

Advancing EEG-based brain-computer interfaces through performance modeling,
decoding validation, and methodological optimization: a multi-study approach

by

Nazmun Nahar Khan

B.Sc., Chittagong University of Engineering and Technology, 2013

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Mike Wieggers Department of Electrical and Computer Engineering
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2026

Abstract

Brain-computer interfaces represent a transformative technology that enables direct communication between neural activity and external devices. By decoding brain signals associated with movement intentions or cognitive states, BCIs offer potential solutions for individuals with severe paralysis or communication impairments. Interest from both researchers and clinicians in BCI applications has accelerated in recent years. Yet, despite this promise and decades of intensive research, BCIs have not transitioned successfully from controlled laboratory environments to everyday practice. Significant implementation barriers continue to limit widespread adoption; such as insufficient speed and accuracy, limited and unstandardized training data, and inconsistent evaluation practices. This dissertation examines critical obstacles hindering practical BCI deployment across two distinct application domains: P300-based communication spellers and EEG-based emotional state recognition. The work addresses these challenges through three strategic approaches focusing on efficient performance estimation in P300 spellers, comprehensive emotion dataset development, and rigorous methodological evaluation.

P300-based BCI spellers enable communication for individuals with severe motor impairments, yet performance validation creates a practical dilemma. Accurate assessment demands extensive testing sessions where users type 20 or more characters, consuming 4-20 minutes. For individuals who are already fatigued, these lengthy procedures are impractical. Moreover, even after substantial time investment, confidence intervals for accuracy estimates can span $\pm 23\%$, making it difficult to determine system reliability for a particular user. The first study addresses one of the key challenges in communication BCIs, particularly the slow validation and calibration of P300-based BCI speller systems. A predictive framework based on Classifier-Based Latency Estimation (CBLE) was used to estimate user accuracy from a small amount of typing data ($\sim 3-8$ characters). This approach provides faster and more reliable performance prediction, allowing quicker adaptation of the system to individual users.

Following this development, the second study expands the scope of BCI research toward affective computing. Current emotion recognition research relies on small, proprietary datasets that typically employ single-modality stimuli, limiting model generalizability across different emotion-elicitation scenarios. To address this data scarcity, a comprehensive multimodal database was created with recordings from 30 participants experiencing 240 stimuli across three categories: emotional pictures, facial expressions, and music. Each participant provided both SAM ratings for valence, arousal, and dominance, and discrete emotion labels for happiness, sadness, fear, disgust, surprise, and anger. This database provides nearly three times more stimuli than comparable datasets and represents the first to integrate all three stimulus modalities within a unified experimental framework per participant.

However, large-scale dataset availability alone cannot ensure progress if methodological practices remain flawed. The insights gained from dataset design motivated the third study, which systematically examined methodological flaws that affect the reliability of emotion recognition results. A systematic review of 101 studies using the widely-adopted DEAP dataset revealed that 87% contained critical errors, including data leakage from improper segmentation, biased feature selection, and flawed hyperparameter optimization. These methodological oversights artificially inflate reported accuracies, creating unrealistic expectations about system capabilities. Experimental validation demonstrated that such errors can boost performance estimates by up to 46 percentage points, explaining why laboratory results fail to translate to real-world deployment. The review establishes rigorous evaluation protocols and provides concrete guidelines to prevent common methodological pitfalls.

Together, these three studies form a continuous progression—from improving BCI speller performance modeling, enabling standardized affective datasets, and establishing rigorous evaluation protocols. Collectively, they contribute to building faster, more reliable, and methodologically sound EEG-based BCI systems for assistive and adaptive real-world applications.

Advancing EEG-based brain-computer interfaces through performance modeling,
decoding validation, and methodological optimization: a multi-study approach

by

Nazmun Nahar Khan

B.S., Chittagong University of Engineering & Technology, Bangladesh, 2013

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Mike Wieggers Department of Electrical and Computer Engineering
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2026

Approved by:

Co-Major Professor
David E. Thompson

Approved by:

Co-Major Professor
Steve Warren

Copyright

© Nazmun Nahar Khan 2026.

Abstract

Brain-computer interfaces represent a transformative technology that enables direct communication between neural activity and external devices. By decoding brain signals associated with movement intentions or cognitive states, BCIs offer potential solutions for individuals with severe paralysis or communication impairments. Interest from both researchers and clinicians in BCI applications has accelerated in recent years. Yet, despite this promise and decades of intensive research, BCIs have not transitioned successfully from controlled laboratory environments to everyday practice. Significant implementation barriers continue to limit widespread adoption; such as insufficient speed and accuracy, limited and unstandardized training data, and inconsistent evaluation practices. This dissertation examines critical obstacles hindering practical BCI deployment across two distinct application domains: P300-based communication spellers and EEG-based emotional state recognition. The work addresses these challenges through three strategic approaches focusing on efficient performance estimation in P300 spellers, comprehensive emotion dataset development, and rigorous methodological evaluation.

P300-based BCI spellers enable communication for individuals with severe motor impairments, yet performance validation creates a practical dilemma. Accurate assessment demands extensive testing sessions where users type 20 or more characters, consuming 4-20 minutes. For individuals who are already fatigued, these lengthy procedures are impractical. Moreover, even after substantial time investment, confidence intervals for accuracy estimates can span $\pm 23\%$, making it difficult to determine system reliability for a particular user. The first study addresses one of the key challenges in communication BCIs, particularly the slow validation and calibration of P300-based BCI speller systems. A predictive framework based on Classifier-Based Latency Estimation (CBLE) was used to estimate user accuracy from a small amount of typing data ($\sim 3-8$ characters). This approach provides faster and more reliable performance prediction, allowing quicker adaptation of the system to individual users.

Following this development, the second study expands the scope of BCI research toward affective computing. Current emotion recognition research relies on small, proprietary datasets that typically employ single-modality stimuli, limiting model generalizability across different emotion-elicitation scenarios. To address this data scarcity, a comprehensive multimodal database was created with recordings from 30 participants experiencing 240 stimuli across three categories: emotional pictures, facial expressions, and music. Each participant provided both SAM ratings for valence, arousal, and dominance, and discrete emotion labels for happiness, sadness, fear, disgust, surprise, and anger. This database provides nearly three times more stimuli than comparable datasets and represents the first to integrate all three stimulus modalities within a unified experimental framework per participant.

