



**NANYANG
TECHNOLOGICAL
UNIVERSITY**

**EEG-BASED ATTENTION RECOGNITION
USING MACHINE LEARNING**

SUN HAOQI

Interdisciplinary Graduate School

2017

EEG-BASED ATTENTION RECOGNITION USING MACHINE LEARNING

SUN HAOQI

Interdisciplinary Graduate School

Energy Research Institute @ NTU (ERI@N)

A thesis submitted to the Nanyang Technological University

in partial fulfillment of the requirement for the degree of

Doctor of Philosophy

2017

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

.....

Date

.....

SUN HAOQI

Acknowledgments

The Ph.D degree is never completed alone. The ability to do high-quality research independently is developed by every discussion and guidance from all the people I met during the years.

I would like to give my great respect to my supervisor Professor Huang Guang-Bin. He constantly offers valuable guidance and suggestions, in both strategic and philosophical levels. His style of thinking, making decision and writing has deeply affected me. He also provides me great opportunities in pursuing an advanced academic career.

I'm really grateful to my co-supervisor Dr. Olga Sourina. It is her who leads me into the field of EEG and further into neuroscience. I sincerely appreciate the patient, warm and invaluable guidance in both detailed and philosophical levels throughout the years. Without her I cannot have the opportunity to finish this thesis.

Special thanks to BMW group to enable the Future Mobility Research Lab. Thanks Dr. Guo Hongliang for the detailed helps from the mathematical aspects. Thanks Dr. Yang Yan for all the instructions on the presentation skills. Thanks Dr. Felix Klanner and Cornelia Denk for the great support from the BMW side. Thanks Dr. Liu Yisi from Cognitive Human-Computer Interaction lab in Fraunhofer IDM@NTU for discussion on EEG. Thanks my lab mates, Cao Zhiguang, Chamara Kasun, Han Wei, Liu Tianchi and Cheng Zihao for all the helpful discussions we had.

Last but importantly, I really thank the love, care and support from my family members and my girl friend Gong Qin. You are the strongest power in my heart that drive us to the bright future.

Table of Contents

Acknowledgments	i
Summary	vii
Table Captions	xi
Figure Captions	xiv
Abbreviations	xv
1 Introduction	1
1.1 Background and Motivation	2
1.2 Research Scope and Questions	3
1.3 Contributions and Novelty	5
1.3.1 EEG Channel Selection with Different Features	6
1.3.2 Vigilance Stage Model based on EEG	6
1.3.3 Sparse Feature Learning with Joint STDP	7
1.4 Thesis Overview	7
2 Literature Review	9

2.1	Experiment Design	10
2.1.1	Psychological Tests	10
2.1.2	Labeling Methods	11
2.2	EEG Artifact Removal	13
2.3	EEG Feature Extraction	14
2.4	EEG Feature Selection and Channel Selection	15
2.5	EEG Feature Learning	16
2.5.1	Restricted Boltzmann Machine	17
2.5.2	Extreme Learning Machine Autoencoder	18
2.6	Studies on Selective Attention	19
2.7	Studies on Vigilance	20
2.8	Public EEG Datasets Related to Attention	21
2.9	Summary	22
3	Recognition of Selective Attention	23
3.1	Background and Motivation	24
3.2	Method	25
3.2.1	Experiment Setup	25
3.2.2	Preprocessing and Feature Extraction	26
3.2.3	Classification	27
3.3	Results	28
3.3.1	Comparison with Results on Public Mental Task Dataset	29
3.3.2	Comparison with A Study using SCT Task	30
3.3.3	Comparison with A Study with the Best Accuracy in the Literature	31

3.3.4	Comparison with SVM	32
3.4	Summary	33
4	EEG Channel Selection with Different Features (CSDF)	35
4.1	Background and Motivation	36
4.2	Method	37
4.2.1	Least Square Linear Discriminant Analysis	37
4.2.2	Channel-Wise Sparseness for Channel Selection	39
4.2.3	Feature-Wise Sparseness for Determining Features in Each Channel	40
4.2.4	Flowchart of the CSDF Algorithm	40
4.2.5	EEG Features	42
4.3	Results	43
4.3.1	Validation of CSDF on Public Mental Task Dataset	43
4.3.2	Comparison of Testing Accuracy for the TOVA Experiment	45
4.3.3	Comparison of Selected Features for the TOVA Experiment	48
4.3.4	Comparison of Selected Channels for the TOVA Experiment	49
4.3.5	Relevant EEG Channels and Features in Selective Attention	50
4.4	Discussion	53
4.4.1	Neurophysiological Interpretations of Selected Channels and Features	53
4.4.2	Limitations	54
4.5	Summary	55
5	Recognition of Sustained Attention	57

5.1	Background and Motivation	58
5.2	Method	59
5.2.1	Experiment Setup	59
5.2.2	Preprocessing and Labelling	60
5.2.3	Machine Learning Model	61
5.3	Results	63
5.3.1	Performance Comparison	63
5.3.2	Vigilance Stages	66
5.3.3	The Learned Hidden Weights and Features	71
5.4	Discussion	77
5.5	Summary	77
6	Joint Weight-Delay Spike-Timing Dependent Plasticity for Sparse EEG Feature Learning	79
6.1	Background and Motivation	80
6.2	Method	81
6.2.1	Neuron Model	81
6.2.2	Joint STDP	82
6.3	Results	84
6.3.1	Learning Predefined Patterns from Synthetic Noisy Spike Trains	85
6.3.2	Validation of Joint STDP on Public Mental Task Dataset	89
6.3.3	Learning Sparse Features from the Vigilance Experiment	93
6.4	Discussion	99
6.4.1	A Bio-inspired Feature Learning Algorithm	99

6.4.2	Related Works in the Literature	99
6.4.3	Limitations	100
6.5	Summary	101
7	Conclusions and Future Work	102
7.1	Conclusions	102
7.2	Future Work	104
	Appendix	106
A	Proof for Sparseness in the Solution in Section 4.2.3	106
B	Proof for Replacing Discrete Labels with Probabilistic Labels as a Generalization of Training Logistic Regression and Naive Bayes	111
C	Proof for Equation (6.8) and (6.9) in Section 6.2	113
C.1	Proof Part 1	114
C.2	Proof Part 2	120
D	Parameter Values in Section 6.3	122
E	Input Spike Trains in Section 6.3.1	123
F	Latency, Pattern Hit Rate and False Alarm Frequency in Section 6.3.1	124
	Publication List	125
	References	127

Summary

Attention is constantly required in many daily life tasks. Attention-related behavior, such as driving distraction, has been reported as a major reason in traffic accidents. Therefore, the recognition of attention can enhance task performance. Electroencephalogram (EEG) is used to study attention, since it provides a direct measure of the brain activity with high temporal resolution at 1–10ms. In this thesis, we study the recognition of selective attention and sustained attention (vigilance) using machine learning.

We start by proposing an experiment which aims to recognize unattended and attended conditions induced by Test of Variables of Attention (TOVA), using EEG features supported by the event-related potential (ERP) literature. However, ERP does not work for real-time applications where the external stimuli (event) required by ERP cannot be controlled. In face of the low signal-to-noise ratio (SNR) in the non-ERP approach, we propose Channel Selection with Different Features (CSDF) algorithm, which selects channels with their own different feature sets, as well as restricts features to as few channels as possible. Using CSDF, 83 out of 868 features are selected to distinguish the unattended and attended conditions. The accuracy 94.3% ($\pm 5.6\%$) is the best compared to other feature selection and channel selection algorithms. Based on CSDF, we find that the first and second order difference in the left parietal and temporal lobes, as well as the Higuchi fractal dimension and mean signal amplitude in the right frontal lobe, are relevant to selective attention.

Unlike selective attention which has discrete conditions such as attended/unattended, the vigilance stages cannot be easily observed. Analogous to sleep stages, we want to define the vigilance stages in open eye and situation-aware state in a subject-independent and

data-driven way. In the literature, there are vigilance stage models defined under closed eye. However, the EEG signals are more complex in open eye and situation-aware state. Extreme learning machine autoencoder (ELMAE) is used to learn the EEG spectral features and define the vigilance stages during simulated driving. Results show that ELMAE is an efficient alternative to restricted Boltzmann machine (RBM) in vigilance recognition: ELMAE achieves root mean square error at $0.189 (\pm 0.049)$, which is better than RBM at $0.195 (\pm 0.046)$; and training speed significantly faster than RBM. Based on ELMAE, we define three vigilance stages in open eye and situation-aware state. Stage I is high vigilance, where the subject is attentive. Stage II is low vigilance, which is further divided into Stage II.1: drowsiness and difficulty in attention allocation; and Stage II.2: distraction instead of falling asleep. A possible explanation for stage II.2 is that, the environment contains not enough external stimuli to keep the open eye and situation-aware state, so that the brain performs vigilance regulation to seek external stimuli, and hence leading to distraction.

A major limitation of ELMAE is that it learns nonsparse hidden weights and features. Analogous to feature selection using CSDF, we want to learn sparse features so that the useful information is restricted to a few nonzero features to achieve better interpretability. To address this issue, a novel bio-inspired algorithm, joint weight-delay spike-timing dependent plasticity (joint STDP), is proposed for learning sparse hidden weights and EEG spectral features. Compared to other nonsparse and sparse algorithms, joint STDP achieves the highest level of sparseness in both learned features and hidden weights significantly. On the other hand, joint STDP has slightly larger root mean square error at $0.206 (\pm 0.061)$, due to the trade-off between performance and sparseness.

Extensive experiments and comparisons are carried out to evaluate all the proposed algorithms. The experimental results confirm the advantages of the proposed algorithms, hence making contribution to EEG-based attention research. This thesis is interdisciplinary and includes several fields such as machine learning, computational neuroscience, brain-computer interface and psychology.

Table Captions

1.1	Frequency bands in EEG signal	3
1.2	Different types of attention and related driving examples	4
2.1	Psychological tests related to attention in the literature	10
2.2	Advantage and disadvantage of experimental labeling methods in Figure 2.1	12
2.3	Common EEG features in the literature	14
2.4	EEG datasets related to attention in the literature	21
2.5	Recognition performance of the mental task dataset in the literature	22
3.1	Testing accuracy using our feature and the state-of-the-art result on mental task dataset	29
3.2	Testing accuracy using Besserve's features, our features and combined features	30
3.3	Testing accuracy using Yeo's features, our features and combined features	32
3.4	Testing accuracy using ELM and SVM	33
4.1	EEG features extracted for selection	43
4.2	Testing accuracy from 5 channel selection algorithms and CSDF on the mental task dataset	44

4.3	Testing accuracy from 5 feature selection algorithms and CSDF on the mental task dataset	45
4.4	Testing accuracy from 5 channel selection algorithms and CSDF on the TOVA dataset	46
4.5	Testing accuracy from 5 feature selection algorithms and CSDF on the TOVA dataset	47
4.6	Testing accuracy of a sequential combination of channel (mRMR_SFFS_C) and then feature selection (mRMR_SFFS) and CSDF on the TOVA dataset	48
4.7	Number of selected features from 5 channel selection algorithms and CSDF on the TOVA dataset	49
4.8	Number of selected features from 5 feature selection algorithms and CSDF on the TOVA dataset	50
4.9	Number of selected channels from 5 channel selection algorithms and CSDF on the TOVA dataset	51
4.10	Number of selected channels from 5 feature selection algorithms and CSDF on the TOVA dataset	51
4.11	Selected channels for each subject	52
4.12	Mostly selected features in all subjects	52
5.1	Terminology in this chapter	63
5.2	The training RMSE, testing RMSE and training time of ELMAE, RBM and predefined features	64
5.3	Vigilance stages in open eye and situation-aware state	70
5.4	Predefined features used in Figure 5.4 and Table 5.2 relevant to the vigilance stages	71
6.1	Terminology in this chapter	81

6.2	Testing accuracy on the mental task dataset	92
6.3	Testing RMSE for vigilance recognition	95
6.4	Vigilance Stages in Open Eye and Situation-Aware State using Joint STDP	98
D.1	Joint STDP parameters	122
D.2	SRM neuron model parameters	122

Figure Captions

1.1	Relationship between attention and other psychological terms	2
2.1	Experimental methods for labeling brain states in the literature	11
3.1	EEG signal processing flowchart	26
3.2	Visualization of the feature space for subject 1	27
3.3	Experiment procedure and data splitting	28
4.1	Illustration of channel-wise sparseness in the weight matrix	40
4.2	Flowchart of EEG based brain state recognition using CSDF	41
4.3	Feature space visualization of subject S3 using PCA	47
4.4	Top 10 features that are selected across all channels and all subjects . .	53
4.5	Channel importance heat map and a related neuroimaging study in the literature	54
5.1	The experiment setup	59
5.2	The machine learning model	62
5.3	The response time, reference vigilance and recognized vigilance using ELMAE and RBM for subject S10	64

5.4	Comparing testing RMSE and training time of ELMAE, RBM and pre-defined features for all subjects	65
5.5	The reference vigilance and recognized vigilance of each cluster	66
5.6	The spectrum-vigilance representation using ELMAE	68
5.7	The spectrum-vigilance representation using RBM	69
5.8	An example in stage I	72
5.9	An example in stage II.1	73
5.10	An example in stage II.2	73
5.11	The Hoyer sparseness measure of hidden weights	74
5.12	The Hoyer sparseness measure of learned features	74
5.13	The hidden weights in channel T8	75
5.14	The hidden weights with the most positive and negative output weight in each channel	76
6.1	The weight and delay STDP windows	83
6.2	Experiment setup for learning from synthetic nosy spike trains	85
6.3	The membrane potential trace, spike arrival times and their weights for readout neuron #10	86
6.4	The response latency, hit rate and false alarm frequency for weight STDP and joint STDP	87
6.5	The learned patterns, weights and delays for weight STDP and joint STDP	88
6.6	Phase encoding	90
6.7	Two stages in training using the joint STDP	91
6.8	The distribution of the weights and delays of all connections from all mental tasks	91

6.9	The network structure for recognizing vigilance using joint STDP	93
6.10	Vigilance recognition for one subject using joint STDP	94
6.11	The distribution of the weights and delays of all connections trained on the vigilance dataset	96
6.12	The comparison of Hoyer sparseness for the learned features and hidden weights	96
6.13	The hidden weight-vigilance representation using joint STDP	97

Abbreviations

ADHD	Attention Deficit Hyperactivity Disorder
AE	Autoencoder
ANOVA	Analysis of Variance
CSDF	Channel Selection with Different Features
EEG	Electroencephalogram
ELM	Extreme Learning Machine
ELMAE	Extreme Learning Machine Autoencoder
EM	Expectation Maximization
ERP	Event-Related Potential
ICA	Independent Component Analysis
LDA	Linear Discriminant Analysis
NB	Naive Bayes
PSP	Postsynaptic Potential
PVT	Psychomotor Vigilance Task
SD	Standard Deviation
RBM	Restricted Boltzmann Machine
RMSE	Root Mean Square Error
RT	Response Time
SVM	Support Vector Machine
STDP	Spike-Timing Dependent Plasticity
TOVA	Test of Variables of Attention
VIGALL	Vigilance Algorithm Leipzig

Chapter 1

Introduction

Chapter 1 begins with the background and motivation for EEG-based attention recognition. Then, the research scope is limited to two attention types: selective attention and sustained attention (vigilance). Three research questions are raised, followed by the proposed algorithms in this thesis and contributions to the field. At last, the organization of the thesis is presented.

1.1 Background and Motivation

Attention is the process by which organisms selectively focus on a subset of available information for enhanced processing and integration [1]. In many daily life tasks, attention is constantly required to deal with various stimuli from the dynamic environment. In this thesis, we take driving as an example, where attentional, visual and motor processing with sustained brain activity are required in face of the dynamic road condition [2]. Attention-related behavior, such as driving distraction, has been reported as a major reason in traffic accidents [3]. The recognition of attention can assist or enhance task performance.

The definition of attention overlaps with other psychological concepts such as arousal, alertness and consciousness [4, 5]. In our research, the relationship between attention and other terms is summarized in Figure 1.1.

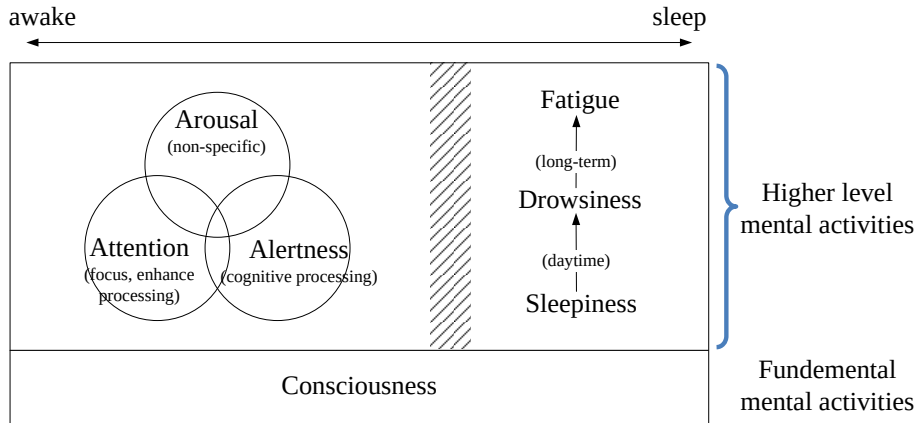


Figure 1.1: Relationship between attention and other psychological terms

Historically, the study of attention originated from philosophy. Then it was introduced into psychology, making use of various cognitive tests (mental chronometry [6]). Later in the 1960s, neurophysiological approaches were used to find the neural correlates of attention [7], where electroencephalogram (EEG) started to become one of the major approaches.

EEG is the recording of electric currents developed in the brain using electrodes attached to the scalp [8], with typical amplitude at 10–100 μV . Each electrode is attached to a predefined location on the scalp (also called channel) based on the international 10–20 system. In Table 1.1 we list the common frequency bands in EEG. Compared to other methods such as video camera, EEG 1) provides a direct measure of the brain since the neuronal activity is directly related to various brain functions; and 2) has relatively high temporal resolution at the scale of 1–10ms.

Table 1.1: Frequency bands in EEG signal

Band name	Frequency	Interpretation
delta δ	<4Hz	deep sleep, infancy, and in some brain disorders
theta θ	4–8Hz	awake, but relaxed or sleepy
alpha α	8–12Hz	idle state, relaxed with the eyes closed
beta β	12–30Hz	wakefulness
gamma γ	>30Hz	sensory binding or short-term memory

Attention recognition provides potential benefits in real-time applications. In manual driving, waiting at traffic lights requires selective attention [9], which can be recognized to provide warning to the driver if not attended. In automated driving, the vehicle should be able to access the sustained attention state of the driver, to check whether the driver is vigilant when the vehicle cannot handle the situation alone, paving the way to driver safety in automated driving.

1.2 Research Scope and Questions

From the clinical point of view, attention can be categorized into five types [10], as shown in Table 1.2. Attention can also be categorized as overt versus covert, and exogenous versus endogenous. In this thesis, we stick to the clinical categorization since it provides relatively detailed categories compared to other approaches. In this thesis, we target at recognizing selective attention and sustained attention (vigilance) from EEG to limit the research scope.

Table 1.2: Different types of attention and related driving examples

Type [10]	Definition	Driving example
Selective attention	The ability to maintain a behavioral or cognitive set in the face of distracting or competing stimuli.	Waiting at traffic lights
Sustained attention (vigilance)	The ability to maintain a consistent behavioral response during continuous and repetitive activity.	Prolonged driving in monotonous environment
Focused attention	The ability to respond discretely to specific visual, auditory or tactile stimuli.	Attentively following the front car
Divided attention	The ability to respond simultaneously to multiple tasks or multiple task demands.	Driving and chatting at the same time
Alternating attention	The ability of mental flexibility that allows individuals to shift their focus of attention between tasks.	Checking rear and side mirrors alternatively

Under this scope, this thesis aims to answer the following three research questions.

(1) What are the relevant brain regions (in terms of EEG channels) and distinguishable EEG features in selective attention? In the literature, the study on selective attention can be based on event-related potential (ERP), which averages multiple EEG signals time-locked to attended/unattended stimuli. The averaging reduces information irrelevant to the stimuli, and improves the signal-to-noise ratio (SNR). Results in the ERP approach mainly suggest that alpha oscillation serves as active suppression of unattended information, while beta and gamma oscillation indicates active processing of attended information [11, 12]. However, the ERP approach does not work for real-time applications where the stimuli comes from the environment, since the stimuli cannot be controlled to be repeated for multiple times to average the EEG signals. There are also studies on selective attention based on non-ERP in the literature. However, in face of lower SNR in non-ERP approach [13], there is no consensus understanding of the relevant channels and distinguishable features in selective attention.

(2) What are the vigilance stages in open eye and situation-aware state? Unlike se-

lective attention which has discrete conditions such as attended/unattended, there is no clear definition of vigilance stages in open eye and situation-aware state. All the existing vigilance stage models are defined for closed eye or task-free condition. For example, VIGALL [14] defines stages from high alertness to sleep onset under closed eye. Alford et al. [15] defines stages from quiet wakefulness to deep sleep under both closed and open eye state, but not engaged in any task. For conditions such as closed eye or task-free, it is known that vigilance decrements and finally reaches sleep stage [14]. However, the vigilance stages in open eye and situation-aware state, particularly during attention-demanding task such as driving, remain unclear. It is desirable to define the vigilance stages in a data-driven and subject-independent way by learning features instead of predefining features.

(3) Analogous to feature selection, can we learn sparse features so that the useful information is restricted to a few nonzero features? Automatically learning sparse features provides additional benefit compared to feature selection. In particular, the features need not to be selected from a predefined set, which is constrained by the current understanding of attention. There are several sparse feature learning algorithms available, such as sparse restricted Boltzmann machine [16] and sparse autoencoder [17]. However, the sparseness is realized by reducing the average activation in the hidden layer, with no attempt on sparse hidden weights. The hidden weights act as templates to match meaningful patterns in the EEG signals. Therefore it is desirable to have both sparse hidden weights and sparse hidden activation, at the same time attaining similar or better performance compared to non-sparse counterparts.

1.3 Contributions and Novelty

To answer the above research questions in EEG-based attention recognition, we made contributions in both algorithmic and neurophysiological aspects: (1) novel EEG channel and feature selection algorithm; (2) novel bio-inspired sparse feature learning algo-

rhythm; and (3) novel vigilance stage model in open eye and situation-aware state. The research is interdisciplinary and includes several fields such as machine learning, computational neuroscience, brain-computer interface and psychology.

1.3.1 EEG Channel Selection with Different Features

EEG channel selection helps to identify the most relevant brain regions, especially in face of the low SNR in non-ERP approaches. However, channel selection selects the same set of features in each channel, ignoring the fact that different features could be important in each channel. For example, alpha power increases in occipital channels while has less increase in the frontal channels during closed eye [18]. For feature selection, it is also a challenge to restrict the useful features to as few channels as possible to allow a clearer interpretation. We proposed Channel Selection with Different Features (CSDF) algorithm to solve this problem. CSDF provides a convex optimization framework for simultaneously channel and feature selection, where the global optimum is solvable. We compared CSDF to various channel selection algorithms and feature selection algorithms. The proposed algorithm achieved similar or better recognition accuracy with significantly fewer number of features and channels. The selected channels turned out to be consistent with a neuroimaging-based selective attention study in the literature [19].

1.3.2 Vigilance Stage Model based on EEG

Compared to the discrete conditions such as attended/unattended in selective attention, there is no clear definition of vigilance stages in open eye and situation-aware state, where the EEG signals are more complex than the closed eye and relaxed state. Based on extreme learning machine autoencoder (ELMAE) for feature learning, we propose three vigilance stages in open eye and situation-aware state in a subject-independent and data-driven way. The comparison between ELMAE and restricted Boltzmann machine

(RBM) indicates that ELMAE is an efficient and rapid alternative to RBM. The proposed vigilance stage model indicates that, at low vigilance level (stage II.2), the brain performs vigilance regulation so that the subject seeks external stimuli to maintain the open eye and situation-aware state.

1.3.3 Sparse Feature Learning with Joint STDP

Although ELMAE can learn features from EEG spectrum, the learned features and hidden weights contains global components (holistic-based) [20]. Therefore the interpretability of the learned features and hidden weights remains difficult. To restrict the useful information to a few nonzero values in the learned features and hidden weights, we propose a novel bio-inspired algorithm named joint STDP for sparse feature learning based on the bimodal convergence property of additive STDP in spiking neural network [21, 22]. After learning, the learned sparse features can be retrieved according to the synaptic weights and axonal delays from input to readout neurons.

1.4 Thesis Overview

The remaining parts of the thesis are organized as follows. Chapter 2 reviews the literature on the recognition of selective attention and sustained attention (vigilance) using EEG, from both experimental and algorithmic points of view.

In Chapter 3 we propose an experiment for the recognition of two states: open eye resting (unattended) and performing Test of Variables of Attention (TOVA) (attended), to fill the gap that there is no non-ERP studies using TOVA to induce selective attention.

Chapter 4 addresses the first research question: What are the relevant brain regions (in terms of EEG channels) and distinguishable EEG features in selective attention? The channel selection with different features (CSDF) algorithm is proposed to simultane-

ously select channels and features in the TOVA test.

In Chapter 5 we answer the second research question: What are the vigilance stages in open eye and situation-aware state? We propose 3 vigilance stages in open eye and situation-aware state in a data-driven and subject-independent way. The nonsparseness in the learned features and hidden weights lead us to Chapter 6.

Chapter 6 answers the last research question: Can we automatically learn sparse features so that the useful information is restricted to a few nonzero features? We propose joint spike-timing dependent plasticity (joint STDP) to reveal the sparse features and hidden weights in vigilance recognition.

Chapter 7 concludes the thesis and outlines several important future directions and applications.

Chapter 2

Literature Review

Chapter 2 is the literature review on EEG-based recognition of selective attention and sustained attention (vigilance). Section 2.1 reviews the psychological tests used to induce different attentional conditions. Section 2.2, 2.3, 2.4 and 2.5 review the artifact removal, feature extraction, feature and channel selection, and feature learning algorithms in EEG signal processing. Section 2.6 and 2.7 focus on state-of-the-art studies on selective attention and vigilance. Section 2.8 discusses the public EEG dataset for benchmark purpose.

2.1 Experiment Design

Psychological experiments are used to induce different attentional types. The quality of experiment design determines the quality of the EEG features and target brain states, which in turn affects attention recognition performance.

2.1.1 Psychological Tests

Table 2.1 shows psychological tests from the test battery of Psychology Experiment Building Language (PEBL) [23] related to selective and sustained attention.

Table 2.1: Psychological tests related to attention in the literature

Test name	Brain state	Description	Ref.
Test of Variables of Attention (TOVA) (a type of go/no-go test)	selective attention	A square flashes on the screen. The participant must respond only to targets, i.e. when the square is at the top.	[24]
Continuous Performance Test (CPT) (a type of go/no-go test)	selective attention	Letters appear on the screen in sequence. The participant must respond to all letters EXCEPT "X".	[25]
Stroop test	selective attention	Tell the color of the word while ignoring the color the word denotes	[26]
Stimulus-response compatibility task (SCT)	selective attention	Tell the direction of the arrow while ignoring the location of the arrow	[27]
Psychomotor Vigilance Task (PVT)	vigilance (sustained attention)	A circle appears with random interval between 2 and 12 seconds, and the participant must response as fast as possible.	[28]
Clock Test	vigilance (sustained attention)	Watch a clock and determine when it skips a position.	[29]

For selective attention, we used TOVA in our experiment. As for CPT, the letter stimuli requires larger amount of perception than the square stimuli in TOVA. So it is difficult

2.1 Experiment Design

to distinguish between brain activity induced by selective attention and perception of the letters (cf. Figure 1.1). As for Stroop test and SCT, they are not compatible with the Go/No-Go test, hence less suited in driving. For vigilance (sustained attention), we used PVT in our experiment. This is because the primary task in the vigilance experiment is driving in a simulator, while the Clock Test requires the subject to look at the clock constantly, which affects the primary driving task.

2.1.2 Labeling Methods

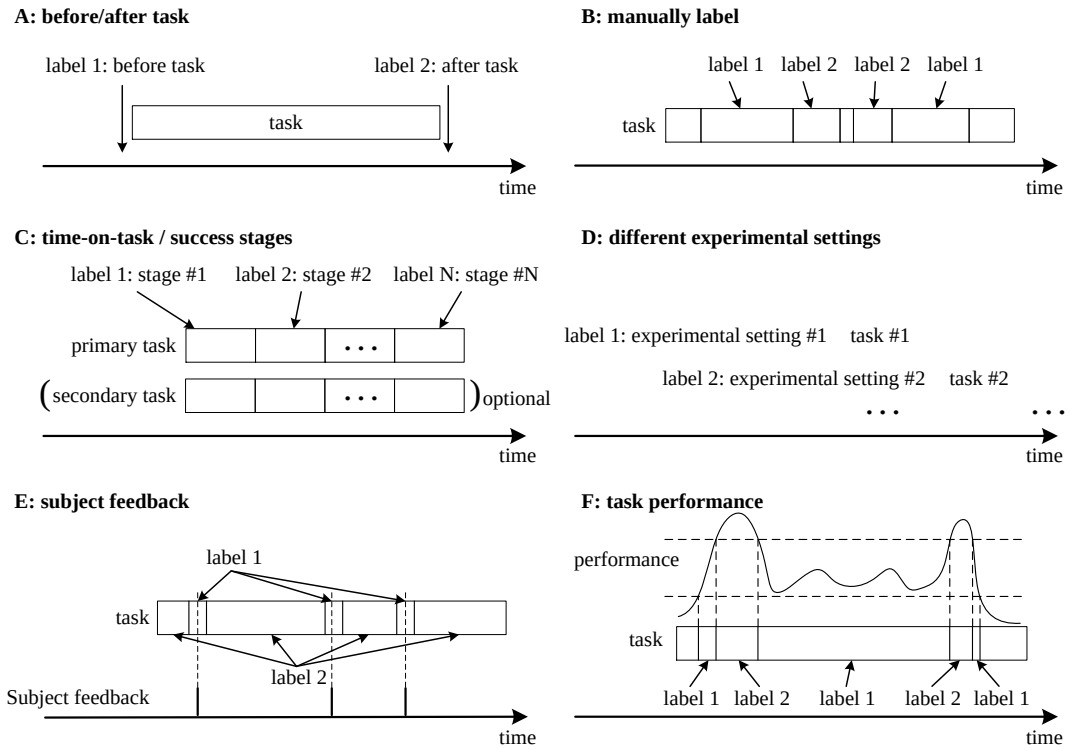


Figure 2.1: Experimental methods for labeling brain states in the literature

Psychological tests are used to induce the target brain state. There are multiple ways to define when the target brain state is induced, i.e. to give “labels” to the EEG signal. In some cases, the exact label is unavailable, e.g. vigilance, we need to estimate the label from some proxy measures, such as task performance. In Figure 2.1 we illustrate six labeling methods in the literature. And in Table 2.2 we briefly describe each method as well as the advantage and disadvantage.

2.1 Experiment Design

Table 2.2: Advantage and disadvantage of experimental labeling methods in Figure 2.1

Id	Method	Advantage	Disadvantage	Ref.
A	EEG signal before and after task are labeled as two different classes	Easy implementation	Cannot guarantee correct labels	[30]
B	Manually label EEG signals by observing the subject from video	Labels are correct for obvious cases such as awake/asleep	Subjective, extensive work	[31]
C	EEG signals are labeled in successive time intervals due to time-on-task effect	Easy implementation	Cannot guarantee correct labels	[32]
D	Label EEG signal recorded under different experimental conditions	Embedded in experiment design	Variability within each experimental condition	[33]
E	Subjective feedback when subject feels himself/herself in the desired brain state	Applicable to experiments related to self-consciousness	Feedback can interrupt the brain state and subject-dependent	[34]
F	Label EEG according to task performance	Objective label	Affected by familiarity with the task	[35]

As a necessary condition of achieving good performance using machine learning methods, an important factor in this research is to guarantee correct brain state labels. Therefore we cannot use method A and C. Method B requires manual labeling, which is subjective and extensive. So it cannot be used either. Subjective feedback cannot be carried out frequently, thus method E cannot provide enough labels to train the classifier.

For selective attention, we use method D to define labels, i.e. two experimental conditions: open eye resting (unattended) and task-performing (attended). The variability of brain state within each experimental condition should be considered as a minor factor due to contrast between two conditions as well as the robustness of machine learning classifiers.

For vigilance (sustained attention), we use method F, since it provides an objective label. The performance measure can be accuracy and/or response time (RT) to some stimuli such as questions, as in situation present assessment method (SPAM) [36] for measuring

situation awareness in human factors study. For continuous performance measure, it is possible to use regression or training with probabilistic labels. The effect of familiarity with the task can be reduced by allowing the subject to practice the task before the experiment.

2.2 EEG Artifact Removal

It is believed that the EEG signals represent a mixture of signals from various sources, including both brain and artifact sources [37]. Blind source separation (BSS) is used to recover the source signals without the aid of information (or with very little information) about the sources or the mixing process, by assuming some statistically properties of the components [37]. BSS includes

- (1) Principle component analysis (PCA) [38]: uncorrelated orthogonal components;
- (2) Independent component analysis (ICA) [37]: statistically independent components;
- (3) Second order blind interference (SOBI) [39]: temporally uncorrelated components.

Principal component analysis (PCA) finds the orthogonal directions of largest variance in the data. However in general, there is no reason to believe that distinct EEG sources are orthogonal to one another in neurophysiology [38]. Independent component analysis (ICA) is used to separate multivariate signals into statistically independent components [37]. It is not unrealistic to assume that the artifacts and the uncontaminated EEG signal are statistically independent, so it is promising to use ICA to remove artifacts. Traditional ICA needs visual inspection of the independent components and determination of which components correspond to artifact. Some kinds of automatic artifact component identification methods are needed to reduce human intervention and enhance efficiency. It turns out that the artifact components usually have some distinct scalp topography patterns, enabling machine learning classifiers to classify artifact and non-artifact components [40]. However, ICA is better at identifying ocular artifacts,

but less sensitive to muscular artifacts, because muscle activity tends to saturate single rather than distributed channel sites [41]. SOBI exploits the time coherence of the source signals [39]. This is achieved by utilizing time delayed cross-correlations to impose an uncorrelated structure on the components at different time delays. There also exist algorithms such as joint cumulant and correlation (JCC) [42], which utilizes both ICA and SOBI algorithms.

In a comparative study of ICA and SOBI algorithms, it turned out that ICA outperforms SOBI on simulated EEG signals [37]. There are also suggestions that the SOBI framework requires further tests on large-scale real EEG datasets [43]. As a result we used ICA as the artifact removal algorithm in the thesis.

2.3 EEG Feature Extraction

After artifact removal, EEG signals are segmented into overlapping segments. To reduce the dimensionality and enhance SNR of the signal, various features are extracted from each segment to describe the signal. In general, EEG features can be categorized according to Table 2.3.

Table 2.3: Common EEG features in the literature

Category		Feature	Ref.
single channel features	temporal	signal amplitude statistics	[44]
		auto-regression coefficients	[45]
	spectral	power spectral density	
	nonlinear dynamics	entropy and Lyapunov exponents	[46]
		fractal dimension	[47]
	wavelet	discrete wavelet transformation	[48]
inter-channel features	linear	cross correlation	[49]
		spectral coherence	
	nonlinear	mutual information	
		transfer entropy	
		phase synchronization	

Within various features, there are some features with clear neurophysiological mean-

ings. For example, the power spectral density (PSD) of each specific frequency band (Table 1.1) is correlated with different brain states/functions, because they represent oscillation and synchronization levels on different scales of neural networks.

2.4 EEG Feature Selection and Channel Selection

The next step is to select features in order to find the most relevant and discriminant features, as well as to reduce dimensionality and avoid the “curse of dimensionality”. Feature selection has two steps: (1) subset selection; and (2) subset relevance evaluation.

Two common subset selection approaches are considered. The first approach is to recursively selecting or eliminating a (subset of) feature(s), where the selected/eliminated feature(s) is determined by optimizing some criteria at the current step. We can see that the recursive feature selection/elimination is a greedy algorithm which does not guarantee the global optimality of the solution [50]. Examples include sequential forward selection (SFS), sequential floating forward selection (SFFS) [51] and support vector machine recursive feature elimination (SVM-RFE) [52], etc. The second approach is based on optimization with sparse solutions. In this case, a model with sparse-inducing parameter prior can be set, which usually leads to optimization problem with ℓ_1 -norm regularization. The solution contains many zeros, while the nonzero values indicate the selected features. If formulated as a convex problem, the solution is guaranteed to be global optimum. Examples include lasso [53] and sparse discriminant analysis [54].

There are in general two subset relevance evaluation approaches, i.e. wrapper and filter. Wrappers utilize the classifier and use classification accuracy as the relevance of a certain feature set, for example SVM-RFE [52]. It is obvious that wrappers are computationally intractable for large feature space, although for small feature space, it can find the exactly the best feature set. Filters are methods which do not require classifiers. They use some computational tractable proxy of classification accuracy, such as Fisher discriminant ratio (FDR) [55], maximum-relevance-minimum-redundancy (mRMR) [56]

and mutual information [57].

For channel selection, a channel relevance measure is computed as a combination of feature relevance measures belonging to the same channel, and then sequential channel selection is done in the similar greedy update manner [58–60]. In fact, the channel selection algorithms can be generalized to feature selection where features in the same channel are selected or eliminated at the same time.

In channel selection, all features in the same channel are treated with equal importance. However, different brain regions have different functions, suggesting different set of important features in each channel. For example, the alpha power is dominant in the occipital channel when closing eye, reflecting idling state or inhibition of task-irrelevant activities [11]; while weak in the frontal areas. However, it remains a challenge to select channels with their own different sets of features, so that the number of irrelevant features can be reduced.

