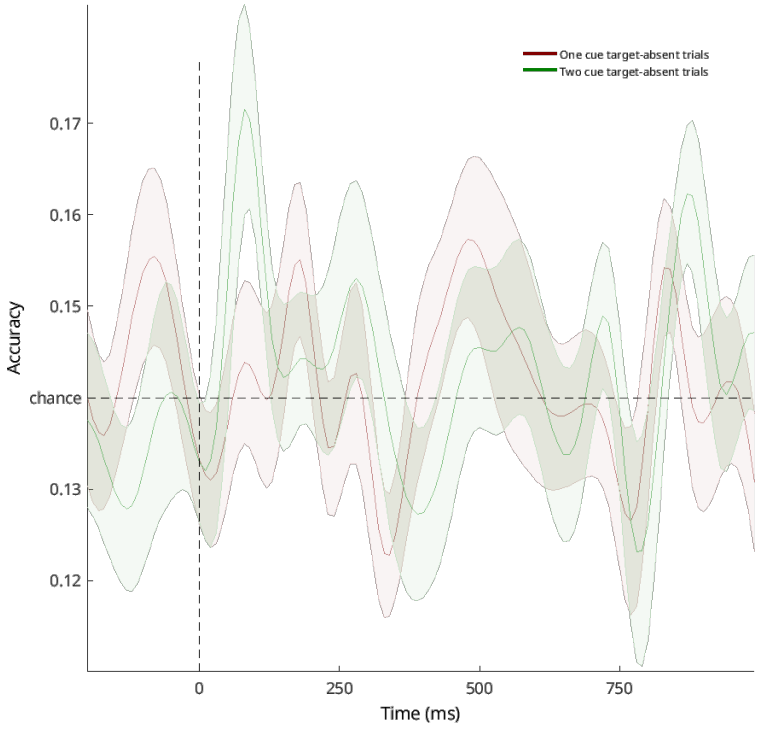

Figure A.1 Decoding target-absent EEG trials. There were no significant decoding clusters for either the one or two-cue target-absent trials.

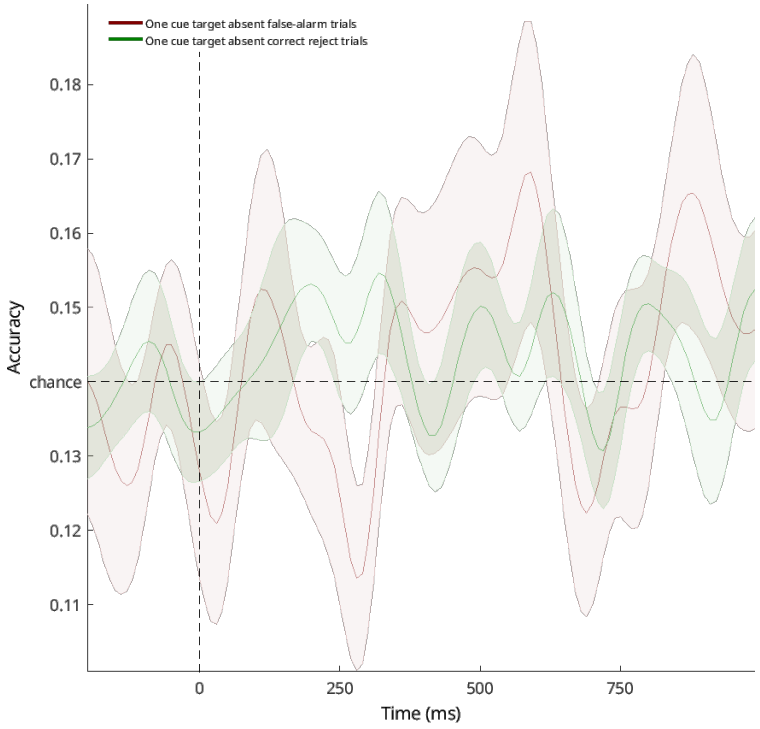

Figure A.2 Decoding one-cue target-absent EEG trials. There were no significant decoding clusters for either the one-cue target-absent false-alarm or correct reject trials.

APPENDIX B

DECODING SIMULATED NOISE

To insure that our preprocessing does not inflate classifier accuracy. We simulated noise by producing a noise trial n_i for every real EEG trial x_i, the label of the noise trial corresponded with the label of the real trial, and for every electrode we simulated the noise n_{ij} by sampling from a Gaussian distribution with the same mean and standard deviation as x_{ij} and smoothing the resulting signal using a moving mean (4-sample window). This is a stringent test, as in non-EEG data the standard deviation and mean could potentially be meaningful for signal classification, however, this should not be the case in our EEG data (especially considering the colors are isoluminant).