However, large-scale dataset availability alone cannot ensure progress if methodological practices remain flawed. The insights gained from dataset design motivated the third study, which systematically examined methodological flaws that affect the reliability of emotion recognition results. A systematic review of 101 studies using the widely-adopted DEAP dataset revealed that 87% contained critical errors, including data leakage from improper segmentation, biased feature selection, and flawed hyperparameter optimization. These methodological oversights artificially inflate reported accuracies, creating unrealistic expectations about system capabilities. Experimental validation demonstrated that such errors can boost performance estimates by up to 46 percentage points, explaining why laboratory results fail to translate to real-world deployment. The review establishes rigorous evaluation protocols and provides concrete guidelines to prevent common methodological pitfalls.

Together, these three studies form a continuous progression—from improving BCI speller performance modeling, enabling standardized affective datasets, and establishing rigorous evaluation protocols. Collectively, they contribute to building faster, more reliable, and methodologically sound EEG-based BCI systems for assistive and adaptive real-world applications.

Table of Contents

List of Figures	xii
List of Tables	xiv
Acknowledgements	xv
Dedication	xvii
1 Introduction	1
1.1 Fundamentals	2
1.1.1 Brain-Computer Interfaces (BCI)	2
1.1.2 P300 BCI Speller	3
1.1.3 Affective Brain-Computer Interface	5
1.1.4 Emotion Model	7
1.2 Motivation	9
1.3 Outline	12
2 P300-Based Brain-Computer Interface Speller Performance Estimation with Classifier- Based Latency Estimation	13
2.1 Introduction	14
2.2 Methods and Materials	16
2.2.1 Classifier	16
2.2.2 Classifier-Based Latency Estimation (CBLE)	17
2.2.3 Experimental Data	18
2.3 Protocol	20

2.4	Representative Results	27
2.5	Discussion	30
3	A Comprehensive EEG Database for Human Emotion Analysis Induced by Multimodal Stimuli	34
3.1	Introduction	35
3.2	Related Work	38
3.3	Experimental Design and Protocol	42
3.3.1	Stimuli Selection	42
3.3.2	Hardware Setup	44
3.3.3	Experimental Design	45
3.3.4	Participants	48
3.4	Data Record	49
3.5	Subjective Rating Analysis	53
3.5.1	SAM Ratings	54
3.5.2	Discrete Ratings	56
3.6	Affective State Estimation	58
3.6.1	Data Preprocessing	58
3.6.2	Feature Extraction	59
3.6.3	Classifier	59
3.6.4	Performance Metrics	61
3.7	Results	62
3.7.1	Binary Classification	62
3.7.2	Discrete Classification	64
3.7.3	Comparative Analysis	66
3.8	Conclusion	67

4	The Role of Review Process Failures in Affective State Estimation: An Empirical Investigation of the DEAP Dataset	70
4.1	Introduction	71
4.2	Materials and Methods	75
4.2.1	Review Paper Strategy	75
4.2.2	DEAP Emotion Dataset	80
4.2.3	Watermelon Dataset	81
4.2.4	Procedure of Experimental Analysis	82
4.3	Results	86
4.3.1	Study Selection	86
4.3.2	Prevalence of Methodological Pitfalls from Reviewed Papers	87
4.3.3	Additional Reporting and Reproducibility Concerns	90
4.3.4	Experimental Analysis: Segmentation	92
4.3.5	Experimental Analysis: Feature Selection Method	93
4.3.6	Experimental Analysis: Hyperparameter Optimization	94
4.4	Discussion	95
5	Conclusion and Future Work	100
5.1	Contribution	101
5.1.1	P300-Based Brain-Computer Interface Speller Performance Estimation	101
5.1.2	A Comprehensive EEG Database of Emotion Recognition	101
5.1.3	Methodological Assessment of Affective State Estimation Studies	102
5.2	Limitations	103
5.3	Future Work	103
	Bibliography	105
A	Graphical User Interface Developed for CBLE Performance Estimation	133

B Reuse Permissions from Publishers	135
---	-----

List of Figures

1.1	P300 BCI speller	4
1.2	P300 speller ERP	5
1.3	Ekman's six basic emotions	8
1.4	Russell's circumplex model of affect	9
2.1	Block diagram of the proposed protocol	21
2.2	Accuracy vs. vCBLE	27
2.3	RMSE of BCI accuracy and vCBLE	28
2.4	Comparison of models using RMSE	29
2.5	RMSE Values of vCBLE Models	30
2.6	Model comparison	31
2.7	Mean RMSE difference from the best model	32
3.1	Mean ratings of selected IAPS images	43
3.2	Mean ratings of selected music	45
3.3	Electrode placements for the International 10-20 system	46
3.4	Flow chart of experimental protocol	47
3.5	Self-assessment using SAMS Manikin	48
3.6	Structure of the DEEAMS dataset repository	52
3.7	Participants' ratings across three continuous dimensions	53
3.8	Mean rating distribution	55
3.9	The distribution of discrete emotion categorical ratings across all stimulus types . .	57
3.10	EEG data power spectral analysis	60

4.1	Systematic review selection process for EEG-based emotion recognition studies . .	76
4.2	EEG data collection and segmentation	81
4.3	Prevalence and distribution of methodological errors in EEG-based emotion recog- nition studies	89
4.4	Performance overestimation due to segmenting trial in EEG-based binary classifi- cation	92

List of Tables

2.1	Standard parameters for BrainInvaders and Michigan datasets	27
3.1	The summary of EEG based emotion databases	39
3.2	Number of POFA images	44
3.3	Self-assessment scale scoring	48
3.4	Summary of demographic data	49
3.5	Mean correlation coefficient between participants' rating and original ratings of the stimuli	56
3.6	Regular accuracy of binary classification	63
3.7	Balanced accuracy and macro F1 score of binary classification	64
3.8	Regular accuracy of discrete classification	65
3.9	F1 score for discrete classification	65
3.10	Comparison chart	66
4.1	Definitions of common methodological pitfalls	78
4.2	Comparison of feature selection strategies in binary classification across DEAP and watermelon datasets	93
4.3	Comparison of hyperparameter optimization strategies in binary classification across DEAP and watermelon datasets	94

Acknowledgments

This dissertation represents the outcomes of my six years of research work. None of this would have been possible without the unwavering support of many amazing people who stood by me every step of the way. Above all, I'm grateful to Almighty Allah for His blessings and guidance that brought me to this moment.

I extend my deepest gratitude to my advisor and co-major professor, Dr. David E. Thompson, for his guidance throughout these years. His knowledge and research experience provided invaluable advice during my academic pursuits. Beyond academics, I am sincerely grateful for his compassion and understanding during difficult moments in my personal life, which made this journey possible.

I sincerely appreciate my co-major professor, Dr. Steve Warren, and my committee members, Dr. Dwight Day and Dr. Doina Caragea, for their encouragement and time. Additionally, I appreciate the Mike Wieggers Department of Electrical and Computer Engineering for providing financial support through teaching assistantships.