2.5 EEG Feature Learning

Despite extracting predefined features from EEG, recent advances in representation learning allow to learn features directly from the various raw biosignals including EEG [61, 62]. Restricted Boltzmann machine (RBM) [63, 64] and autoencoder (AE) [65, 66] are widely used to automatically extract features from the raw data in an unsupervised manner. The training of RBM and AE rely on iterative algorithms, which imposes longer training time. Extreme learning machine autoencoder (ELMAE) [20, 67] represents another branch of autoencoders, where random neurons are used as the hidden nodes and the decoding weights are untied from encoding weights, so that there is a closed form solution for the decoding weights, hence faster.

2.5.1 Restricted Boltzmann Machine

Restricted Boltzmann machine (RBM) is an energy-based model which has a bipartite graph structure and consists of visible variables x and hidden variables h . The objective is to maximize the log-likelihood with an intractable partition function $Z(\theta)$

$$\begin{aligned} \underset{\theta}{\text{maximize}} \quad \ell(\theta) &= \frac{1}{N} \sum_n \log p(x^{(n)}; \theta) \\ &= -\frac{1}{N} \sum_n F(x^{(n)}; \theta) - \log Z(\theta) \\ &= -\frac{1}{N} \sum_n F(x^{(n)}; \theta) + \mathbb{E}_{x'}[F(x'); \theta] , \end{aligned} \quad (2.1)$$

where θ includes the weight matrix $W \in \mathbb{R}^{d \times L}$, the bias of input nodes $b^x \in \mathbb{R}^d$, and the bias of hidden nodes $b^h \in \mathbb{R}^L$. For RBM with binary hidden variables (Bernoulli-RBM), the free energy $F(x; \theta)$ is

$$\begin{aligned} F(x; \theta) &= -\log \sum_h \exp(-E(x, h; \theta)) \\ &= -b^{x^\top} x - \sum_j \log(1 + \exp(b_j^h + W_{:j}^\top x)) . \end{aligned} \quad (2.2)$$

The expectation term in Equation (2.1) is intractable since it is an expectation over the entire distribution of x . To handle the intractability, various approximation algorithms are proposed, such as contrastive divergence (CD) [64] and stochastic maximum likelihood (SML [68], also known as persistent contrastive divergence, PCD [69]). Compared to CD, SML directly approximates the likelihood function and shows better performance in the literature [69]. The approximation is

$$\mathbb{E}_{x'}[F(x'); \theta] \approx \frac{1}{M} \sum_{m=1}^M F(\bar{x}^{(m)}; \theta) , \quad (2.3)$$

where $\bar{x}^{(m)}$ is obtained by Gibbs sampling from M “fantasy particles” using updated θ in each iteration without restarting the chain.

The learned features are embedded in the weight matrix W , which lead to higher activation in the hidden variables if similar patterns present in the input. This is achieved by taking dot product, i.e. $H = \text{sigmoid}(XW + \mathbf{1}b_h^\top)$, where $\mathbf{1} \in \mathbb{R}^N$ is a column vector with all ones to apply bias to each sample.

2.5.2 Extreme Learning Machine Autoencoder

Extreme learning machine autoencoder (ELMAE) [20, 67] is an autoencoder with random weights at the encoding stage and untied weights at the decoding stage. It learns to reconstruct input data $X = \{x^{(n)}\}_{n=1}^N \in \mathbb{R}^{N \times d}$ from a nonlinearly corrupted input $H \in \mathbb{R}^{N \times L}$ obtained from the random neurons in the hidden layer. The objective of ELMAE is

$$\underset{\beta}{\text{minimize}} \quad \frac{1}{2} \|H\beta - X\|_F^2 + \frac{C}{2} \|\beta\|_F^2, \quad (2.4)$$

where

$$H = g(XW + \mathbf{1}b^\top); \quad W^\top W = I; \quad b^\top b = 1,$$

and $W \in \mathbb{R}^{d \times L}$ and $b \in \mathbb{R}^L \times 1$ are the encoding weight and bias respectively which are randomly generated, orthogonalized and then fixed; $\mathbf{1} \in \mathbb{R}^N$ is a column vector with all ones; C is the regularization parameter; and $g(\cdot)$ is any piecewise continuous function, which is a sigmoid function here.

Similar to RBM, the learned features are embedded in the decoding weight $\beta \in \mathbb{R}^{L \times d}$, which has a closed-form solution

$$\beta = (H^\top H + CI)^{-1} H^\top X. \quad (2.5)$$

The decoding weight β finds input patterns similar to the weights by taking dot product, i.e. $H = \text{sigmoid}(X\beta^\top)$. According to Theorem III.1 in Kasun et al. [20], the projection

$X\beta^\top$ decreases the within-class distance, so that better class discriminability is obtained.

2.6 Studies on Selective Attention

In the literature, some studies are based on event-related potential (ERP), which requires to average multiple EEG responses time-locked to the attended/unattended stimuli. The research focus is to find the difference in the ERP, rather than the recognition accuracy. Results in ERP study mainly suggest that alpha oscillation serves as active suppression of unattended information, while gamma oscillation indicates active processing of attended information [11, 12]. However, the ERP approach does not work for real-time applications where the stimuli comes from the environment, so that the stimuli cannot be repeated for multiple times to average EEG signals. Instead, non-ERP approaches should be used to remove the need for the external stimuli, although with lower signal-to-noise ratio [13].

There are also some studies not based on ERP. Besserve et al. [27] recognizes the performance level during a spatial stimulus-response compatibility task (SCT), where an arrow appears randomly on the left or right side of the screen, and the subject only responds to the direction of the arrow while ignoring the location of the arrow. The band power and phase locking value are extracted and SVM is used as the classifier. They achieved average recognition accuracy of 71% from 12 subjects. Ghimire et al. [26] utilizes Stroop test and compares different EEG band powers under open eye resting, reading congruent words (red printed in red ink) and reading incongruent words (red printed in blue ink). The research focus of this study is *not* reporting the classification accuracy over the 3 conditions. Instead, using statistical tests, they report that during selective attention, there is more theta power at frontal-central regions and less alpha power at parietal regions. Nazari et al. [25] recorded EEG signals in open eye resting baseline state and performing a continuous performance test (CPT) for children with and without ADHD. Similarly, the research focus is to find the changes in EEG features

between these two states, hence there is no classification accuracy available. The results suggested that CPT induced increased alpha power in the posterior site in children with ADHD, while decreased in children without ADHD.

2.7 Studies on Vigilance

Borghini et al. [70] provides a comprehensive review for assessing vigilance-related brain states under open eye and situation-aware state (in driving and flying tasks), where they summarize that theta, alpha, as well as delta bands are characteristic in these brain states. However, they point out that there is no consensus about the best features.

Some studies focus on subject-dependent model. Shi et al. [35] compares the vigilance recognition performance of various modifications of ELM and support vector machine (SVM) in a traffic sign color recognition task. It turns out that ELM with elastic net penalty achieves the lowest RMSE at 0.16 normalized vigilance level. However, there has been no analysis on vigilance stages, or the description of the characteristic features in different vigilance levels. Lin et al. [71] applies support vector regression on the EEG spectral features to predict response time (RT) in a lane keeping task. They obtain root mean square error (RMSE) at 0.12ms. They also proposed “vigilance degree”, similar to a vigilance stage model. However, the “vigilance degree” is based on RT, which is easier to be affected by other factors such as unexpectedly touching the button.

There are also works targeting at subject-independent models. Peiris et al. [72] detects task lapse to assess vigilance based on 15 subjects. The lapse is defined as out of track in a visuomotor tracking task or a low vigilance stage determined manually from video. To achieve subject-independence, they train a “level-1” model using weighted sum of “level-0” models trained in each subject. The lapse detection has moderate accuracy where sensitivity is 73.5% and selectivity is 25.5%. In a related work on EEG-based workload recognition, Wang et al. [73] introduces a hidden variable $H (\in \{1, 2, \dots, K\})$ to account for the difference between 8 subjects. It turns out that the accuracy on 8

subjects is the best when H has $K = 18$ states, of which the classification accuracy is about 80.3%, similar or better than subject-dependent models at 80.0%.

2.8 Public EEG Datasets Related to Attention

Table 2.4: EEG datasets related to attention in the literature

Database	# Subject	Description	Ref.
Five mental task classification	6	10-second trials of 5 pre-defined mental tasks	[74]
Rapid serial visual presentation (RSVP) study	8	decide whether the satellite image includes a airplane	[75]
Five-box task	60	press button whenever a disc appears in the attended location	[76]
Oddball	12	respond to specific tones when listening mixture of songs	[77]
categorization and recognition tasks under go/no-go paradigm	14	Two tasks alternated: a categorization task and a recognition task. Give response following a go/no-go paradigm	[78]

The public EEG datasets related to attention are shown in Table 2.4. Here we use the mental task dataset as the benchmark dataset. The mental task dataset consists of five mental tasks: baseline, counting, letter-composing, multiplication and rotation. 6 subjects carries out five mental tasks. There are 11 sessions in total. However, the first session of subject 3 has overwhelming artifacts in the EEG signal. And the third session of subject 5 is not available in our downloaded dataset. Therefore we took 9 out of the 11 sessions. Each session consists of 5 trials and each trial lasts 10 seconds. EEG signals are recorded at 250Hz sampling frequency from 6 EEG channels: C3, C4, P3, P4, O1 and O2; and 1 EOG channel.

The reason of using the mental task dataset as the benchmark dataset is threefold: (1) There is no benchmark EEG dataset specifically for comparing attention recognition performance; (2) Mental task requires attention in general; (3) Recognition performance is available in the literature, for all 5 mental tasks and 6 subjects, the overall accuracy

2.9 Summary

is $54.8\% \pm 6.7\%$ before smoothing and $71.4\% \pm 9.8\%$ after smoothing using 6-order autoregression coefficients as the feature [79] (Table 2.5).

Table 2.5: Recognition performance of the mental task dataset in the literature

#subject	Mental task	Performance	Ref.
1	baseline, letter and multiplication	70%	[80]
6	all pairs, binary classification	89%	[81]
6	all 5 mental tasks	no smoothing: $54.8\% \pm 6.7\%$; smoothing: $71.4\% \pm 9.8\%$	[79]

2.9 Summary

The research questions and topics in the following chapters are based on the literature. First, what are the relevant brain regions and distinguishable EEG features in selective attention? The ERP approach is not applicable to real-time applications where the stimuli comes from the environment, so that the stimuli cannot be repeated for multiple times to average EEG signals. As a result, the SNR is lower in the non-ERP approach. In Chapter 4 we propose channel selection with different features (CSDF) to find the relevant brain regions and features in the recognition of selection attention without external stimuli.

Second, what are the vigilance stages in open eye and situation-aware state? According to the literature review, various vigilance stage models are proposed for closed eye resting condition, while no vigilance stage model under open eye and situation-aware state. In Chapter 5 we describe the vigilance stages during simulated driving. The vigilance stages are defined in a subject-independent way, achieved by learning features from the EEG spectrum using ELMAE.

Finally, analogue to selecting features from a predefined set of features, can we automatically learn sparse features so that the useful information is restricted to a few nonzero features? In Chapter 6 we propose joint spike-timing dependent plasticity (joint STDP) to learn sparse EEG spectral features to achieve better interpretability.

Chapter 3

Recognition of Selective Attention

We propose an experiment which aims to recognize open eye resting state (unattended) and performing TOVA test (attended), since there is no non-ERP studies using TOVA to induce selective attention. Such non-ERP experiment enables the recognition of selective attention without referring to external stimuli. Compared to other psychological tests, TOVA uses geometric stimuli which minimizes the side effects in the perception of stimuli.

Section 3.2 describes the feature extraction and classification algorithm. In Section 3.3 we compare the recognition performance on the public mental task dataset, a study using SCT task, a study with the best accuracy claimed in the literature, and with different classifiers.

The experiment proposed in this chapter is used in Chapter 4 for proposing new feature selection algorithm.

3.1 Background and Motivation

The Test of Variables of Attention (TOVA) is a standardized psychological assessment of attention for attentional disorders such as Attention Deficit Hyperactive Disease (ADHD) [24]. In TOVA, a square flashes on the screen with random interval. There are two types of stimuli: target and non-target, where target means that the square appears at the top and non-target means the square appears at the bottom. The subject should respond only to the target. TOVA contains two sections: target infrequent and target frequent: the target infrequent section is 11.2 minutes long and the target vs. non-target ratio is 1:3.5; the target frequent section is also 11.2 minutes long and the target vs. non-target ratio is 3.5:1.

In the literature, there are ERP studies which study EEG segments time-locked to the stimuli in TOVA test [82]. However, TOVA has not been used in non-ERP studies, compared to other psychological tests related to selective attention [25–27].

Here we propose an experiment which classifies EEG signals between open eye resting state (unattended) and performing TOVA test (attended). Such non-ERP experiment enables the recognition of selective attention without referring to external stimuli, which is advantageous in real-time applications where the external stimuli cannot be controlled. For example, selective attention induced in TOVA test is analogous to that in a driving scenario: ignoring traffic lights (unattended) and attentively looking at traffic lights (attended). The TOVA test is also advantageous in a EEG-based attention study compared to other psychological tests, since it uses geometric stimuli (up and down) which minimizes the side effects in the perception of stimuli (Table 2.1).

In this chapter, the machine learning algorithm for recognizing selective attention is based on the common features and widely used classifiers in the literature. This chapter serves as a starting point for the novel algorithms in the later chapters.

3.2 Method

3.2.1 Experiment Setup

The whole experiment consisted of two parts, i.e. open eye resting (baseline, unattended) and performing TOVA test (attended). The baseline state is defined as subject sitting in a comfortable chair in front of the computer with eyes open without large body movements for 10 minutes, during which the subject was asked to relax and not pay attention to any particular objects for long time. Such open eye baseline is appropriate for tasks involving visual processing [83]. In the other part, the subject performed the target infrequent section of TOVA test for 10 minutes. This is because the two sections of TOVA are designed to distinguish different subtypes of ADHD, which is not studied in this experiment. Between baseline and TOVA test, the subject had a 1 minute break with minor body movements but cannot leave the seat. The order of these two parts was counterbalanced for different subjects to minimize the effect of ordering.

9 subjects (2 females and 7 males) participated the experiment (age from 23 to 35, mean = 26, SD = 4). They were recruited from Nanyang Technological University by word of mouth and informed 5 days in advance before each test run. All of them are right-handed and without reported history of mental deficiency. The experiment time was controlled between 11:00am to 12:00am before lunch without consuming tea, coffee or alcohol before or during the experiment. All instructions were given and all experiment data were collected by the same researcher.

We used Emotiv EEG headset (Emotiv Systems Inc., San Francisco, CA, USA) with 14 channels located at AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8 and AF4 according to the international 10-20 system. The sampling rate of EEG signals is 128Hz, notch filtered at 50Hz and 60Hz and bandpass filtered from 0.2Hz to 45Hz.

3.2.2 Preprocessing and Feature Extraction

In Figure 3.1 we show the flowchart of EEG signal processing from raw signal to the final classification.

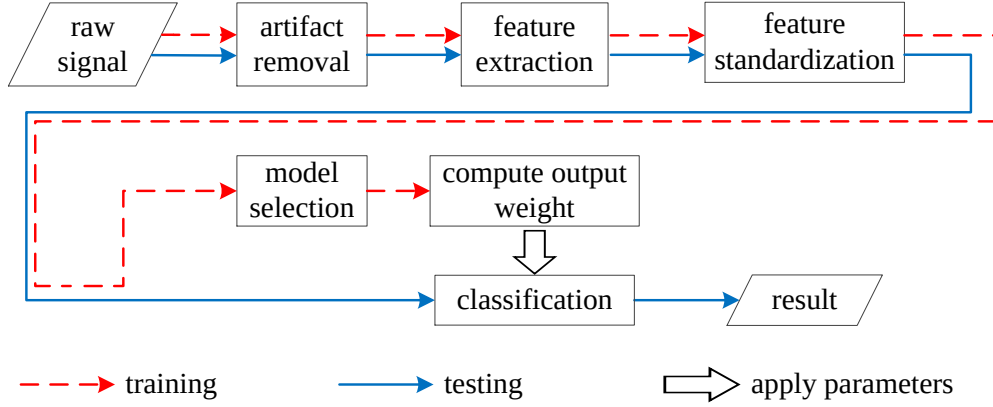


Figure 3.1: EEG signal processing flowchart

EEG data were first pre-processed to remove artifacts. These artifacts include blink, eyeball movement (horizontal, vertical and saccade), muscular artifacts, electrocardiogram (ECG) artifacts and loss of electrode contact with scalp. Blink artifacts usually exhibit peak-like shape in the signal, which is most dominant in frontal channels. Eyeball movement can bring staircase-like shape into the signal. Muscular artifacts resemble short burst of high frequency signals. ECG is represented as rhythmic fluctuation. Loss of electrode contact with scalp leads to dramatic change in the signal amplitude. It is expected that artifacts are statistically independent from the EEG signal from brain. Independence component analysis (ICA, Section 2.2) was used for the whole data (including training and testing data) to get 14 components (same with the number of channels), followed by manual identification of components corresponding to artifacts.

For the training set, EEG signals from each channel were first segmented into overlapping windows with length of 4 seconds and overlap of 3 seconds. EEG features were extracted from each segment. Three features were extracted from the time domain: the signal mean, standard deviation and line length ($\frac{1}{N-1} \sum_n |x_{n+1} - x_n|$) [44, 84]. We also used 4 band power features from θ (4–8Hz), α (8–12Hz), β (12–30Hz) and γ (30–42Hz)

bands. The reason of using band power features is supported by the ERP literature where the band powers are found to be functionally related to behavioral efficiency and perceptual accuracy [11, 12]. In total, 98 features were generated from all EEG signals (7 features per channel and 14 channels in total). The feature space of subject 1 is visualized in Figure 3.2. The feature space for other subjects share a similar pattern, where the extent of separability is related to the classification accuracy. For the training set, features were standardized by subtracting mean and dividing standard deviation to have zero mean and unit variance across samples. For the testing set, we standardized features in the same way by eliminating mean and dividing standard deviation from the training features.

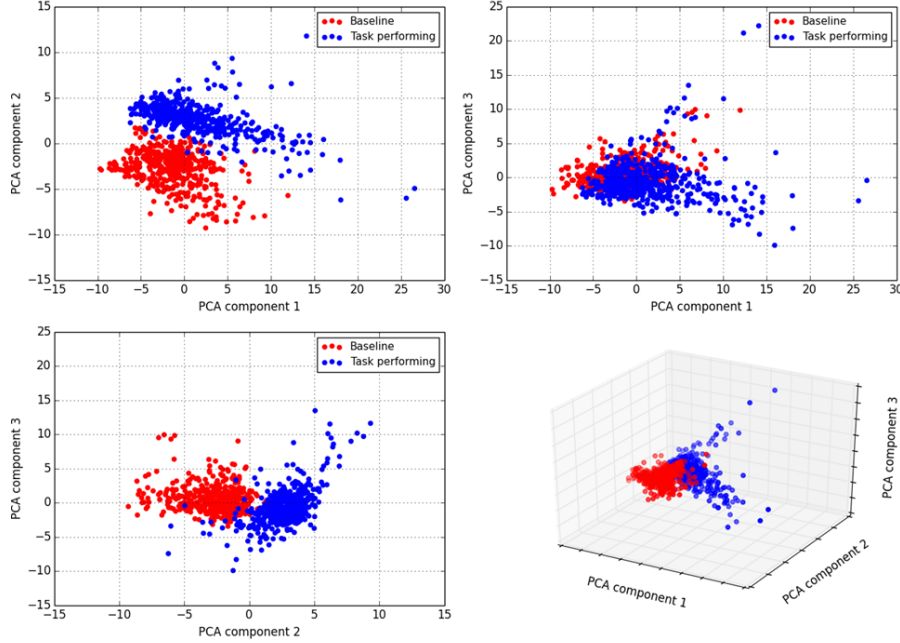


Figure 3.2: Visualization of the feature space for subject 1

3.2.3 Classification

We used extreme learning machine (ELM) [85, 86] to classify the EEG signals into baseline and task-performing states. SVM was used as a benchmark algorithm to compare the accuracy. Five-fold cross validation was used to select classifier hyperparameters to reduce over-fitting. The training dataset, which lasts for 8 minutes long, was partitioned

into 5 nonoverlapping folds with equal time length, i.e. 1.6 minutes for each segment (Figure 3.3). Each fold was used as validation set in turn and the other four folds were used as training set. Different combinations of hyperparameters were explored to find the optimal cross validation accuracy averaged from the 5 folds. For the ELM classifier with sigmoid activation function, the range of hyperparameter C was $[10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^0, 10^1, 10^2, 10^3, 10^4]$; the range of hidden layer neuron number h was $[500, 1000, 2000]$. For the SVM classifier with sigmoid kernel, the range of hyper parameter C was the same with ELM; the range of γ is $[10^{-3}, 10^{-2}, 10^{-1}, 10^0]$; the range of bias term in the kernel was $[-10^1, -10^0, 0, 10^0, 10^1]$.

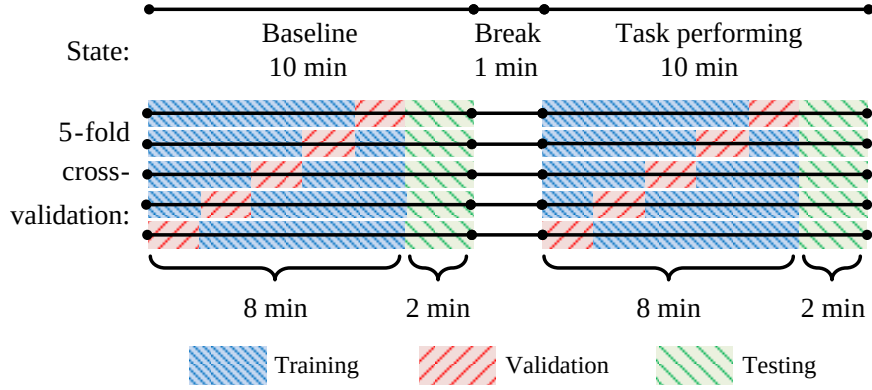


Figure 3.3: Experiment procedure and data splitting

3.3 Results

Three types of comparison are introduced in the following four subsections. First, the only public EEG dataset recorded during psychological test, including resting baseline, and with published recognition accuracy is the mental task dataset (Section 2.8), based on which we compare our accuracy with state-of-the-art result [79]. Second, since the data recorded during other psychological tests are not available, we compare the recognition performance between using our features and using their features on our TOVA dataset. This includes a study using SCT task which shares common features with ours [27]; and a study recognizing drowsiness [31], which claims the best recognition

accuracy among studies recording EEG during psychological tests, but not necessarily on selective attention. Finally, we also compare the performance on our TOVA dataset using different classifiers.

3.3.1 Comparison with Results on Public Mental Task Dataset

The mental task dataset (details in Section 2.8) [74, 79–81] is used for performance comparison. For each session, both EEG and EOG signals were segmented using 1s window size and 0.75s overlap, where EEG with amplitude more than $\pm 100\mu V$ and EOG more than $\pm 50\mu V$ were removed due to possible artifacts. For each session, we did nested leave-one-trial-out cross validation, resulting in 5 testing accuracy per session, where the average value is used as the testing accuracy of that session.

Table 3.1: Testing accuracy using our feature and the state-of-the-art result on mental task dataset

Subject	Our features	State-of-the-art result
S1_1	66.0%	60.9%
S1_2	56.6%	64.0%
S2_1	64.6%	61.4%
S3_2	61.3%	51.3%
S5_1	53.7%	48.5%
S5_2	41.8%	43.2%
S6_1	69.9%	52.3%
S6_2	61.3%	50.8%
S7_1	71.6%	60.5%
Mean	60.8%	54.8%
SD	8.6%	6.7%

From Table 3.1 we see that in 7 out of 9 sessions we had higher accuracy than the state-of-the-art result using 6-order autoregression coefficients [79]. On average, we achieved recognition accuracy $60.8\% \pm 8.6\%$, significantly higher compared to state-of-the-art accuracy $54.8\% \pm 6.7\%$ according a paired t-test (p-value 0.04).

We need to point out that in Liang et al. [79], they used offline smoothing to improve the

accuracy to $71.4\% \pm 9.8\%$, which however makes use of future EEG segments. In real-time applications, we cannot use information from the future EEG segments. Therefore the improvement in accuracy cannot be guaranteed using smoothing.

3.3.2 Comparison with A Study using SCT Task

For psychological tests other than TOVA, Besserve et al. [27] recognizes the performance level (from percentile of reaction times) during a spatial stimulus-response compatibility task (SCT) based on EEG. SCT is a selective attention task where an arrow appears randomly on the left or right side of the screen, and the subject only responds to the direction of the arrow while ignoring the location of the arrow (Table 2.1). Same with our study, power spectral density features from θ , α and β bands were also extracted. Different with our study, the phase locking value between each pair of channels were also extracted as features. SVM was used as the classifier. They achieved recognition accuracy of 71% averaged from 12 subjects.

Table 3.2: Testing accuracy using Besserve’s features, our features and combined features

Subject	Besserve’s (B)	Ours (O)	Combined (C)	Significance
1	82.9%	94.4%	91.5%	BC*
2	98.3%	96.2%	97.4%	n.s.
3	91.9%	99.1%	97.4%	n.s.
4	71.8%	95.3%	73.9%	BO**, OC**
5	82.9%	89.7%	87.6%	n.s.
6	77.8%	95.3%	92.3%	BO**
7	91.9%	98.7%	100%	n.s.
8	81.6%	85.9%	90.6%	n.s.
9	76.9%	98.3%	89.7%	BO*, OC**, BC**
Mean	84.0%	94.8%	91.2%	n.s.
SD	8.0%	4.1%	7.2%	
n.s. $p > 0.05$ * $p < 0.05$ ** $p < 0.01$				

Since their data is not publicly available, we applied Besserve’s features on our data, and kept SVM as the classifier. From Table 3.2 we observe that in terms of the average accu-

racy over all subjects, none of our features, Besserve’s features and combined features yields significant difference. Since the recognized label is known in all cases, we used McNemar’s test to determine whether there is marginal homogeneity on the paired recognized labels. The insignificance between the results in the average case indicates that our approach has the recognition ability similar with studies having common features in the literature.

3.3.3 Comparison with A Study with the Best Accuracy in the Literature

We also compare the study with the best claimed accuracy, among studies recording EEG during psychological tests, but not necessarily on selective attention. Yeo et al. [31] recognizes drowsiness when watching a driving video for 1 hour. The accuracy was 99.3% for “drowsy” and “alert” averaged from 20 subjects, which is the best accuracy in the literature with similar background using EEG. The data are manually labelled by two independent raters using criteria combining EEG spectrum and eye blink. They use four types of features derived from EEG spectrum: dominant frequency, average power of dominant frequency peak, center of gravity frequency and frequency variability. No feature or channel selection is used in their work. SVM is used as the classifier.

Due to the fact that the data in Yeo et al. [31] is not publicly available, we applied Yeo’s features on our data, and kept SVM as the classifier. The result is shown in Table 3.3 with significance from McNemar’s test.

For the comparison between our features and Yeo’s features in Table 3.3 we can see that our features outperformed Yeo’s features on 6 out of 9 subjects. No significant improvement is found for Yeo’s features in any subject. On average, the accuracy obtained using our features ($94.8\% \pm 4.1\%$) outperformed Yeo’s features ($80.3\% \pm 8.6\%$) significantly (McNemar’s test p -value 0.034), indicating the better performance using our features in our dataset.

Table 3.3: Testing accuracy using Yeo’s features, our features and combined features

Subject	Yeo’s (Y)	Ours (O)	Combined (C)	Significance
1	87.6%	94.4%	95.7%	n.s.
2	97.9%	96.2%	97.0%	n.s.
3	75.2%	99.1%	96.2%	YO**, YC**, OC**
4	70.9%	95.3%	83.3%	YO*, OC**
5	73.9%	89.7%	82.5%	OC*
6	69.2%	95.3%	92.7%	YO**, YC**
7	84.6%	98.7%	100%	YO**, YC**
8	82.9%	85.9%	88.9%	YO*
9	80.8%	98.3%	85.5%	YO*, YC*, OC**
Mean	80.3%	94.8%	91.3%	YO*, YC*
SD	8.6%	4.1%	6.1%	
n.s. $p > 0.05$ * $p < 0.05$ ** $p < 0.01$				

For the comparison between Yeo’s features and combined features in Table 3.3, 4 out of all 9 subjects showed significant improvement for the combined features. No significant improvement is found for Yeo’s features in any subject. On average, the accuracy of using combined features ($91.3\% \pm 6.1\%$) outperformed Yeo’s features ($80.3\% \pm 8.6\%$) significantly confirmed by McNemar’s test with p-value 0.025, indicating the better performance using the combined features.

For the comparison between our features and combined features in Table 3.3, 4 out of all 9 subjects showed significant improvement for our features. No significant improvement is found for combined features in any subject. On average, the accuracy of using our features ($94.8\% \pm 4.1\%$) did not show statistically significant difference confirmed by McNemar’s test with p-value 0.969. In other words, the combination of Yeo’s feature did not add new information to our features.

3.3.4 Comparison with SVM

We also compared the performance between ELM and SVM, both being widely used in the literature. The testing accuracy of both ELM and SVM are shown in Table 3.4. We achieved 94.4% average testing accuracy ($SD = 5.0\%$) for ELM and 94.8% average

3.4 Summary

testing accuracy (SD = 4.1%) for SVM. On average, the difference between ELM and SVM is not statistically significant confirmed by McNemar’s test (p-value 0.8). The 0.4% difference in the testing accuracy is about 1 sample out of 234 testing samples.

Table 3.4: Testing accuracy using ELM and SVM

Subject	ELM	SVM	significance
1	96.2%	94.4%	n.s.
2	95.7%	96.2%	n.s.
3	97.9%	99.1%	n.s.
4	97.0%	95.3%	n.s.
5	88.9%	89.7%	n.s.
6	92.7%	95.3%	n.s.
7	100%	98.7%	n.s.
8	83.3%	85.9%	n.s.
9	97.9%	98.3%	n.s.
Mean	94.4%	94.8%	n.s.
SD	5.0%	4.1%	
n.s. p > 0.05 * p < 0.05 ** p < 0.01			

3.4 Summary

TOVA has not been used in non-ERP studies, compared to other psychological tests related to selective attention. Here we propose an experiment which classifies EEG signals between open eye resting state (unattended) and performing TOVA test (attended). Such non-ERP experiment enables the recognition of selective attention without referring to external stimuli, which is advantageous in real-time applications where the external stimuli cannot be controlled.

We first compared to the results on EEG-based mental task recognition, which is the only public EEG dataset recorded during psychological test, including resting baseline, and with published recognition accuracy. On average, we achieved $60.8\% \pm 8.6\%$ accuracy, significantly higher compared to the state-of-the-art accuracy $54.8\% \pm 6.7\%$ according to a paired t-test.

There are studies recognizing brain states during other psychological tests except TOVA test, while their data are not available. As a result, we compare the recognition performance between using our features and using their features on our TOVA dataset. First, there is a study recognizing performance level during a SCT task, which shares common features with ours. It turned out that in terms of the average accuracy over all subjects, none of our features, their features and the combined features yielded significant difference. As a result, our approach had the recognition ability similar with the study having some common features in the literature, which ensured the validity of our research. Second, we compared with the one claimed the highest accuracy on their dataset, among studies recording EEG during psychological tests, but not necessarily on selective attention (Yeo et al. [31]). On our dataset, for both cases of using our features ($94.8\% \pm 4.1\%$) and using combined features ($91.3\% \pm 6.1\%$), the accuracy was higher than Yeo's ($80.3\% \pm 8.6\%$) with statistical significance.

To validate the classifier, we compared the performance of ELM and SVM. For the 9 subjects, we obtained $94.4\% \pm 5.0\%$ average test accuracy for ELM and $94.8\% \pm 4.1\%$ average test accuracy for SVM, where the difference was not statistically significant (McNemar's test, about 1 out of 234 testing samples).

There is strong subject dependence in the recognition of selection attention, so the result ranged from 83.3% to 100% (about 39 out of 234 testing samples, Table 3.4). During the experiment, we noticed that some subjects tended to think carefully before making the response, which led to slower reaction; other subjects tended to make response without much thinking, which led to faster reaction. Subject-independent algorithms will be introduced in Chapter 5 and Chapter 6.

Chapter 4

EEG Channel Selection with Different Features (CSDF)

In the previous chapter, we propose an experiment which aims to recognize unattended and attended states induced by TOVA test, using EEG features from ERP studies without selection. However, ERP does not work for real-time applications where the stimuli cannot be controlled. In face of low SNR in the non-ERP approach, what are the relevant brain regions and distinguishable EEG features in selective attention?

In this chapter, we propose Channel Selection with Different Features (CSDF) algorithm for EEG-based recognition of selective attention, as described in Section 4.2. In Section 4.3 we compare the performance of CSDF on a public EEG dataset, as well as on the TOVA dataset in Chapter 3. Section 4.4 describes the relevant EEG channels and distinguishable EEG features: the first and second order difference in the left parietal and temporal lobes, and the Higuchi fractal dimension and mean signal amplitude in the right frontal lobe. The selected channels are consistent with a PET-rCBF neuroimaging study using ERP in the literature.

4.1 Background and Motivation

EEG signals come with multiple channels and features but not all of them contribute to the target brain state, especially in the non-ERP approach with low SNR. In EEG based attention recognition, channel selection provides evidence for the optimal electrode positions for recording and understanding of the brain regions involved in different attentional states; feature selection finds the discriminant features which improves recognition performance as well as reduces computational burden. As a result, feature selection and channel selection help to identify the important features and channels in selective attention.

In existing channel selection methods, all features in the same channel are treated with equal importance. However, different brain regions have different functions, therefore having different important features in each channel. For example, the alpha power is dominant in the occipital lobe, reflecting idling state or inhibition of task-irrelevant activities [12]. It remains a challenge to select channels with their own different sets of features, as well as to restrict the useful features to as few channels as possible to allow a clearer interpretation.

Here we address the question of selecting an optimal subset of channels with their own different best features. We propose Channel Selection with Different Features (CSDF), based on least square linear discriminant analysis (LS-LDA) [87] and group lasso. It provides an optimization framework where features are selected in both channel-wise and feature-wise manner, allowing for different features being selected in channels and also the features being grouped into channels as much as possible.

4.2 Method

4.2.1 Least Square Linear Discriminant Analysis

The formulation of CSDF consists of two parts. The first part, i.e. the loss function, is based on least square linear discriminant analysis (LS-LDA) [87], which transforms the original trace ratio problem of multiple discriminant analysis (MDA, a multiclass extension of LDA) into a least square problem under a mild condition.

First we introduce the MDA problem. Given a dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the i -th feature vector (sample), $y_i \in \{1, 2, \dots, k\}$ is the class label of the i -th sample, n is the total number of samples, d is the number of features and k is the total number of classes. The optimal solution of MDA, i.e. $\mathbf{W}_{\text{MDA}}^*$ is obtained from

$$\underset{\mathbf{W}}{\text{minimize}} \quad J_0 = \text{tr} \left((\mathbf{W}^\top \mathbf{S}_b \mathbf{W}) (\mathbf{W}^\top \mathbf{S}_t \mathbf{W})^\dagger \right), \quad (4.1)$$

where $\mathbf{W} \in \mathbb{R}^{d \times l}$ maps $\mathbf{x}_i$ from d -dimensional space to l -dimensional space ($l < d$); $\mathbf{A}^\dagger$ denotes the pseudo-inverse of matrix $\mathbf{A}$. The three scatter matrices, namely within-class $\mathbf{S}_w$, between-class $\mathbf{S}_b$ and total scatter matrix $\mathbf{S}_t$ are

$$\mathbf{S}_w = \frac{1}{n} \sum_{j=1}^k \sum_{\mathbf{x}_i \in \mathbf{X}_j} (\mathbf{x}_i - \mu_j) (\mathbf{x}_i - \mu_j)^\top, \quad (4.2)$$

$$\mathbf{S}_b = \frac{1}{n} \sum_{j=1}^k n_j (\mu_j - \mu) (\mu_j - \mu)^\top, \quad (4.3)$$

$$\mathbf{S}_t = \mathbf{S}_w + \mathbf{S}_b = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \mu) (\mathbf{x}_i - \mu)^\top, \quad (4.4)$$

where n_j is the number of samples belonging to class j ; $\mathbf{X}_j \in \mathbb{R}^{n_j \times d}$ denotes the samples belonging to class j ; $\mu \in \mathbb{R}^d$ is the mean of all samples; $\mu_j \in \mathbb{R}^d$ is the mean of samples belong to class j .

Next we briefly review LS-LDA based on [87]. The optimal solution of LS-LDA, i.e. $\mathbf{W}_{\text{LS-LDA}}^*$ is solved from

$$\underset{\mathbf{W}}{\text{minimize}} \quad J_1 = \frac{1}{2} \left\| \hat{\mathbf{X}}\mathbf{W} - \mathbf{Y} \right\|_{\text{F}}^2, \quad (4.5)$$

where $\hat{\mathbf{X}} \in \mathbb{R}^{n \times d}$ is the centered data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]^\top \in \mathbb{R}^{n \times d}$, $\mathbf{W} \in \mathbb{R}^{d \times k}$ and the label indicator matrix $\mathbf{Y} \in \mathbb{R}^{n \times k}$ is defined as

$$Y_{ij} = \begin{cases} \sqrt{\frac{n}{n_j}} - \sqrt{\frac{n_j}{n}} & \text{if } y_i = j, \\ -\sqrt{\frac{n_j}{n}} & \text{otherwise.} \end{cases} \quad (4.6)$$

$\mathbf{W}_{\text{LS-LDA}}^*$ and $\mathbf{W}_{\text{MDA}}^*$ are related by

$$\mathbf{W}_{\text{LS-LDA}}^* = \left[\mathbf{W}_{\text{MDA}}^* \Sigma_{bq}^{\frac{1}{2}} \quad \mathbf{0}_{d \times (k-q)} \right] \mathbf{Q}^\top, \quad (4.7)$$

where $\mathbf{Q}$ is a $k \times k$ orthogonal matrix and $\Sigma_{bq}^{\frac{1}{2}} = \mathbf{I}_q$ under a mild condition that $\text{rank}(\mathbf{S}_b) + \text{rank}(\mathbf{S}_w) = \text{rank}(\mathbf{S}_t)$, and $\mathbf{I}_q$ is the identity matrix with size $q = \text{rank}(\mathbf{S}_b) < k$.