As can be seen in figure B.1, there were no significant classification clusters for noise data generated using target-present correct one-cue trials.

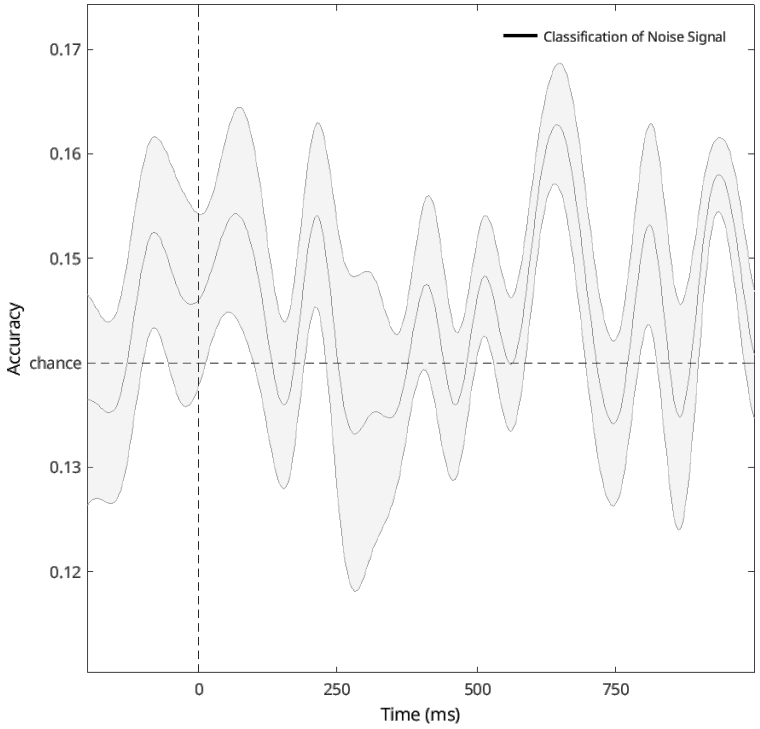

Figure B.1 Classification of noise data generated using target-present correct one-cue trials.

APPENDIX C

MAHALANOBIS DISTANCE DECODING RESULTS

In the literature, EEG trials are always averaged before classification using the Mahalanobis distance methods [184, 190]. For instance, the researchers might divide the data into five equal parts, each with an equal number of trials from each label. For every label, the trials are averaged to produce five trials, five-fold (leave-one-out) classification is used to get decoding accuracy. While x-fold iterations are the default for the ADAM LDA classifier, the number of trials in each fold is greater than one, resulting in less noisy classification performance (Figure c left). We produced a new division of trials before every six-fold loop, adding another layer of selection, and increasing the number of classifiers we train from 10 to 50 (10 different data divisions, each producing 5-folds).

While the overall performance is comparable between the default LDA classifier and the Mahalanobis distance classifier we implemented. Moreover, the overall trend in the data remained the same. For instance, target-present correct decoding was significant earlier (and lasted longer) in the one-cue condition in comparison to the two-cue condition (Figure A.2 right). However, the added layer of averaging seems to mask the transient effects, and there was no longer a significant cluster of difference between the two cues.