I want to acknowledge my current and former labmates, Taylor Sweet, Chase A. Harvey, and Calder Knapp, for many helpful discussions and for making our lab meetings a place of learning and collaboration. Those lab meetings will be missed. A special thanks to Tegan Brandt and Quinlan Seeberger for their help with data collection. I also thank all the participants who volunteered for my BCI experiments over the years.

To my friends and the K-State community, thank you for everything. You showed up for me and my little one time and time again. All of you made K-State feel like home, even when I was thousands of miles from my own. I especially want to mention Syeda Tasnuva, a friend who lovingly cared for my daughter and was always just a phone call away whenever I needed it.

My family always stood by me and encouraged me. My mother, Ainun Nahar Khanam, was my first teacher and shaped the foundation of who I am today. My academic success is not due to my efforts, but to hers from the time I started kindergarten. My father, Md. Ismail, has been the pillar

of my life with his constant support and belief in me. All of my achievements are theirs, coming from their dedication, sacrifice, and blessings.

I would love to mention my two sisters, Sayema Sharmin and Nusrat Afrin, and my brother, Sharif Imtiaz, for their boundless love and encouragement throughout this long journey. I am also grateful to my in-laws, who spent several months with us, for their presence and blessings.

To my husband, Tareq Hossen, thank you for always being there for me. There were times when I felt like giving up, thinking I couldn't make this journey. You were there to encourage me and believe in me. Thank you for all those hours you spent with our daughter, so I could focus on my work. And to my beloved daughter, Numairah, your smiles and hugs kept me going. I am so grateful for your support during the three years since your birth, when my studies kept me away from you.

Dedication

To my parents, my spouse, and my daughter—
for their love, patience, and inspiration.

Chapter 1

Introduction

Brain-computer interfaces (BCIs) have demonstrated impressive capabilities in controlled laboratory environments. However, translating these systems into practical, everyday assistive devices remains challenging. The real-world success of BCIs depends on several critical factors, such as achieving sufficient speed and accuracy for practical operation, obtaining adequate validation data with appropriate sample sizes for training, addressing the complexity of data processing and decoding algorithms, and demonstrating valid and reproducible results across diverse datasets and user populations¹⁻³. This dissertation aims to address these implementation challenges and advance BCIs toward real-world applications. The investigation focuses on two prominent BCI domains: P300-based BCI spellers for communication and EEG-based affective state estimation systems.

This chapter provides background and context for the dissertation. Section 1.1 discusses the fundamental principles of P300-based BCI spellers and affective BCIs, including their potential applications. Section 1.2 presents the motivation and research objectives of this dissertation. Finally, the organization of this dissertation is outlined in Section 1.3.

1.1 Fundamentals

1.1.1 Brain-Computer Interfaces (BCI)

A brain-computer interface (BCI) is a communication system that allows people to interact with computers or external devices using only their brain signals⁴. Unlike traditional assistive technologies that rely on peripheral nerves and muscles, BCIs form direct communication pathways from brain activity. This makes BCIs particularly valuable for individuals with severe motor disabilities, including amyotrophic lateral sclerosis (ALS), disorders of consciousness (DoC), brainstem stroke, spinal cord injury, and cerebral palsy⁵⁻⁷, who may have lost the ability to speak or move but retain cognitive function.

The fundamental principle behind BCIs is straightforward: the system records brain activity, extracts meaningful patterns from these signals, and translates them into commands that control external devices. For example, a person who cannot speak might use a BCI to select letters on a computer screen simply by thinking about them, without any physical movement. The BCI accomplishes this by monitoring specific brain signals that reflect the user's intentions and converting these signals into actionable commands.

Among various methods for recording brain activity, including functional magnetic resonance imaging (fMRI), magnetoencephalography (MEG), and electrocorticography (ECoG), electroencephalography (EEG) has become the most widely used approach for BCI applications⁸. EEG is non-invasive and measures electrical activity on the scalp surface using sensors placed at specific locations on the head. This means it does not require surgery or penetration through the skull to access brain tissue. This makes EEG significantly safer and more accessible than invasive recording methods (e.g. ECoG).

EEG-based BCIs can utilize different types of brain signals as input. Four main categories of brain activity patterns have been commonly employed in BCI systems. First, visual evoked potentials (VEP) are brain responses triggered by visual stimuli⁹. Second, slow cortical potentials represent gradual voltage shifts in brain activity¹⁰. Third, sensorimotor rhythms, particularly mu and beta frequency bands, reflect brain activity related to motor planning and execution¹¹. Finally, event-related potentials (ERPs)¹², including the P300 component, are brain responses to specific events or stimuli.

1.1.2 P300 BCI Speller

Among various BCI speller approaches, the P300-based BCI speller introduced by Farwell and Donchin (1988)¹² has become one of the most widely adopted communication protocols.

The P300, also called the P3 component, is one of the most studied brain signals in neuroscience and BCI research. First discovered by Sutton et al. in 1965¹³, the P300 appears as a positive deflection in the EEG signal when someone sees something unexpected or important. This electrical response appears in the brain roughly 300 milliseconds after seeing a rare or meaningful event^{12;14;15}.

The P300 signal is usually strongest at certain locations on the scalp, particularly at sites called Fz, Cz, and Pz (located along the center and back of the head). At these locations, the brain produces a clear positive peak when someone notices a target item, but shows much smaller responses to non-target items. Researchers typically use something called the oddball paradigm to trigger P300 responses¹⁴. In this approach, rare target items appear occasionally among frequent non-target items. The brain naturally responds more strongly to these rare, important events.

However, capturing P300 signals is quite difficult. The P300 itself is very weak, producing only 5 to 10 microvolts of electrical activity¹⁶. Meanwhile, normal background brain activity is

A	B	C	D	E	F
G	H	I	J	K	L
M	N	O	P	Q	R
S	T	U	V	W	X
Y	Z	1	2	3	4
5	6	7	8	9	0

Figure 1.1: Matrix layout of the classic Farwell and Donchin row-column P300 speller¹². The grid contains all 26 English letters and 10 digits (0–9), where rows and columns are flashed sequentially to identify the target symbol based on P300 responses.

much stronger, measuring 10 to 100 microvolts. This makes the P300 hard to spot in a single recording. Therefore, averaging across multiple repetitions is commonly employed to improve the signal-to-noise ratio¹⁷. This averaging process strengthens the consistent P300 pattern while reducing random noise¹⁶.