Equation (4.7) shows an essentially equivalence relationship between $\mathbf{W}_{\text{MDA}}^*$ and $\mathbf{W}_{\text{LS-LDA}}^*$ up to a rotation $\mathbf{Q}$ in Euclidean space and additional zero columns to compensate for the rank [87]. The rotation $\mathbf{Q}$ does not affect classifier performance since it preserves the relative position of samples in the feature space. The additional zero columns can be avoided by taking the first $r = \text{rank}(\mathbf{W}_{\text{LS-LDA}}^*)$ columns of $\mathbf{W}_{\text{LS-LDA}}^*$. So the MDA problem in Equation (4.1) can be reformulated as a least square problem in Equation (4.5) with label indicator matrix $\mathbf{Y}$ defined in Equation (4.6). And by solving the least square problem, we map the data into an optimal discriminant subspace so that the irrelevant features are reduced.

4.2.2 Channel-Wise Sparseness for Channel Selection

The second part of CSDF is the regularization terms. There are two parts: a channel-wise group lasso term for channel selection (subsection 4.2.2) and an $\ell_{1,2}$ -norm term for determining features in each channel (subsection 4.2.3).

Since we have already mapped data into an optimal l -dimensional discriminant subspace ($l = q = \text{rank}(\mathbf{S}_b) < k$) using LS-LDA, the next step is to introduce channel-wise sparseness in the rows of the transformation matrix $\mathbf{W} \in \mathbb{R}^{d \times k}$ that belongs to the same channel, so that the features tend to be grouped in channels. The formulation is

$$\underset{\mathbf{W}}{\text{minimize}} \quad J_2 = \frac{1}{2} \left\| \hat{\mathbf{X}}\mathbf{W} - \mathbf{Y} \right\|_{\text{F}}^2 + C_1 \|\mathbf{W}\|_{\text{ch},2}, \quad (4.8)$$

where C_1 is a hyperparameter which indicates the trade-off between class discrimination and channel-wise sparseness. We define $\|\mathbf{W}\|_{\text{ch},2}$ as

$$\|\mathbf{W}\|_{\text{ch},2} := \sum_{j=1}^g \|\mathbf{W}^j\|_{\text{F}} = \sum_{j=1}^g \sqrt{\sum_{i \in \text{Channel}_j} \|\mathbf{w}_i\|_2^2}, \quad (4.9)$$

where g is the number of channels, $\mathbf{W}^j \in \mathbb{R}^{d_j \times k}$ is the matrix consisting of rows corresponding to Channel_j , d_j is the number of features belonging to Channel_j and $\mathbf{w}_i$ is the i -th row of $\mathbf{W}$. The reason of designing such regularization is illustrated in Figure 4.1. It is desirable that in the solution, all entries in the block of $\mathbf{W}^j$ become zero all together, or becomes nonzero all together. If all entries in $\mathbf{W}^j$ become zero, i.e. $\|\mathbf{W}^j\|_{\text{F}} = \sqrt{\sum_{i \in \text{Channel}_j} \|\mathbf{w}_i\|_2^2} = 0$, Channel_j can be removed since no feature of that channel is selected in the next step. From optimization point of view, this is equivalent to imposing group lasso regularization to blocks of rows in $\mathbf{W}$, where different groups (i.e. channels) are combined in an ℓ_1 -norm manner, i.e. $\sum_{j=1}^g \|\mathbf{W}^j\|_{\text{F}} = \|\mathbf{W}\|_{\text{ch},2}$.

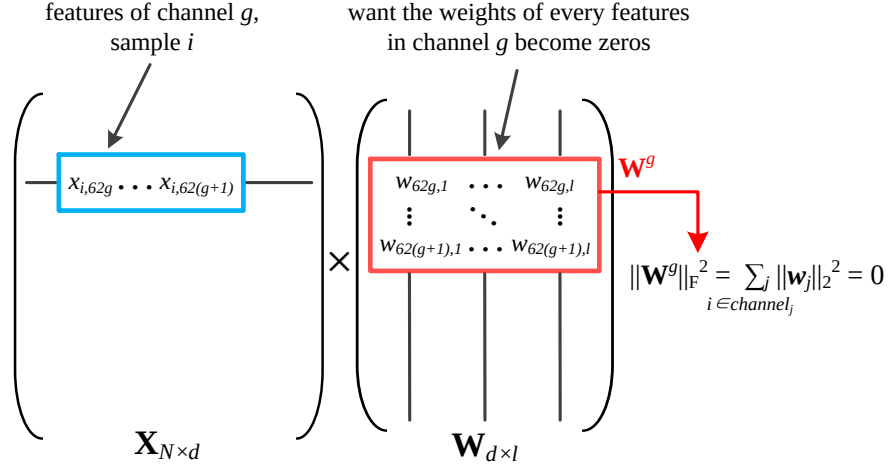


Figure 4.1: Illustration of channel-wise sparseness in the weight matrix $\mathbf{W}$. The blocks in matrices indicate values corresponding to Channel_j . 62 is the number of features in each channel.

4.2.3 Feature-Wise Sparseness for Determining Features in Each Channel

Similar to introducing channel-wise sparseness, we extend our formulation by adding the second regularization term $\|\mathbf{W}\|_{1,2}$ to have feature-wise sparseness in the each channel

$$\underset{\mathbf{W}}{\text{minimize}} \quad J_3 = \frac{1}{2} \left\| \hat{\mathbf{X}}\mathbf{W} - \mathbf{Y} \right\|_F^2 + C_1 \|\mathbf{W}\|_{\text{ch},2} + C_2 \|\mathbf{W}\|_{1,2}, \quad (4.10)$$

where C_2 is another hyperparameter, $\|\mathbf{W}\|_{1,2} = \sum_i \|\mathbf{w}_i\|_2$ and $\mathbf{w}_i$ is the i -th row of $\mathbf{W}$. Equation (4.10) is the final formulation of CSDF. After solving the optimal weight matrix, only channels with nonzero blocks and features with nonzero rows are selected. For the proof of sparseness in the solution, refer to Appendix A.

4.2.4 Flowchart of the CSDF Algorithm

As shown in Equation (4.10), CSDF provides an optimization framework where channels are selected with different features by tuning the hyperparameters C_1 and C_2 . In

Figure 4.2 we show the whole flowchart of data processing including CSDF.

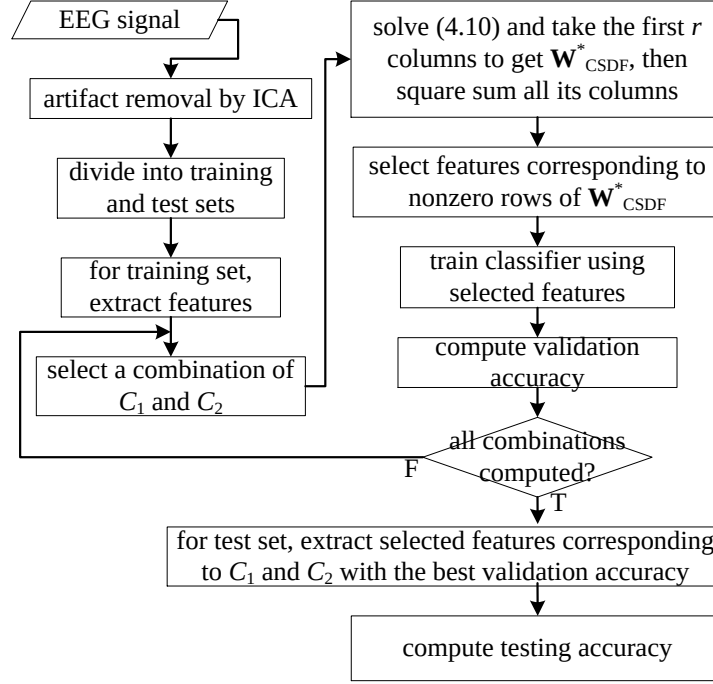


Figure 4.2: Flowchart of EEG based brain state recognition using CSDF

We iterated over all combinations of C_1 and C_2 selected from a list $[0, 2^{-10}, 2^{-9}, \dots, 2^9, 2^{10}]$ and solved Equation (4.10) for each combination. Note that for $C_1 = 0$ and $C_2 \neq 0$, Equation (4.10) reduces to feature selection only and is essential equivalent to sparse LDA [54]; likewise for $C_1 \neq 0$ and $C_2 = 0$, Equation (4.10) reduces to channel selection only. We used CVX, a package for specifying and solving convex programs [88, 89] in Matlab to solve Equation (4.10). Assume that the number of features d is much larger than the number of classes k ($d \gg k$), according to Equation (4.7), we have the following relationship regarding the rank of $\mathbf{W}_{\text{LS-LDA}}^*$ and $\mathbf{W}_{\text{MDA}}^*$,

$$\begin{aligned}
 r &= \text{rank}(\mathbf{W}_{\text{LS-LDA}}^*) \\
 &\leq \min \{ \text{rank}(\mathbf{W}_{\text{MDA}}^*), \text{rank}(\mathbf{Q}^\top) \} \\
 &= \min \{ \text{rank}(\mathbf{W}_{\text{MDA}}^*), k \} && (\mathbf{Q} \text{ is a } k \times k \text{ orthogonal matrix}) \quad (4.11) \\
 &= \text{rank}(\mathbf{W}_{\text{MDA}}^*) && (\text{rank}(\mathbf{W}_{\text{MDA}}^*) \leq q < k) \\
 &\leq q < k \ll d.
 \end{aligned}$$

Since $\mathbf{W}_{\text{LS-LDA}}^*$ is a $d \times k$ matrix, it is not full column rank. Therefore we only take the first $r = \text{rank}(\mathbf{W}_{\text{LS-LDA}}^*)$ columns to get $\mathbf{W}_{\text{CSDF}}^*$

$$\mathbf{W}_{\text{CSDF}}^* = \mathbf{W}_{\text{LS-LDA}}^* \times \begin{bmatrix} \mathbf{I}_r \\ \mathbf{0}_{(k-r) \times r} \end{bmatrix} \in \mathbb{R}^{d \times r}. \quad (4.12)$$

Rows with all zeros in $\mathbf{W}_{\text{CSDF}}^*$ indicate that the corresponding features are not used; while the nonzero rows correspond to features that are selected. A convenient way of evaluating the values is to square sum all columns of $\mathbf{W}_{\text{CSDF}}^*$ so that each row is represented by one number.

For classification, we used ELM with sigmoid activation function as the classifier, where the regularization parameter C were chosen from $[10^{-4}, 10^{-3}, \dots, 10^4]$ and number of hidden nodes fixed at 1000. After iterating over all C_1 and C_2 combinations, we recorded the values and selected features corresponding to the best validation accuracy. We applied these values to testing set, where only the selected features are extracted. Testing accuracy was computed on the 234 testing samples as the measure of algorithm performance.

4.2.5 EEG Features

For feature selection, one should first extract an abundant feature set, so that the selected set can represent the optimal combination of features. We extracted 62 features in each channel from five categories, as shown in Table 4.1. In total there were $62 \times 14 = 868$ features. The EEG device is the same with Section 3.2.1.

Table 4.1: EEG features extracted for selection

category	feature	# / channel
statistics	mean & standard deviation of amplitude [90]	2
	1st, 2nd order diff [90] (i.e. line length [84])	2
	normalized 1st, 2nd order difference [90]	2
	skewness, kurtosis	2
frequency domain	$\theta, \alpha, \beta, \gamma$ absolute band power	4
	$\theta, \alpha, \beta, \gamma$ relative band power	4
	$\theta, \alpha, \beta, \gamma$ dominant frequency [31]	4
	$\theta, \alpha, \beta, \gamma$ average peak power [31]	4
	$\theta, \alpha, \beta, \gamma$ center of gravity [31]	4
	$\theta, \alpha, \beta, \gamma$ band variance [31]	4
	$(\alpha + \theta)/\beta, \alpha/\beta, (\alpha + \theta)/(\alpha + \beta), \theta/\beta$ [91]	4
time domain	6-order auto-regression [79]	6
	higher order crossing (HOC) [92]	16
entropy	Shannon entropy [93]	1
	approximate entropy [93]	1
	sample entropy [93]	1
fractal dimension	Higuchi fractal dimension [94]	1
#feature / channel		62

4.3 Results

4.3.1 Validation of CSDF on Public Mental Task Dataset

CSDF was first applied to a public EEG dataset — mental task dataset (details in Section 2.8) [74, 79–81] — for performance evaluation. The preprocessing is already introduced in Section 3.3.1.

For channel selection, CSDF was compared with other 5 channel selection algorithms: Fisher’s criterion based sequential forward selection [55] for channels (Fisher_SFS_C), Fisher’s criterion based sequential floating forward selection [51] for channels (Fisher_SFFS_C), maximum-relevance-minimum-redundancy based sequential forward selection [56] for channels (mRMR_SFS_C), maximum-relevance-minimum-redundancy based sequential floating forward selection [56] for channels (mRMR_SFFS_C) and SVM recursive

4.3 Results

channel elimination (SVM.RCE) [58]. For each algorithm, the cross-trial validation accuracy of 1, 2, $\dots$, 5 channel(s) from the total 6 channels were calculated where the best one was used to evaluate the final testing accuracy.

Table 4.2 shows that CSDF has the best accuracy for 5 out of 9 sessions. For the overall accuracy, CSDF is the best among no channel selection and other channel selection algorithms. However, Kruskal-Wallis H test (nonparametric version of ANOVA one-way test) rejects statistical difference among all cases at $\alpha = 0.05$ level.

Table 4.2: Testing accuracy from 5 channel selection algorithms and CSDF on the mental task dataset

Subject_Session	Channel selection algorithm						
	None	C1	C2	C3	C4	C5	CSDF
S1_1	73.7%	73.7%	73.7%	74.4%	74.4%	72.9%	72.2%
S1_2	71.9%	75.6%	75.6%	76.3%	76.3%	73.7%	79.1%
S2_1	72.3%	65.4%	65.4%	67.7%	67.7%	66.0%	70.4%
S3_2	53.2%	54.3%	54.3%	43.9%	43.9%	54.3%	60.1%
S5_1	59.5%	53.0%	53.0%	53.0%	53.0%	53.4%	59.5%
S5_2	50.5%	47.2%	47.2%	42.9%	42.9%	50.6%	41.6%
S6_1	71.9%	71.3%	71.3%	71.3%	71.3%	70.4%	74.8%
S6_2	58.3%	55.0%	55.0%	60.5%	60.5%	58.7%	61.9%
S7_1	72.4%	71.4%	71.4%	71.4%	71.4%	74.6%	74.4%
Mean	64.8%	63.0%	63.0%	62.4%	62.4%	63.8%	66.0%
SD	8.8%	10.0%	10.0%	12.2%	12.2%	9.1%	10.9%

None: no channel selection; C1: Fisher_SFS_C; C2: Fisher_SFFS_C;
C3: mRMR_SFS_C; C4: mRMR_SFFS_C; C5: SVM_RCE.

For feature selection, CSDF was compared with other 5 feature selection algorithms, namely Fisher's criterion based sequential forward selection [55] (Fisher_SFS), Fisher's criterion based sequential floating forward selection [51] (Fisher_SFFS) maximum-relevance-minimum-redundancy based sequential forward selection [56] (mRMR_SFS), maximum-relevance-minimum-redundancy based sequential floating forward selection [56] (mRMR_SFFS) and SVM recursive feature elimination (SVM_RFE) [52]. For each algorithm, the cross-trial validation accuracy of $[2^3, 2^4, \dots, 2^8]$ features were calculated where the best one was used to evaluate the final testing accuracy. Table 4.3 shows that CSDF has the best accuracy for 3 out of 9 sessions. For the overall accuracy, CSDF is the second best

4.3 Results

among no feature selection and other feature selection algorithms. However, Kruskal-Wallis H test again reveals no statistical difference among all cases at $\alpha = 0.05$ level.

Table 4.3: Testing accuracy from 5 feature selection algorithms and CSDF on the mental task dataset

Subject_Session	Feature selection algorithm						CSDF
	None	F1	F2	F3	F4	F5	
S1_1	73.7%	77.7%	77.7%	71.5%	71.5%	80.5%	72.2%
S1_2	71.9%	72.5%	72.5%	61.3%	61.3%	74.4%	79.1%
S2_1	72.3%	78.0%	78.0%	76.7%	76.7%	74.6%	70.4%
S3_2	53.2%	50.9%	50.9%	63.9%	63.9%	67.3%	60.1%
S5_1	59.5%	51.7%	51.7%	49.1%	49.1%	52.6%	59.5%
S5_2	50.5%	50.6%	50.6%	49.4%	49.4%	39.3%	41.6%
S6_1	71.9%	74.5%	74.5%	66.7%	66.7%	79.8%	74.8%
S6_2	58.3%	58.3%	58.3%	58.3%	58.3%	59.9%	61.9%
S7_1	72.4%	78.9%	78.9%	72.8%	72.8%	77.8%	74.4%
Mean	64.8%	65.9%	65.9%	63.3%	63.3%	67.4%	66.0%
SD	8.8%	12.0%	12.0%	9.3%	9.3%	13.3%	10.9%

None: no feature selection; F1: Fisher_SFS; F2: Fisher_SFFS;
F3: mRMR_SFS; F4: mRMR_SFFS; F5: SVM_RFE.

In conclusion, CSDF obtained average recognition accuracy $66.0\% \pm 10.9\%$ on the public mental task dataset, which outperforms the state-of-the-art results in Table 2.5, where the accuracy is $54.8\% \pm 6.7\%$.

4.3.2 Comparison of Testing Accuracy for the TOVA Experiment

The data were taken from the TOVA experiment in the last chapter (Section 3.2.1). The EEG signals were segmented using an overlapping window of length of 4s and overlap of 3s, from which $62 \text{ feature/channel} \times 14 \text{ channel} = 868$ features were extracted.

Comparing with channel selection algorithms, in Table 4.4 we show that CSDF has the best accuracy for 8 out of 9 subjects. For all subjects, CSDF achieved improvement comparing to the accuracy without channel selection. For the average case, CSDF was the highest ($94.3\% \pm 5.6\%$), mRMR_SFFS_C ($91.8\% \pm 6.1\%$) was the second and SVM_RCE lowest ($91.2\% \pm 6.9\%$), although the difference was not significant.

Table 4.4: Testing accuracy from 5 channel selection algorithms and CSDF on the TOVA dataset

Subject	Channel selection algorithm						CSDF
	None	C1	C2	C3	C4	C5	
S1	91.5%	91.0%	91.0%	93.2%	93.2%	91.0%	95.7%
S2	94.9%	98.7%	100%	98.7%	99.6%	98.7%	100%
S3	94.0%	95.3%	95.3%	93.6%	93.6%	95.3%	99.1%
S4	76.1%	81.2%	82.1%	86.3%	86.3%	84.2%	88.5%
S5	80.8%	85.0%	85.0%	85.0%	85.0%	88.5%	90.2%
S6	74.8%	88.0%	87.2%	81.6%	81.6%	77.4%	83.8%
S7	98.7%	99.1%	99.1%	100%	100%	99.1%	100%
S8	81.6%	90.6%	91.0%	90.6%	90.6%	89.3%	91.9%
S9	94.9%	94.9%	94.9%	96.6%	96.6%	97.4%	99.1%
Mean	87.5%	91.5%	91.7%	91.7%	91.8%	91.2%	94.3%
SD	8.6%	5.7%	5.8%	6.0%	6.1%	6.9%	5.6%

None: no channel selection; C1: Fisher_SFS_C; C2: Fisher_SFFS_C;
C3: mRMR_SFS_C; C4: mRMR_SFFS_C; C5: SVM_RCE.

Comparing with feature selection algorithms, in Table 4.5 we show that CSDF has the best accuracy for 6 out of 9 subjects. For the average case, CSDF was the highest ($94.3\% \pm 5.6\%$), with mRMR_SFFS ($94.0\% \pm 4.4\%$) ranked second. The visualization of feature space of subject S3 before and after CSDF is shown in Figure 4.3, where 64 out of 868 features from 6 out of 14 channels were selected and accuracy improved from 94.0% to 99.1%. The feature space for other subjects share a similar pattern, where the improvement of separability is related to the improvement in classification accuracy before and after CSDF.

Next, CSDF was compared with a sequential combination of “first channel and then feature selection” in Table 4.6. The channel selection algorithm, mRMR_SFFS_C, was chosen since it had the best average testing accuracy besides CSDF (Table 4.4); for the feature selection algorithm, we used mRMR_SFFS, the one with the best average testing accuracy besides CSDF (Table 4.5). The combination was done in a sequential way, i.e. features were selected from the selected channels. For all 9 subjects, CSDF outperformed the sequential combination in terms of testing accuracy, where 6 of them were significantly different. The average testing accuracy for CSDF and sequential combina-

4.3 Results

Table 4.5: Testing accuracy from 5 feature selection algorithms and CSDF on the TOVA dataset

Subject	Feature selection algorithm						CSDF
	None	F1	F2	F3	F4	F5	
S1	91.5%	92.7%	92.7%	93.2%	94.0%	91.5%	95.7%
S2	94.9%	100%	100%	99.6%	98.7%	99.6%	100%
S3	94.0%	96.2%	96.2%	92.7%	92.7%	98.7%	99.1%
S4	76.1%	77.4%	77.4%	91.0%	91.9%	97.4%	88.5%
S5	80.8%	88.5%	88.5%	87.2%	87.2%	86.8%	90.2%
S6	74.8%	80.8%	88.0%	88.0%	87.2%	76.1%	83.8%
S7	98.7%	100%	100%	98.7%	100%	99.1%	100%
S8	81.6%	87.6%	87.6%	96.2%	96.2%	88.0%	91.9%
S9	94.9%	97.0%	98.3%	97.9%	97.9%	99.1%	99.1%
Mean	87.5%	91.1%	92.1%	93.8%	94.0%	92.9%	94.3%
SD	8.6%	7.7%	7.0%	4.3%	4.4%	7.6%	5.6%

None: no feature selection; F1: Fisher_SFS; F2: Fisher_SFFS;

F3: mRMR_SFS; F4: mRMR_SFFS; F5: SVM_RFE.

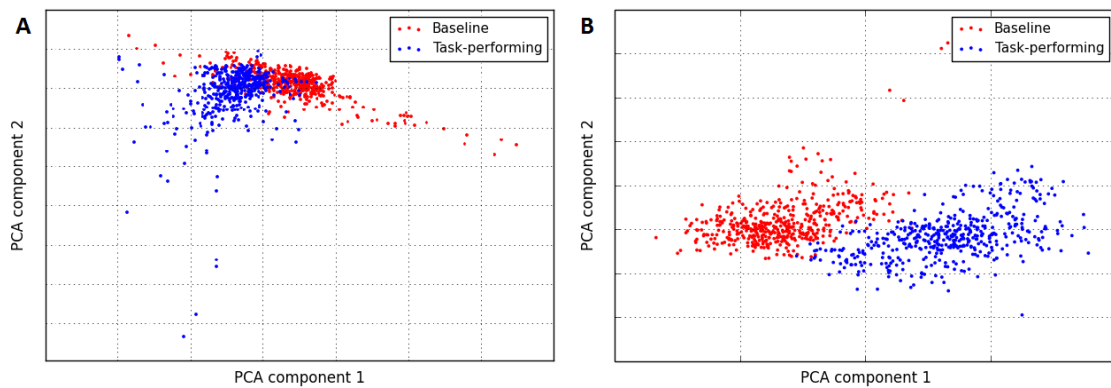


Figure 4.3: Feature space visualization of subject S3 using principal component analysis (PCA). (A) Before feature/channel selection. (B) After CSDF.

4.3 Results

tion were $94.3\% \pm 5.6\%$ and $80.8\% \pm 11.1\%$ respectively. We used McNemar’s test on paired classification results to test the null hypothesis of marginal homogeneity. It turned out that for the average accuracy, CSDF outperformed the sequential combination significantly (p-value < 0.01).

Table 4.6: Testing accuracy of a sequential combination of channel (mRMR_SFFS_C) and then feature selection (mRMR_SFFS) and CSDF on the TOVA dataset

Subject	Sequential	CSDF	significance
1	89.7%	95.7%	n.s.
2	94.0%	100%	**
3	82.9%	99.1%	**
4	73.5%	88.5%	n.s.
5	87.6%	90.2%	n.s.
6	62.8%	83.8%	**
7	93.2%	100%	**
8	64.5%	91.9%	**
9	78.6%	99.1%	**
Mean	80.8%	94.3%	**
SD	11.1%	5.6%	
n.s. p > 0.05 * p < 0.05 ** p < 0.01			

4.3.3 Comparison of Selected Features for the TOVA Experiment

In the following subsections 4.3.3 and 4.3.4, we compare the number of selected features/channels with other feature/channel selection algorithms. It should be noted that the purpose of channel/feature selection is not simply selecting as few channels/features as possible, rather, any channels which contain discriminant features should be selected. As a result, in the following 4 tables (Table 4.7, Table 4.8, Table 4.9 and Table 4.10) we shall emphasize the algorithm with least number of selected features/channels is not necessarily the best one, hence no bold font is used.

In Table 4.7, we compare the number of selected features with other channel selection algorithms. For 8 out of 9 subjects (except subject S2), CSDF achieved least feature number. For subject S2, CSDF selected second least number of features after SVM_RCE. On

average, CSDF reduced the original 868 features to 83 features, reducing about 90% of total features. In Table 4.8, we compare the number of selected features with other feature selection algorithms. CSDF selected the second lowest number of features, ranked after SVM_RFE. SVM_RFE is a greedy feature selection algorithm which requires to train a sparse linear SVM in each iteration. On the other hand, CSDF is not a greedy algorithm which guarantees the global optimum under its formulation.

Table 4.7: Number of selected features from 5 channel selection algorithms and CSDF on the TOVA dataset

Subject	Channel selection algorithm					
	C1	C2	C3	C4	C5	CSDF
S1	806	806	372	372	620	49
S2	744	682	310	372	186	283
S3	744	806	806	806	744	64
S4	248	682	186	186	496	117
S5	434	434	434	434	682	49
S6	310	372	496	496	558	61
S7	806	434	558	558	744	38
S8	682	620	434	434	682	53
S9	248	248	186	186	682	33
Mean	558	565	420	427	599	83

C1: Fisher_SFS_C; C2: Fisher_SFFS_C; C3: mRMR_SFS_C;
C4: mRMR_SFFS_C; C5: SVM_RCE

4.3.4 Comparison of Selected Channels for the TOVA Experiment

In Table 4.9 we show the number of selected channels from CSDF and other channel selection algorithms. It turned out that CSDF selected similar numbers of channels with other algorithms in terms of the average case. We also compared the number of selected channels with feature selection algorithms, as shown in Table 4.10, where SVM_RFE selected the least channels (7) and CSDF selected the second least channels (9). In contrast, the feature selection algorithms which achieved similar accuracy with CSDF in Table 4.5 — mRMR_SFS (F3, 93.8% $\pm$ 4.3%) and mRMR_SFFS (F4, 94.0% $\pm$ 4.4%) — required 12 channels. This indicates that CSDF selects features in a way that

Table 4.8: Number of selected features from 5 feature selection algorithms and CSDF on the TOVA dataset

Subject	Feature selection algorithm					CSDF
	F1	F2	F3	F4	F5	
S1	512	512	128	256	256	49
S2	256	256	256	128	2	283
S3	64	128	512	512	4	64
S4	128	128	32	32	32	117
S5	32	32	16	16	2	49
S6	32	16	128	128	4	61
S7	32	32	128	128	32	38
S8	64	64	1	1	4	53
S9	256	512	128	128	16	33
Mean	153	187	148	148	39	83

F1: Fisher_SFS; F2: Fisher_SFFS; F3: mRMR_SFS;
F4: mRMR_SFFS; F5: SVM_RFE.

the features are grouped into channels as much as possible.

4.3.5 Relevant EEG Channels and Features in Selective Attention

In this subsection, we investigate the selected channels and features in all subjects. In Table 4.11 we show the selected channels for each subject. The channels were ordered according to how many times it was selected across all subjects. It can be observed that the channels selected varied between subjects and there was no single channel that was selected for all subjects. The reason is that our algorithm is subject-dependent and data-driven, i.e. the model was automatically fitted to the most discriminating features and channels, and thus reflecting the difference in the brain dynamics of each subject.

Table 4.12 shows the features which were selected in more than half of all subjects. In total, 6 features met this requirement, i.e. first/second order difference (or line length in seizure detection literature [84]) in T7 and P7, Higuchi fractal dimension [94] in F8 and mean signal amplitude in AF4.

4.3 Results

Table 4.9: Number of selected channels from 5 channel selection algorithms and CSDF on the TOVA dataset

Subject	Channel selection algorithm					CSDF
	C1	C2	C3	C4	C5	
S1	13	13	6	6	10	12
S2	12	11	5	6	3	11
S3	12	13	13	13	12	6
S4	4	11	3	3	8	13
S5	7	7	7	7	11	12
S6	5	6	8	8	9	1
S7	13	7	9	9	12	4
S8	11	10	7	7	11	12
S9	4	4	3	3	11	7
Mean	9	9	7	7	10	9

C1: Fisher_SFS_C; C2: Fisher_SFFS_C; C3: mRMR_SFS_C;
C4: mRMR_SFFS_C; C5: SVM_RCE.

Table 4.10: Number of selected channels from 5 feature selection algorithms and CSDF on the TOVA dataset

Subject	Feature selection algorithm					CSDF
	F1	F2	F3	F4	F5	
S1	14	14	14	14	14	12
S2	14	14	14	14	2	11
S3	13	14	14	14	4	6
S4	14	14	10	10	12	13
S5	12	12	9	9	2	12
S6	11	7	14	14	3	1
S7	11	11	14	14	13	4
S8	12	12	1	1	3	12
S9	14	14	14	14	11	7
Mean	13	12	12	12	7	9

F1: Fisher_SFS; F2: Fisher_SFFS; F3: mRMR_SFS;
F4: mRMR_SFFS; F5: SVM_RFE.

4.3 Results

Table 4.11: Selected channels for each subject

channel	subject									count
	1	2	3	4	5	6	7	8	9	
O1	x	x	x	x	x			x	x	7
F8	x	x	x	x	x		x	x		7
T7	x	x	x	x	x			x		6
P7	x	x	x	x	x			x		6
O2	x	x		x	x		x		x	6
P8	x	x	x	x				x	x	6
AF4	x			x	x		x	x	x	6
AF3		x		x	x	x			x	5
F7			x	x			x	x	x	5
F3	x	x		x	x			x		5
FC5	x	x			x			x	x	5
T8	x	x		x	x			x		5
FC6	x	x		x	x			x		5
F4	x			x	x			x		4
# selected channels	12	11	6	13	12	1	4	12	7	
# selected features	49	283	64	117	49	61	38	53	33	

Table 4.12: Mostly selected features in all subjects

feature	channel	subjects									count
		1	2	3	4	5	6	7	8	9	
first order difference (line length)	T7	x	x	x	x	x					5
second order difference	T7	x	x	x	x	x					5
first order difference (line length)	P7	x	x	x		x			x		5
second order difference	P7	x	x	x		x			x		5
Higuchi fractal dimension	F8	x	x	x	x			x			5
mean amplitude	AF4	x				x	x		x	x	5

Next, we examined how frequent each feature was selected across all channels and subjects to evaluate its reliability. Figure 4.4 shows the top 10 reliable features. We see that the top 4 features were consistent with the features that were mostly selected in all subjects from Table 4.12. It also shows that all the values were much smaller than the maximum possible value $14 \text{ channels} \times 9 \text{ subjects} = 126$, reflecting large subject dependence.

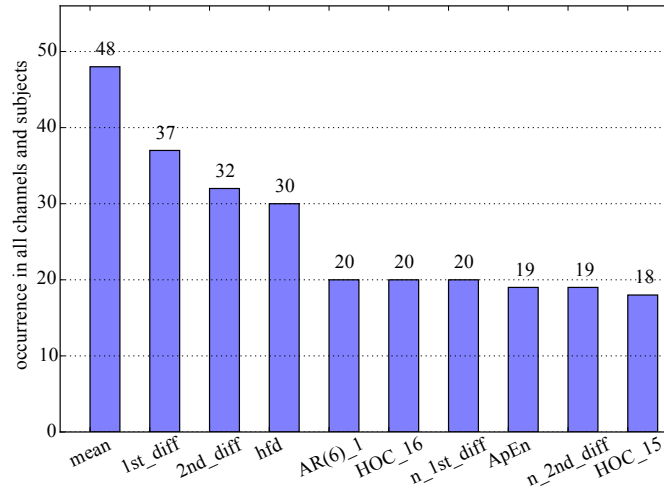


Figure 4.4: Top 10 features that are selected across all channels and all subjects

4.4 Discussion

4.4.1 Neurophysiological Interpretations of Selected Channels and Features

For the selected channels, we can compare our study to a similar study in the literature [19], where selective attention is studied in visual cognitive tasks. In Figure 4.5A we show the heat map plotted based on how many times each channel is selected in all subjects (Table 4.11) during the TOVA selective attention task. In Figure 4.5B we show the regional cerebral blood flow (rCBF) and positron emission tomography (PET) imaging of the brain activity during a perceptual matching task, where the highlighted

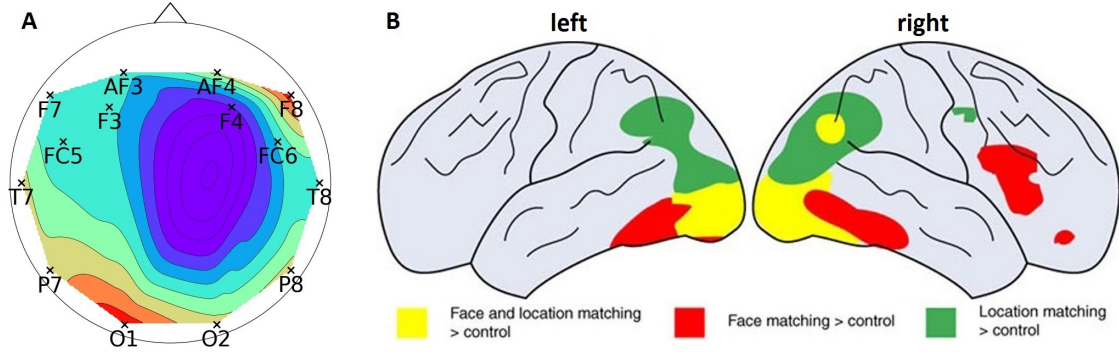


Figure 4.5: (A) The importance of each channel during the TOVA selective attention task. This is a top view of the head, where the triangle refers to the nose. (B) PET-rCBF brain imaging during selective attention to faces and locations. Adapted from [19, 95].

cortical regions are associated with selective attention [19]. Comparing Figure 4.5A and 4.5B, we observe consistent patterns. According to the “two-streams visual pathway hypothesis” [95, 96], the TOVA task in our study enhances the activity in the occipital, parietal and temporal lobes, which are involved in object recognition and motion processing. We suggest that the selected channels are consistent with neurophysiological interpretations.

Combining the selected features and channels, from Table 4.12 we know that the first and second order difference (or line length [84]) in the left parietal and temporal lobes are important in selective attended and unattended baseline states. The Higuchi fractal dimension and mean signal amplitude in the right frontal lobe are also important.

4.4.2 Limitations

CSDF, like the other algorithms compared in this chapter, is a subject-dependent algorithm where the choice of feature and channel is different for each subject (Table 4.11 and Table 4.12). Subject-independent methods are described in Chapter 5 and Chapter 6. In addition, we do not consider inter-channel features. There are evidences showing synchronization between different brain regions on different frequency bands play important roles [27, 97]. Including theses features into CSDF requires novel inter-channel-wise lasso terms in the formulation in Equation (4.10).

During the research of CSDF, there was a similar study [98] on the simultaneous feature and channel selection using logistic regression and group lasso. Our method is developed independently of this study and uses LS-LDA instead of logistic regression. Here we compare LS-LDA and logistic regression. In terms of complexity, since LS-LDA has least square formulation, the gradient is $X^\top(XW - Y)$; the gradient for logistic regression is $X^\top(Y - \text{sigmoid}(XW))$. Therefore the computation of the gradient for logistic regression has one more step — applying sigmoid function to each element of matrix XW . In terms of model assumption, LS-LDA assumes multivariate normal distribution in the feature space and similar within-class covariance; while logistic regression does not assume any distribution. Although the stronger assumption in LS-LDA could not be strictly met, models with stronger assumption generally requires 30% less number of training samples [99], which is suitable for EEG-based attention recognition.

4.5 Summary

We introduced Channel Selection with Different Features (CSDF) for the recognition of selective attention in the non-ERP approach with low SNR. CSDF selects channels with their own different feature sets, as well as restricts the useful features to as few channels as possible to allow a clearer interpretation. CSDF is based on least square LDA and group lasso, which forms a convex optimization framework where the results are guaranteed to be global optimal.

On the public mental task dataset, CSDF obtained average recognition accuracy $66.0\% \pm 10.9\%$, where the accuracy in the literature is $54.8\% \pm 6.7\%$ [79]. On the TOVA experiment, CSDF obtained similar or better recognition accuracy at $94.3\% \pm 5.6\%$, with significantly less features and comparably less channels, compared to other 5 channel selection and 5 feature selection algorithms.

The answer to the first research question in Section 1.2: “What are the relevant brain regions (in terms of EEG channels) and distinguishable EEG features in selective atten-

4.5 Summary

tion?” is, the first and second order difference (or line length [84]) in the left parietal and temporal lobes, as well as the Higuchi fractal dimension and mean signal amplitude in the right frontal lobe, are relevant to selective attention. The selected channels are consistent with a PET-rCBF neuroimaging study using ERP in the literature [19].

Chapter 5

Recognition of Sustained Attention

In the previous chapter, we aim to find relevant brain regions and distinguishable EEG features between unattended and attended conditions, in a subject-dependent way. However, for another type of attention: sustained attention (vigilance), it cannot be easily described by discrete conditions. Can we define the vigilance stages in open eye and situation-aware state in a subject-independent way, analogous to sleep stages in closed eye?