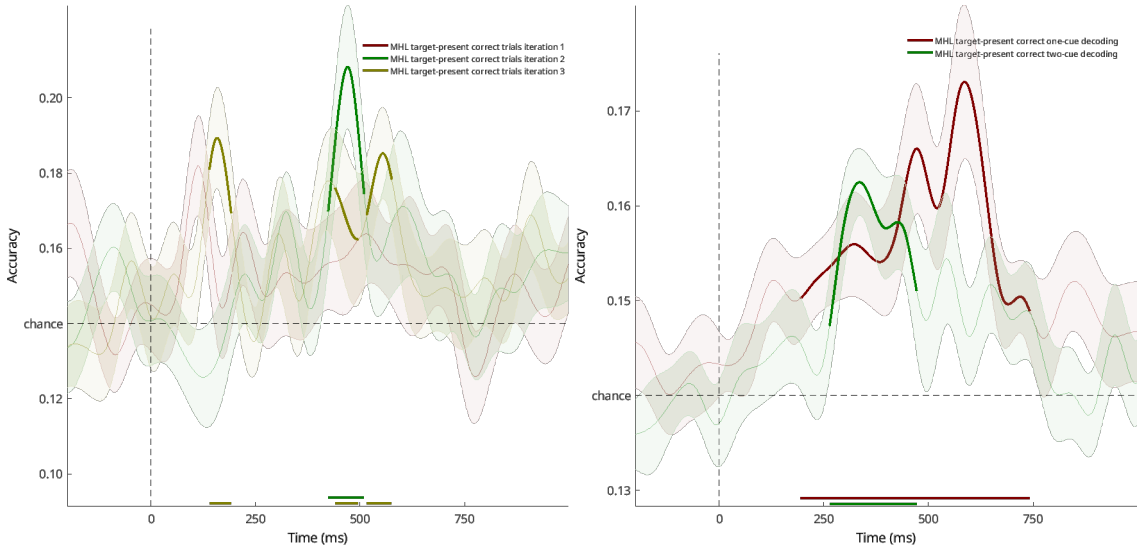

Figure C.1 Left: Three Iterations of Mahalanobis Classification, each is the result of a five-fold classification, in each iteration the trial were randomly assigned to a different fold, and trials for each label were averaged, producing five trials per label in total. Right: Target-present correct results using 10 iterations of the 5-fold Mahalanobis classifiers.

APPENDIX D

BAYESIAN ANALYSIS

Default JASP prior of 0.707 on the Cauchy scale with 95% credibility interval was used for all Bayesian paired t-test. The null hypotheses were that the hit rate, false-alarm rate, and d-prime did not differ across three and no-cue trials. Bayesian factors, error percentages, median effect sizes, and effect size confidence intervals are provided in the table below. Follow up robustness check demonstrates that the null hypothesis remains more likely than the alternative for a wide range of priors.

Measure being tested	BF_{01}	error %	Median effect size	Effect size confidence interval
Hit rate	4.628	0.024	−0.022	[−0.398, 0.353]
False-Alarm rate	2.879	0.026	0.187	[−0.191, 0.574]
d-prime	2.122	0.026	−0.242	[−0.634, 0.14]

Table D.1 Results of Bayesian paired t-test for hit rate, false-alarm, and d-prime equality in three and no-cue trials.

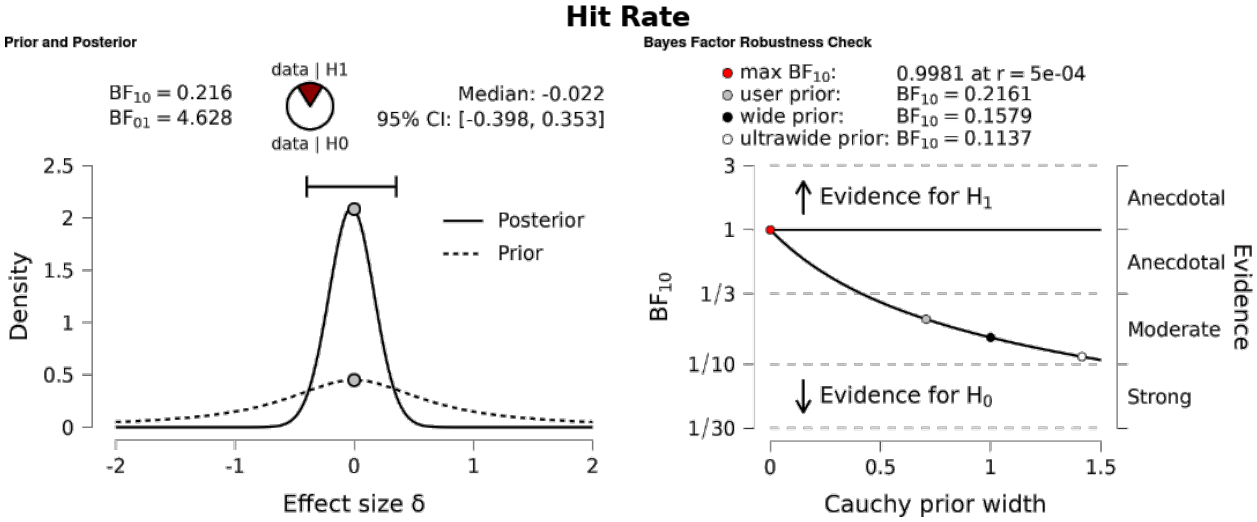

Figure D.1 Bayes factor robustness check for hit rate Bayesian paired t-test analysis.