In the first BCI speller system¹², a matrix of symbols (typically comprising all 26 letters of the alphabet and 10 digits, from 0 to 9, arranged in a 6×6 grid) is displayed on a computer screen (see Fig.1.1). During operation, the rows and columns of this matrix flash or intensify one at a time in a random sequence. Each flash typically lasts 100 milliseconds, followed by a brief blank interval of approximately 75 milliseconds before the next intensification occurs¹².

The user focuses attention on the specific symbol they wish to communicate and mentally counts how many times that symbol flashes. This focused attention creates an oddball paradigm¹⁴: the flashes containing the target symbol are rare and task-relevant events, while all other flashes are frequent non-target stimuli. When the row or column containing the target character intensifies, this infrequent target event elicits a characteristic P300 wave in the user’s EEG signal, occurring

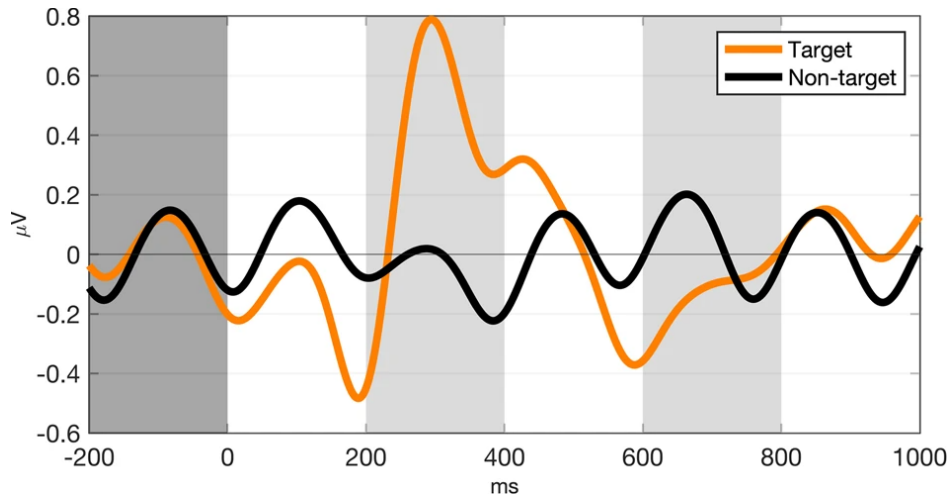


Figure 1.2: P300 speller ERP¹⁸. The grand-averaged ERP response at the midline (Fz, Cz, and Pz) during P300 speller test sessions.

approximately 300 milliseconds after the flash. In contrast, the remaining 10 rows and columns that do not contain the target symbol produce significantly weaker or no P300 responses.

By detecting and analyzing these P300 patterns across multiple intensifications, the BCI system can identify which row and which column consistently generate the strongest responses. The intersection of the detected row and column reveals the user's intended character, enabling communication without any physical movement or muscle control.

Fig.1.2 illustrates an example of a grand-averaged P300 waveform recorded during a P300 speller experiment¹⁸. The graph displays brain activity recorded from central scalp locations, comparing responses to target characters versus non-target characters. When flashes contain the user's intended character (shown in orange), a distinct positive peak emerges between 200 and 400 milliseconds. Flashes without the target character (shown in black) remain near baseline.

1.1.3 Affective Brain-Computer Interface

Emotions, or affective states, represent complex combinations of physiological, behavioral, and cognitive responses to internal or external stimuli. They play a vital role in human decision-

making, perception, and communication¹⁹. The growing need to build systems that can sense, interpret, and respond to human emotions has driven research in this area. Emotion-sensitive systems have gained attention in assistive technology, adaptive learning, neurorehabilitation, and human-computer interaction. By integrating emotion recognition, machines can adjust responses based on user emotional states, making interactions more natural and personalized. An adaptive learning platform could simplify tasks when detecting frustration. A therapeutic system could monitor emotional well-being to guide mental health interventions.

To achieve such adaptability, emotions must be measured objectively and reliably. Early attempts relied on behavioral cues such as facial expressions^{20;21}, vocal tone^{22;23}, pupil dilation²⁴, speech²⁵, and posture^{21;26}. Studies later expanded to peripheral physiological signals like electrodermal activity, heart rate variability (HRV), PPG, blood volume pulse, respiration, and skin conductance¹⁹. These approaches provided valuable information but faced limitations. For example, facial expressions and vocal patterns could be consciously controlled or masked, making them unreliable indicators of true emotional states. Peripheral physiological signals lacked specificity since similar patterns could occur under different conditions. A racing heart could indicate excitement, fear, or physical exertion.

These limitations drove researchers toward neurophysiological signals that could directly capture brain activity associated with emotional processing. The brain, after all, is where emotions originate²⁷. The introduction of electroencephalography (EEG) provided a way to observe neural oscillations linked to emotion with millisecond precision. Compared to imaging methods like functional magnetic resonance imaging (fMRI), EEG offered greater temporal resolution, reliability, lower cost, portability, and easier integration into interactive systems²⁸. Most importantly, EEG signals directly measure electrical activity generated by neurons and are difficult to disguise,

providing a more accurate reflection of an individual's real emotional state²⁹. Over the last two decades, this transition marked a fundamental shift from behavioral and peripheral physiological measures toward brain-based emotion decoding.

The integration of emotional decoding into BCI systems gave rise to **affective brain-computer interfaces (aBCI)**. Unlike traditional BCIs that focus on motor imagery or cognitive control, affective BCIs aim to detect and interpret a user's emotional state directly from brain signals. These systems enable adaptive interaction by linking neural activity to emotional feedback, allowing computers or assistive devices to respond appropriately to the user's affective condition. A typical aBCI system follows a structured process. The user is presented with emotional or neutral stimuli, and EEG signals are recorded during the experience. The data are then preprocessed to remove noise and artifacts. Features are extracted from specific frequency bands or connectivity patterns, and classification algorithms are applied to identify the user's emotional state. The output can be used to adjust interfaces, regulate environments, or deliver neurofeedback for emotion regulation training. Common features include power spectral density, asymmetry indices between hemispheres, and connectivity measures that capture how different brain regions interact³⁰. Machine learning methods like support vector machines, neural networks, and deep learning approaches have all been applied to classify these features into emotional states³⁰.