In this chapter, we first describe the experiment — simulated driving with psychomotor vigilance test (PVT), as well as the reference vigilance labelling and feature learning algorithm in Section 5.2. Then we compare the recognition performance in Section 5.3.1. Based on the results, we propose the vigilance stages in open eye and situation-aware state in Table 5.3. In Section 5.3.3, it turns out that the learned hidden weights are nonsparse and holistic-based, which will lead us to Chapter 6.

5.1 Background and Motivation

Vigilance (sustained attention) is an important brain state when performing prolonged and attention-demanding tasks such as driving. Therefore, recognizing vigilance has been an important topic in passive BCI [100–102]. Here, “vigilance” refers to the ability of human operators to maintain their focus of attention and remain alerted to stimuli over prolonged time [103]. There are two other definitions of vigilance in the literature [4]: (1) the arousal level on the sleep-wake continuum; and (2) the attention to threats or dangers. While the second definition is mostly used in animal studies, the first one is closely related but often confused with the definition of vigilance in this chapter. On the one hand, they are closely related since the vigilance in this chapter can be mapped to the wake side of the sleep-wake continuum. On the other hand, the main difference is that the vigilance in this chapter is assessed in a psychological test under open eye and situation-aware state; while the sleep-wake continuum definition of vigilance is assessed under closed eye and relaxed state.

For the sleep-wake continuum definition, there are many stage models which categorize vigilance into different stages between sleep and wakefulness. For example, VIGALL [14] defines vigilance stages from high alertness to sleep onset under closed eye. According to VIGALL, under closed eye and relaxed state, the brain regulates the vigilance which declines to lower vigilance stages with time, and finally to the sleep onset.

However, there is no vigilance stage model defined under our definition, where the subject is performing a task, i.e. in open eye and situation-aware state. In this case, the EEG signals are more complex compared to those in closed eye or sleeping such as slow waves and sleep spindles. In fact, the agreement between human raters in vigilance scoring is lower (71.8% [104]) than that in the sleep stage scoring (87.5% [105]).

Here we apply extreme learning machine autoencoder (ELMAE) [20, 67] to learn the EEG spectral features and compute vigilance levels, based on which we can define a subject-independent vigilance stage model in open eye and situation-aware state.

5.2 Method

5.2.1 Experiment Setup

The subject performs a primary task and a secondary task at the same time (Figure 5.1). The primary task is to drive in a simulator for 30 minutes in a monotonous highway without any traffic, and to keep the speed around 60 km/h. The secondary task is psychomotor vigilance task (PVT) from the Psychology Experiment Building Language [106] where a visual stimulus (a red dot) repeatedly appears at a fixed position on the screen. The appearance interval is uniformly random, with minimum 1 second and maximum 10 seconds. The subject is asked to press a button on the steering wheel as soon as the red dot appears. The red dot disappears once the subject presses a button on the steering wheel, and then the response time (RT) is recorded for further analysis (Section 2.1).

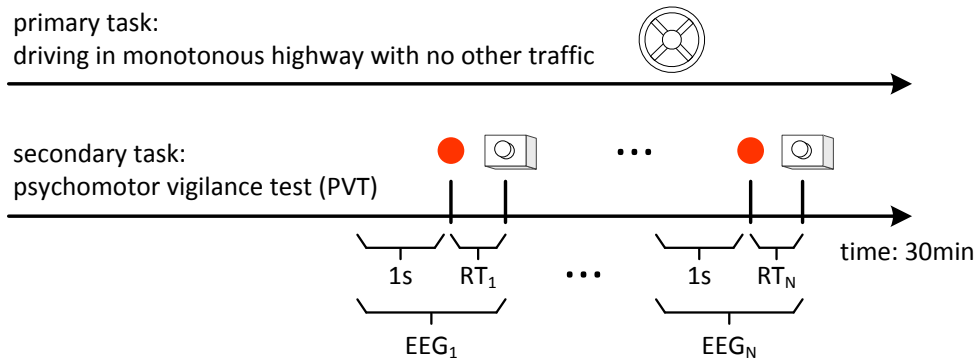


Figure 5.1: The experiment setup. See text for detailed description.

Twelve subjects are recruited (2 females and 10 males), aged from 20 to 40 (mean = 26, standard deviation = 5). All of them are informed five days before the experiment. All participants are right-handed and healthy with no history of neurological diseases. The subjects were told not to consume coffee, tea, tobacco or alcohol before and during the experiment. The EEG device is the same with Section 3.2.1.

5.2.2 Preprocessing and Labelling

We filter subjects' responses in the following way. Responses with response time (RT) less than 100ms or the button pressed before the stimulus are omitted, since they are considered as guessing or accidentally pressing the button. Responses with more than 5 seconds RT are also omitted, since they are considered to be sleep onset and the research question here is to study vigilance in open eye and situation-aware state. EEG signals from 1 second before the stimulus to the response are taken, and segmented using 1 second window and 0.75 second overlap. Segments with amplitude larger than 600uV are omitted, since they are considered to be artifacts caused by body or electrode movement. All together, 9054 spectrum samples are obtained from all 12 subjects.

For each EEG segment, the signal is first detrended, and then the spectrum is computed using multitaper power spectral density estimation [107], which is robust to the random fluctuation due to the finite observations of the EEG signal. In open eye and situation-aware state, the spectral power mainly exists in theta (4–8Hz), alpha (8–12Hz), beta (12–30Hz) and gamma (>30Hz) bands. In our case the gamma band cannot be reliably detected because (1) the sampling rate is relatively low (128Hz); and (2) in open eye and situation-aware state, there is much EMG artifacts in the gamma band [4]. Therefore we computed the spectrum from 4Hz to 30Hz.

We map RTs to probabilities of having high vigilance as probabilistic reference vigilance levels. This is done in the following four steps. First, all RTs from all subjects are fitted to an exponentially modified Gaussian distribution (ExGaussian) [108]. Second, for each response, the probability of having longer than its RT is obtained from the cumulative density function (CDF) of ExGaussian, which equates to $1 - \text{CDF}$. Third, we approximate the unnormalized vigilance levels $\overline{\text{vig}}(t)$ using Gaussian kernel densities according to Equation (5.2), so that the vigilance level at each response is the sum of Gaussian kernels with means at 4 most recent response times TT_i , standard deviation σ at two times the mean response interval, and amplitudes weighted by the probabilities from the second step. Last, the vigilance level $\text{vig}(t)$ is normalized across all responses

and all subjects to $[0, 1]$. Formally, the probabilistic reference vigilance is computed as

$$\text{vig}(t) = \frac{\overline{\text{vig}}(t) - \min_{t'} \overline{\text{vig}}(t')}{\max_{t'} \overline{\text{vig}}(t') - \min_{t'} \overline{\text{vig}}(t')}, \quad (5.1)$$

where for each subject

$$\overline{\text{vig}}(t) = \sum_{\text{TT}_i < t} \frac{P_{\text{ExGaus}}(t \geq \text{RT}_i)}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t - \text{TT}_i)^2}{2\sigma^2}\right), \quad (5.2)$$

$$\sigma = \frac{2}{\text{NT} - 1} \sum_{i=1}^{\text{NT}-1} |\text{TT}_{i+1} - \text{TT}_i|, \quad (5.3)$$

and NT is the total number of responses of that subject.

5.2.3 Machine Learning Model

The machine learning model is a single layer feed-forward neural network with channel-specific connection from input to hidden layer (Figure 5.2), which resembles the multimodal structure in Liu et al. [62]. For a channel c ($1 \leq c \leq C = 14$), there are $d = 26$ input nodes to receive the EEG spectrum from [4Hz, 5Hz, ..., 29Hz], and L hidden nodes fully connected to the input nodes with weight $W^{(c)} \in \mathbb{R}^{d \times L}$ and bias $b^{(c)} \in \mathbb{R}^{L \times 1}$. There is no connection between different channels, i.e. from input nodes of channel c_i to hidden nodes of channel c_j ($c_i \neq c_j$). $W^{(c)}$ and $b^{(c)}$ are obtained by training an ELMAE or a RBM for each channel separately.

The details of ELMAE and RBM are described in Section 2.5.2 and Section 2.5.1 respectively. For training ELMAE, 5-fold cross validation is used to find the optimal hyperparameters C and L . The regularization parameter C is selected from $[10^{-3}, 10^{-2}, 10^{-1}, 10^0, 10^1, 10^2, 10^3]$. The number of hidden nodes L for each channel is selected from $[20, 50]$. The input spectra are normalized to have zero mean and unit standard deviation across all EEG segments.

Before training RBM, b^x and b^h are initialized to 0, and W is initialized using β from

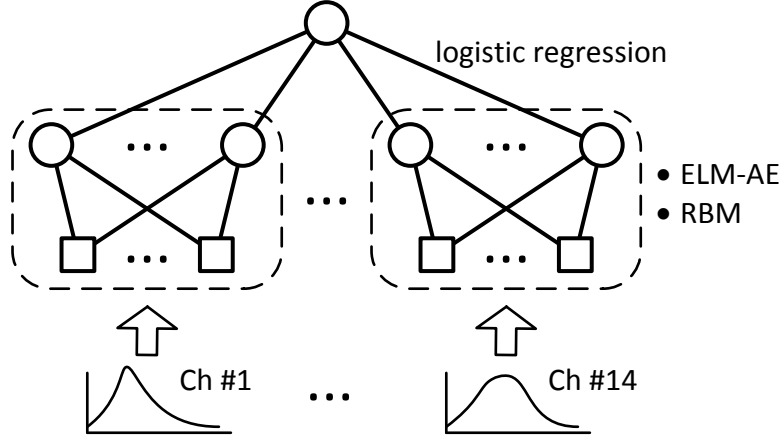


Figure 5.2: The machine learning model. See text for detailed description.

ELMAE with regularization parameter $C = 1$ to bring it closer to certain local minimum. The number of hidden nodes L for each channel is selected from $[20, 50]$. The learning rate is selected from $[10^{-2}, 10^{-1}]$. The momentum is fixed at 0.9. The weight decay is selected from $[10^{-3}, 10^{-2}, 10^{-1}, 10^0]$. The mini-batch size is fixed at 200. The maximum epoch number is fixed at 200. The input spectra are normalized to $[0, 1]$, where the 98% percentile is mapped to 1 to eliminate outliers.

All the hidden nodes from all channels are connected to a single output sigmoid node with output weight $W^{\text{out}} \in \mathbb{R}^{CL \times 1}$. The output weight is obtained by training a logistic regression classifier with the probabilistic reference labels defined in Equation (5.1), by replacing the target values from $\{0, 1\}$ to $[0, 1]$ in the log-likelihood function. This is because in case of recognizing vigilance, the vigilance labels are estimated as the probability of being high vigilance level from the response times, as in Section 5.2.2. The replacement turns out to be a proper generalization of logistic regression, where the proof is given in Appendix B.

The output of the network is the probability of having high vigilance, which is compatible with the probabilistic reference vigilance level. For both ELMAE and RBM, the regularization parameter in the logistic regression is selected from $[10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^0, 10^1]$. After training, the vigilance output is smoothed using moving average for each subject, where the window size is optimized so that the smoothed

training output has minimum root mean square error (RMSE) compared to the reference vigilance levels.

The terminology about feature learning is clarified in Table 5.1.

Table 5.1: Terminology in this chapter

Term	Meaning
hidden weights	the weights from the input nodes to the hidden nodes, which act as templates to find meaningful patterns in the EEG spectra.
learned features	the activation level of the hidden nodes (hidden activation), which are actually the inner product between EEG spectra and hidden weights, followed by the activation function.

5.3 Results

5.3.1 Performance Comparison

To ensure subject-independence, for each subject, the model is trained on all other subjects and tested on that subject for 5 repetitions with different random seeds, and the RMSE is obtained as the mean value of the 5 RMSEs between the reference vigilance and the recognized vigilance. Figure 5.3 shows the result of subject S010, who has a testing RMSE of 0.21 for ELMAE and 0.24 for RBM. Table 5.2 shows the training RMSE, testing RMSE and training time of ELMAE and RBM for all subjects. Note that the reference vigilance level is a rough estimate of the unknown ground truth. For example, since the reference vigilance is obtained as the sum of Gaussian kernels on each response, the reference vigilance decreases when there is no response near 500 seconds in Figure 5.3, which does not faithfully reflect the ground truth vigilance. Table 5.2 also includes the RMSE using 11 predefined features which are important for distinguishing different vigilance levels based on the spectrum-vigilance representation. The details of the 11 predefined features will be explained later in Table 5.4.

5.3 Results

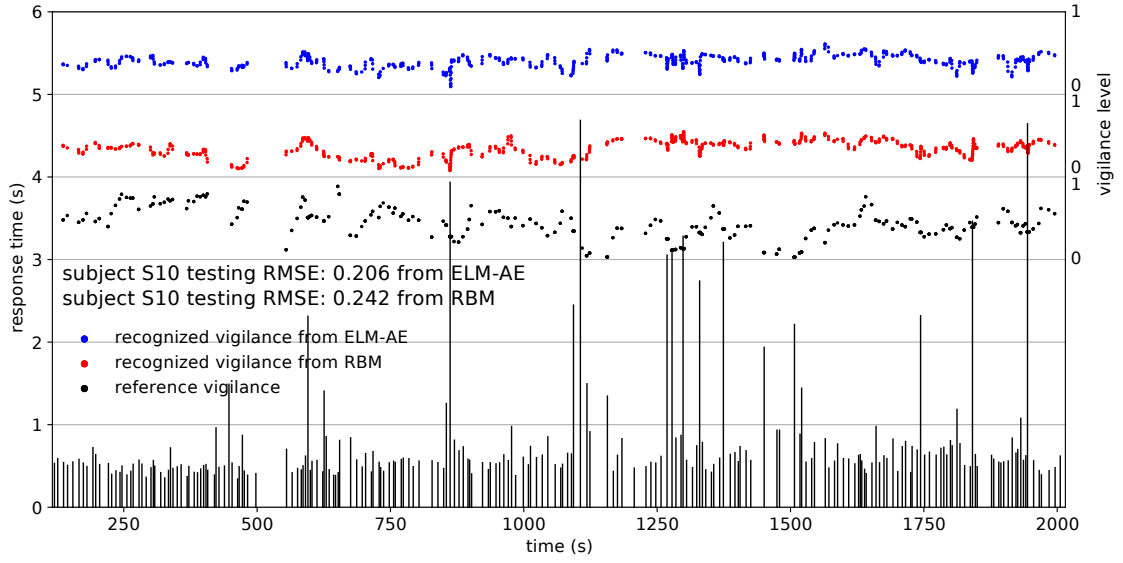


Figure 5.3: The response time, reference vigilance and recognized vigilance using ELM-AE and RBM for subject S10. The reference vigilance is computed from the response times according to Equation (5.1), in which longer response time roughly corresponds to lower reference vigilance. The displayed result is the one with the lowest testing RMSE for each algorithm over the 5 repetitions.

Table 5.2: The training RMSE, testing RMSE and training time of ELMAE, RBM and predefined features

Subject	Training RMSE			Testing RMSE			Training time (s)	
	ELMAE	RBM	feature	ELMAE	RBM	feature	ELMAE	RBM
S1	0.158	0.164	0.193	0.182	0.183	0.266	1.5	54
S2	0.161	0.167	0.199	0.134	0.138	0.168	1.5	52.3
S3	0.154	0.167	0.196	0.208	0.215	0.203	1.5	52.6
S4	0.155	0.163	0.195	0.181	0.19	0.217	1.5	53.6
S5	0.155	0.163	0.197	0.182	0.183	0.191	1.5	50.4
S6	0.158	0.164	0.198	0.184	0.193	0.187	1.5	51.5
S7	0.151	0.158	0.191	0.294	0.286	0.27	1.6	55.9
S8	0.159	0.168	0.188	0.322	0.312	0.379	1.5	48.9
S9	0.158	0.165	0.198	0.184	0.185	0.192	1.5	52.8
S10	0.155	0.162	0.197	0.216	0.249	0.187	1.6	52.1
S11	0.158	0.165	0.197	0.194	0.196	0.191	1.5	53.7
S12	0.153	0.161	0.195	0.208	0.207	0.223	1.5	54.4
Mean	0.156	0.164	0.195	0.208	0.211	0.223	1.5	52.7
Median	0.157	0.164	0.197	0.189	0.195	0.198	1.5	52.7
SD	0.003	0.003	0.003	0.049	0.046	0.056	0	1.8

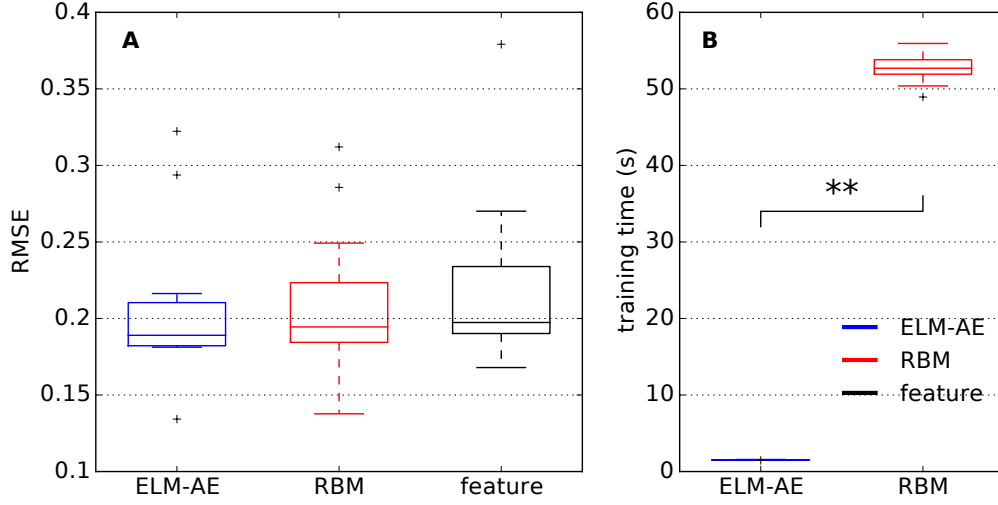


Figure 5.4: (A) The boxplot of the testing RMSE of ELMAE, RBM and predefined features for all subjects. (B) The boxplot of the training times of ELMAE and RBM for all subjects.

The comparison of the testing RMSE and training time of all subjects are presented as boxplots in Figure 5.4. For testing RMSEs of all subjects, we use Friedman test, which is a nonparametric version of repeated-measures ANOVA, since there is no assumption that the values follow normal distribution. It turns out that ELMAE has the median testing RMSE of 0.189 ± 0.049 , which is however not significantly different with RBM (0.195 ± 0.046) and predefined features (0.198 ± 0.056).

We then compare the training time of ELMAE and RBM. For both ELMAE and RBM, the training time includes the time for cross-validation and refitting the model using the best hyperparameters. To normalize the training time for different number of hyperparameter combinations, the training time is divided by the number of hyperparameter combinations for each algorithm. Each fit includes data normalization, training ELMAEs/RBMs for all 14 channels, and training the output layer. We run the training for 5 repeats with different random seeds. Finally the average training time is displayed in Table 5.2. It turns out that ELMAE needs 1.5 ± 0 seconds, which is significantly faster than 52.7 ± 1.8 seconds for RBM (Wilcoxon test p-value 0.0022).

5.3.2 Vigilance Stages

In order to define the vigilance stages, we create a spectrum-vigilance representation. First, to control the subject heterogeneity in the EEG, we apply adaptive affinity propagation clustering [109, 110] to obtain 453 clusters out of 9054 EEG spectra samples. We only take clusters with more than 3 samples, which leads to 230 valid clusters. The 9054 samples are reduced to 230 samples by only taking the centers of the 230 clusters as the respective representation. We then obtain a recognized vigilance for each cluster by taking the mean of the recognized vigilance of all spectra samples in that cluster. All the following results come from the model which is trained on all subjects and has the lowest training RMSE over 5 repetitions.

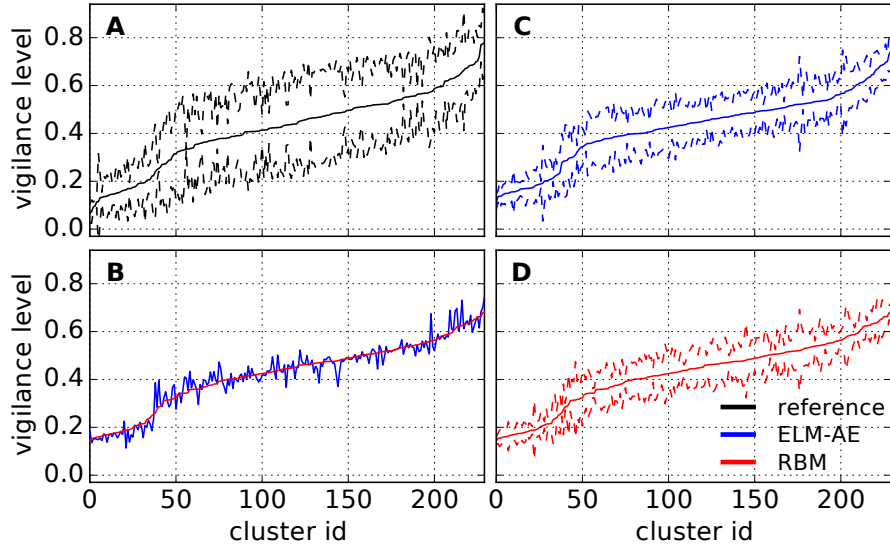


Figure 5.5: (A) The solid line is the reference vigilance of each cluster sorted in ascending order. The dashed lines indicate the standard deviation of the reference vigilance levels of each cluster. (B) The recognized vigilance of each cluster using ELM-MAE and RBM, sorted by RBM in ascending order. (C) The recognized vigilance from ELM-MAE. (D) The recognized vigilance from RBM.

Figure 5.5 shows the reference vigilance as well as the recognized vigilance from ELM-MAE and RBM of each cluster in ascending order. The dashed lines indicate the standard deviation of the vigilance levels in each cluster. Comparing the mean vigilance level in each cluster, it turns out that the reference vigilance, and the recognized vig-

ilance for ELMAE and RBM do not exhibit significant difference (Friedman test, 230 samples). On the other hand, comparing the standard deviation of the vigilance level in each cluster, the recognized vigilance from ELMAE (0.083) and RBM (0.069) are both significantly lower than that from reference vigilance (0.15) (Wilcoxon test, 230 samples, ELMAE p-value 5.2×10^{-32} , RBM p-value 5.0×10^{-37}). In summary, the recognized vigilance have consistent value with the reference vigilance in each cluster, but the standard deviation is smaller, indicating that the recognized vigilance is a *refinement* of the reference vigilance. This is because the reference vigilance is a rough estimate of the ground truth vigilance, which only contains information from task performance, i.e. the response time; while the recognized vigilance combines information from both the EEG signals and the task performance.

Having confirmed that the recognized vigilance is a refinement of the reference vigilance, we obtain the spectrum-vigilance representation by plotting the cluster center EEG spectra versus the mean recognized vigilance of clusters in Figure 5.6 for ELMAE, and Figure 5.7 for RBM, which show a similar pattern. There are several observations: (1) For vigilance between 0.25 and 0.4, there are light areas in the alpha frequency mainly in the occipital and parietal areas. (2) T8 has a clear pattern in different vigilance levels, i.e. high beta power for vigilance < 0.25 and low beta power for vigilance > 0.25 . (3) For vigilance > 0.4 , there are light areas in the low frequency area mainly in the frontal areas.

We can divide the vigilance levels into three stages as in Table 5.3. Stage I (high vigilance) involves dominant eyeball movement and blinking as reflected in the theta power in frontal area [111], indicating that the subject is attentively following the driving scenario. In the literature, Duc et al. [112] also reveals the close relationship between (visual) attention and eye movements including saccade and blinking. Stage II (low vigilance) is further divided into 2 stages. Stage II.1 has dominant alpha power in the occipital and parietal areas. This is a sign of difficulty in attention allocation [18] as well as drowsiness [4, 32], indicating that the subject cannot continue paying attention to the driving and PVT task. Stage II.2 involves muscular movements reflected by the beta

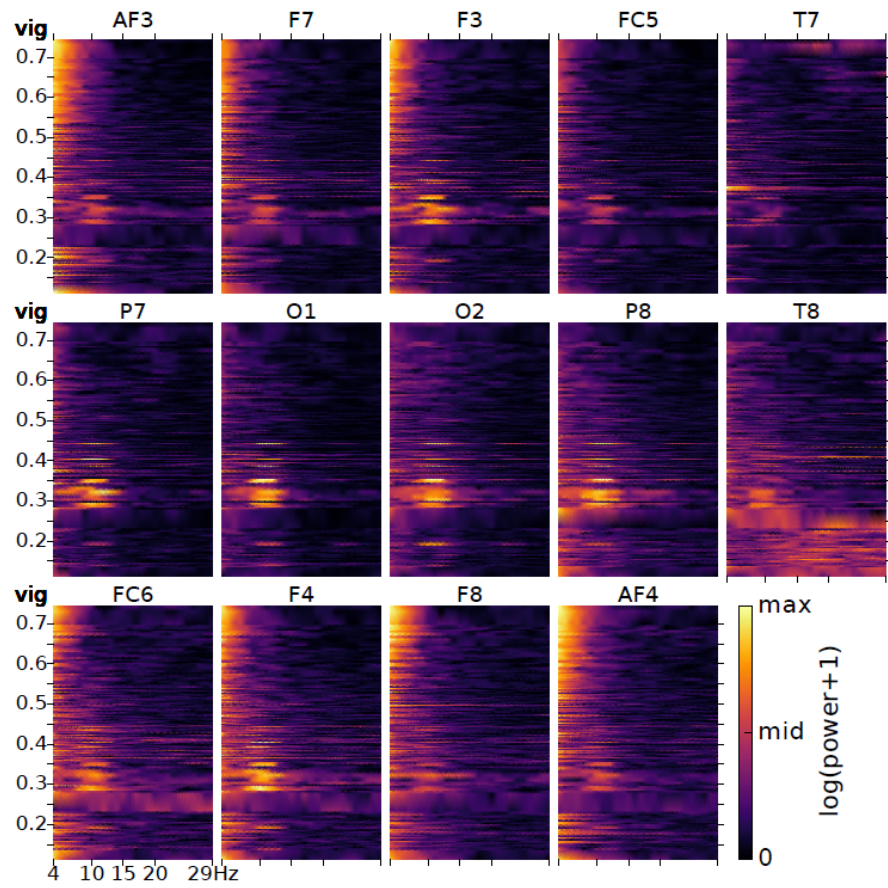


Figure 5.6: The spectrum-vigilance representation using ELMAE, i.e. the cluster center spectra of each channel versus the mean vigilance of each cluster.

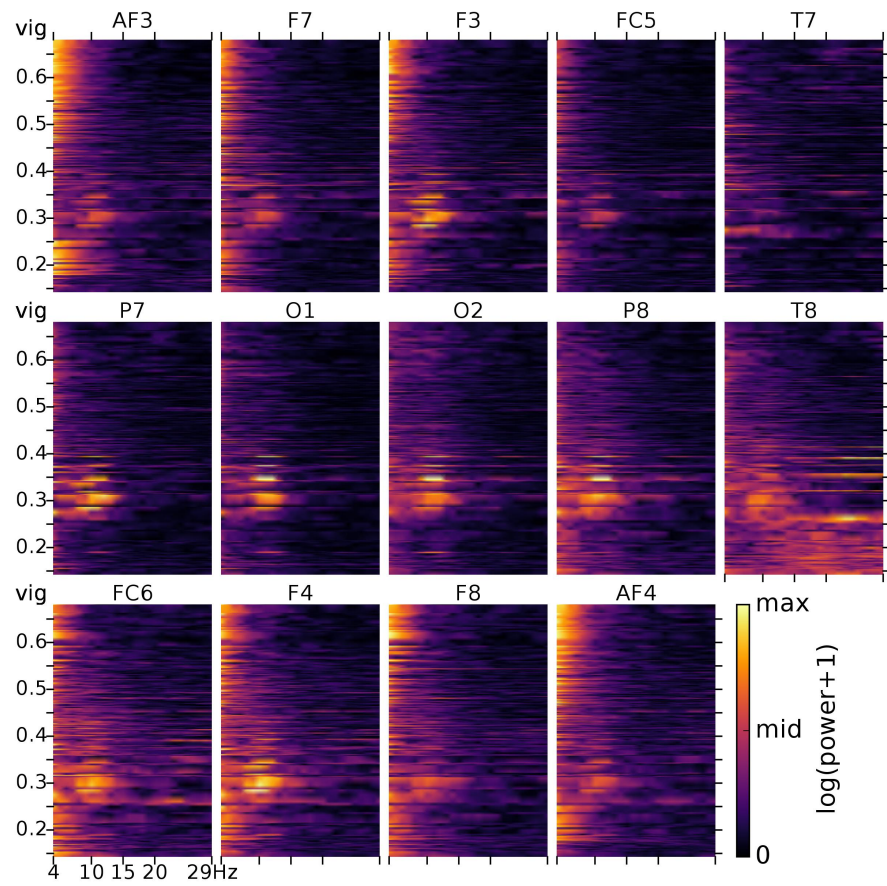


Figure 5.7: The spectrum-vigilance representation using RBM, i.e. the cluster center spectra of each channel versus the mean vigilance of each cluster.

power in right temporal regions [113] and some eye movements reflected by the theta power in right frontal regions, indicating distraction [114, 115]. It reflects the vigilance regulation to maintain the open eye and situation-aware state [116].

The reason of Stage II.2 having lower vigilance level than Stage II.1, although being not drowsy any more, is due to the definition of vigilance in this chapter, where the reference vigilance labels for training are obtained from RTs (Section 5.2.2). In Stage II.2, when the subject is distracted, the RT is longer than being drowsy, leading to lower vigilance level.

Table 5.3: Vigilance stages in open eye and situation-aware state

Stage	Vigilance	EEG Characteristics	Interpretation
I	High	Theta power (4–8Hz) in the frontal areas (AF3, AF4, F7, F8, F3 and F4)	Paying attention to driving scenario induced by eye movement [111, 112]
II.1	Low	Alpha power (8–12Hz) in the occipital and parietal areas (O1, O2, P7 and P8)	Drowsiness [4, 32], difficulty in attention allocation [18]
II.2		Beta power (12–30Hz) in T8 and theta power (4–8Hz) in FC6, F4, F8 and AF4	Distraction to maintain the open eye and situation-aware state [114, 115] induced by muscular movements [113] and some eye movement

Based on Table 5.3 we can define 11 features relevant to the vigilance stages, as shown in Table 5.4. The performance comparison with ELMAE and RBM are combined in Table 5.2 and Figure 5.4. It turns out that the testing RMSE for these extracted 11 features is the largest. However, the difference is not statistically significant. It indicates that these 11 features indeed capture some information related to EEG-vigilance, however, it does not contain more information than the feature learning approach. In addition, manually defining features requires domain knowledge, which cannot be efficiently transferred to other domains, or benefit from the availability of big data.

Table 5.4: Predefined features used in Figure 5.4 and Table 5.2 relevant to the vigilance stages

Id	Feature	Id	Feature
1	T8 beta relative power	7	F4 theta relative power
2	AF3 theta relative power	8	P7 alpha relative power
3	AF4 theta relative power	9	P8 alpha relative power
4	F7 theta relative power	10	O1 alpha relative power
5	F8 theta relative power	11	O2 alpha relative power
6	F3 theta relative power		

5.3.3 The Learned Hidden Weights and Features

Having defined the vigilance stages, here we examine the learned hidden weights and features from ELMAE, and compare to RBM. The learned features, i.e. outputs of the hidden nodes, are computed as the inner product between normalized EEG spectra and columns of hidden weights $W^{(c)}$, plus bias $b^{(c)}$, and followed by a sigmoid function (monotonic). The spectra more aligned to particular columns of $W^{(c)}$ have larger inner product at those particular hidden nodes. So the columns of hidden weights $W^{(c)}$ act as templates to find characteristic EEG spectra and to map them to vigilance levels. The positive values in $W^{(c)}$ indicate where the EEG spectra should have high power; and the negative values indicated where the EEG spectra should have low power.

We first compare the hidden weights in different vigilance stages. In Figure 5.8, Figure 5.9 and Figure 5.10, we show one example from each vigilance stage respectively, including EEG signals (panel A), spectra (panel B), and the weights from the maximally activated hidden node from ELMAE and RBM (panel C and D). Comparing the hidden weights between ELMAE and RBM, we can make several observations: (1) The hidden weights from ELMAE have more local maximums/minimums (more peaky) than RBM. This is confirmed by a comparison of the Hoyer sparseness measure (normalized ℓ_1/ℓ_2 norm ratio in range $[0, 1]$, the larger the more sparse/peaky [117]) of the hidden weights in Figure 5.11. (2) The hidden weights from ELMAE are diverse in different vigilance stages, while the hidden weights from RBM are similar in different stages. (3) There is

no blank rectangle for ELMAE while there are some blank rectangles for RBM, indicating that all channels in ELMAE have maximum hidden activation larger than 0.5 while this is not true for RBM. Therefore the learned features in ELMAE are more sparse than those in RBM. This is confirmed by a comparison of the Hoyer sparseness measure of the learned features in Figure 5.12.

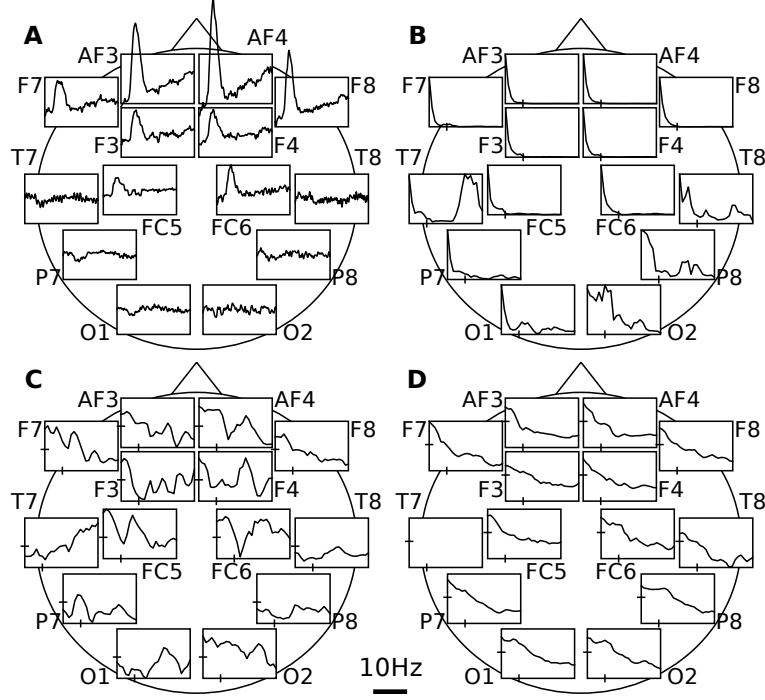


Figure 5.8: An example in stage I including EEG signals (A), spectra (B) and the weights from the maximally activated hidden node for ELMAE (C) and RBM (D). Each rectangle in (A) represents 1 second signal (128 points). Each rectangle in (B,C,D) ranges from 4Hz to 30 Hz, where the tick at the bottom indicates 10Hz. The tick at the left in the rectangles in (C) and (D) indicates 0. In (C) and (D), only the maximally activated hidden nodes with activation larger than 0.5 are shown, since this implies the hidden activation before sigmoid function is positive.

Next we compare the hidden weights associated with different output weights. The output weight W^{out} between the hidden nodes and output node reflects the importance in the recognized vigilance. Hidden nodes with positive output weights mean that their hidden weights serve as templates for high vigilance, while others for low vigilance. Hidden nodes with close to zero output weights meant that the hidden weights are irrelevant to the vigilance.