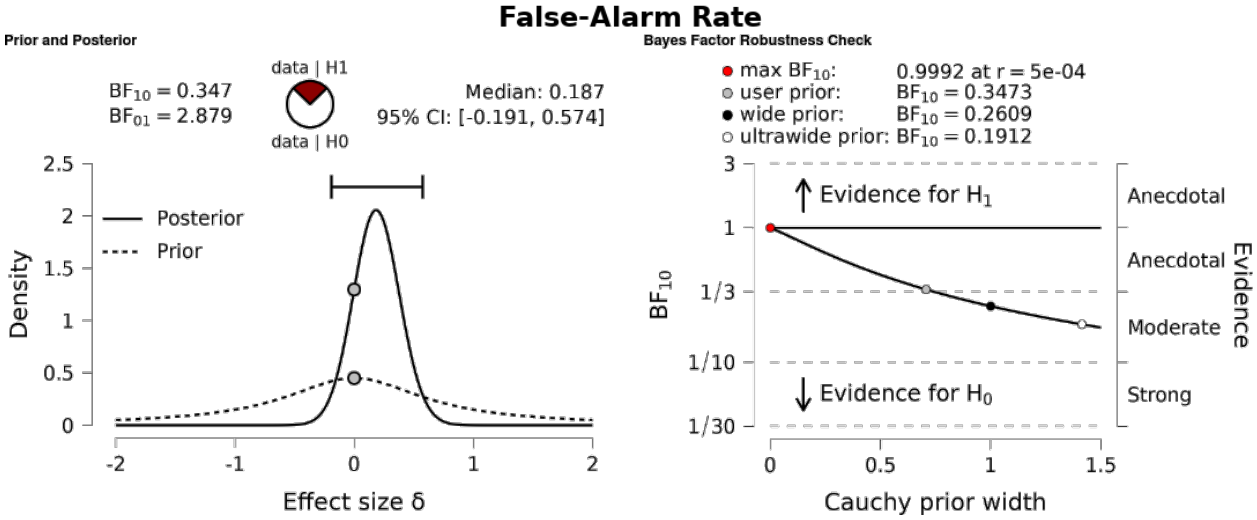

Figure D.2 Bayes factor robustness check for false-alarm rate Bayesian paired t-test analysis.

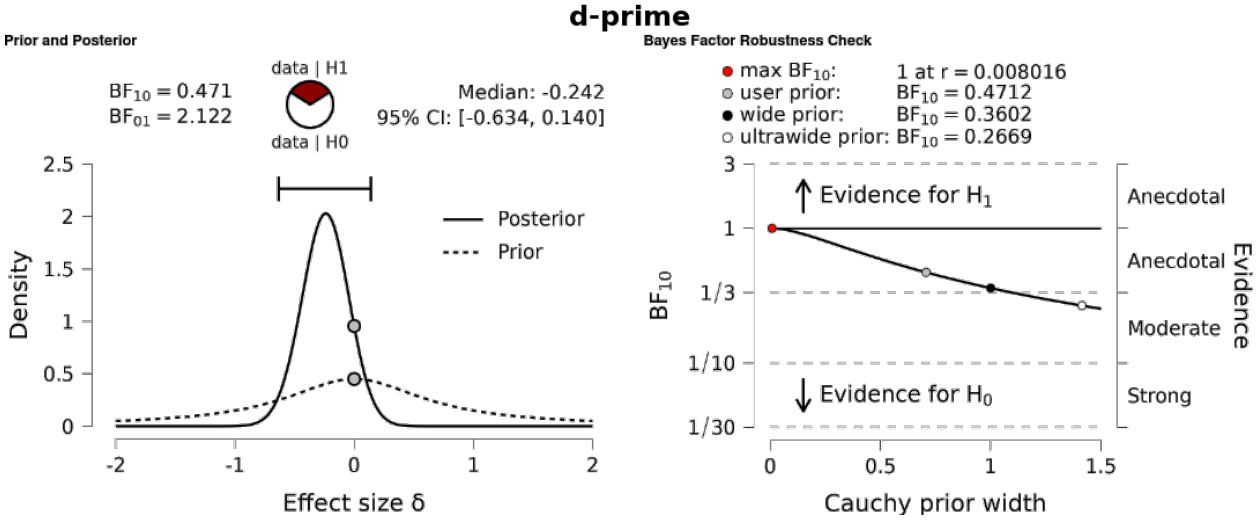

Figure D.3 Bayes factor robustness check for d-prime Bayesian paired t-test analysis.