1.1.4 Emotion Model

Emotions have been described using two primary theoretical perspectives: discrete models and dimensional models. The discrete emotion model, originally proposed by Paul Ekman^{32;33}, classifies human emotions into a small set of fundamental categories that are universally recognizable across cultures. Ekman's framework identifies six basic emotions: happiness, sadness, anger,

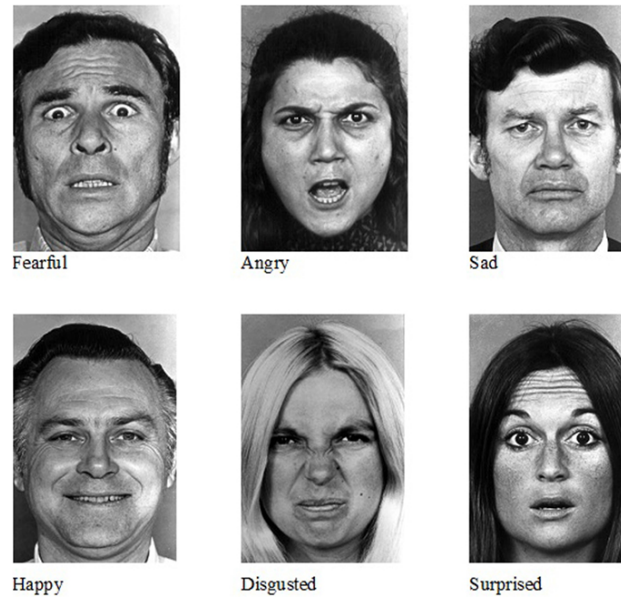


Figure 1.3: Ekman's six basic emotions³¹.

fear, disgust, and surprise. They serve as the foundation for most emotion recognition research based on facial expression or behavioral cues. This categorical approach simplifies emotion classification and aligns with intuitive understanding. However, it struggles to capture the full spectrum of emotional experiences and nuanced transitions between states.

To address this limitation, researchers have developed dimensional models, such as Russell's Circumplex Model of Affect³⁴, which represent emotional experiences within continuous dimensions. In this model, emotions are placed on a two-dimensional plane defined by valence and arousal. Valence ranges from unpleasant (negative) to pleasant (positive). Arousal ranges from passive (low) to active (high), describing energy intensity. For example, happiness occupies the high arousal and positive valence quadrant. Sadness falls in the low arousal and negative valence. Anger represents high arousal with negative valence. A third dimension, dominance, is sometimes added to describe the degree of control or influence a person feels in a situation. Both fear and anger are negative, high-arousal emotions, but they differ in dominance. A fearful person feels less control, while an angry person feels more empowered. Together, these axes allow for a more

flexible representation of emotional states, making dimensional models particularly suitable for affective BCI applications. The choice between discrete and dimensional models depends on the specific application and required specificity.

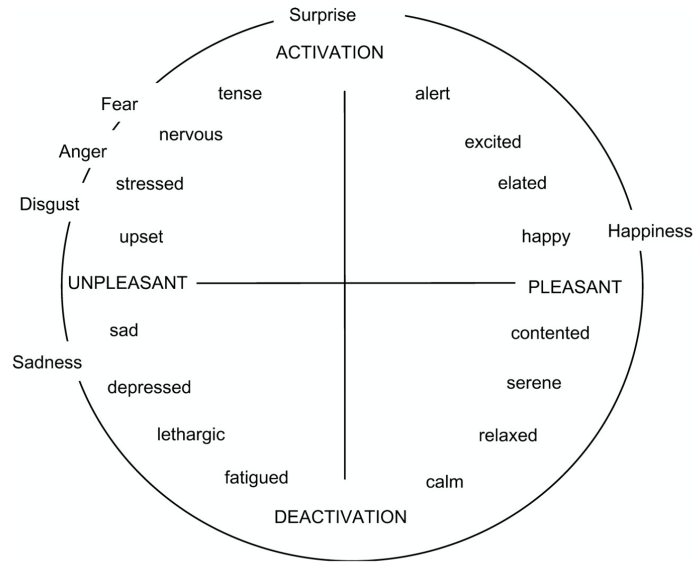


Figure 1.4: Russell's circumplex model of affect³⁴. This figure illustrates the valence (pleasant to unpleasant) and arousal (activation to deactivation) axes that organize common emotions in a continuous space.

1.2 Motivation

The development of effective brain-computer interface systems requires careful offline analysis and optimization before deployment in real-time applications. This preliminary phase involves initial calibration or adjustment of the BCI, data analysis, parameter tuning, and performance validation, which are essential to ensure reliable online BCI operation¹. P300-based BCI spellers represent one of the most promising assistive communication technologies for individuals with severe motor impairments. Despite considerable research progress, P300 speller systems face a critical performance validation dilemma. Obtaining high accuracy and sufficient speed is essential

for practical real-world communication. However, EEG-based BCI spellers typically achieve information transfer rates (ITR) of approximately 1.0 bits per second (bps) due to computational processing limitations^{4,35}. For example, the P300 speller developed by Farwell and Donchin (1988) achieves spelling rates of up to five characters per minute (~ 0.5 bps)¹². Other recent improvements have achieved rates of approximately 5.45 symbols per minute³⁶. This slow operation creates a significant burden during the validation phase, when obtaining reliable performance estimates requires extensive testing sessions. Users who need these systems the most cannot maintain long validation sessions. This dissertation addresses this challenge by developing a protocol that predicts P300 BCI speller accuracy from substantially fewer typed characters.

While P300 spellers face validation challenges, another critical area of BCI research, affective state estimation, presents its own set of implementation barriers. Research on EEG-based aBCI is growing rapidly because of their potential to support adaptive communication tools, mental-health assessment and intervention, and more responsive human–computer interaction. Despite growing research interest, several fundamental challenges hinder the translation of aBCIs from laboratory demonstrations to practical real-world applications.

First, acquiring high-quality EEG data requires specialized equipment, making data collection both expensive and time-consuming. The scarcity of standardized, large-scale datasets has made objective performance comparison across different methodological approaches nearly impossible, as researchers often work with small, proprietary datasets that vary widely in experimental design and quality.

Second, the type of stimulus used to evoke emotions may influence the resulting neural responses and, consequently, the generalizability of emotion recognition models. Researchers have employed diverse stimuli, including static images, video clips, music, and virtual reality scenarios, to elicit

emotional responses. However, a model trained to recognize emotions evoked by a single modality stimulus (such as pictures from the International Affective Picture System) may fail when presented with emotions elicited through different modalities like music or real-world social interactions.

Third, and perhaps most critically, affective BCIs suffer from unstable and often unreliable performance. Most existing emotion recognition algorithms operate in an offline mode, where EEG data are analyzed after collection³⁷. While reported classification accuracies for distinguishing two or more discrete emotional states typically range from 70–90% in controlled studies³⁸, these results often do not translate to real-world applications. This performance degradation stems from methodological issues throughout the emotion recognition pipeline, from data pre-processing and feature extraction to model validation. Many published studies inadvertently introduce bias during the training phase through improper preprocessing, feature extraction, and selection, or cross-validation procedures. Models that appear successful in offline analysis frequently fail when deployed in online, real-time applications. Because their training process contaminated the evaluation with information that would not be available during real-world deployment.