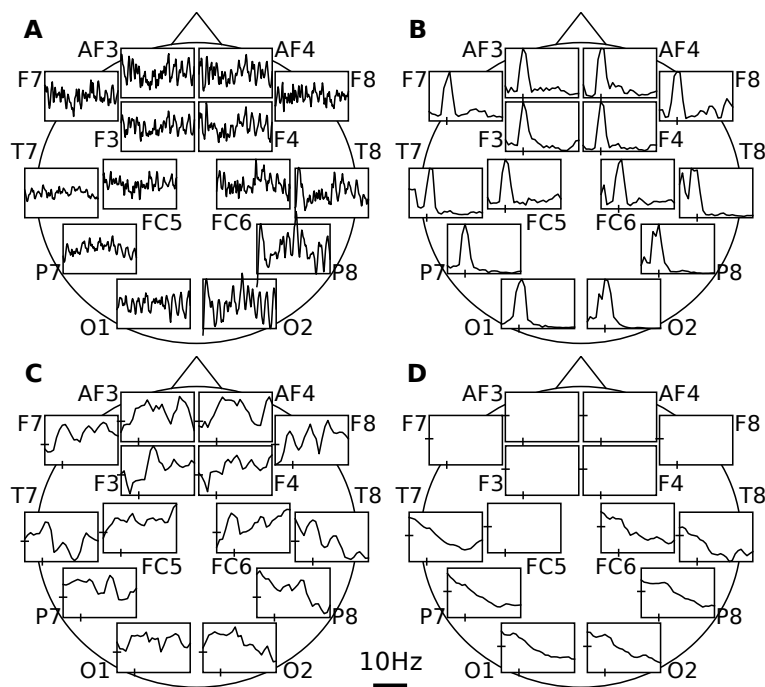


Figure 5.9: An example in stage II.1

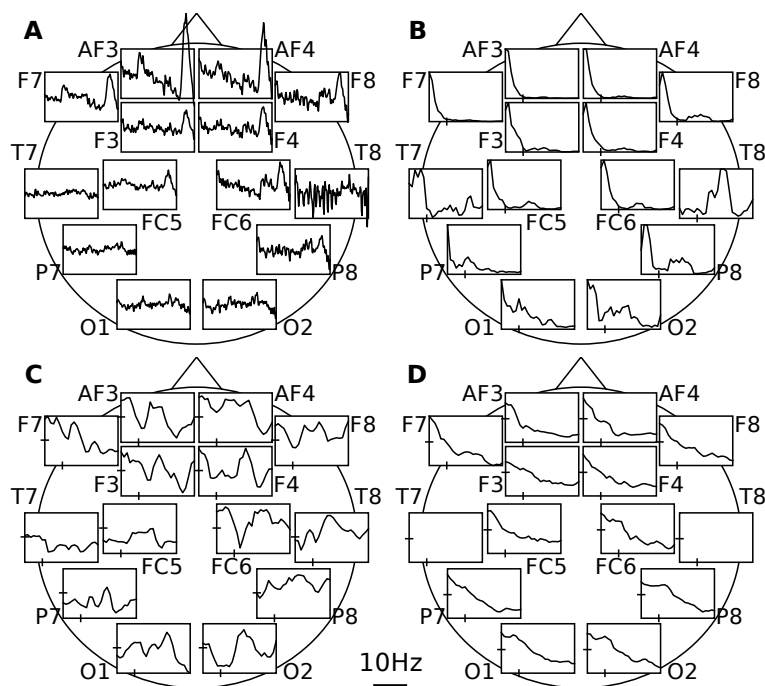


Figure 5.10: An example in stage II.2

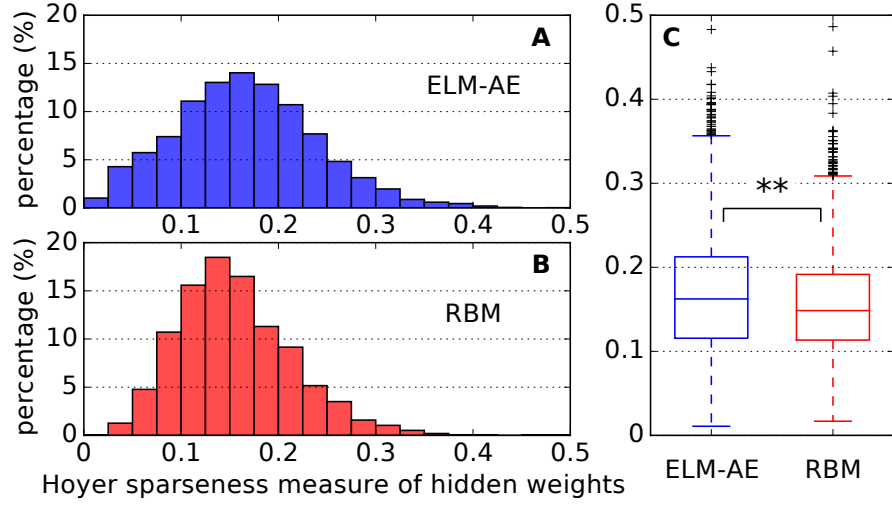


Figure 5.11: (A) The distribution of Hoyer sparseness measure (normalized ℓ_1/ℓ_2 norm ratio) of hidden weights for ELMAE. (B) Hoyer sparseness measure for RBM. (C) The comparison of the Hoyer sparseness measure for ELMAE and RBM. The median value of ELMAE is statistically larger than that of RBM (Mann-Whitney U test, p-value 4.2×10^{-11} , 3500 samples for ELMAE and 3080 samples for RBM).

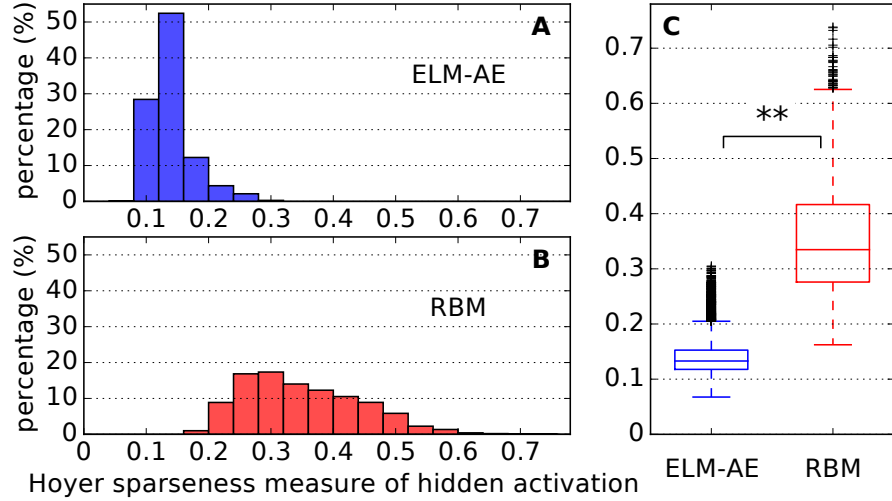


Figure 5.12: (A) The distribution of Hoyer sparseness measure of learned features for ELMAE. (B) Hoyer sparseness measure for RBM. (C) The comparison of the Hoyer sparseness measure for ELMAE and RBM. The median value of RBM is statistically larger than that of ELMAE (Mann-Whitney U test, p-value 0, 9054 samples).

In Figure 5.13 we show the hidden weights in channel T8 sorted by the output weights in descending order. As a result, the first one has the most positive output weight and is the most strongly associated with high vigilance, and the last one has the most negative output weight and is most strongly associated with low vigilance. Only hidden weights with associated large enough (positive, blue) or small enough (negative, red) output weights are shown in Figure 5.13, so that the displayed hidden weights are closely related to the recognized vigilance. The criteria of being displayed in Figure 5.13 are: (1) the corresponding output weight is larger than the mean of all positive output weights or smaller than the mean of all negative output weights, and (2) the hidden weights have at least one positive value, since a hidden node with all nonpositive weights does not learn any useful features. It turns out that for all the hidden nodes of ELMAE, 688 out of 700 hidden nodes (98.3%) have weights with at least one positive value; for RBM, 208 out of 280 hidden nodes (74.3%) have weights with at least one positive value. Therefore more percentage of hidden nodes in ELMAE learned features. On the other hand, the learned features are sparser for RBM (Figure 5.12).

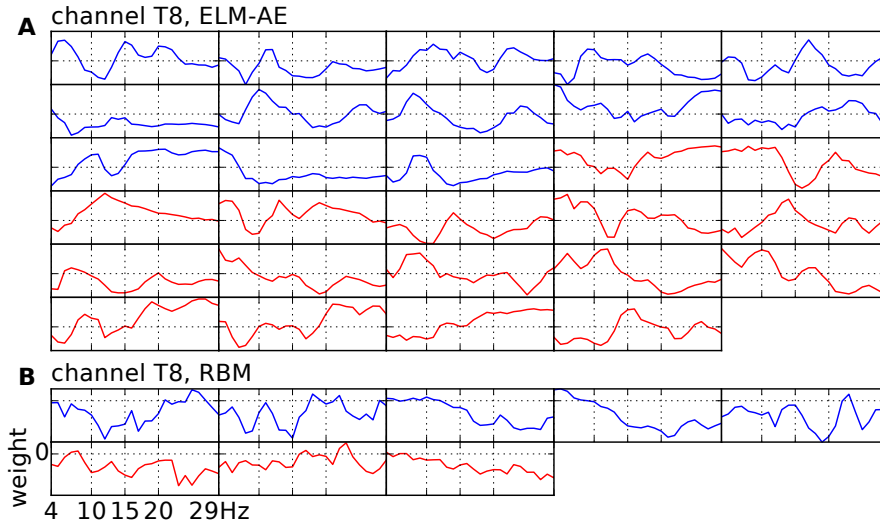


Figure 5.13: The hidden weights in channel T8. Blue means hidden nodes associated with positive output weights, i.e. they are the templates for high vigilance; while red means negative output weights, i.e. templates for low vigilance. They are sorted by their output weights in descending order, so that the first one is most strongly associated with high vigilance, and the last one is most strongly associated with low vigilance.

In Figure 5.14 we show the hidden weights of nodes associated with the most posi-

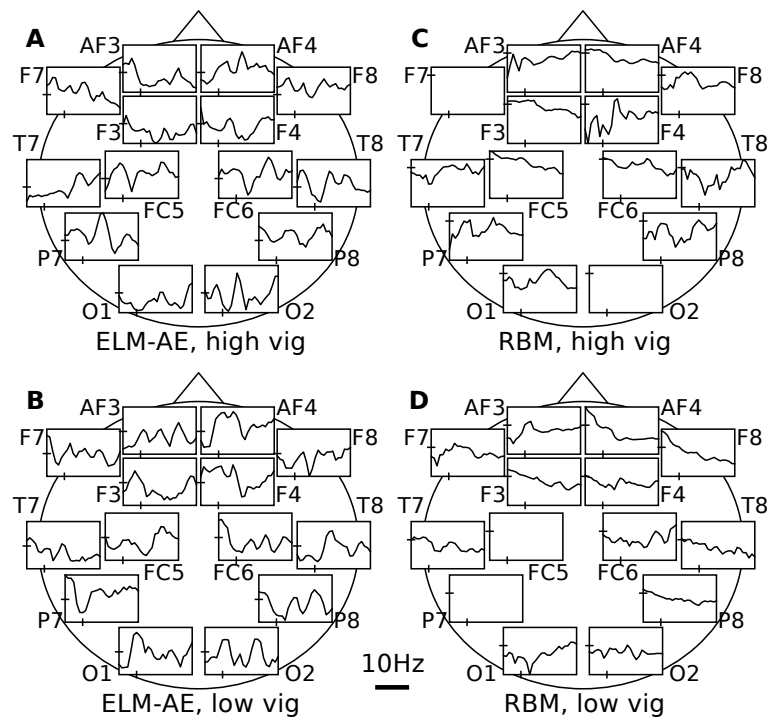


Figure 5.14: (A) The hidden weights with the most positive output weight in each channel for ELMAE. (B) The hidden weights with the most negative output weight in each channel for ELMAE. (C,D) Similar to (A,B), but for RBM. Each rectangle ranges from 4Hz to 30 Hz, where the tick at the bottom indicates 10Hz and the tick at the left indicates 0.

tive/negative output weight in all channels, representing the templates of the learned EEG spectral features for high/low vigilance. All the shown hidden weights are selected according to the criteria in Figure 5.13. In case of RBM, there are some channels with no hidden nodes satisfying the criteria (blank), since there are more percentage of hidden nodes which do not learn useful features compared to ELMAE.

5.4 Discussion

From Figure 5.13 we see that the hidden weights, i.e. the templates of the learned features, show a varying pattern, being nonsparse and holistic-based [20]. For humans it is not easy to understand the underlying information, and eventually the neurophysiological meanings. This is also the reason that we characterize the EEG features in different stages based on the spectrum-vigilance representation in Figure 5.6, rather than directly from the learned features or hidden weights. One approach to mitigate the difficulty in understanding the learned features is to introduce sparseness to the learned features and/or hidden weights, as introduced in Chapter 6.

Another limitation is that the current approach does not explicitly learn EEG waveform patterns, such as sleep spindles [118]. This is because these patterns can occur at any position in the EEG segment, while ELMAE and RBM are not shift-invariant. Advanced techniques such as downsampling or pooling should be used to enable shift-invariant feature learning.

5.5 Summary

We recognized vigilance from EEG signals using ELMAE. Experimental results show that ELMAE had median testing RMSE of 0.189 ± 0.049 , which was not significantly different compared to 0.195 ± 0.046 for RBM (Wilcoxon test p-value 0.18, 12 samples).

The training speed of ELMAE was 1.5 ± 0 seconds, which was significantly faster than 52.7 ± 1.8 seconds for RBM (Wilcoxon test p-value 0.0022, 12 samples).

We examined several properties of the learned hidden weights and features, leading to three conclusions. (1) ELMAE has more diverse hidden weights than RBM (Figure 5.8, Figure 5.9 and Figure 5.10), and more percentage of hidden nodes in ELMAE (98.3%) learn features with at least one positive value than RBM (74.3%). (2) The hidden weights from ELMAE have more local maximums/minimums (more peaky) than RBM (Figure 5.11). (3) ELMAE learns denser features than RBM (Figure 5.12). The comparison indicates that ELMAE is an efficient alternative to RBM.

We answer the second research question in Section 1.2: “What are the vigilance stages in open eye and situation-aware state?” Based on the recognized vigilance levels from ELMAE, the vigilance stages are defined in Table 5.3 in a subject-independent way. Stage I is high vigilance, where we observe eyeball movement and blinking in the frontal area, implying that the subject is attentively following the driving scenario. Stage II is low vigilance, divided into two stages. In stage II.1, we see the spreading alpha power around occipital and parietal areas, which is related to difficulty in attention allocation and drowsiness in open eye state. Stage II.2 involves muscular movements, indicating distraction. Stage II.2 is of particular interest since it is the key difference compared to the existing models [14, 15, 104, 118, 119], where in the existing models the vigilance continues decreasing after stage I and stage II.1. A possible explanation for stage II.2 is that, the environment contains not enough external stimuli to keep the open eye and situation-aware state, so that the brain performs vigilance regulation to seek external stimuli, which leads to distraction [116]. Therefore stage II.2 provides a possible mechanism of distracted driving, which is a major reason of driving accidents [120].

Chapter 6

Joint Weight-Delay Spike-Timing Dependent Plasticity for Sparse EEG Feature Learning

In the previous chapter, we define vigilance stages in open eye and situation-aware state in a subject-independent way using ELMAE. However, the learned hidden weights are nonsparse, making the interpretation difficult. Analogous to feature selection using CSDF, can we learn sparse features so that the useful information is restricted to a few nonzero features?

In this chapter, we propose joint weight-delay spike-timing dependent plasticity (joint STDP) for learning both sparse features and hidden weights. We validate joint STDP on synthetic spike trains in Section 6.3.1, and on the public mental task EEG dataset in Section 6.3.2. We apply joint STDP to recognize vigilance and present a graphical representation of the sparse hidden weights in Section 6.3.3. The similarities to biological learning are discussed in Section 6.4.

6.1 Background and Motivation

An innovative approach to do sparse feature learning is inspired by neuroscience. It is hypothesized that the spike pattern in a cell assembly can selectively activate downstream readout neurons [121]. The connection weights between cell assembly and readout neurons are learned based on spike-timing dependent plasticity (STDP, or weight STDP) [122], where the weight is updated based on the relative spike timing difference between pre- and postsynaptic neurons. It is known that if the amplitude of weight update is independent of the current weight (additive), STDP leads to bimodal weight distribution at the boundary values at convergence [21], due to a positive feedback. If the lower bound of weight is 0, the weights become sparse. The firing state of the readout neurons (learned features) is also sparse since each neuron either fires or not fires.

In fact, learning spike patterns with STDP has already been studied under discrete spike volleys [123], continuous spike trains [124] and oscillatory drive [125]. The results are consistent, i.e. some readout neurons learn to fire if a particular pattern is presented in the spike trains. However, the readout spikes only indicate presence of particular spike patterns, while there is no way to know the shape of these patterns using STDP.

In this chapter, we propose joint weight-delay spike-timing dependent plasticity (joint STDP) to learn the shape of the spike patterns, which encode EEG spectral patterns. Using joint STDP, we can obtain sparse readout weights as well as sparse learned features (firing state of the readout neurons). Joint STDP is in line with John Hopfield’s hypothesis [126], where “a set of features ... can be recognized by ... time delays ... which are organized so that the features ... produce signals arriving simultaneously at a neuron which responds maximally.” In joint STDP, the readout neurons adapt their incoming weights and delays jointly by strengthening and pulling together spikes arriving before the postsynaptic spike, and weakening and delaying others. Because of the delay plasticity, the delays become complementary to the spike pattern to maximize the arrival synchrony. In this way, the identity of neurons involved in the pattern and their ex-

pected spike timing, at millisecond scale, can be recovered from the incoming weights and delays of the readout neurons. From the spike patterns we can recover EEG spectral patterns.

Throughout this chapter, all delays are assumed to be axonal. The terminology is clarified in Table 6.1.

Table 6.1: Terminology in this chapter

Term	Meaning
STDP or weight STDP	the conventional STDP which changes synaptic weights only.
joint STDP	the proposed STDP which changes synaptic weights and delays jointly.
readout weights/delays	the weights/axonal delays from the input sources to the readout neurons.
hidden weights/delays	same with readout weights/delays.
learned (spike) patterns	the EEG spectral patterns (encoded in spike patterns) recovered from readout weights and delays according to the convergence property of joint STDP.
learned features	the firing state of the readout neurons, represented by a vector of binary values (1/0, fired or not fired).

6.2 Method

6.2.1 Neuron Model

All simulations were done using NEural Simulation Tool (NEST) [127]. The input nodes are spike sources which emit spikes at certain time. The readout neuron is modeled by Spike Response Model (SRM [128]), which is implemented as “iaf_chs_2007”

model in NEST [129]. The membrane potential $U_i(t)$ of neuron i is expressed as

$$U_i(t) = U_{\text{spike}}(t - t_i^f) + \sum_j w_{ij} U_{\text{syn}}(t - t_j^f - d_{ij}) , \quad (6.1)$$

$$U_{\text{syn}}(t) = \frac{t}{\tau_{\text{PSP}}} \exp \left(1 - \frac{t}{\tau_{\text{PSP}}} \right) , \quad (6.2)$$

$$U_{\text{spike}}(t) = \delta(t) + U_{\text{reset}} \exp \left(\frac{-t}{\tau_{\text{reset}}} \right) , \quad (6.3)$$

where $U_{\text{syn}}(t)$ is the postsynaptic potential (PSP), $U_{\text{spike}}(t)$ is the spike and the following after-hyperpolarization, t_i^f is the time of spike generated by neuron i , t_j^f is the spike time of input spikes from synapse j , and w_{ij} and d_{ij} is the weight and axonal delay from neuron j to neuron i respectively. Here w_{ij} is a dimensionless scaling factor of PSP.

6.2.2 Joint STDP

The joint weight-delay spike-timing dependent plasticity learns weights indicating which neurons are contained in the pattern, and delays complementary to the spike timing in the pattern to maximize the arrival synchrony. Like weight STDP, the learning rule is local, which only requires the information from the pair of pre- and postsynaptic neurons. It is also unsupervised, which strengthens the synapses carrying causal spikes while weakens the others without supervision.

Formally, the joint STDP is additive and uses nearest spike approximation.

$$\Delta w_{ij} = \begin{cases} w_{\text{in}} & \text{if } t = t_j^f + d_{ij} ; \\ w_{\text{out}} & \text{if } t = t_i^f ; \\ W \left(t_j^f + d_{ij} - t_i^f \right) & \text{if } t = \max \left(t_j^f + d_{ij}, t_i^f \right) ; \\ 0 & \text{otherwise,} \end{cases} \quad (6.4)$$

$$\Delta d_{ij} = \begin{cases} D \left(t_j^f + d_{ij} - t_i^f, w_{ij} \right) & \text{if } t = \max \left(t_j^f + d_{ij}, t_i^f \right) ; \\ 0 & \text{otherwise,} \end{cases} \quad (6.5)$$

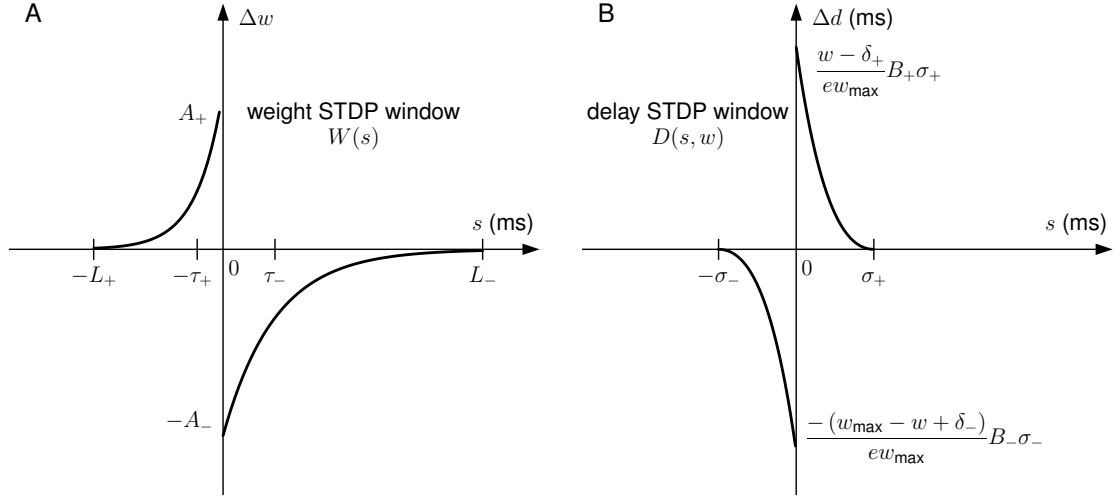


Figure 6.1: (A) The weight STDP window $W(s)$ in Equation (6.6), which is exactly the window in weight STDP. (B) The delay STDP window $D(s, w)$ in Equation (6.7). Note that the amplitudes are controlled by weight-related gains, and the effective length of the delay STDP window is different with the one of the weight window.

where $0 \leq w_{\text{in}} \ll w_{\text{max}}$ and $0 \leq -w_{\text{out}} \ll w_{\text{max}}$ are the nonhebbian homeostatic terms [130].

The learning windows in Equation (6.4) and (6.5) are (Figure 6.1)

$$W(s) = \begin{cases} A_+ \exp\left(\frac{s}{\tau_+}\right) & \text{if } -L_+ \leq s < 0; \\ -A_- \exp\left(\frac{-s}{\tau_-}\right) & \text{if } 0 \leq s \leq L_-; \\ 0 & \text{otherwise,} \end{cases} \quad (6.6)$$

$$D(s, w) = \begin{cases} \frac{-(w_{\max} - w + \delta_-)B_-}{w_{\max}} \left[\frac{\sigma_-}{e} + s \exp\left(\frac{s}{\sigma_-}\right) \right] & \text{if } -\sigma_- \leq s < 0; \\ \frac{(w - \delta_+)B_+}{w_{\max}} \left[\frac{\sigma_+}{e} - s \exp\left(\frac{-s}{\sigma_+}\right) \right] & \text{if } 0 \leq s \leq \sigma_+; \\ 0 & \text{otherwise,} \end{cases} \quad (6.7)$$

where τ_+ and τ_- are the weight learning window time constants; A_+ and A_- are the weight learning window amplitudes ($A_- \tau_- > A_+ \tau_+$); L_+ and L_- are the effective weight learning window lengths; σ_+ and σ_- are the delay learning window time con-

stants on the same scale ($\sigma_+ \approx \sigma_-$); B_+ and B_- are the delay learning window amplitudes on the same scale ($B_+ \approx B_-$); and $0 \leq \delta_- \approx -\delta_+ \ll w_{\max}$ are the weight-related gain offsets on the same scale. The weights are clipped at the boundaries $[0, w_{\max}]$, and the delays at $[d_{\min}, d_{\max}]$.

Under the dynamics, for readout neuron i which learns a pattern involving neuron set P , its incoming weights and delays arrive at the following quasi-equilibrium state, with T becoming smaller slowly with speed proportional to $\delta_-/w_{\max} \ll 1$ (quasi-equilibrium) unless $\delta_- = \delta_+ = 0$ (fixed point) (proof in Appendix C).

$$w_{ij}^* = \begin{cases} w_{\max} & \text{if } j \in Q \subseteq P ; \\ 0 & \text{if } j \notin Q, j \in P ; \\ 0 & \text{if } j \notin P , \end{cases} \quad (6.8)$$

$$d_{ij}^* = \begin{cases} T - t_j^P & \text{if } j \in Q \subseteq P ; \\ \in (T - t_j^P, d_{\max}] , \text{ indeterminate} & \text{if } j \notin Q, j \in P ; \\ d_{\min} & \text{if } j \notin P . \end{cases} \quad (6.9)$$

Here j runs over all the presynaptic neurons of neuron i ; t_j^P is the spike timing of the j -th neuron in the pattern; $Q = \{j | t_j^P + d_{ij}^* < t_i^{f*}, j \in P\} \subseteq P$ is a subset of P containing synapses with spikes arriving before t_i^{f*} , defined as the postsynaptic firing time under w_{ij}^* and d_{ij}^* ; and T is an indeterminate variable to show that the delay d_{ij}^* is complementary to the spike timing t_j^P in the pattern to maximize the arrival synchrony, i.e. $t_j^P + d_{ij}^* = T$.

6.3 Results

The parameter values of the joint STDP as well as the neuron models in this section are shown in Appendix D.

6.3.1 Learning Predefined Patterns from Synthetic Noisy Spike Trains

In the first experiment, we predefined 3 spike patterns, embedded in continuous noisy spike trains to simulate EEG features. By knowing the “ground truth”, we can validate the joint STDP by comparing the learned patterns, as well as computing hit rate and false alarm frequency of the patterns.

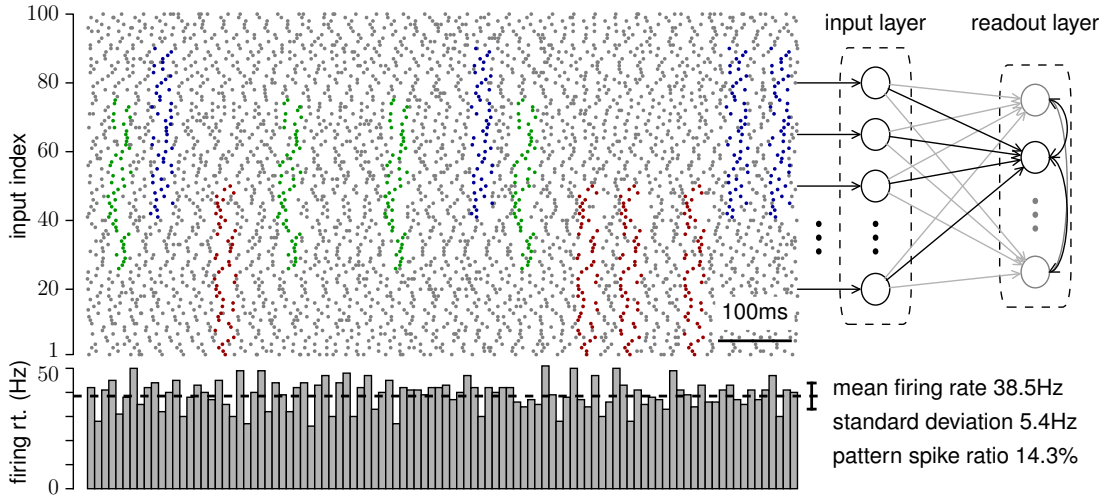


Figure 6.2: One second of the input continuous spike trains. Three patterns, indicated by red, green and blue, were embedded in the spike train, while non-pattern spikes are gray. The bottom panel shows the population-averaged firing rates over 10 ms bins. The spike trains had firing rate at 38.5Hz and standard deviation 5.4Hz. From the firing rate alone, there was nothing characteristic to indicate the patterns. So the information was encoded in the spike timing. The ratio of spikes involved in patterns was 14.3%, lower than 16.7% ($1/3$ pattern duration ratio $\times 1/2$ sources in patterns) in [124].

Figure 6.2 shows the network structure. In the readout layer, there were 10 readout neurons. There was nonplastic lateral inhibition between readout neurons with no delay. To compare the performance between the weight STDP and joint STDP, we created two networks of the same structure, but the connections from spike sources to readout neurons were subject to different STDP types. For the joint STDP, the weights were initially $w_{\max}/2$ and delays at $(d_{\min} + d_{\max}^{\text{readout}})/2$. For the weight STDP, the weights were initially $w_{\max}/2$ and the delays were constants at $d_{\min}$. The continuous spike trains were synthesized similar to [124], which had 100 spike sources and were 405s long. The details of generating the input noisy spike trains are available in Appendix E.

6.3 Results

For hyperparameter selection, the readout firing threshold θ was selected from a list $[0.1, 0.15](\times 100\text{mV})$; readout lateral inhibition weight from $[0.1, 0.15, 0.2, 0.25, 0.3](\times \theta)$; nonhebbian homeostatic term w_{in} from $[2^{-11}, 2^{-10}, 2^{-9}]$; and ratio between delay and weight learning rates B_+/A_- from $[2^{-2}, 2^{-1}, 2^0, 2^1, 2^2]$. For each combination of hyperparameter values, the pattern hit rate H and false alarm frequency F were computed over the last 60s of the simulation for all 10 readout neurons and for 3 predefined patterns. The hyperparameter combination with highest average $(H - f)$ over all predefined patterns is selected.

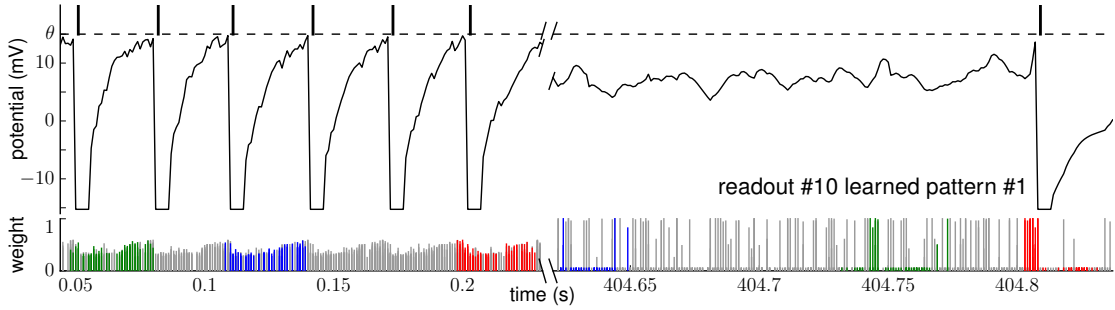


Figure 6.3: (Top) The membrane potential trace of readout neuron #10 at the beginning and end of the simulation. The top black vertical bars are the readout spikes. (Bottom) The spike arrival times $t_j^f + d_{ij}$ vs. their weights. For each half of the figure, synaptic weights measured at the timing of the last readout spike in the respective half of the figure were used. At the beginning of the simulation, the arriving spikes from each pattern (color) were indistinguishable with non-pattern spikes (gray), leading to nonselective periodical firing of the readout neuron. At the end of the simulation, for this neuron which learned pattern #1, the spike arrival times for pattern #1 (red) were pulled together due to the delay STDP, leading to postsynaptic firing. In contrast, other patterns did not trigger postsynaptic firing due to the scattered arrival times.

At the beginning of the simulation, one of the readout neurons first reached its firing threshold, where its incoming weights and delays were then updated by the joint STDP. For the weights, the update strengthened synapses with spikes arriving before the postsynaptic spike, while weakened others. For the delays, spikes arriving after the postsynaptic spike were pulled afterward to arrive even later, making it less possible to contribute to the postsynaptic spike; on the other hand, spikes arriving before the postsynaptic spike were pulled forward and closer to each other to increase the synchrony, making it more possible to contribute to the postsynaptic spike. The pulling dynamics stemmed

from the negative side of the delay learning window, which made the relative difference between spike arrival times of any synapses j and k smaller ($j \neq k, j \in Q$ and $k \in Q$, see the proof (1.2) in Appendix C).

At the end of the simulation, each pattern was optimally learned by some readout neuron(s), such as pattern #1 by readout #10 in Figure 6.3. Due to the joint STDP, the spikes from pattern #1 were pulled together and maximally strengthened. While for other patterns, their spikes were scattered. Therefore readout #10 could fire exclusively for pattern #1 (Figure 6.3 right).

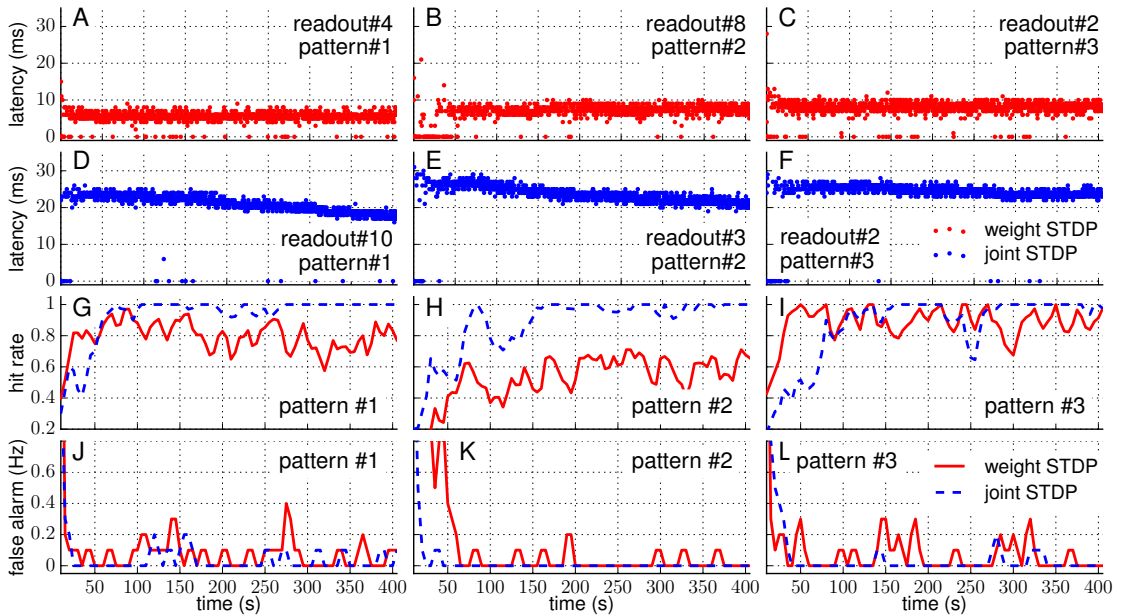


Figure 6.4: (A–C) The readout response latency of the three patterns for the weight STDP. The latency was computed with respect to the pattern occurrence times, with false alarm indicated by 0. (D–F) Similar to (A–C), but for the joint STDP. We can observe a slow decreasing trend due to the quasi-equilibrium state. The latency for the weight STDP was lower because their delays were fixed at $d_{\min}$. (G–I) The hit rates of the three patterns under the weight and joint STDP, along with time. The values were computed using a 10s sliding window with 5s overlap. (J–L) The false alarm frequencies of the three patterns.

The response latency, hit rate and false alarm frequency of the three patterns for both STDP types are shown in Figure 6.4, where the implementation details are available in Appendix F. It turned out that the joint STDP had higher pattern hit rate and lower false alarm frequency than the weight STDP for all three patterns. The reason again lay

in the delay learning part. For the weight STDP, from Figure 6.5A we can see that it actually implemented delay selection, where synapses with similar spike timing were strengthened (with $d = d_{\min}$), indicated by the horizontal lines. However, only a few synapses satisfied this condition. As a result, fewer synapses were strengthened for the weight STDP. This effect was more obvious for pattern #2 in Figure 6.5A in green. Due to fewer strengthened synapses, the firing of the readout neuron became less reliable, leading to lower pattern hit rate and higher false alarm frequency. For the joint STDP, it explicitly pulled the spikes closer to maximize the arrival synchrony. Therefore more synapses were strengthened and the readout response became more reliable.

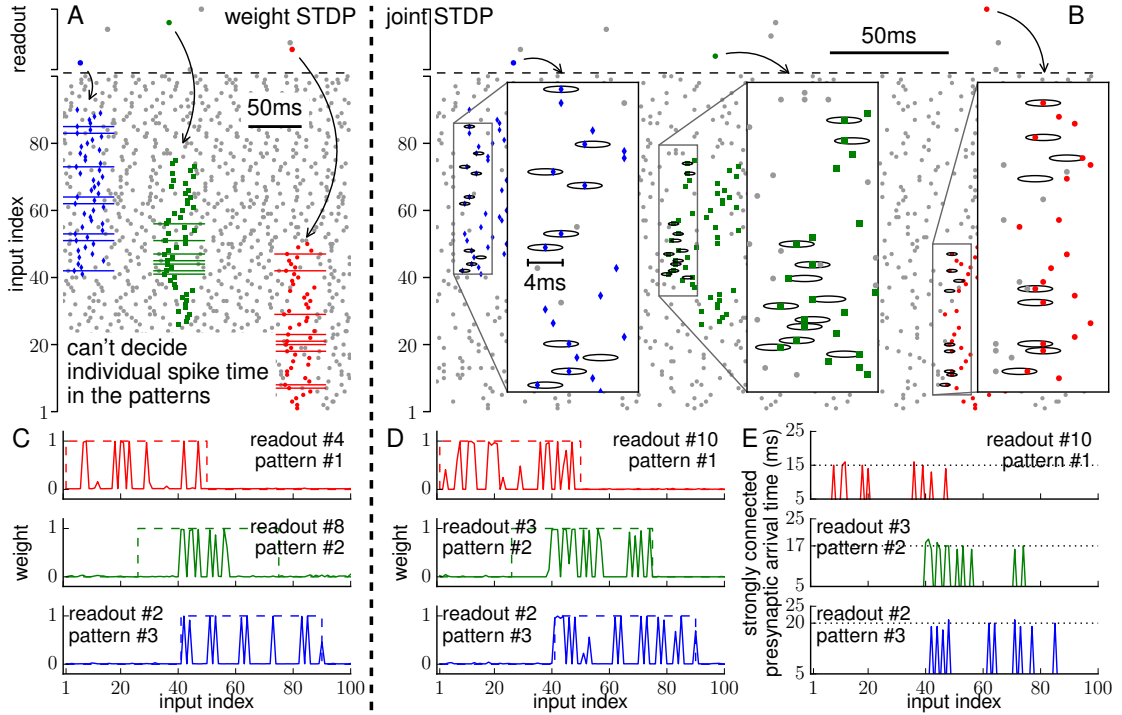


Figure 6.5: (A) Input and readout spike trains under the weight STDP. The readout spikes from the optimally learned readout neurons for each pattern are highlighted by colors and arrows. The horizontal lines in the input spike trains indicate the strengthened spike sources. (B) Input and readout spike trains under the joint STDP. The spike trains are the same with (A), but with a different time scale. The insets show the expected spike timing, represented by receptive fields with ± 2 ms precision with respect to the colored readout spikes. (C, D) The weights at the end of the simulation under the weight and joint STDP respectively. The dashed lines indicate the spike sources containing the pattern. (E) The pattern presynaptic arrival times $t_j^P + d_{ij}$ for the strengthened synapses ($w_{ij}/w_{\max} \geq \Theta = 0.95$) under the joint STDP at the end of the simulation.

The weights at the end of the simulation are shown in Figure 6.5C and 6.5D. Due to the additive rule, the weights became bimodally distributed for both weight STDP and joint STDP, exhibiting the sparse property. We can identify which spike sources were contained in the pattern based on the sparseness pattern. The spike arrival times relative to each pattern occurrence under the joint STDP are shown in Figure 6.5E. Since the delays were complementary to the spike timing in the pattern, their arrival times should be the same ($t_j^P + d_{ij} = T$), up to the spike jitter.

Using the joint STDP, we recovered the identity of neurons in the pattern and their expected spike timing in the patterns, at millisecond scale, from the incoming weights and delays of the readout neurons, where the spike timing is unavailable in the weight STDP. The learned patterns are represented using receptive fields in Figure 6.5B.

6.3.2 Validation of Joint STDP on Public Mental Task Dataset

Having confirmed the improved capability of joint STDP on the synthetic noisy spike trains than weight STDP, we validate joint STDP on the public EEG dataset — mental task dataset [74, 79–81] — for performance comparison. The details of the mental task dataset are described in Section 2.8.

To be comparable to the works in previous chapters, for the mental task dataset, we extracted the same set of features from the EEG signals as in Section 4.2.5, and then selected features according to CSDF. The feature learning approach will be described in the next section.

To use joint STDP, the features were encoded into spikes using phase encoding [125, 126]. Phase encoding assumes that the encoding (input) neuron receives two sources: (1) an oscillatory current input, generated by background activity or intrinsic cellular function; and (2) the input related to the value to be encoded. From Figure 6.6 we see that neuron encoding larger value fires earlier relative to the maximum point of the oscillation. In our case, we first used features selected by CSDF. Then the spike time

for feature value x was encoded as a spike at $t = T - f(\text{sign}(x) \log(|x| + 1))$ [126], where function $f(\cdot)$ normalized the value into $[-35, 35]$ ms. We used 10Hz background oscillation (100ms interval between samples), which was long enough compared to τ_{PSP} , ensuring no inter-sample interaction.

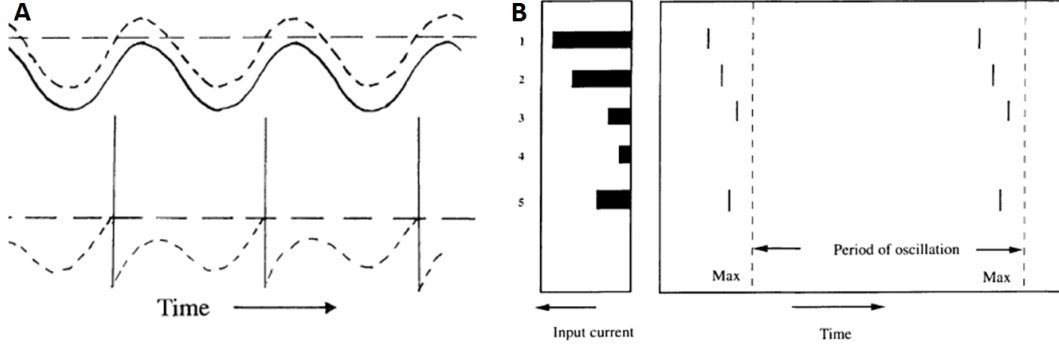


Figure 6.6: (A) The encoding (input) neuron receives an oscillatory current and the input positively related to the value to be encoded, resulting in periodic firing. (B) An illustrative example showing larger/smaller input is encoded as earlier/later spike relative to the oscillatory reference time. Both panels are modified from Hopfield et al. [126].

The training was split into two stages, as shown in Figure 6.7. In the first stage, we trained a network model with 100 readout neurons for each of the five mental tasks using all samples belonging to that mental task. The connection from input to readout neurons were subject to the joint STDP. The connection between readout neurons is called “lateral inhibition”, which connects all pairs of readout neurons and is inhibitory without transmission delay. In this stage, the training is unsupervised, where the features are learned to represent the input.

In the second stage, after training all mental tasks, the readout neurons are combined to generate combined learned features. The joint STDP is turned off, and the firing state of the readout neurons are trained using naive Bayes classifier. In this stage, the training is supervised, where the output weights are modified based on the mental tasks. Figure 6.8 shows the distribution of the weights and delays of connections from all mental tasks at the end of training. As expected, after training, the weights exhibit bimodal distribution at $w_{\min} = 0$ and $w_{\max} = 1$, hence being sparse. The delays are concentrated at two value ranges, where the delays around 18ms are those do not have much departure from the

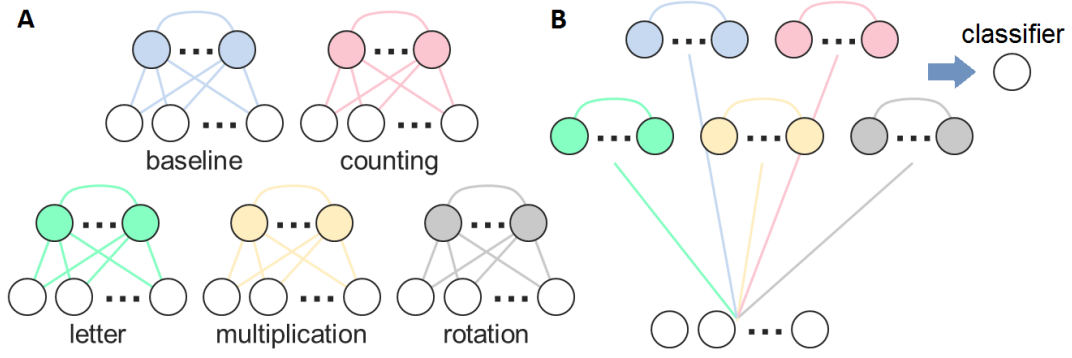


Figure 6.7: Two stages in training using the joint STDP. (A) Stage 1. The filled circles are the readout neurons. The unfilled circles are the input neurons which receive encoded features. There are lateral inhibition connections between readout neurons. (B) Stage 2: The readout neurons are combined and joint STDP is turned off. In this stage, we only train the output weights from readout neurons to the output weights.

original delay value, while the delays around 5ms are those shortened due to the spike activity.

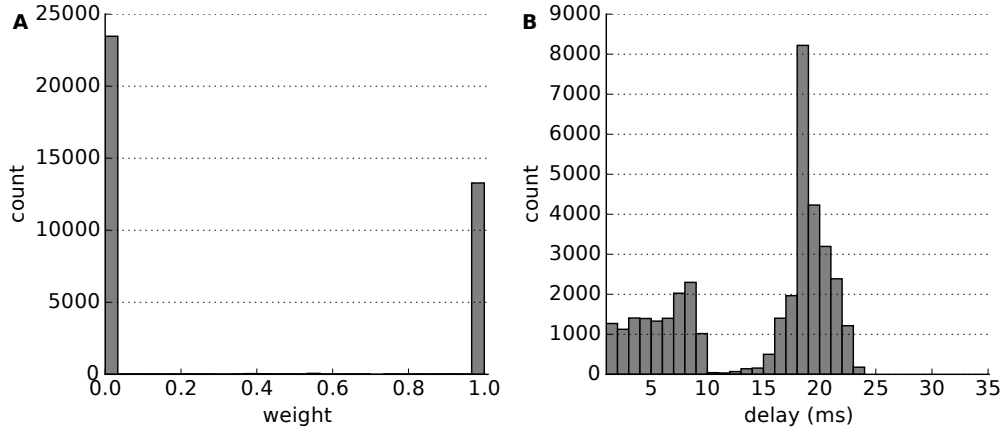


Figure 6.8: The distribution of the weights and delays of all connections from all mental tasks

The joint STDP using features selected by CSDF in Chapter 4 was compared to (1) ELM using features selected by CSDF; (2) sparse RBM [16] using features selected by CSDF; (3) sparse autoencoder [17] using features selected by CSDF; and (4) NeuCube [131] based on raw EEG signals. The reason of comparing to ELM is that ELM serves as a baseline algorithm in traditional machine learning. The reason of comparing to sparse RBM and sparse autoencoder is that they represent algorithms in the artificial neural network domain which share sparse-inducing ability with joint STDP. In these algorithms,

sparseness is achieved as encourage low average activation in the hidden layer, indicating low firing rate, and hence sparse spikes. In joint STDP, the sparseness is achieved by additive rule as well as lateral inhibition to inhibit overwhelming firing response. The reason of comparing to NeuCube is that NeuCube is a spiking neural network architecture specialized for EEG data. NeuCube takes raw EEG signals as the input, which are then fed into a reservoir of spiking neurons. The evolution of the connections between spiking neurons is driven by the input, so that after training, one can investigate the connectivity between spiking neurons to help understand the functional pathways in the EEG data [131].

Table 6.2: Testing accuracy on the mental task dataset

Subject_Session	Algorithm				
	ELM	S-RBM	S-AE	NeuCube	Joint STDP
S1_1	72.2%	69.2%	70.8%	43.2%	73.8%
S1_2	79.1%	72.5%	60.5%	43.8%	73.8%
S2_1	70.4%	74.3%	58.2%	41.6%	66.4%
S3_2	60.1%	76.9%	73.9%	24.9%	44.2%
S5_1	59.5%	60.3%	62.9%	37.8%	68.4%
S5_2	41.6%	45.5%	45.5%	30.8%	23.8%
S6_1	74.8%	70.4%	74.0%	32.4%	66.0%
S6_2	61.9%	60.5%	55.6%	28.1%	43.4%
S7_1	74.4%	73.5%	68.6%	47.0%	67.1%
Mean	66.0%	67.0%	63.3%	36.6%	58.5%
Median	70.4%	70.4%	62.9%	37.8%	66.4%
SD	10.9%	9.4%	9.0%	7.4%	16.3%

S-RBM: sparse RBM; S-AE: sparse autoencoder.

In Table 6.2 we show the comparison of the testing accuracy in mental task recognition using these algorithms. Kruskal-Wallis one-way analysis of variance (KW-ANOVA, nonparametric version of ANOVA) suggested that NeuCube performed significantly lower than other algorithms ($p\text{-value } 3.7 \times 10^{-14}$). The performance of NeuCube could be improved by (1) applying cross-validation on its hyperparameters. In this experiment, we did not cross-validate NeuCube since the training time is significantly longer than other algorithms; and (2) NeuCube encodes the raw EEG signals as input, which have low signal-to-noise ratio compared to the features selected by CSDF used in other algorithms. On the other hand, KW-ANOVA on the rest 4 algorithms revealed no significant

difference (p-value 0.58). In this case, joint STDP achieved similar testing accuracy with ELM, sparse RBM and sparse autoencoder.

6.3.3 Learning Sparse Features from the Vigilance Experiment

Having validated the performance of joint STDP on the public mental task dataset, we proceed to apply joint STDP to recognize vigilance in a subject-independent way, to check the recognition performance as well as the readout weights and learned features.

The network structure is shown in Figure 6.9. For each channel, we constructed a network with input and readout neurons. The connections from input to readout neurons were subject to joint STDP. There were lateral inhibition connections between readout neurons in each channel, which was nonplastic and had no transmission delay. There were output connections from readout neurons in all channels to one single output node. The output weights were obtained by training a naive Bayes classifier. When used for recognizing vigilance, the discrete class labels $y \in \{1, \dots, K\}$ were replaced by probabilistic labels $y \in [0, 1]^K$. This is because in our case, the vigilance labels are estimated as the probability of being high vigilance level from the response times, as in Section 5.2.2. The replacement turns out to be a proper generalization of naive Bayes, where the proof is given in Appendix B.

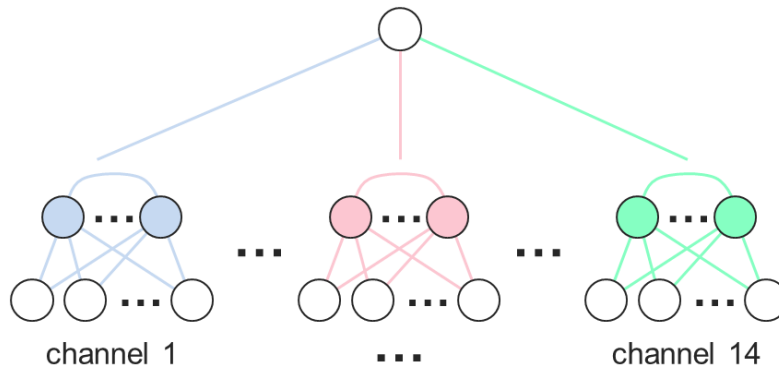


Figure 6.9: The network structure for recognizing vigilance using joint STDP. The filled circles are the readout neurons. The unfilled circles are the input neurons which receive encoded features. There are lateral inhibition connections between readout neurons in each channel. The output weights are trained using a naive Bayes classifier.

The input to the network was the EEG spectral power at $[4, 5, \dots, 29]$ Hz from each of the 14 EEG channels. For the joint STDP algorithm, they were phase-encoded into spikes (Figure 6.6), where larger power value indicates earlier spikes and vice versa. For other algorithms, the EEG spectral power values are scaled to $[0, 1]$. Five-fold cross-validation was used to tune the hyperparameters. For joint STDP, we selected the amplitude ratio between weight STDP and delay STDP $B_{\pm}/A_{-}$ from $[4, 8]$; the amplitude ratio A_{+}/A_{-} from $[0.6, 0.8]$; and the readout firing threshold from $[3, 5]$ mV.

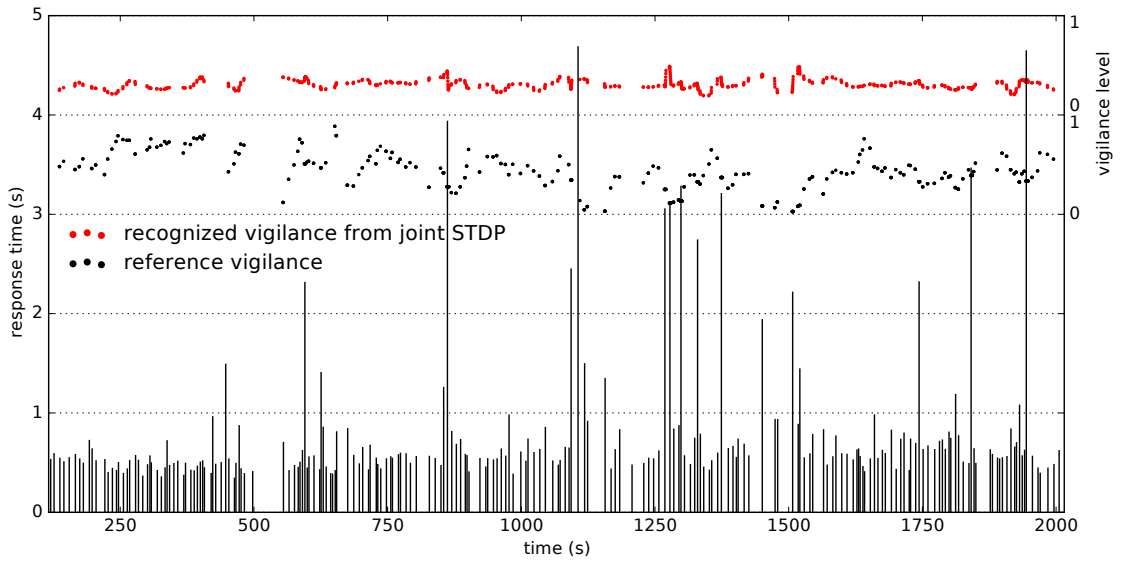


Figure 6.10: The response time, reference vigilance and recognized vigilance using joint STDP for subject S10. The reference vigilance is computed from the response times according to Equation (5.1), in which longer response time approximately corresponds to lower reference vigilance. The displayed result is the one with the lowest testing RMSE for each algorithm over the 5 repetitions.

In Figure 6.10 we show the vigilance recognition result for one subject using joint STDP. It is observed that in trials with significantly longer response times around 900s, 1100s and 1900s, the recognized vigilance drops. In Table 6.3 we show the testing root mean square error (RMSE) for all subjects. All algorithms had similar performance with no significant difference, confirmed by Kruskal-Wallis H-test at p-value 0.34. ELMAE had slightly lower average RMSE, while joint STDP had slightly higher average RMSE.

Comparing sparse algorithms (sparse RBM, sparse AE and Joint STDP) to nonsparse algorithms (ELMAE, RBM and AE), the sparse algorithms had slightly worse perfor-

mance. The loss in performance is due to the sparseness, which learned features with major contribution to vigilance, while ignoring the minor ones. In this way, the sparse algorithms can have a clearer interpretation of the hidden weights and learned features, in the price of slight performance loss.

Table 6.3: Testing RMSE for vigilance recognition

Subject	Algorithm					
	ELMAE	RBM	AE	S-RBM	S-AE	Joint STDP
S1	0.182	0.183	0.184	0.215	0.195	0.219
S2	0.134	0.138	0.136	0.153	0.145	0.137
S3	0.208	0.215	0.215	0.219	0.216	0.196
S4	0.181	0.19	0.193	0.193	0.198	0.207
S5	0.182	0.183	0.182	0.183	0.180	0.187
S6	0.184	0.193	0.188	0.220	0.187	0.181
S7	0.294	0.286	0.297	0.283	0.298	0.298
S8	0.322	0.312	0.329	0.334	0.304	0.385
S9	0.184	0.185	0.178	0.200	0.193	0.187
S10	0.216	0.249	0.23	0.207	0.203	0.223
S11	0.194	0.196	0.201	0.211	0.203	0.227
S12	0.208	0.207	0.209	0.207	0.217	0.206
Mean	0.208	0.211	0.212	0.219	0.212	0.221
Median	0.189	0.195	0.197	0.209	0.200	0.206
SD	0.049	0.046	0.051	0.045	0.044	0.061

S-RBM: sparse RBM; S-AE: sparse autoencoder.

In Figure 6.11 we plot the distribution of the hidden weights and delays at the end of training. Similar to Figure 6.8, the weights exhibited bimodal distribution at $w_{\min} = 0$ and $w_{\max} = 1$ due to the additive STDP rule. For the delay, the connections with shorter delay than the initial delay $d_{\text{init}} = 18\text{ms}$ represent input containing patterns. They were shortened because the input spikes contributed to the readout spike, hence were earlier than the readout spike, falling into the negative branch of the delay STDP window (Figure 6.1B). On the other hand, the connections with long delay than the initial delay values correspond to inputs with pattern arriving later than other patterns (smaller feature values), or not involved in any pattern. In either case, the input spikes arrived after readout spike, hence increasing the delay value according to the positive branch of the delay STDP (Figure 6.1B).

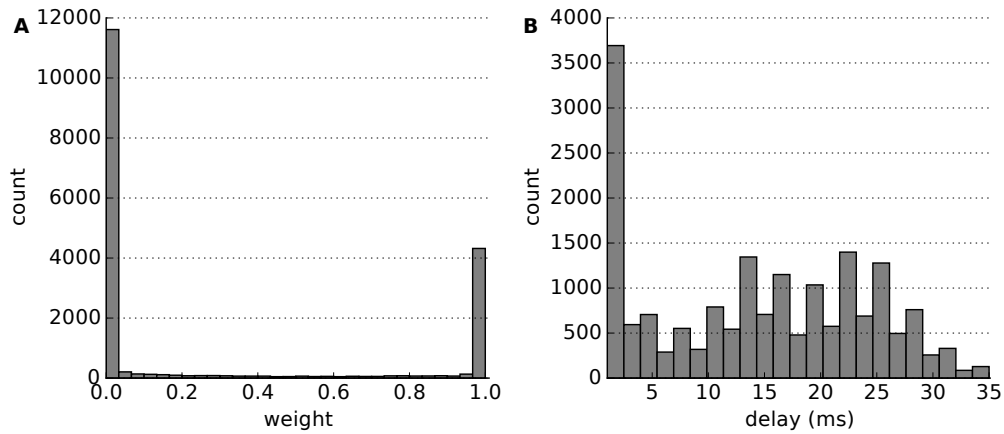


Figure 6.11: The distribution of the weights and delays of all connections trained on the vigilance dataset

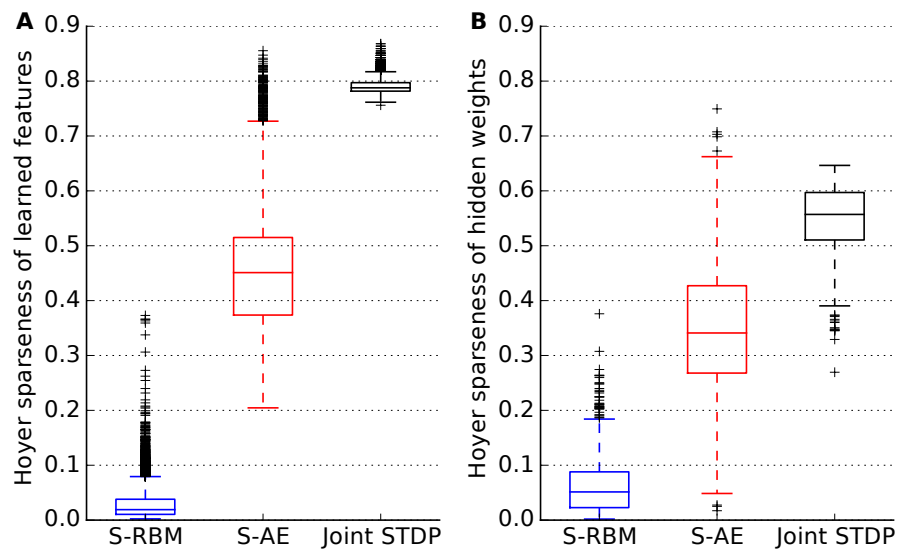


Figure 6.12: (A) The comparison of Hoyer sparseness for the learned features among sparse RBM, sparse autoencoder and joint STDP. The difference is significant between all pairs of algorithms, confirmed by Kruskal-Wallis ANOVA with Dunn's multiple comparison test. (B) The comparison of Hoyer sparseness for the hidden weights. The difference is still significant between all pairs of algorithms

We next examine the sparseness among the three sparse algorithms including joint STDP. Similar to Figure 5.11 and Figure 5.12, we computed Hoyer’s sparseness measure [117]. The comparison was done from two aspects: sparseness of the learned features and sparseness of the hidden weights. In Figure 6.12A we show the Hoyer’s sparseness of the learned features, where joint STDP had significantly higher sparseness than the other two algorithms, confirmed by Kruskal-Wallis ANOVA with Dunn’s multiple comparison test. This is because for joint STDP, the learned features are represented by readout spikes, which has binary states: fired or not fired. While for sparse RBM and sparse autoencoder, the learned features are represented by the activation level of the readout neurons, analogous to the firing rate. In Figure 6.12B we show the hidden weights, where joint STDP again had significantly higher Hoyer’s sparseness than the other two algorithms. This is a result of the additive STDP rule, as mentioned earlier in this section, which leads to bimodal distribution of the hidden weights, hence more sparse than other two algorithms. In contrast, sparse RBM and sparse autonecoder do not have explicit regularization on the sparseness of the hidden weights.

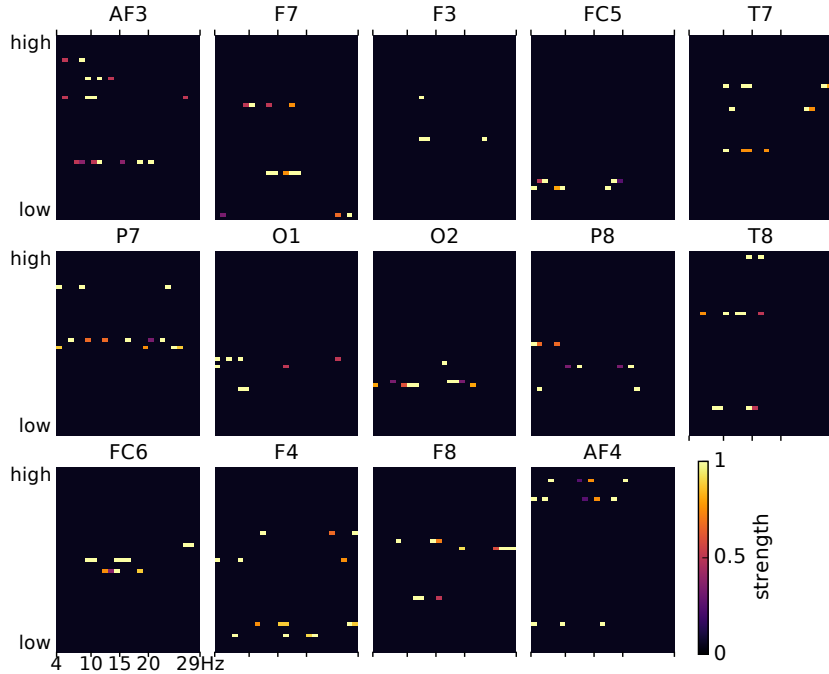


Figure 6.13: The hidden weight vs. its contribution to high vigilance, obtained using joint STDP. Each row in each channel represents the hidden weights from one readout neuron i . The rows are sorted according to $\pi_{i,2} - \pi_{i,1}$ in the output naive Bayes classifier.

Finally we present the hidden weights learned by joint STDP for vigilance recognition. Figure 6.13 is analogue to Figure 5.6, where in that figure we plot the spectrum versus the vigilance level, after clustering the spectra and computing the average vigilance level in each cluster. However, Figure 5.6 does not directly tell how the hidden weights and learned features are related to different vigilance levels, because the learned features and hidden weights do not have a sparse pattern, as described in Section 5.4. Here, using joint STDP, we are able to directly plot hidden weights due to the sparseness. In Figure 6.13 we plot the hidden weights of the readout neurons from each channel. Each row in each channel represents the hidden weights of a readout neuron i . The rows are sorted by $\pi_{i,2} - \pi_{i,1}$ according to the output naive Bayes classifier (Appendix B). As a result, the vertical axis in Figure 6.13 is not vigilance level itself, but represents the importance / degree of which the readout neuron contributes to high vigilance.

There are many details in Figure 6.13. In Table 6.4 we conclude the important characteristics in the hidden weights, as well as compare to the vigilance stages derived in Table 5.3.

Table 6.4: Vigilance Stages in Open Eye and Situation-Aware State using Joint STDP

Output weight	Hidden weights characteristics	Consistent part with Table 5.3
high	majority in frontal regions (AF3 and AF4), some in temporal and parietal regions (T7, T8, P7)	activity in the AF3 and AF4 are consistent with eye movement in stage I
medium	non-zero values in the occipital and parietal regions (O1, O2, P7, P8)	it is consistent that the occipital and parietal regions are best at recognizing vigilance stage II.1
low	non-zero values in FC5, T8, F4 and AF4	T8, F4 and AF4 are consistent with the muscular and eye movements in distraction in stage II.2

6.4 Discussion

6.4.1 A Bio-inspired Feature Learning Algorithm

The sparse EEG feature learning using the proposed joint STDP is inspired by spike-timing dependent plasticity, a type of neuronal plasticity observed in biological experiments [122]. Therefore, joint STDP shares several properties with biological learning: (1) spike-based computation, so that information is encoded in spike events instead of the firing rate; (2) the change in the weights and delays are local, i.e. only the information from the pair of pre- and postsynaptic neurons is needed to update the synapse; (3) the sparseness in hidden weights is guaranteed by the additive STDP rule; (4) lateral inhibitory connections are used to prohibit excessive firing of the readout neurons which further encourages sparseness; and (5) similar to the organization in neural system, the spikes from readout neurons represent a higher level of features, which can also be fed into another layer of readout neurons, thus forming hierarchical feature learning. In this way, joint STDP allows to combine lower levels of representations into a more abstract neural syntax [121].

6.4.2 Related Works in the Literature

In the literature, Huning et al. [132] also maximized the arrival synchrony by delay plasticity. Our work differs with them from two aspects. 1) In their work, the delay learning happens continuously during or at the end of an interval when the sum of PSPs is higher than a learning threshold; In our work, the delay learning happens at the beginning of that interval, i.e. when the PSP rises across the threshold and triggered a postsynaptic spike; 2) In their work, the learning rule is to maximize the peak of the sum of PSPs by gradient descent, so that the spikes are pulled into synchrony towards the mean arrival time; In our work, the joint STDP maximizes arrival synchrony by minimizing the relative difference between spike arrival times and pull them forward since $D(s, w) < 0$ for

$$s = t_j^f + d_{ij} - t_i^f < 0.$$

Eurich et al. [133] also suggested a joint use of weight and delay STDP. However, their learning windows are different from us. They stated that synchronous spike arrival (i.e. complementary delays) is a stable solution when $D(s) > 0$ for $s < 0$ and $D(s) < 0$ for $s > 0$. In contrast, our delay learning window is reversed with $D(s, w) < 0$ for $s < 0$ and $D(s, w) > 0$ for $s \geq 0$. In fact, without the weight-related gains, our delay learning is indeed unstable. With the help of weight-related gains, it actually goes into quasi-equilibrium states where the delays become smaller slowly with speed proportional to $\delta_-/w_{\max} \ll 1$. Before reaching $d_{\min}$, the relative differences between delays are kept (see the appendix). In fact, δ_- is small enough compared to $w_{\max}$ and it is not possible for biological brains to encounter an exactly same pattern for many times since STDP constantly shapes the network. As a result, for realistic cases, the dynamics is at quasi-equilibrium. In addition, $D(s, w) < 0$ for $s < 0$ leads to reduced latency for strengthened synapses, a functional consequence consistent with the weight STDP.

There are some biological evidences to support delay STDP. Researches in the calyx of Held in the mammalian auditory central nervous system reveal that activity-dependent modification of the synaptic transmission delay exists together with weight plasticity [134, 135], which is found to depend on the strength of local Ca^{2+} transients [136], although on a smaller scale compared to upstream factors [137].

6.4.3 Limitations

There are two major limitations in our approach. First, there is currently no principled way to do model selection for joint STDP, except cross-validation. Several attempts were made to understand the neural network dynamics as probabilistic sampling, called “neural sampling” [138], where the neural activity is understood as MCMC sampling of a posterior distribution.

Second, the readout neurons have a preference to learn patterns involving more spikes

(larger pattern) when multiple patterns with different sizes are presented. And the read-out neurons could fire unreliably or become silent for small patterns. This is because larger patterns bring more spikes to the readout neuron, making it more possible to fire and learn. Therefore, the readout firing threshold plays an important role in the joint STDP. An adaptive firing threshold should be used to address this limitation.

6.5 Summary

In this chapter, we propose joint STDP for learning sparse EEG features. It is motivated by the last chapter, where ELM autonencoder leads to nonsparse features, so that the interpretability of the learned features and hidden weights remains a problem.

We first validated joint STDP on synthetic noisy spike trains similar to the literature ([124] and Appendix E). It turned out that the joint STDP had higher pattern hit rate and lower false alarm frequency than the weight STDP (Figure 6.4). We next validated the joint STDP on a public mental task dataset. The median testing accuracy was $66.4\% \pm 16.3\%$, which was not significantly different with ELM, sparse RBM and sparse AE, while significantly higher than NeuCube (Table 6.2).

After validating joint STDP, we applied it to vigilance recognition during the driving task (Section 5.2.1), in order to answer the last research question in Section 1.2: “Analogous to feature selection, can we automatically learn sparse features so that the useful information can be restricted to a few nonzero features?” It turned out that compared to other 3 nonsparse algorithms and 2 sparse algorithms, the joint STDP achieved the highest level of sparseness in both learned features and hidden weights (Figure 6.12). We plotted the hidden weights versus the output weight to show the learned pattern with higher interpretability (Figure 6.13). In terms of recognition performance, joint STDP had slightly larger testing RMSE (Table 6.3) at 0.206 ± 0.061 . The 3 sparse algorithms including joint STDP have slightly larger testing RMSE than the 3 nonsparse algorithms, due to the trade-off between performance and sparseness.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

This thesis studies the recognition of two types of attention: selective attention and sustained attention (vigilance) from EEG signals using machine learning. This thesis answers the following 3 research questions.

- (1) What are the relevant brain regions (in terms of EEG channels) and distinguishable EEG features in selective attention?

As a starting point, in Chapter 3 we propose an experiment which aims to recognize open eye resting state (unattended) and performing TOVA test (attended), since there is no non-ERP studies using TOVA test to induce selective attention. The average recognition accuracy is $94.4\% \pm 5.0\%$ (ELM) or $94.8\% \pm 4.1\%$ (SVM) using 98 features on 9 subjects. However, the features are based on event-related potential (ERP) studies, which does not work for real-time applications where the external stimuli required by ERP cannot be controlled. In face of the low signal-to-noise ratio in the non-ERP approach, in Chapter 4 we propose Channel Selection with Different Features (CSDF)

algorithm to select EEG features from an extensive set of features. CSDF is able to select channels with their own different feature sets, as well as restrict the useful features to as few channels as possible. Using CSDF, few features (83 out of 868) are selected, and the recognition accuracy $94.3\% \pm 5.6\%$ is the best compared to other 5 feature selection and 5 channel selection algorithms (not statistically significant). Based on CSDF, the answer to this research question is that, the first and second order difference in the left parietal and temporal lobes, as well as the Higuchi fractal dimension and mean signal amplitude in the right frontal lobe, are relevant to selective attention. The selected channels are consistent with a neuroimaging study using ERP in the literature [19,95].

(2) What are the vigilance stages in open eye and situation-aware state?

Unlike selective attention which has discrete conditions such as attended/unattended, the vigilance stages cannot be easily observed. In Chapter 5 we use ELMAE to recognize vigilance, learn the EEG spectral features and define the vigilance stages in a driving task, in a data-driven and subject-independent way. For the recognition performance, experimental results show that ELMAE is an efficient alternative to RBM: ELMAE achieves RMSE at 0.189 ± 0.049 , which is better than RBM at 0.195 ± 0.046 (not statistically significant); and the training speed of ELMAE is 1.5 ± 0 seconds, significantly faster than 52.7 ± 1.8 seconds for RBM. For the learned hidden weights and features, ELMAE is able to learn diverse hidden weights, however being nonsparse and holistic-based. To answer the research question, the vigilance stages in open eye and situation-aware state are described in Table 5.3. Stage I is high vigilance, where we observe eyeball movement and blinking in the frontal area, implying that the subject is attentively following the driving scenario. Stage II is low vigilance, which can be divided into two stages. In stage II.1, the spreading alpha power around occipital and parietal areas, which is related to drowsiness and difficulty in attention allocation in open eye state. In stage II.2, instead of falling sleep, this stage involves muscular movements, which indicates distraction. A possible explanation for stage II.2 is that, the environment contains not enough external stimuli to keep the open eye and situation-aware state, so

that the brain performs vigilance regulation to seek external stimuli, which leads to distraction.

- (3) Analogous to feature selection in CSDF, can we learn sparse features so that the useful information is restricted to a few nonzero features?

ELMAE learns nonsparse hidden weights and features, making the interpretation difficult. In the literature, there are several sparse feature learning algorithms such as sparse RBM and sparse AE. However, the sparseness is realized by reducing the average activation in the hidden layer, with no attempt on sparse hidden weights. To address the limitation, we propose a novel bio-inspired algorithm, joint STDP, for learning sparse hidden weights and EEG features. The joint STDP is first validated on synthetic noisy spike trains containing predefined patterns, which achieves higher pattern hit rate and lower false alarm frequency than the conventional weight STDP. The joint STDP is also validated on the public mental task dataset, where the accuracy $66.4\% \pm 16.3\%$ is not significantly different with ELM, sparse RBM and sparse AE. After validation, the joint STDP is applied to recognize vigilance and learn sparse hidden weights and EEG spectral features. It turns out that compared to other 3 nonsparse algorithms and 2 sparse algorithms, the joint STDP achieved the highest level of sparseness in both learned features and hidden weights, with statistical significance. In terms of recognition performance, joint STDP has slightly larger RMSE at 0.206 ± 0.061 , due to the trade-off between performance and sparseness.

7.2 Future Work

From experimental point of view, parallel to selective attention and sustained attention (vigilance) in this thesis, the other types of attention, i.e. focused attention, divided attention and alternating attention, can be studied. The experiment should be designed to fit the definition of these attention types. Similar machine learning algorithms can

be used to recognize these attention types. Predefined features can be extracted, where feature selection and channel selection are used to identify important features and brain regions in these attention types. Alternatively, sparse feature learning can be used to automatically learn and select features in a data-driven and subject-independent way.

From theoretical point of view, there are 3 important future directions. First, all the proposed algorithms in this thesis assume the independence between EEG segments. However, this does not hold due to the complicated dynamics of the brain. Machine learning algorithms which consider the interaction between EEG segments, such as hidden Markov model and recurrent neural network, should be investigated.

Second, in the EEG feature learning algorithms such as ELMAE and joint STDP, we only learn spectral features, and do not learn features from the EEG waveform, such as alpha spindles. This is because these waveform patterns can occur at any position within the EEG segment, while ELMAE and joint STDP are not shift-invariant. Advanced techniques such as downsampling or pooling should be combined in the proposed algorithms to enable shift-invariant feature learning.

Third, joint STDP is not supported by appropriate mathematics/probability theory. For example, if we could fit the spiking neural network dynamics to the Bayesian framework, it becomes possible to find the posterior distribution sampled by the spikes, to use STDP to implement different learning schemes such as maximum likelihood estimation (MLE) or maximum a posterior estimation (MAP), as well as automatically determine the hyperparameters using type-II MLE (evidence method). Initial research in this direction can be found in “neural sampling” [138].

From application point of view, EEG-based attention recognition can be used for real-time assessment of new human-machine interface (HMI) designs, for example, how different styles of driving panels affect drivers’ attention. The real-time ability enables the assessor to receive feedback from the subject at high frequency, without interrupting the subject such as using questionnaires [28].