Given these interconnected challenges, advancing EEG-based emotion recognition toward practical implementation requires more than incremental improvements in classification algorithms. Success demands: (1) comprehensive, publicly available datasets with sufficient sample sizes and multiple stimulus modalities to enable robust model development and fair performance comparison; and (2) rigorous, standardized methodological practices that ensure reported performance metrics genuinely reflect system capabilities in practical deployment scenarios. This dissertation takes steps toward addressing these challenges and limitations. First, a comprehensive multimodal emotion database is developed to provide the research community with standardized, large-scale data across multiple stimulus modalities. Second, a systematic evaluation of methodological practices

in affective BCI research is conducted to identify common errors and establish rigorous evaluation protocols.

1.3 Outline

The remainder of this dissertation is organized into four chapters, each addressing a distinct yet complementary component of advancing EEG-based BCIs through performance modeling, dataset design, and methodological validation. The first research focus, presented in Chapter 2, investigates the estimation of P300-based BCI speller accuracy using relatively small test samples. The chapter details the experimental protocol, signal processing pipeline, validation methods, and implementation results. Chapter 3 introduces a new multimodal EEG dataset developed to study human emotional responses. This dataset integrates visual and auditory stimuli to elicit a wide range of affective states. The chapter describes participant recruitment, experimental design, data recording, preprocessing, and feature extraction procedures. Statistical analyses of participant ratings and classification results across the valence-arousal dimensions are also presented to establish the reliability of the proposed dataset. The next phase, presented in Chapter 4, explores methodological reliability in EEG-based emotion recognition studies. Focusing on studies that rely on the DEAP dataset, this chapter conducts a structured analysis to identify recurring methodological flaws in affective state estimation. By applying multiple commonly used experimental approaches, the analysis reveals common methodological flaws that can inflate performance metrics and proposes practical guidelines for reproducible affective BCI studies. Chapter 5 summarizes the major contributions of the dissertation. It also discusses the current limitations and proposes future work that can contribute to implementing real-time BCI systems more effectively and reliably in practical applications.

Chapter 2

P300-Based Brain-Computer Interface Speller Performance Estimation with Classifier-Based Latency Estimation

Copyright notice: This chapter substantially reproduces content from "P300-Based Brain-Computer Interface Speller Performance Estimation with Classifier-Based Latency Estimation", published in the Journal of Visualized Experiments (JoVE) in 2024³⁹. The material is included in this dissertation with authorization from JoVE, which owns all copyright to the work. While the text has been revised for dissertation format, the core methodology, results, and analysis remain consistent with the published work.

This chapter focuses on implementing a method for estimating P300-based BCI speller accuracy using a small testing dataset. Modern P300 speller systems remain slow, making data collection for validation time-consuming for both end users and experimenters. P300 spellers typically operate around 1–5 characters per minute, and estimating accuracy with 5% resolution requires 20 characters (4–20 minutes). Even with this substantial time investment, confidence bounds can reach $\pm 23\%$

depending on observed accuracy. To address this challenge, Classifier-Based Latency Estimation (CBLE) was applied to predict a user's P300 speller accuracy from relatively few characters ($\sim 3-8$) of typing data.

The content of this chapter is organized into five sections. Section 2.1 provides an introduction to the performance estimation challenge, and Section 2.2 presents the methods and materials. The step-by-step protocol for implementing the CBLE Performance Estimation GUI is detailed in Section 2.3. Section 2.4 demonstrates the representative results obtained from both datasets. Concluding remarks and implications of the findings are given in Section 2.5.

2.1 Introduction

Brain-computer interfaces (BCIs) are a noninvasive technology that allows individuals to communicate through machines directly without regard for physical limitations imposed by the body. BCI can be utilized as an assistive device operated directly by the brain. BCI uses the brain activity of a user to determine if the user intends to choose a certain key (letter, number, or symbol) displayed on the screen⁴⁰. In a typical computer system, a user physically presses the intended key on a keyboard. However, in a BCI system with a visual display, the user needs to focus on the desired key. Then, BCI will select the intended key by analyzing the measured brain signals¹. The activity of the brain can be measured using various techniques. Though there are competing BCIs technologies, electroencephalogram (EEG) is considered a leading technique due to its noninvasive nature, high temporal resolution, reliability, and relatively low cost²⁸.

Applications of BCI include communication, device control, and also entertainment⁴¹⁻⁴⁴. One of the most active BCI application areas is the P300 speller, which was introduced by Farwell and Donchin¹². The P300 is an event-related potential (ERP) produced in response to the recognition of

a rare but relevant stimulus⁴⁵. When a person recognizes their target stimulus, they automatically produce a P300. The P300 is an effective signal for a BCI because it conveys the participant’s recognition of the target event without requiring an outward response⁴⁶.

The P300 BCI has attracted researchers from computer science, electrical engineering, psychology, human factors, and various other disciplines. Advances have been made in signal processing, classification algorithms, user interfaces, stimulation schemes, and many other areas^{47–52}. However, regardless of the research area, the common thread in all of this research is the necessity of measuring the BCI system performance. This task typically requires the generation of a test data set. This necessity is not limited to research; eventual clinical application as an assistive technology will likely require individual validation sets for each end user to ensure the system can generate reliable communication.

Despite the considerable research applied toward the P300 BCI, the systems are still quite slow. While the majority of people are able to use a P300 BCI⁵³, most P300 Spellers produce text on the order of 1–5 characters per minute. Unfortunately, this slow speed means that generating test data sets requires substantial time and effort for participants, experimenters, and eventual end users. Measuring BCI system accuracy is a binomial parameter estimation problem, and many characters of data are necessary for a good estimate.

In earlier research, we argued that Classifier-Based Latency Estimation (CBLE) can be used to estimate BCI accuracy^{54–56}. Here, this work presents a protocol that estimates BCI performance from a small dataset using CBLE. As a representative analysis, the number of characters was varied to make predictions of the overall performance of an individual. In this work, we applied and evaluated the complete methodology on two datasets—an in-lab Michigan P300 speller dataset⁵⁵ and the publicly available Brain Invaders (bi2014a) dataset⁵⁷. For both example datasets, the root

mean square error (RMSE) for vCBLE and actual BCI accuracy were computed. The results indicate that the RMSE from vCBLE predictions, using its fitted data, was consistently lower than the accuracy derived from 1 to 7 tested characters.