Appendix

A Proof for Sparseness in the Solution in Section 4.2.3

The sparseness-inducing ability of ℓ_1 -norm and group lasso is well-known. Here we prove that the desired channel-wise and feature-wise sparseness can be obtained by solving Equation (4.10). We first clarify the definitions: ℓ_2 -norm is defined for a vector $\|x\|_2 = \sqrt{\sum_i x_i^2}$, and Frobenius-norm is defined for a matrix $\|W\|_F = \sqrt{\sum_i \sum_j W_{ij}^2}$.

To start, Equation (4.10) can be rewritten as

$$\begin{aligned} \underset{\mathbf{W}}{\text{minimize}} \quad J'_3 &= \frac{1}{2} \left\| \hat{\mathbf{X}}\mathbf{W} - \mathbf{Y} \right\|_F^2 + C_1 \sum_{j=1}^g \left\| \mathbf{W}^j \right\|_F + C_2 \sum_{i=1}^d \left\| \mathbf{w}_i \right\|_2 \\ &= \frac{1}{2} \left\| \hat{\mathbf{X}}\mathbf{W} - \mathbf{Y} \right\|_F^2 + C_1 \sum_{j=1}^g \left\| \mathbf{M}^j \mathbf{W} \right\|_F + C_2 \sum_{i=1}^d \left\| \mathbf{N}^i \mathbf{W} \right\|_2, \end{aligned} \quad (\text{A.1})$$

where $\mathbf{M}^j \in \mathbb{R}^{d_j \times d}$ and $\mathbf{N}^i \in \mathbb{R}^{1 \times d}$ are mask matrices for channels and features respectively.

$$\mathbf{M}^j = \begin{bmatrix} \mathbf{0}_{d_j \times (j-1)d_j} & \mathbf{I}_{d_j} & \mathbf{0}_{d_j \times (g-j)d_j} \end{bmatrix}, \quad (\text{A.2})$$

$$\mathbf{N}^i = \begin{bmatrix} 0 & \cdots & 0 & 1_{(\text{at } i\text{-th column})} & 0 & \cdots & 0 \end{bmatrix}. \quad (\text{A.3})$$

Since J'_3 is a convex function w.r.t. $\mathbf{W}$, we can take the derivative of J'_3 w.r.t. $\mathbf{W}$, which

is pointing at the global minimum,

$$\frac{dJ'_3}{d\mathbf{W}} = \hat{\mathbf{X}}^\top (\hat{\mathbf{X}}\mathbf{W} - \mathbf{Y}) + C_1 \sum_{j=1}^g \frac{\mathbf{M}^{j\top} \mathbf{M}^j \mathbf{W}}{\|\mathbf{M}^j \mathbf{W}\|_F} + C_2 \sum_{i=1}^d \frac{\mathbf{N}^{i\top} \mathbf{N}^i \mathbf{W}}{\|\mathbf{N}^i \mathbf{W}\|_2}. \quad (\text{A.4})$$

Next we examine the value of $\left\| \frac{dJ'_3}{d\mathbf{W}} \Big|_{\mathbf{W}=\mathbf{W}^*} \right\|_F$. To evaluate that, we first derive two lemmas.

Lemma A.1 *Given $\mathbf{M}^j$, $\mathbf{N}^i$ and $\mathbf{W}$ as defined above, matrix $\frac{\mathbf{M}^{j\top} \mathbf{M}^j \mathbf{W}}{\|\mathbf{M}^j \mathbf{W}\|_F}$ only has nonzero rows which belong to $Channel_j$ and matrix $\frac{\mathbf{N}^{i\top} \mathbf{N}^i \mathbf{W}}{\|\mathbf{N}^i \mathbf{W}\|_2}$ only has one nonzero row at i -th row.*

Proof: We have

$$\begin{aligned} \frac{\mathbf{M}^{j\top} \mathbf{M}^j \mathbf{W}}{\|\mathbf{M}^j \mathbf{W}\|_F} &= \frac{\begin{bmatrix} \mathbf{0}_{d_j \times (j-1)d_j} & \mathbf{I}_{d_j} & \mathbf{0}_{d_j \times (g-j)d_j} \end{bmatrix}^\top \mathbf{M}^j \mathbf{W}}{\|\mathbf{M}^j \mathbf{W}\|_F} \\ &= \frac{1}{\|\mathbf{M}^j \mathbf{W}\|_F} \begin{bmatrix} \mathbf{0}_{(j-1)d_j \times k} \\ \mathbf{M}^j \mathbf{W} \\ \mathbf{0}_{(g-j)d_j \times k} \end{bmatrix}. \end{aligned} \quad (\text{A.5})$$

So $\frac{\mathbf{M}^{j\top} \mathbf{M}^j \mathbf{W}}{\|\mathbf{M}^j \mathbf{W}\|_F}$ only has nonzero rows from row number $(j-1)d_j + 1$ to $j \cdot d_j$, which correspond to $Channel_j$. Similarly,

$$\begin{aligned} \frac{\mathbf{N}^{i\top} \mathbf{N}^i \mathbf{W}}{\|\mathbf{N}^i \mathbf{W}\|_2} &= \frac{\begin{bmatrix} 0 & \cdots & 0 & 1_{(\text{at } i\text{-th column})} & 0 & \cdots & 0 \end{bmatrix}^\top \mathbf{N}^i \mathbf{W}}{\|\mathbf{N}^i \mathbf{W}\|_2} \\ &= \frac{1}{\|\mathbf{N}^i \mathbf{W}\|_2} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \mathbf{N}^i \mathbf{W}_{(\text{at } i\text{-th row})} \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \end{aligned} \quad (\text{A.6})$$

So $\frac{\mathbf{N}^{i\top} \mathbf{n}^i \mathbf{W}}{\|\mathbf{N}^i \mathbf{W}\|_2}$ only has nonzero row at i -th row. ■

Lemma A.2 Given $\mathbf{M}^j$, $\mathbf{N}^i$, $\mathbf{W}$, g , d , $1 \leq j \leq g$ and $1 \leq i \leq d$ as defined above, we have

$$\left\| \sum_{j=1}^g \frac{\mathbf{M}^{j\top} \mathbf{M}^j \mathbf{W}}{\|\mathbf{M}^j \mathbf{W}\|_F} \right\|_F = \sqrt{g} \text{ and } \left\| \sum_{i=1}^d \frac{\mathbf{N}^{i\top} \mathbf{N}^i \mathbf{W}}{\|\mathbf{N}^i \mathbf{W}\|_2} \right\|_F = \sqrt{d}. \quad (\text{A.7})$$

Proof: We have

$$\begin{aligned} \left\| \sum_{j=1}^g \frac{\mathbf{M}^{j\top} \mathbf{M}^j \mathbf{W}}{\|\mathbf{M}^j \mathbf{W}\|_F} \right\|_F &= \sqrt{\sum_{j=1}^g \left\| \frac{\mathbf{M}^{j\top} \mathbf{M}^j \mathbf{W}}{\|\mathbf{M}^j \mathbf{W}\|_F} \right\|_F^2} \quad (\text{according to Lemma A.1}) \\ &= \sqrt{\sum_{j=1}^g \left(\frac{\|\mathbf{M}^{j\top} \mathbf{M}^j \mathbf{W}\|_F}{\|\mathbf{M}^j \mathbf{W}\|_F} \right)^2} \\ &= \sqrt{\sum_{j=1}^g \left(\frac{\left\| \begin{bmatrix} \mathbf{0} & \mathbf{I} & \mathbf{0} \end{bmatrix}^\top \mathbf{M}^j \mathbf{W} \right\|_F}{\|\mathbf{M}^j \mathbf{W}\|_F} \right)^2} \quad (\text{A.8}) \\ &= \sqrt{\sum_{j=1}^g \left(\frac{\left\| \begin{bmatrix} \mathbf{0} & (\mathbf{M}^j \mathbf{W})^\top & \mathbf{0} \end{bmatrix} \right\|_F}{\|\mathbf{M}^j \mathbf{W}\|_F} \right)^2} \\ &= \sqrt{\sum_{j=1}^g \left(\frac{\|\mathbf{M}^j \mathbf{W}\|_F}{\|\mathbf{M}^j \mathbf{W}\|_F} \right)^2} = \sqrt{g}. \end{aligned}$$

Similarly,

$$\left\| \sum_{i=1}^d \frac{\mathbf{N}^{i\top} \mathbf{N}^i \mathbf{W}}{\|\mathbf{N}^i \mathbf{W}\|_2} \right\|_F = \sqrt{\sum_{i=1}^d \left(\frac{\|\mathbf{N}^i \mathbf{W}\|_2}{\|\mathbf{N}^i \mathbf{W}\|_2} \right)^2} = \sqrt{d}. \quad (\text{A.9})$$

■

Now we start the proof for sparseness in the solution to Equation (4.10).

Let $\mathbf{W}^*$ be the optimal solution that minimizes J'_3 ($J'_3 = J_3$), where the sparseness pattern is unknown. Let $\Delta\mathbf{W}$ be a small deviation from $\mathbf{W}^*$, i.e. $0 < \|\Delta\mathbf{W}\|_F = \epsilon \ll 1$ is small. Let us first consider the Frobenius norm of each right hand side (RHS) term of $\left. \frac{dJ'_3}{d\mathbf{W}} \right|_{\mathbf{W}=\mathbf{W}^*+\Delta\mathbf{W}}$ according to Equation (A.4).

The Frobenius norm of the first RHS term of Equation (A.4) is

$$\begin{aligned}
 & \left\| \hat{\mathbf{X}}^\top \left(\hat{\mathbf{X}} (\mathbf{W}^* + \Delta\mathbf{W}) - \mathbf{Y} \right) \right\|_F \\
 &= \left\| \hat{\mathbf{X}}^\top \left(\hat{\mathbf{X}} \mathbf{W}^* - \mathbf{Y} \right) + \hat{\mathbf{X}}^\top \hat{\mathbf{X}} \Delta\mathbf{W} \right\|_F \\
 &\leq \left\| \hat{\mathbf{X}}^\top \left(\hat{\mathbf{X}} \mathbf{W}^* - \mathbf{Y} \right) \right\|_F + \left\| \hat{\mathbf{X}}^\top \hat{\mathbf{X}} \Delta\mathbf{W} \right\|_F \\
 &\leq \left\| \hat{\mathbf{X}}^\top \left(\hat{\mathbf{X}} \mathbf{W}^* - \mathbf{Y} \right) \right\|_F + \left\| \hat{\mathbf{X}}^\top \hat{\mathbf{X}} \right\|_F \|\Delta\mathbf{W}\|_F \\
 &= \left\| \hat{\mathbf{X}}^\top \left(\hat{\mathbf{X}} \mathbf{W}^* - \mathbf{Y} \right) \right\|_F + \left\| \hat{\mathbf{X}}^\top \hat{\mathbf{X}} \right\|_F \epsilon \\
 &= a\epsilon + b \approx b,
 \end{aligned} \tag{A.10}$$

where $a = \left\| \hat{\mathbf{X}}^\top \hat{\mathbf{X}} \right\|_F$ is a constant, $b = \left\| \hat{\mathbf{X}}^\top \left(\hat{\mathbf{X}} \mathbf{W}^* - \mathbf{Y} \right) \right\|_F$ is another constant.

According to Lemma A.2, the Frobenius norm of the second RHS term of Equation (A.4) is

$$\left\| C_1 \sum_{j=1}^g \frac{\mathbf{M}^{j\top} \mathbf{M}^j (\mathbf{W}^* + \Delta\mathbf{W})}{\|\mathbf{M}^j (\mathbf{W}^* + \Delta\mathbf{W})\|_F} \right\|_F = \sqrt{g} C_1 = \text{constant}. \tag{A.11}$$

Similarly, according to Lemma A.2, the Frobenius norm of the third RHS term of Equation (A.4) is

$$\left\| C_2 \sum_{i=1}^d \frac{\mathbf{N}^{i\top} \mathbf{N}^i (\mathbf{W}^* + \Delta\mathbf{W})}{\|\mathbf{N}^i (\mathbf{W}^* + \Delta\mathbf{W})\|_2} \right\|_F = \sqrt{d} C_2 = \text{constant}. \tag{A.12}$$

Putting all together,

$$\left\| \left. \frac{dJ'_3}{d\mathbf{W}} \right|_{\mathbf{W}=\mathbf{W}^*+\Delta\mathbf{W}} \right\|_F \leq a\epsilon + b + \sqrt{g} C_1 + \sqrt{d} C_2 \approx b + \sqrt{g} C_1 + \sqrt{d} C_2. \tag{A.13}$$

We discuss 4 cases here:

- (1) If $\sqrt{g}C_1 \gg b$, $\sqrt{g}C_1$ dominates Equation (A.13). When close to the optimum, i.e. $\mathbf{W} = \mathbf{W}^* + \Delta\mathbf{W}$, the dominant driving force of the gradient drives $\mathbf{W}$ along direction $\sum_{j=1}^g \frac{\mathbf{M}^j \top \mathbf{M}^j \mathbf{W}}{\|\mathbf{M}^j \mathbf{W}\|_F}$. According to Lemma A.1, each matrix $\frac{\mathbf{M}^j \top \mathbf{M}^j \mathbf{W}}{\|\mathbf{M}^j \mathbf{W}\|_F}$ only has nonzero rows belonging to $Channel_j$. As a result, this dominant term drives the solution towards channel-wise sparseness;
- (2) Similarly, if $\sqrt{d}C_2 \gg b$, $\sqrt{d}C_2$ dominates Equation (A.13). When close to the optimum, i.e. $\mathbf{W} = \mathbf{W}^* + \Delta\mathbf{W}$, the dominant driving force of the gradient drives $\mathbf{W}$ along direction $\sum_{i=1}^d \frac{\mathbf{N}^i \top \mathbf{N}^i \mathbf{W}}{\|\mathbf{N}^i \mathbf{W}\|_2}$. According to Lemma A.1, each matrix $\frac{\mathbf{N}^i \top \mathbf{N}^i \mathbf{W}}{\|\mathbf{N}^i \mathbf{W}\|_2}$ only has one nonzero row at i -th row. As a result, this dominant term drives the solution towards feature-wise (row-wise) sparseness;
- (3) If both $\sqrt{g}C_1 \gg b$ and $\sqrt{d}C_2 \gg b$ hold, both of them dominate Equation (A.13), which lead to a combination of (1) and (2), and also result in both channel-wise and feature-wise sparseness;
- (4) If none of $\sqrt{g}C_1 \gg b$ or $\sqrt{d}C_2 \gg b$ holds, there will be a balance between these 3 forces in Equation (A.13), which leads to a nonsparse solution as result of compromise.

In conclusion,

- (1) If we select C_1 such that $\sqrt{g}C_1 \gg \left\| \hat{\mathbf{X}}^\top (\hat{\mathbf{X}}\mathbf{W}^* - \mathbf{Y}) \right\|_F$, the optimal solution $\mathbf{W}_{CSDF}^*$ is channel-wise sparse in its rows;
- (2) If we select C_2 such that $\sqrt{d}C_2 \gg \left\| \hat{\mathbf{X}}^\top (\hat{\mathbf{X}}\mathbf{W}^* - \mathbf{Y}) \right\|_F$, the optimal solution $\mathbf{W}_{CSDF}^*$ is feature-wise sparse in its rows;
- (3) If (1) and (2) both hold, the optimal solution $\mathbf{W}_{CSDF}^*$ is both channel-wise and feature-wise sparse in its rows;
- (4) If none of (1) or (2) holds, the optimal solution $\mathbf{W}_{CSDF}^*$ is nonsparse in its rows.

B Proof for Replacing Discrete Labels with Probabilistic Labels as a Generalization of Training Logistic Regression and Naive Bayes

Proof: Replacing discrete labels $y \in \{1, \dots, K\}$ with probabilistic labels $p \in [0, 1]^K$ is a generalization of training logistic regression and naive Bayes.

We are given features $X = \{x_i\}_{i=1}^N \in \mathbb{R}^{N \times D}$ and probabilistic labels $P = \{p_i\}_{i=1}^N \in [0, 1]^{N \times K}$ ($\sum_k p_{ik} = 1$). The objective function is

$$\underset{\theta}{\text{maximize}} \quad J(\theta) = -D_{\text{KL}}(P \parallel p(Z \mid X; \theta)) , \quad (\text{B.1})$$

where $Z = \{z_i\}_{i=1}^N \in \{1, \dots, K\}^N$ is the latent label and θ is the model parameter.

The objective function cannot be directly optimized due to the latent variable Z . We apply expectation maximization (EM) to maximize the lower bound of the objective function,

$$\begin{aligned} Q(\theta; \theta^{(t)}) &= \mathbb{E}_{Z \mid X; \theta^{(t)}} [J(\theta)] + \text{constant} \\ &= \sum_{i=1}^N \sum_{k=1}^K \delta_{ik} \sum_{l=1}^K p_{il} \log p(z_i = l \mid x_i; \theta) \\ &= \sum_{i=1}^N \left(\sum_{k=1}^K \delta_{ik} \right) \sum_{l=1}^K p_{il} \log p(z_i = l \mid x_i; \theta) \\ &= \sum_{i=1}^N \sum_{k=1}^K p_{ik} \log p(z_i = k \mid x_i; \theta) , \end{aligned} \quad (\text{B.2})$$

where $\delta_{ik} = p(z_i = k \mid x_i; \theta^{(t)})$ and $\sum_k \delta_{ik} = 1$. Therefore $Q(\theta; \theta^{(t)})$ is independent on δ and hence $\theta^{(t)}$. So EM can be solved in one step without iteration.

For logistic regression, $\theta = \{W \in \mathbb{R}^{D \times K}\}$, we have

$$\begin{aligned}
 Q(W; W^{(t)}) &= \sum_{i=1}^N \sum_{k=1}^K p_{ik} \log p(z_i = k | x_i; W) \\
 &= \sum_{i=1}^N \sum_{k=1}^K p_{ik} [W_{:,k}^\top x_i - \text{logsumexp}(W^\top x_i)] \\
 &= \sum_{i=1}^N \left[\left(\sum_{k=1}^K p_{ik} W_{:,k}^\top x_i \right) - \text{logsumexp}(W^\top x_i) \right]. \tag{B.3}
 \end{aligned}$$

Comparing Equation (B.3) with logistic regression trained with discrete labels $Y = \{y_i\}_{i=1}^N \in \{1, \dots, K\}^N$ as below,

$$\begin{aligned}
 \ell(W) &= \sum_{i=1}^N \sum_{k=1}^K y_{ik} [W_{:,k}^\top x_i - \text{logsumexp}(W^\top x_i)] \\
 &= \sum_{i=1}^N \left[\left(\sum_{k=1}^K y_{ik} W_{:,k}^\top x_i \right) - \text{logsumexp}(W^\top x_i) \right]. \tag{B.4}
 \end{aligned}$$

Equation (B.3) and Equation (B.4) have the same form where the only difference is replacing the discrete labels Y with the probabilistic labels P .

For Bernoulli naive Bayes, $X = \{x_i\}_{i=1}^N \in \{0, 1\}^{N \times D}$ follows Bernoulli distribution given its latent label $Z = \{z_i\}_{i=1}^N \in \{1, \dots, K\}^N$, i.e. $p(x_{ij} | z_i; \pi) = \pi_{jz_i}^{x_{ij}} (1 - \pi_{jz_i})^{1-x_{ij}}$. In this case, $\theta = \{\pi, p(z)\}$, $\pi \in [0, 1]^{D \times K}$ and $p(z) \in [0, 1]^K$ ($\sum_k p(z = k) = 1$). We have

$$\begin{aligned}
 Q(\theta; \theta^{(t)}) &= \sum_{i=1}^N \sum_{k=1}^K p_{ik} \log p(z_i = k | x_i; \theta) \\
 &= \sum_{i=1}^N \sum_{k=1}^K p_{ik} \left[\log p(z_i = k) + \sum_{j=1}^D [x_{ij} \log \pi_{jk} + (1 - x_{ij}) \log(1 - \pi_{jk})] \right]. \tag{B.5}
 \end{aligned}$$

Maximizing $Q(\theta; \theta^{(t)})$ with respect to π_{jk} , we obtain

$$\pi_{jk} = \arg \max_{\pi_{jk}} Q(\theta; \theta^{(t)}) = \frac{\sum_{i=1}^N p_{ik} x_{ij}}{\sum_{i=1}^N p_{ik}}. \quad (\text{B.6})$$

Maximizing $Q(\theta; \theta^{(t)})$ with respect to $\phi_k = p(z = k)$ (using Lagrange multiplier to handle constraint $\sum_k \phi_k = 1$), we obtain

$$\phi_k = \arg \max_{\phi_k} Q(\theta; \theta^{(t)}) = \frac{\sum_{i=1}^N p_{ik}}{\sum_{l=1}^K \sum_{i=1}^N p_{il}} = \frac{\sum_{i=1}^N p_{ik}}{N}. \quad (\text{B.7})$$

Comparing Equation (B.6) and (B.7) with Bernoulli naive Bayes trained with discrete labels $Y = \{y_i\}_{i=1}^N \in \{1, \dots, K\}^N$ as below

$$\pi_{jk} = \frac{\sum_{i=1}^N y_{ik} x_{ij}}{\sum_{i=1}^N y_{ik}}, \quad (\text{B.8})$$

$$\phi_k = p(y = k) = \frac{\sum_{i=1}^N y_{ik}}{N}. \quad (\text{B.9})$$

They have the same form where the only difference is replacing the discrete labels Y with the probabilistic labels P . This completes the proof. $\blacksquare$

As a remark, this proof is closely related to the common sense that maximizing log-likelihood is equivalent to minimizing KL-divergence [139],

$$\begin{aligned} & \underset{\theta}{\text{minimize}} \quad D_{\text{KL}}(P \parallel p(X, Z; \theta)) \\ & \iff \underset{\theta}{\text{maximize}} \quad p(X, Y; \theta). \end{aligned} \quad (\text{B.10})$$

C Proof for Equation (6.8) and (6.9) in Section 6.2

We give a proof for the quasi-equilibrium weights and delays $\{w_{ij}^*, d_{ij}^*\}_{j=1}^N$ in Equation (6.8) and (6.9) under the dynamics of (6.6) and (6.7) and the following conditions.

- (1) There is only one pattern involving neuron set $P \subseteq \{1, \dots, N\}$ with spike timing at t_j^P ;
- (2) There is only one readout neuron i , so that there is no lateral inhibition;
- (3) We only analyze the case where the readout spike is caused by the pattern, not by the noise.

The proof also holds for multiple patterns and readout neurons. This is because with the lateral inhibition, some readout neurons only have the chance to learn one particular pattern, thus reducing to the case of single pattern and readout neuron. The proof also holds for mediate level of random noise between multiple pattern activations. The random noise can make the weights slowly drift to 0 with speed proportional to $\int W(s)ds$, and the delays slowly drift to $d_{\min}$ with speed proportional to $\int D(s, 0)ds$. If the noise density and duration are mediate so that the update caused by the patterns can overcome these drifts, the readout dynamics is mainly pattern-driven. These cases are validated experimentally in Section 6.3.1.

The proof is divided into two parts.

- (1) $\{w_{ij}^*, d_{ij}^*\}_{j=1}^N$ is a quasi-equilibrium state;
- (2) All quasi-equilibrium states are of the form of $\{w_{ij}^*, d_{ij}^*\}_{j=1}^N$.

C.1 Proof Part 1

Proof: $\{w_{ij}^*, d_{ij}^*\}_{j=1}^N$ is a quasi-equilibrium state.

Suppose $\{w_{ij}^*, d_{ij}^*\}_{j=1}^N$ is distorted by a small amount,

$$w'_{ij} = w_{ij}^* + \psi_{ij} ; \tag{C.1}$$

$$d'_{ij} = d_{ij}^* + \phi_{ij} , \tag{C.2}$$

where ψ_{ij} and ϕ_{ij} are small compared to w_{ij}^* and d_{ij}^* respectively. The postsynaptic spike

time is changed to $t_i^{f'}$ under the distorted weights and delays $\{w'_{ij}, d'_{ij}\}_{j=1}^N$.

We show that under a small distortion, $\{w'_{ij}, d'_{ij}\}_{j=1}^N$ recover to $\{w^*_{ij}, d^*_{ij}\}_{j=1}^N$, but with T becoming smaller slowly with speed proportional to $\delta_-/w_{\max} \ll 1$ (quasi-equilibrium, not fixed point) unless $\delta_- = \delta_+ = 0$ (fixed point).

1) When $j \in Q \subseteq P$.

1.1) For the weight, we have

$$\Delta w_{ij} = W \left(t_j^P + d'_{ij} - t_i^{f'} \right) = W \left(T + \phi_{ij} - t_i^{f'} \right). \quad (\text{C.3})$$

Under the assumption that the distortion is small, we have $T + \phi_{ij} - t_i^{f'} \approx T - t_i^{f*} < 0$, therefore $\Delta w_{ij} > 0$. And w'_{ij} recovers to $w^*_{ij} = w_{\max}$ for $j \in P$. We note that $T + \phi_{ij} - t_i^{f'} \approx T - t_i^{f*}$ only holds for small distortion.

1.2) For the delay, we note that the following four statements are equivalent.

- d_{ij} is of the form $T - t_j^P$.
- The delay d_{ij} is complementary to the spike timing t_j^P in the pattern.
- The spike arrival times of any synapses j and k are the same, i.e. $t_j^P + d_{ij} = t_k^P + d_{ik} = T$, for $j \neq k, j \in Q$ and $k \in Q$.
- The relative difference between spike arrival times of any synapses j and k is 0, i.e. $(t_j^P + d_{ij}) - (t_k^P + d_{ik}) = 0$ for $j \neq k, j \in Q$ and $k \in Q$.

Therefore we equivalently prove the last statement. Formally, after the distortion, $(t_j^P + d'_{ij}) - (t_k^P + d'_{ik})$ recovers to 0.

Without loss of generality, assuming $t_j^P + d'_{ij} \geq t_k^P + d'_{ik}$, we have

$$\begin{aligned}
 & (t_j^P + d'_{ij}) - (t_k^P + d'_{ik}) \tag{C.4} \\
 & \rightarrow (t_j^P + d'_{ij} + \Delta d_{ij}) - (t_k^P + d'_{ik} + \Delta d_{ik}) \quad (\text{after one update}) \\
 & = (t_j^P + d'_{ij} - t_i^{f'}) - (t_k^P + d'_{ik} - t_i^{f'}) + (\Delta d_{ij} - \Delta d_{ik}) \\
 & = s_j - s_k - \left[M_j s_j \exp\left(\frac{s_j}{\sigma_-}\right) - M_k s_k \exp\left(\frac{s_k}{\sigma_-}\right) \right] \\
 & = \left[s_j - M_j s_j \exp\left(\frac{s_j}{\sigma_-}\right) \right] - \left[s_k - M_k s_k \exp\left(\frac{s_k}{\sigma_-}\right) \right] \\
 & \approx f(s_j) - f(s_k) \\
 & = \int_{s_k}^{s_j} f'(s) ds \\
 & = \int_{s_k}^{s_j} \left[1 - M \left(1 + \frac{s}{\sigma_-} \right) \exp\left(\frac{s}{\sigma_-}\right) \right] ds ,
 \end{aligned}$$

where

$$M_j = \frac{w_{\max} - w'_{ij} + \delta_-}{w_{\max}} B_- = \frac{-\psi_{ij} + \delta_-}{w_{\max}} B_- \gtrsim 0 ; \tag{C.5}$$

$$M_k = \frac{w_{\max} - w'_{ik} + \delta_-}{w_{\max}} B_- = \frac{-\psi_{ik} + \delta_-}{w_{\max}} B_- \gtrsim 0 ; \tag{C.6}$$

$$M = M_j \approx M_k ; \tag{C.7}$$

$$-\sigma_- \leq s_j = t_j^P + d'_{ij} - t_i^{f'} = T + \phi_{ij} - t_i^{f'} \approx T - t_i^{f*} < 0 ; \tag{C.8}$$

$$-\sigma_- \leq s_k = t_k^P + d'_{ik} - t_i^{f'} = T + \phi_{ik} - t_i^{f'} \approx T - t_i^{f*} < 0 ; \tag{C.9}$$

$$f(s) = s - M s \exp\left(\frac{s}{\sigma_-}\right) . \tag{C.10}$$

So $0 < f'(s) \leq 1$.

Continuing Equation (C.4),

$$(t_j^P + d'_{ij}) - (t_k^P + d'_{ik}) \quad (\text{C.11})$$

$$\begin{aligned} &\rightarrow (t_j^P + d'_{ij} + \Delta d_{ij}) - (t_k^P + d'_{ik} + \Delta d_{ik}) \quad (\text{after one update}) \\ &\approx \int_{s_k}^{s_j} \left[1 - M \left(1 + \frac{s}{\sigma_-} \right) \exp \left(\frac{s}{\sigma_-} \right) \right] ds \end{aligned} \quad (\text{C.12})$$

$$\begin{aligned} &\leq \int_{s_k}^{s_j} 1 ds = s_j - s_k \\ &= (t_j^P + d'_{ij}) - (t_k^P + d'_{ik}) . \end{aligned} \quad (\text{C.13})$$

After the update, $(t_j^P + d'_{ij}) - (t_k^P + d'_{ik})$ becomes smaller and finally recovers to 0, therefore d'_{ij} recovers to the form $d_{ij}^* = T - t_j^P$. The recovery speed is faster if $f'(s)$ is smaller, i.e. the distortion is larger or δ_- is larger. Once again, if the distortion becomes large enough so that $T + \phi_{ij} - t_i^{f'} \approx T - t_i^{f*}$ in Equation (C.8) and (C.9) does not hold, the dynamics goes to another state. Note that in Equation (C.13), the equality only holds when $f'(s) = 1$ holds for all $s \in [s_k, s_j]$, which corresponds to no distortion $\psi_{ij} = 0$ and no weight-related gain offset $\delta_- = 0$.

However, d_{ij}^* is not a fixed point because

$$\begin{aligned} \Delta d_{ij} &= D \left(t_j^P + d'_{ij} - t_i^{f'}, w'_{ij} \right) \\ &\approx D \left(t_j^P + d_{ij}^* - t_i^{f*}, w_{ij}^* \right) \\ &= \frac{-\delta_-}{w_{\max}} B_- \left[\frac{\sigma_-}{e} + s \exp \left(\frac{s}{\sigma_-} \right) \right] \propto \frac{\delta_-}{w_{\max}} \lesssim 0 , \end{aligned} \quad (\text{C.14})$$

where $s = t_j^P + d_{ij}^* - t_i^{f*} < 0$.

In this case, d_{ij}^* becomes smaller slowly with speed proportional to $\delta_-/w_{\max} \ll 1$ and keeps the form $T - t_j^P$, or equivalently T becomes smaller slowly. After learning many activations of the same pattern, d_{ij}^* reaches $d_{\min}$, at which the $T - t_j^P$ form cannot be kept anymore. However, δ_- is small enough compared to $w_{\max}$ and it is not possible for biological brains to encounter an exactly same input for many times since STDP constantly

shapes the network. On the other hand, if $\delta_- = 0$, it shuts down the dynamics with $\Delta d_{ij} = 0$, leading to a fixed point. Similar conclusion that d_{ij} is at quasi-equilibrium unless $\delta_+ = 0$ also holds for $j \notin Q$ and $j \in P$. We note that without the weight-related gains, $\Delta d_{ij} < 0$, so d_{ij}^* becomes smaller faster and reaches $d_{\min}$ soon after a few updates, so the form $T - t_j^P$ cannot be kept.

To conclude 1), for $j \in Q \subseteq P$, under a small distortion, w'_{ij} recovers to $w_{ij}^* = w_{\max}$ and d'_{ij} recovers to the form $d_{ij}^* = T - t_j^P$. However, d_{ij}^* becomes smaller slowly with speed proportional to $\delta_-/w_{\max} \ll 1$ (quasi-equilibrium, not fixed point) unless $\delta_- = 0$ (fixed point).

2) When $j \notin Q$ and $j \in P$.

2.1) For the weight, under the assumption that the distortion is small, we have $t_j^P + d'_{ij} - t_i^{f'} \approx t_j^P + d_{ij}^* - t_i^{f*}$. For $j \notin Q$ and $j \in P$, we have $j \in \{j \mid t_j^P + d_{ij}^* \geq t_i^{f*}, j \in P\}$, so $t_j^P + d_{ij}^* - t_i^{f*} \geq 0$, and

$$\Delta w_{ij} = W \left(t_j^P + d'_{ij} - t_i^{f'} \right) \approx W \left(t_j^P + d_{ij}^* - t_i^{f*} \right) < 0. \quad (\text{C.15})$$

Therefore $\Delta w_{ij} < 0$. And w'_{ij} recovers to $w_{ij}^* = 0$ for $j \notin Q$ and $j \in P$. Note that $t_j^P + d'_{ij} - t_i^{f'} \approx t_j^P + d_{ij}^* - t_i^{f*}$ only holds for small distortion, otherwise the dynamics goes to another state.

2.2) For the delay, we do not check whether it is of the form $T - t_j^P$ because these synapses do not contribute to the postsynaptic spike ($w_{ij}^* = 0$).

Similar to 2.1), under the assumption that the distortion is small, we have $t_j^P + d'_{ij} - t_i^{f'} \approx$

$t_j^P + d_{ij}^* - t_i^{f*} \geq 0$, therefore

$$\begin{aligned} \Delta d_{ij} &= D(t_j^P + d'_{ij} - t_i^{f'}, w'_{ij}) \\ &\approx D(t_j^P + d_{ij}^* - t_i^{f*}, w_{ij}^*) \\ &= \frac{-\delta_+}{w_{\max}} B_+ \left[\frac{\sigma_+}{e} - s \exp\left(\frac{-s}{\sigma_+}\right) \right] \propto \frac{-\delta_+}{w_{\max}} \gtrsim 0, \end{aligned} \quad (\text{C.16})$$

where $s = t_j^P + d_{ij}^* - t_i^{f*} \geq 0$.

Similar to 1.2), d_{ij}^* becomes larger slowly with speed proportional to $-\delta_+/w_{\max} \ll 1$ (quasi-equilibrium, not fixed point) unless $\delta_+ = 0$ (fixed point). The exact value is indeterminate, which depends on the ratio between delay-weight learning rates B_-/A_+ . But we can still give a range, i.e. d_{ij}^* is always larger than $T - t_j^P$ because $t_j^P + d_{ij}^* \geq t_i^{f*} > T$, and upper bounded by $d_{\max}$.

3) When $j \notin P$.

For the case of no spike for $j \notin P$, w'_{ij} recovers to $w_{ij}^* = 0$ if $w_{\text{out}} < 0$ or keeps its initial value if $w_{\text{out}} = 0$; the delays are kept at the initial value. In the following parts we only consider the case with random noise spikes for $j \notin P$.

3.1) For the weight, the spikes arrive randomly before or after the postsynaptic spike at t_i^{f*} . In this case,

$$\begin{aligned} \Delta w_{ij} &\propto \int_{-L_+}^{L_-} W(s) ds \\ &= A_+ \tau_+ \left[1 - \exp\left(\frac{-L_+}{\tau_+}\right) \right] - A_- \tau_- \left[1 - \exp\left(\frac{-L_-}{\tau_-}\right) \right] \\ &\approx A_+ \tau_+ - A_- \tau_- < 0. \end{aligned} \quad (\text{C.17})$$

So $\Delta w_{ij} < 0$. And w'_{ij} recovers to $w_{ij}^* = 0$ for $j \notin P$.

3.2) For the delay, we have

$$\begin{aligned}
 \Delta d_{ij} &\propto \int_{-\sigma_-}^{\sigma_+} D(s, w'_{ij}) ds \\
 &\approx \int_{-\sigma_-}^{\sigma_+} D(s, w^*_{ij}) ds = \int_{-\sigma_-}^{\sigma_+} D(s, 0) ds \\
 &= \left(\frac{3}{e} - 1 \right) \left[\frac{-\delta_+}{w_{\max}} B_+ \sigma_+^2 - \frac{w_{\max} + \delta_-}{w_{\max}} B_- \sigma_-^2 \right].
 \end{aligned} \tag{C.18}$$

Meanwhile the variables satisfy the following conditions: $0 \leq \delta_- \approx -\delta_+ \ll w_{\max}$, $\delta_- + \delta_+ \approx 0$, $B_+ \approx B_-$, and $\sigma_+ \approx \sigma_-$. Therefore Equation (C.18) continues as

$$\begin{aligned}
 \Delta d_{ij} &\propto \left(\frac{3}{e} - 1 \right) \left[\frac{-\delta_+}{w_{\max}} B_+ \sigma_+^2 - \frac{w_{\max} + \delta_-}{w_{\max}} B_- \sigma_-^2 \right] \\
 &\approx - \left(\frac{3}{e} - 1 \right) B \sigma^2 < 0,
 \end{aligned} \tag{C.19}$$

where B is a variable on the same scale of B_+ and B_- ; and σ is a variable on the same scale of σ_+ and σ_- . Therefore $\Delta d_{ij} < 0$. And d'_{ij} recovers to $d^*_{ij} = d_{\min}$ for $j \notin P$.

In conclusion, combining 1), 2) and 3), under a small distortion, $\{w'_{ij}, d'_{ij}\}_{j=1}^N$ recover to $\{w^*_{ij}, d^*_{ij}\}_{j=1}^N$, with T becoming smaller slowly with speed proportional to $\delta_-/w_{\max} \ll 1$ (quasi-equilibrium, not fixed point) unless $\delta_- = \delta_+ = 0$ (fixed point). ■

C.2 Proof Part 2

Proof: All quasi-equilibrium states are of the form of $\{w^*_{ij}, d^*_{ij}\}_{j=1}^N$.

We equivalently prove its contrapositive, i.e. if the weights and delays are not of the form of $\{w^*_{ij}, d^*_{ij}\}_{j=1}^N$, it is not a quasi-equilibrium state.

First, there is no equilibrium state when $0 < w^*_{ij} < w_{\max}$. This is because the weight learning window is half negative and half positive with no zero point within its effective length. When the weight is larger, the latency is reduced, so the weight is further

strengthened, which forms a positive feedback and finally leads to $w_{ij}^* = w_{\max}$. On the other hand, the weakened weight also follows a positive feedback and leads to $w_{ij}^* = 0$.

Next, for the delays, we show that $\{w_{ij}^*, d_{ij}\}_{j=1}^N$ is not a quasi-equilibrium state and converges to the form of $\{w_{ij}^*, d_{ij}^*\}_{j=1}^N$.

1) For synapses in the set $Q' = \{j \mid w_{ij} = w_{ij}^* = w_{\max}\}$, their weights are maximized. The only reason is that the spikes arrive regularly before the postsynaptic spike, i.e. $t_j^P + d_{ij} < t_i^f$ (assume negligible w_{in}). Similar to Equation (C.11), the relative difference between spike arrival times of any synapses j and k is diminishing to zero when $j \neq k$, $j \in Q$ and $k \in Q$, so d_{ij} cannot be kept and evolves to the form $d_{ij}^* = T - t_j^P$, where $T < t_i^f$.

2) For synapses in the set $\{j \mid w_{ij} = w_{ij}^* = 0\}$, their weights are minimized. There are two possible reasons (assume negligible w_{out}), i.e. the spikes arrive regularly after the postsynaptic spike, or arrive randomly.

In case of the spikes arriving regularly after the postsynaptic spike, i.e. $t_j^P + d_{ij} \geq t_i^f > T$, the synapse satisfies $j \notin Q'$, $j \in P$. Meanwhile d_{ij} becomes larger slowly due to $\Delta d_{ij} \gtrsim 0$ and satisfies $d_{ij} = d_{ij}^* \in (T - t_j^P, d_{\max}]$.

In case of the spikes arriving randomly, the synapse is not involved in the pattern, i.e. $j \notin P$. Since the integral of $\int_{-\sigma_-}^{\sigma_+} D(s, w_{ij}^* = 0) ds$ is negative, d_{ij} cannot be kept and evolves to $d_{ij} = d_{ij}^* = d_{\min}$.

In conclusion, if the weights and delays are not of the form of $\{w_{ij}^*, d_{ij}^*\}_{j=1}^N$, it is not a quasi-equilibrium state. Equivalently, all quasi-equilibrium states are of the form of $\{w_{ij}^*, d_{ij}^*\}_{j=1}^N$, where $Q = Q' = \{j \mid w_{ij} = w_{\max}\}$. Since Q is not unique, the quasi-equilibrium states can have different Q s. ■

D Parameter Values in Section 6.3

Table D.1: Joint STDP parameters

Symbol	Description	Value	Unit
τ_+, τ_-	weight learning window time const.	16.8, 33.7	ms
A_+, A_-	weight learning window amplitude	Section 6.3.1: 0.85×2^{-5} , 2^{-5} Section 6.3.2 and 6.3.3: from cross-validation, 2^{-6}	–
L_+, L_-	effective weight window length	$5\tau_+, 5\tau_-$	ms
$w_{\text{in}}, w_{\text{out}}$	nonhebbian homeostatic term	Section 6.3.1: $2^{-10}, 0$ Section 6.3.2 and 6.3.3: $0, -2^{-11}$	–
σ_+, σ_-	delay learning window time const.	16.8, 16.8	ms
B_+, B_-	delay learning window amplitude	Section 6.3.1: $0.5A_+, 0.5A_-$ Section 6.3.2 and 6.3.3: $B_+ = B_-$, from cross-validation	–
δ_+, δ_-	weight-related gain offset	Section 6.3.1: $-w_{\text{in}}, w_{\text{in}}$ Section 6.3.2 and 6.3.3: $0, 0$	–
$w_{\text{max}}, w_{\text{min}}$	maximum/minimum weight	1, 0	–
$d_{\text{min}}, d_{\text{max}}^{\text{readout}}$	minimum/maximum delay	1, 35	ms

Table D.2: SRM neuron model parameters

Symbol	Description	Value	Unit
$\tau_{\text{EPSP}}, \tau_{\text{IPSP}}$	PSP time constant	4, 4	ms
τ_{reset}	refractory time constant	12	ms
t_{ref}	absolute refractory period	5	ms
$U_{\text{rest}}, U_{\text{reset}}$	resting and reset potential	0, -5	mV
θ	firing threshold	from cross-validation	–

E Input Spike Trains in Section 6.3.1

The way to generate the input spike trains in Section 6.3.1 was similar to [124]. The spike trains had 100 spike sources and were 405s long. Because of memory constraints, for each spike source, a 135s sequence was first generated and repeated three times. This repetition had no impact on the results.

Each spike source independently emitted spikes according to an inhomogeneous Poisson process with the instantaneous firing rate r within $[0, 20]$ Hz and maximum change rate allowing to increase from 0 to 20Hz in 30ms and vice versa. For each spike source at each time step $dt = 1\text{ms}$, 1) it had a probability of $r \times dt$ to emit a spike at a random time within this dt time bin; 2) then r was updated by $dr \times dt$ and clipped in $[0, 20]$ Hz, where dr was updated by a random number from $[-133.3, 133.3]$ Hz/s. Such second order update was to ensure the smoothness of r with time. In addition, to make sure the patterns defined in the following paragraphs contained at least one spike from each spike source, a spike was added whenever a source had been silent for more than 30 ms, which resulted in a minimum firing rate of 33.3Hz.

Three patterns were embedded in the spike trains. The pattern occurrence times were randomly placed throughout the sequence, so that (1) each pattern had the equal occurrence possibility; (2) the total duration of patterns was 1/3 of the whole simulation; and (3) the minimum interval was 60ms. Finally the number of instantiations of the first pattern was 1434, second 1569 and third 1458, all together 4461. The spike sources in each pattern were arbitrarily chosen and grouped together for better visualization.

The major difference from [124] was that each pattern contained only one spike from each spike source. In our case, the patterns were defined by first randomly selecting a 30ms period from the spike train and then randomly taking one spike of each source within this period (“wavefront”). The reason for using 30ms time span was that it was shorter than $d_{\text{max}}^{\text{readout}} = 35\text{ms}$ so that the complementary delays were possible to cover the whole pattern. The reason for using “wavefront” patterns was that the delays com-

plementary to the spike timing can be obtained without ambiguity.

Once the patterns were defined, they were copy-pasted according to the pattern occurrence times. A Gaussian jitter with 1ms standard deviation was added before pasting. To keep firing rate stable, the closest spike in the pasted location was removed before pasting. Finally, the spike times were rounded to the precision of $dt = 1\text{ms}$ with duplicate spikes removed.

F Latency, Pattern Hit Rate and False Alarm Frequency in Section 6.3.1

Before computing the latency, we first defined a response interval for each pattern so that readout spike falling into it was treated as a valid response. For readout neurons under weight STDP, since the delays from spike sources to readout neurons were constant at $d_{\min} = 1\text{ms}$, the response interval was $[d_{\min}, d_{\min} + 30]\text{ms}$ after each pattern. For readout neurons under joint STDP, since the delays were variant, the response interval started at the $\lceil \theta \rceil$ -th early spike of pattern i plus $d_{\min} = 1\text{ms}$ after each pattern, where $\lceil \theta \rceil = 3$ was the minimum number of maximally strengthened spikes to trigger the postsynaptic spike; the response interval ended at the half of pattern duration ($30/2 = 15\text{ms}$) plus $d_{\max}^{\text{readout}} = 35\text{ms}$ after each pattern.

The latency was computed for each readout spike, by first finding its pattern response interval and then subtracting the pattern occurrence time. If it was outside any response interval, it was counted as a false alarm, with the latency denoted as 0. The false alarm frequency (Hz) was computed using the number of false alarms divided by a period T . The pattern hit rate was computed as the fraction of patterns with readout spike within its response interval over all patterns during the period T . Figure 6.4 was generated using a sliding window of $T = 10\text{s}$ and 5s overlap.

Publication List

Journal Papers:

1. Sun, H., Sourina, O., & Huang, G.-B. (2016). Learning polychronous neuronal groups using joint weight-delay spike-timing-dependent plasticity. *Neural Computation*, 28(10), 2181-2212.
2. Sun, H., Sourina, O., Huang, G.-B., Guo, H., Klanner, F., Denk, C., & Rasshofer, R.H. (2017). Assessing EEG-vigilance in open eye and situation-aware state using extreme learning machine autoencoder. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*. [major revision]
3. Sun, H., Jia, J., Goparaju, B., Huang, G.-B., Sourina, O., Bianchi, M., & Westover, M.B. (2017). Large-Scale Automated Sleep Staging. *Sleep*. [major revision]
4. Sun, H., Sourina, O., Huang, G.-B., Klanner, F., & Denk, C. (2016). Using joint weight-delay spike-timing dependent plasticity to find polychronous neuronal groups. *BMC Neuroscience*, 17(Suppl 1), 54.

Conference Papers:

1. Sun, H., Yang, Y., Sourina, O., & Huang, G.-B. (2015, July). Runtime detection of activated polychronous neuronal group towards its spatiotemporal analysis. In *2015 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

-
2. Sun, H., Sourina, O., Yang, Y., Huang, G.-B., Denk, C., & Klanner, F. (2015). Machine learning reveals different brain activities in visual pathway during TOVA test. In *Proceedings of ELM-2014 Volume 1* (pp. 245-262). Springer International Publishing.
 3. Sun, H., Yang, Y., Sourina, O., Huang, G.-B., Klanner, F., Denk, C., & Rasshofer, R. H. (2015, December). Estimating vigilance from EEG using manifold clustering guided by instantaneous lapse rate. In *Information, Communications and Signal Processing (ICICS), 2015 10th International Conference on* (pp. 1-5). IEEE.

References

- [1] L. M. Ward, “Attention,” *Scholarpedia*, vol. 3, no. 10, p. 1538, 2008.
- [2] H. J. Spiers and E. A. Maguire, “Thoughts, behaviour, and brain dynamics during navigation in the real world,” *Neuroimage*, vol. 31, no. 4, pp. 1826–1840, 2006.
- [3] Y. Dong, Z. Hu, K. Uchimura, and N. Murayama, “Driver inattention monitoring system for intelligent vehicles: A review,” *IEEE transactions on intelligent transportation systems*, vol. 12, no. 2, pp. 596–614, 2011.
- [4] B. Oken, M. Salinsky, and S. Elsas, “Vigilance, alertness, or sustained attention: physiological basis and measurement,” *Clinical Neurophysiology*, vol. 117, no. 9, pp. 1885–1901, 2006.
- [5] N. Tsuchiya and C. Koch, “Attention and consciousness,” *Scholarpedia*, no. 5, p. 4173, 2008.
- [6] A. R. Jensen, *Clocking the mind: Mental chronometry and individual differences*. Elsevier, 2006.
- [7] R. H. Wurtz and M. E. Goldberg, “The primate superior colliculus and the shift of visual attention,” *Invest Ophthalmol*, vol. 11, no. 6, pp. 441–450, 1972.
- [8] P. L. Nunez and R. Srinivasan, “Electroencephalogram,” *Scholarpedia*, vol. 2, no. 2, p. 1348, 2007.
- [9] M. N. Lees, J. Sparks, J. D. Lee, and M. Rizzo, “Change blindness, attention, and driving performance,” in *Proceedings of the International Driving Symposium on*

- Human Factors in Driving Assessment, Training, and Vehicle Design*, Portland, OR, 2007.
- [10] M. M. Sohlberg and C. A. Mateer, *Cognitive rehabilitation: An integrative neuropsychological approach*. Guilford Press, 2001.
- [11] T. A. Rihs, C. M. Michel, and G. Thut, “Mechanisms of selective inhibition in visual spatial attention are indexed by α -band EEG synchronization,” *European Journal of Neuroscience*, vol. 25, no. 2, pp. 603–610, 2007.
- [12] T. Womelsdorf and P. Fries, “The role of neuronal synchronization in selective attention,” *Current opinion in neurobiology*, vol. 17, no. 2, pp. 154–160, 2007.
- [13] K.-R. Müller, M. Tangermann, G. Dornhege, M. Krauledat, G. Curio, and B. Blankertz, “Machine learning for real-time single-trial EEG-analysis: From brain-computer interfacing to mental state monitoring,” *Journal of Neuroscience Methods*, vol. 167, no. 1, pp. 82–90, 2008.
- [14] U. Hegerl and T. Hensch, “The vigilance regulation model of affective disorders and adhd,” *Neuroscience & Biobehavioral Reviews*, vol. 44, pp. 45–57, 2014.
- [15] C. Alford, N. Rombaut, J. Jones, S. Foley, C. Idzikowski, and I. Hindmarch, “Acute effects of hydroxyzine on nocturnal sleep and sleep tendency the following day: A C-EEG study,” *Human Psychopharmacology: Clinical and Experimental*, vol. 7, no. 1, pp. 25–35, 1992.
- [16] H. Lee, C. Ekanadham, and A. Y. Ng, “Sparse deep belief net model for visual area v2,” in *Advances in neural information processing systems*, pp. 873–880, 2008.
- [17] A. Ng, “Sparse autoencoder,” *CS294A Lecture notes*, vol. 72, pp. 1–19, 2011.
- [18] E. Ben-Simon, I. Podlipsky, H. Okon-Singer, M. Gruberger, D. Cvetkovic, N. Intrator, and T. Hendler, “The dark side of the alpha rhythm: fMRI evidence for

- induced alpha modulation during complete darkness,” *European Journal of Neuroscience*, vol. 37, no. 5, pp. 795–803, 2013.
- [19] J. V. Haxby, B. Horwitz, L. G. Ungerleider, J. M. Maisog, P. Pietrini, and C. L. Grady, “The functional organization of human extrastriate cortex: a PET-rCBF study of selective attention to faces and locations,” *The Journal of Neuroscience*, vol. 14, no. 11, pp. 6336–6353, 1994.
- [20] C. L. L. Kasun, Y. Yang, G.-B. Huang, and Z. Zhang, “Dimension reduction with extreme learning machine,” *IEEE Transactions on Image Processing*, vol. 25, pp. 3906–3918, Aug 2016.
- [21] R. Gütig, R. Aharonov, S. Rotter, and H. Sompolinsky, “Learning input correlations through nonlinear temporally asymmetric hebbian plasticity,” *The Journal of neuroscience*, vol. 23, no. 9, pp. 3697–3714, 2003.
- [22] M. Bayati, A. Valizadeh, A. Abbassian, and S. Cheng, “Self-organization of synchronous activity propagation in neuronal networks driven by local excitation,” *Frontiers in computational neuroscience*, vol. 9, 2015.
- [23] S. T. Mueller and B. J. Piper, “The psychology experiment building language (pebl) and pebl test battery,” *Journal of neuroscience methods*, vol. 222, pp. 250–259, 2014.
- [24] L. M. Greenberg and I. D. Waldmant, “Developmental normative data on the test of variables of attention (TOVA),” *Journal of Child Psychology and Psychiatry*, vol. 34, no. 6, pp. 1019–1030, 1993.
- [25] M. A. Nazari, F. Wallois, A. Aarabi, and P. Berquin, “Dynamic changes in quantitative electroencephalogram during continuous performance test in children with attention-deficit/hyperactivity disorder,” *International journal of psychophysiology*, vol. 81, no. 3, pp. 230–236, 2011.

- [26] N. Ghimire, B. H. Paudel, R. Khadka, P. N. Singh, and A. Das, "Electroencephalographic changes during selective attention," *Asian Journal of Medical Sciences*, vol. 6, no. 2, pp. 51–56, 2014.
- [27] M. Besserve, M. Philippe, G. Florence, F. Laurent, L. Garnerio, and J. Martinerie, "Prediction of performance level during a cognitive task from ongoing EEG oscillatory activities," *Clinical neurophysiology*, vol. 119, no. 4, pp. 897–908, 2008.
- [28] K. Kaida, M. Takahashi, T. Åkerstedt, A. Nakata, Y. Otsuka, T. Haratani, and K. Fukasawa, "Validation of the karolinska sleepiness scale against performance and EEG variables," *Clinical Neurophysiology*, vol. 117, no. 7, pp. 1574–1581, 2006.
- [29] K. Kaida, T. ÅKERSTEDT, G. Kecklund, J. P. Nilsson, and J. Axelsson, "Use of subjective and physiological indicators of sleepiness to predict performance during a vigilance task," *Industrial health*, vol. 45, no. 4, pp. 520–526, 2007.
- [30] C. Zhao, M. Zhao, J. Liu, and C. Zheng, "Electroencephalogram and electrocardiograph assessment of mental fatigue in a driving simulator," *Accident Analysis & Prevention*, vol. 45, pp. 83–90, 2012.
- [31] M. V. Yeo, X. Li, K. Shen, and E. P. Wilder-Smith, "Can SVM be used for automatic EEG detection of drowsiness during car driving?," *Safety Science*, vol. 47, no. 1, pp. 115–124, 2009.
- [32] M. Simon, E. A. Schmidt, W. E. Kincses, M. Fritzsche, A. Bruns, C. Aufmuth, M. Bogdan, W. Rosenstiel, and M. Schrauf, "EEG alpha spindle measures as indicators of driver fatigue under real traffic conditions," *Clinical Neurophysiology*, vol. 122, no. 6, pp. 1168–1178, 2011.
- [33] G. S. Larue, A. Rakotonirainy, and A. N. Pettitt, "Driving performance impairments due to hypovigilance on monotonous roads," *Accident Analysis & Prevention*, vol. 43, no. 6, pp. 2037–2046, 2011.

- [34] C. Braboszcz and A. Delorme, “Lost in thoughts: neural markers of low alertness during mind wandering,” *Neuroimage*, vol. 54, no. 4, pp. 3040–3047, 2011.
- [35] L.-C. Shi and B.-L. Lu, “EEG-based vigilance estimation using extreme learning machines,” *Neurocomputing*, vol. 102, pp. 135–143, 2013.
- [36] F. T. Durso, A. R. Dattel, S. Banbury, and S. Tremblay, “Spam: The real-time assessment of sa,” *A cognitive approach to situation awareness: Theory and application*, vol. 1, pp. 137–154, 2004.
- [37] A. Delorme, T. Sejnowski, and S. Makeig, “Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis,” *Neuroimage*, vol. 34, no. 4, pp. 1443–1449, 2007.
- [38] T.-P. Jung, C. Humphries, T.-W. Lee, S. Makeig, M. J. McKeown, V. Iragui, and T. J. Sejnowski, “Removing electroencephalographic artifacts: comparison between ica and pca,” in *Neural Networks for Signal Processing VIII, 1998. Proceedings of the 1998 IEEE Signal Processing Society Workshop*, pp. 63–72, IEEE, 1998.
- [39] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, “A blind source separation technique using second-order statistics,” *IEEE Transactions on signal processing*, vol. 45, no. 2, pp. 434–444, 1997.
- [40] L. Shoker, S. Sanei, and J. Chambers, “Artifact removal from electroencephalograms using a hybrid BSS-SVM algorithm,” *IEEE Signal Processing Letters*, vol. 12, no. 10, pp. 721–724, 2005.
- [41] C. A. Joyce, I. F. Gorodnitsky, and M. Kutas, “Automatic removal of eye movement and blink artifacts from EEG data using blind component separation,” *Psychophysiology*, vol. 41, no. 2, pp. 313–325, 2004.
- [42] I. Gorodnitsky and A. Belouchrani, “Joint cumulant and correlation based signal separation with application to EEG data analysis,” in *Proc. 3-rd Int. Conf. on*

- Independent Component Analysis and Signal Separation, San Diego, California*, pp. 475–480, Citeseer, 2001.
- [43] K. Blinowska and J. Zygierecz, *Practical Biomedical Signal Analysis Using MATLAB®*. Series in Medical Physics and Biomedical Engineering, Taylor & Francis, 2011.
- [44] R. W. Picard, E. Vyzas, and J. Healey, “Toward machine emotional intelligence: Analysis of affective physiological state,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 10, pp. 1175–1191, 2001.
- [45] A. Raj, M. K. Bandil, D. Singh, and A. Wadhwani, “Auto-regressive based feature extraction and classification of epileptic EEG using artificial neural network,” *International Journal of Advanced Research in Computer Science*, vol. 4, no. 9, 2013.
- [46] C. J. Stam, “Nonlinear dynamical analysis of EEG and MEG: review of an emerging field,” *Clinical Neurophysiology*, vol. 116, no. 10, pp. 2266–2301, 2005.
- [47] V. Kulish, A. Sourin, and O. Sourina, “Human electroencephalograms seen as fractal time series: Mathematical analysis and visualization,” *Computers in biology and medicine*, vol. 36, no. 3, pp. 291–302, 2006.
- [48] H. Ocak, “Automatic detection of epileptic seizures in EEG using discrete wavelet transform and approximate entropy,” *Expert Systems with Applications*, vol. 36, no. 2, pp. 2027–2036, 2009.
- [49] T. Kreuz, “Measures of neuronal signal synchrony,” *Scholarpedia*, vol. 6, no. 12, p. 11922, 2011.
- [50] I. Koprinska, “Feature selection for brain-computer interfaces,” in *New frontiers in applied data mining*, pp. 106–117, Springer, 2010.
- [51] P. Pudil, F. Ferri, J. Novovicova, and J. Kittler, “Floating search methods for feature selection with nonmonotonic criterion functions,” in *In Proceedings of*

- the Twelveth International Conference on Pattern Recognition, IAPR*, Citeseer, 1994.
- [52] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.
- [53] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.
- [54] Q. Gu, Z. Li, and J. Han, "Linear discriminant dimensionality reduction," in *Machine Learning and Knowledge Discovery in Databases*, pp. 549–564, Springer, 2011.
- [55] A. Gupta, R. Agrawal, and B. Kaur, "Performance enhancement of mental task classification using EEG signal: a study of multivariate feature selection methods," *Soft Computing*, pp. 1–14, 2014.
- [56] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 8, pp. 1226–1238, 2005.
- [57] J. Jeong, J. C. Gore, and B. S. Peterson, "Mutual information analysis of the EEG in patients with alzheimer's disease," *Clinical Neurophysiology*, vol. 112, no. 5, pp. 827–835, 2001.
- [58] T. N. Lal, M. Schroder, T. Hinterberger, J. Weston, M. Bogdan, N. Birbaumer, and B. Scholkopf, "Support vector channel selection in BCI," *Biomedical Engineering, IEEE Transactions on*, vol. 51, no. 6, pp. 1003–1010, 2004.
- [59] T. Lan, D. Erdogmus, A. Adami, M. Pavel, and S. Mathan, "Salient EEG channel selection in brain computer interfaces by mutual information maximization," in *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*, pp. 7064–7067, IEEE, 2006.

- [60] T. Alotaiby, F. E. A. El-Samie, S. A. Alshebeili, and I. Ahmad, “A review of channel selection algorithms for EEG signal processing,” *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 1–21, 2015.
- [61] H. P. Martinez, Y. Bengio, and G. N. Yannakakis, “Learning deep physiological models of affect,” *IEEE Computational Intelligence Magazine*, vol. 8, no. 2, pp. 20–33, 2013.
- [62] W. Liu, W.-L. Zheng, and B.-L. Lu, “Multimodal emotion recognition using multimodal deep learning,” *arXiv preprint arXiv:1602.08225*, 2016.
- [63] P. Smolensky, “Information processing in dynamical systems: Foundations of harmony theory,” tech. rep., DTIC Document, 1986.
- [64] G. E. Hinton, “Training products of experts by minimizing contrastive divergence,” *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.
- [65] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
- [66] Y. Bengio, “Learning deep architectures for ai,” *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.
- [67] C. L. L. Kasun, H. Zhou, G.-B. Huang, and C. M. Vong, “Representational learning with extreme learning machine for big data,” *IEEE Intelligent Systems*, vol. 28, pp. 31–34, Dec 2013.
- [68] L. Younes, “Parametric inference for imperfectly observed gibbsian fields,” *Probability theory and related fields*, vol. 82, no. 4, pp. 625–645, 1989.
- [69] T. Tieleman, “Training restricted boltzmann machines using approximations to the likelihood gradient,” in *Proceedings of the 25th International Conference on Machine Learning*, pp. 1064–1071, ACM, 2008.
- [70] G. Borghini, L. Astolfi, G. Vecchiato, D. Mattia, and F. Babiloni, “Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of

- mental workload, fatigue and drowsiness,” *Neuroscience & Biobehavioral Reviews*, vol. 44, pp. 58–75, 2014.
- [71] C.-T. Lin, C.-H. Chuang, C.-S. Huang, S.-F. Tsai, S.-W. Lu, Y.-H. Chen, and L.-W. Ko, “Wireless and wearable EEG system for evaluating driver vigilance,” *Biomedical Circuits and Systems, IEEE Transactions on*, vol. 8, no. 2, pp. 165–176, 2014.
- [72] M. T. Peiris, P. R. Davidson, P. J. Bones, and R. D. Jones, “Detection of lapses in responsiveness from the EEG,” *Journal of neural engineering*, vol. 8, no. 1, p. 016003, 2011.
- [73] Z. Wang, R. M. Hope, Z. Wang, Q. Ji, and W. D. Gray, “Cross-subject workload classification with a hierarchical Bayes model,” *NeuroImage*, vol. 59, no. 1, pp. 64–69, 2012.
- [74] C. W. Anderson and M. Kirby, “Classification of electroencephalogram (EEG) signals for brainmachine interfaces,” *Computer Science Department, Colorado State University, USA*, 2003.
- [75] N. Bigdely-Shamlo, A. Vankov, R. R. Ramirez, and S. Makeig, “Brain activity-based image classification from rapid serial visual presentation,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 16, no. 5, pp. 432–441, 2008.
- [76] A. Delorme, M. Westerfield, and S. Makeig, “Medial prefrontal theta bursts precede rapid motor responses during visual selective attention,” *The Journal of Neuroscience*, vol. 27, no. 44, pp. 11949–11959, 2007.
- [77] C.-G. Bénar, D. Schön, S. Grimault, B. Nazarian, B. Burle, M. Roth, J.-M. Badier, P. Marquis, C. Liegeois-Chauvel, and J.-L. Anton, “Single-trial analysis of odd-ball event-related potentials in simultaneous EEG-fMRI,” *Human brain mapping*, vol. 28, no. 7, pp. 602–613, 2007.

- [78] A. Delorme, G. A. Rousselet, M. J.-M. Macé, and M. Fabre-Thorpe, “Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes,” *Cognitive Brain Research*, vol. 19, no. 2, pp. 103–113, 2004.
- [79] N.-Y. Liang, P. Saratchandran, G.-B. Huang, and N. Sundararajan, “Classification of mental tasks from EEG signals using extreme learning machine,” *International Journal of Neural Systems*, vol. 16, no. 01, pp. 29–38, 2006.
- [80] S. M. Hosni, M. E. Gadallah, S. F. Bahgat, and M. S. AbdelWahab, “Classification of EEG signals using different feature extraction techniques for mental-task BCI,” in *Computer Engineering & Systems, 2007. ICCES’07. International Conference on*, pp. 220–226, IEEE, 2007.
- [81] A. Gupta, R. Agrawal, and B. Kaur, “Performance enhancement of mental task classification using EEG signal: a study of multivariate feature selection methods,” *Soft Computing*, vol. 19, no. 10, pp. 2799–2812, 2015.
- [82] L.-C. Ku, T.-W. Shen, and S.-T. Chen, “Quantization on eeg covariance matrix images during tova attention test for depression disorder classification,” in *Systems, Signals and Image Processing (IWSSIP), 2012 19th International Conference on*, pp. 488–491, IEEE, 2012.
- [83] R. J. Barry, A. R. Clarke, S. J. Johnstone, C. A. Magee, and J. A. Rushby, “Eeg differences between eyes-closed and eyes-open resting conditions,” *Clinical Neurophysiology*, vol. 118, no. 12, pp. 2765–2773, 2007.
- [84] R. Esteller, J. Echauz, T. Tchong, B. Litt, and B. Pless, “Line length: an efficient feature for seizure onset detection,” in *Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE*, vol. 2, pp. 1707–1710, IEEE, 2001.
- [85] G.-B. Huang, L. Chen, and C.-K. Siew, “Universal approximation using incremental constructive feedforward networks with random hidden nodes,” *Neural Networks, IEEE Transactions on*, vol. 17, no. 4, pp. 879–892, 2006.

- [86] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 42, no. 2, pp. 513–529, 2012.
- [87] J. Ye, "Least squares linear discriminant analysis," in *Proceedings of the 24th international conference on Machine learning*, pp. 1087–1093, ACM, 2007.
- [88] M. Grant and S. Boyd, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control* (V. Blondel, S. Boyd, and H. Kimura, eds.), Lecture Notes in Control and Information Sciences, pp. 95–110, Springer-Verlag Limited, 2008. http://stanford.edu/~boyd/graph_dcp.html.
- [89] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1." <http://cvxr.com/cvx>, Mar. 2014.
- [90] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 10, pp. 1175–1191, 2001.
- [91] B. T. Jap, S. Lal, P. Fischer, and E. Bekiaris, "Using EEG spectral components to assess algorithms for detecting fatigue," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2352–2359, 2009.
- [92] P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion recognition from EEG using higher order crossings," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 14, no. 2, pp. 186–197, 2010.
- [93] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 278, no. 6, pp. H2039–H2049, 2000.
- [94] Y. Liu, O. Sourina, and M. K. Nguyen, "Real-time EEG-based human emotion recognition and visualization," in *Cyberworlds (CW), 2010 International Conference on*, pp. 262–269, IEEE, 2010.

- [95] L. G. Ungerleider and L. Pessoa, “What and where pathways,” *Scholarpedia*, vol. 3, no. 11, p. 5342, 2008.
- [96] M. A. Goodale and A. D. Milner, “Separate visual pathways for perception and action,” *Trends in neurosciences*, vol. 15, no. 1, pp. 20–25, 1992.
- [97] P. Sauseng and W. Klimesch, “What does phase information of oscillatory brain activity tell us about cognitive processes?,” *Neuroscience & Biobehavioral Reviews*, vol. 32, no. 5, pp. 1001–1013, 2008.
- [98] J.-J. Wang, F. Xue, and H. Li, “Simultaneous channel and feature selection of fused EEG features based on sparse group lasso,” *BioMed research international*, vol. 2015, 2015.
- [99] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York, 2009.
- [100] T. O. Zander and C. Kothe, “Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general,” *Journal of neural engineering*, vol. 8, no. 2, p. 025005, 2011.
- [101] L. He, D. Hu, M. Wan, Y. Wen, K. M. von Deneen, and M. Zhou, “Common Bayesian network for classification of EEG-based multiclass motor imagery BCI,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 46, no. 6, pp. 843–854, 2016.
- [102] M. J. Khan and K.-S. Hong, “Passive BCI based on drowsiness detection: an fNIRS study,” *Biomedical optics express*, vol. 6, no. 10, pp. 4063–4078, 2015.
- [103] J. S. Warm, R. Parasuraman, and G. Matthews, “Vigilance requires hard mental work and is stressful,” *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 50, no. 3, pp. 433–441, 2008.
- [104] J. Hasan, K. Hirvonen, A. Värri, V. Häkkinen, and P. Loula, “Validation of computer analysed polygraphic patterns during drowsiness and sleep onset,” *Elec-*

- troencephalography and clinical neurophysiology*, vol. 87, no. 3, pp. 117–127, 1993.
- [105] H. Danker-hopfe, P. Anderer, J. Zeitlhofer, M. Boeck, H. Dorn, G. Gruber, E. Heller, E. Loretz, D. Moser, S. Parapatics, *et al.*, “Interrater reliability for sleep scoring according to the rechtschaffen & kales and the new aasm standard,” *Journal of sleep research*, vol. 18, no. 1, pp. 74–84, 2009.
- [106] S. T. Mueller and B. J. Piper, “The psychology experiment building language (pebl) and pebl test battery,” *Journal of Neuroscience Methods*, vol. 222, pp. 250–259, 2014.
- [107] D. J. Thomson, “Spectrum estimation and harmonic analysis,” *Proceedings of the IEEE*, vol. 70, no. 9, pp. 1055–1096, 1982.
- [108] E. M. Palmer, T. S. Horowitz, A. Torralba, and J. M. Wolfe, “What are the shapes of response time distributions in visual search?,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 37, no. 1, p. 58, 2011.
- [109] K. Wang, J. Zhang, D. Li, X. Zhang, and T. Guo, “Adaptive affinity propagation clustering,” *arXiv preprint arXiv:0805.1096*, 2008.
- [110] B. J. Frey and D. Dueck, “Clustering by passing messages between data points,” *Science*, vol. 315, no. 5814, pp. 972–976, 2007.
- [111] K. McEvoy, K. Hasenstab, D. Senturk, A. Sanders, and S. S. Jeste, “Physiologic artifacts in resting state oscillations in young children: methodological considerations for noisy data,” *Brain imaging and behavior*, vol. 9, no. 1, pp. 104–114, 2015.
- [112] A. H. Duc, P. Bays, and M. Husain, “Eye movements as a probe of attention,” *Progress in brain research*, vol. 171, pp. 403–411, 2008.
- [113] S. Muthukumaraswamy, “High-frequency brain activity and muscle artifacts in meg/eeg: a review and recommendations,” *Frontiers in human neuroscience*, vol. 7, p. 138, 2013.

- [114] H. Almahasneh, W.-T. Chooi, N. Kamel, and A. S. Malik, “Deep in thought while driving: An eeg study on drivers cognitive distraction,” *Transportation research part F: traffic psychology and behaviour*, vol. 26, pp. 218–226, 2014.
- [115] C.-T. Lin, S.-A. Chen, L.-W. Ko, and Y.-K. Wang, “Eeg-based brain dynamics of driving distraction,” in *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pp. 1497–1500, IEEE, 2011.
- [116] L. F. Van Dillen and E. K. Papies, “From distraction to mindfulness: Psychological and neural mechanisms of attention strategies in self-regulation,” in *Handbook of biobehavioral approaches to self-regulation*, pp. 141–154, Springer, 2015.
- [117] P. O. Hoyer, “Non-negative matrix factorization with sparseness constraints,” *Journal of Machine Learning Research*, vol. 5, no. Nov, pp. 1457–1469, 2004.
- [118] C. Iber, *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*. American Academy of Sleep Medicine, 2007.
- [119] A. Rechtschaffen and A. Kales, “A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects,” 1968.
- [120] K. Young and M. Regan, “Driver distraction: A review of the literature,” *Distracted driving*. Sydney, NSW: Australasian College of Road Safety, pp. 379–405, 2007.
- [121] G. Buzsáki, “Neural syntax: cell assemblies, synapsembles, and readers,” *Neuron*, vol. 68, no. 3, pp. 362–385, 2010.
- [122] G.-q. Bi and M.-m. Poo, “Synaptic modification by correlated activity: Hebb’s postulate revisited,” *Annual review of neuroscience*, vol. 24, no. 1, pp. 139–166, 2001.
- [123] R. Guyonneau, R. VanRullen, and S. J. Thorpe, “Neurons tune to the earliest spikes through STDP,” *Neural Computation*, vol. 17, no. 4, pp. 859–879, 2005.

- [124] T. Masquelier, R. Guyonneau, and S. J. Thorpe, “Competitive STDP-based spike pattern learning,” *Neural computation*, vol. 21, no. 5, pp. 1259–1276, 2009.
- [125] T. Masquelier, E. Hugues, G. Deco, and S. J. Thorpe, “Oscillations, phase-of-firing coding, and spike timing-dependent plasticity: an efficient learning scheme,” *The Journal of neuroscience*, vol. 29, no. 43, pp. 13484–13493, 2009.
- [126] J. J. Hopfield *et al.*, “Pattern recognition computation using action potential timing for stimulus representation,” *Nature*, vol. 376, no. 6535, pp. 33–36, 1995.
- [127] M.-O. Gewaltig and M. Diesmann, “NEST (NEural Simulation Tool),” *Scholarpedia*, vol. 2, no. 4, p. 1430, 2007.
- [128] W. Gerstner, W. M. Kistler, R. Naud, and L. Paninski, *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press, 2014.
- [129] M. Carandini, J. C. Horton, and L. C. Sincich, “Thalamic filtering of retinal spike trains by postsynaptic summation,” *Journal of vision*, vol. 7, no. 14, pp. 20–20, 2007.
- [130] M. Gilson, T. Masquelier, and E. Hugues, “STDP allows fast rate-modulated coding with poisson-like spike trains,” *PLoS computational biology*, vol. 7, no. 10, p. e100223, 2011.
- [131] N. Kasabov, N. M. Scott, E. Tu, S. Marks, N. Sengupta, E. Capecci, M. Othman, M. G. Doborjeh, N. Murli, R. Hartono, *et al.*, “Evolving spatio-temporal data machines based on the neucube neuromorphic framework: design methodology and selected applications,” *Neural Networks*, vol. 78, pp. 1–14, 2016.
- [132] H. Hünig, H. Glünder, and G. Palm, “Synaptic delay learning in pulse-coupled neurons,” *Neural Computation*, vol. 10, no. 3, pp. 555–565, 1998.
- [133] C. W. Eurich, K. Pawelzik, U. Ernst, A. Thiel, J. D. Cowan, and J. G. Milton, “Delay adaptation in the nervous system,” *Neurocomputing*, vol. 32, pp. 741–748, 2000.

- [134] S. Tolnai, B. Englitz, J. Scholbach, J. Jost, and R. Rübsamen, “Spike transmission delay at the calyx of held in vivo: rate dependence, phenomenological modeling, and relevance for sound localization,” *Journal of neurophysiology*, vol. 102, no. 2, pp. 1206–1217, 2009.
- [135] D. J. Bakkum, Z. C. Chao, and S. M. Potter, “Long-term activity-dependent plasticity of action potential propagation delay and amplitude in cortical networks,” *PLOS one*, vol. 3, no. 5, pp. e2088–e2088, 2008.
- [136] J. H. Bollmann and B. Sakmann, “Control of synaptic strength and timing by the release-site ca^{2+} signal,” *Nature neuroscience*, vol. 8, no. 4, pp. 426–434, 2005.
- [137] J. A. Lorteije and J. G. G. Borst, “Contribution of the mouse calyx of held synapse to tone adaptation,” *European Journal of Neuroscience*, vol. 33, no. 2, pp. 251–258, 2011.
- [138] L. Buesing, J. Bill, B. Nessler, and W. Maass, “Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons,” *PLoS Comput Biol*, vol. 7, no. 11, p. e1002211, 2011.
- [139] T. M. Cover and J. A. Thomas, “Elements of information theory,” 2006.