We also developed a Graphical User Interface (GUI) called “CBLE Performance Estimation” for the implementation of the proposed methodology. Example code is also provided that operates on the MATLAB platform. The example code performs all of the steps applied in the GUI, but the steps are provided to assist the reader with adapting to a new dataset. This code employs a publicly available dataset “Brain Invaders calibration-less P300-based BCI using dry EEG electrodes Dataset (bi2014a)” to evaluate the proposed method⁵⁷.

2.2 Methods and Materials

2.2.1 Classifier

To estimate the presence or absence of the P300 ERP, most classifiers use a binary classification model, which involves assigning a binary label (e.g., “presence” or “absence”) to each trial or epoch of EEG data. The general equation used by most classifiers can be expressed as:

$$\hat{y}(x) = w^T \cdot f(x) + b \quad (2.1)$$

where $\hat{y}$ is called the classifier’s score, which represents the probability of the P300 response being present, x is the feature vector extracted from the EEG signal, and b is a bias term⁵⁴. The function f is a decision function that maps the input data to the output label, and is learned from a set of labeled training data using a supervised learning algorithm⁵⁴. During training, the classifier is trained on a labeled dataset of EEG signals, where each signal is labeled as either having a P300

response or not. The weight vector and bias term are optimized to minimize the error between the predicted output of the classifier and the true label of the EEG signal. Once the classifier is trained, it can be used to predict the presence of the P300 response in new EEG signals.

Different classifiers can use different decision functions, such as linear discriminant analysis (LDA), step-wise linear discriminant analysis (SWLDA), least squares (LS), logistic regression, support vector machines (SVM), or neural networks (NNs). The least squares classifier is a linear classifier that minimizes the sum of squared errors between the predicted class labels and the true class labels. This classifier predicts the class label of a new test sample using the following equation:

$$\hat{y}(x) = \text{sign}(\hat{w} \cdot x) \quad (2.2)$$

where the sign function returns +1 if the product is positive and −1 if it is negative, and the weight vector $\hat{w}$ is obtained from the feature set of the training data (x) and class labels (y) using the following equation:

$$\hat{w} = (X^T X)^{-1} X^T y \quad (2.3)$$

2.2.2 Classifier-Based Latency Estimation (CBLE)

CBLE is a strategy for evaluating latency variation by exploiting the classifier’s temporal sensitivity⁵⁵. While the conventional approach to P300 classification involves using a single time window that is synchronized with each stimulus presentation, the CBLE method involves creating multiple time-shifted copies of the post-stimulus epochs. Then it detects the time shift that results in the maximum score in order to estimate the latency of the P300 response^{54;55}.

As mentioned, traditional P300 classification relies on a single fixed time window, typically ranging from 0 to 800 ms following stimulus presentation⁴⁷. The CBLE method employs a different

strategy. For instance, with a 1kHz sampling rate, one window might span from -1 to 799 ms, while another spans from 1 to 801 ms⁵⁸. The classifier is applied to each time-shifted window independently. This produces a separate score for each time shift. The time shift that produces the maximum classifier score is identified as the classifier-based latency estimate (CBLE) for that trial. The statistical variance of these latency estimates across multiple target stimuli is called vCBLE⁵⁵. This metric provides a quantitative measure of P300 latency jitter for a given session or participant.

Previous research has shown that latency jitter significantly impacts P300 speller performance. The CBLE method provides a way to quantify this jitter. It also enables investigation of the relationship between jitter and classification accuracy. Thompson et al. (2012) have revealed a strong negative correlation between vCBLE and BCI accuracy⁵⁵. Participants with higher latency variability tend to achieve lower spelling accuracy. This relationship suggests that CBLE-derived metrics can predict BCI performance. Such prediction capability potentially enables faster assessment of system reliability without requiring extensive typing sessions.

2.2.3 Experimental Data

Michigan Dataset

The CBLE approach was also implemented on a Michigan dataset, which contained data from 40 participants⁵⁵. Here, the data from eight participants had to be discarded because their tasks were incomplete. EEG was recorded with a 16-channel g.USBamp sampled at 256 Hz (after hardware filtering). Electrode locations were F3, Fz, F4, T7, C3, Cz, C4, T8, CP3, CP4, P3, Pz, P4, PO7, Oz, and PO8 (10–20 system). The right and left mastoids served as reference and ground, respectively. The entire study required three visits from each participant. On the first day, each participant typed a 19-character training sentence, followed by three 23-character testing sentences on Days 1, 2, and

3. Online classification was performed using least squares (LS). The training file was used to create a participant-specific classifier that was used in all three sessions. In this paradigm, a row-column P300 speller (6×6) was run in BCI2000. The keyboard included 36 characters, which were grouped into six rows and six columns. Each row or column was flashed for 31.25 milliseconds with an interval of 125 milliseconds between flashes. Between characters, a 3.5 s pause was provided.

BrainInvaders

The Brain Invaders (bi2014a)⁵⁷ dataset includes EEG data from 64 participants. Recordings used a research-grade g.USBamp (g.tec) with 16 dry g.Sahara electrodes. The electrodes were placed at Fp1, Fp2, F5, AFz, F6, T7, Cz, T8, P7, P3, Pz, P4, P8, O1, Oz, and O2 following the 10–10/10–20 system. The reference was the right earlobe, and the ground was Fz. Signals were sampled at 512 Hz, with no online digital filtering. Event triggers from the Brain Invaders software were synchronized with the EEG via a USB digital-to-analog converter connected to the g.USBamp trigger channel. Participants played up to three game sessions of Brain Invaders, each session having 9 levels in the game. Gameplay proceeded level-by-level: once the target symbol was “destroyed,” a reward screen appeared and the next level started. Participants had up to eight attempts per level, after which the level restarted if all attempts were missed. Level duration was variable (on average about five minutes). To avoid frustration, the experimenter could terminate a level if no control was achieved after 10 minutes. Data collection continued until all levels were completed or the participant lost all control over the BCI system. The Brain Invaders interface included 36 symbols that flashed in 12 groups of six aliens. According to the Brain Invaders P300 paradigm, a repetition was created by 12 flashes, one for each group. Out of these 12 flashes, two flashes contained the target symbol (known as target flashes), while the remaining 10 flashes did

not contain the target symbol (known as non-target flashes). More information on this paradigm can be found in the original reference⁵⁷.

2.3 Protocol

The “CBLE Performance Estimation” GUI was applied to two datasets: the “BrainInvaders” dataset and the Michigan dataset. For the “BrainInvaders” dataset, the data collection was approved by the Ethical Committee of the University of Grenoble Alpes⁵⁷. Michigan data were collected under the University of Michigan Institutional Review Board approval⁵⁶. Data were analyzed under the Kansas State University exempt protocol 7516. If collecting new data, follow the user’s IRB-approved process for collecting informed consent. Here, the proposed protocol is evaluated using offline analysis of previously-recorded, de-identified data and therefore did not require additional informed consent.

The graphical user interface (GUI) included in this paper is proficient in managing two distinct dataset formats. The first format is associated with the BCI2000 software, while the second format is referred to as the “BrainInvaders” dataset. In order to utilize the “Brain Invaders” format, data must be preprocessed as described in step 1 of the protocol section. However, when dealing with the “BCI2000” dataset format, step 1 can be omitted. Fig. 2.1 shows the flow chart of the proposed method. The detailed procedure is described below.

1. Data preparation

- 1.1. BrainInvaders only: Generate the input data file in “.mat” file format that can be used with the “CBLE performance Estimation” graphical user interface (GUI). For a sample script, refer to Supplementary Coding File 2.

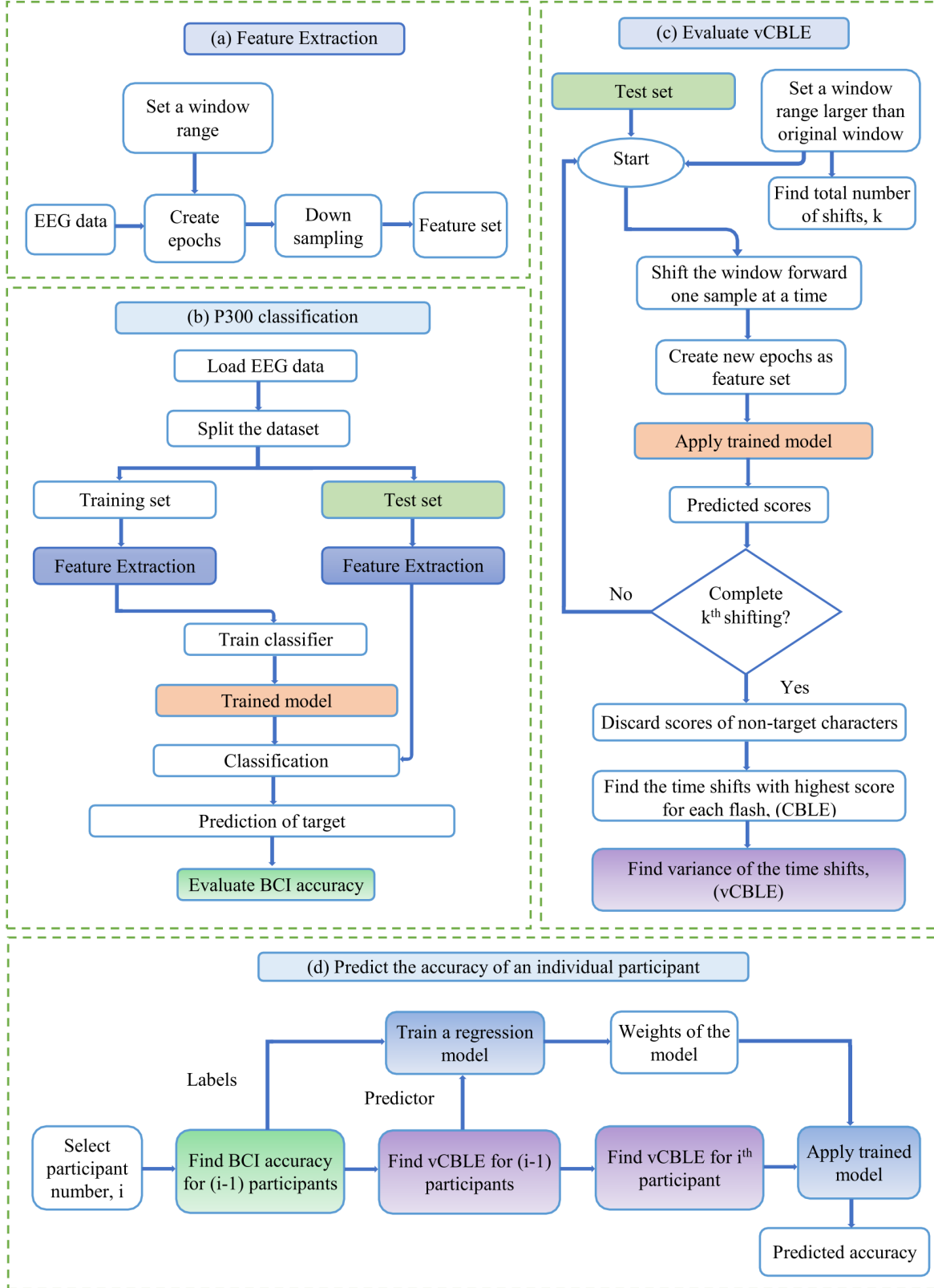


Figure 2.1: **Block diagram of the proposed protocol.** (a) Data preprocessing and feature extraction. (b) P300 classification. (c) Evaluation of vCBLE. (d) Predicting the accuracy of an individual.

NOTE: Each data file consists of a two-dimensional matrix comprised of rows that represent observations recorded at distinct time samples. The matrix columns numbered from 2 to 17 are recordings derived from 16 EEG electrodes. The first column of the matrix denotes the timestamp of each observation, while column 18 encompasses information related to experimental events. In column 19, there are mostly zeros, but when a non-Target (or Target) flash starts, the numbers change to one (or two) at that specific time. A detailed description can be found in the reference⁵⁷.

2. Download and install the GUI package

2.1. Download and install the “CBLE Performance Estimation” GUI.

3. Store the dataset in a subfolder of the GUI location

3.1. Ensure that the dataset folder remains within the same directory as the GUI.

3.2. For instance, create a new folder and place the “CBLE Performance Estimation” GUI inside it. Keep all the datasets in a subfolder within “CBLE GUI” named “Dataset.”

4. Open the installed GUI

4.1. Open MATLAB, change the current directory to folder where you placed the GUI, click “APPS” tab and select “MY APPS”.

4.2. Under the “MY APPS” tab, select “CBLE Performance Estimation”.

5. Choose dataset format

5.1. Select a dataset format from the dropdown “Select dataset format”.

6. Load EEG data file

6.1. Click on the "Select input folder" button to choose the directory where the dataset is located.

6.2. Observe the count of data files present in that selected folder.

NOTE: In the "Brain Invaders" format, each participant is represented by a single data file. Therefore, the total number of data files indicates the number of participants in the study. However, this is not the case for the "BCI2000" format, as each participant may have multiple train and test files.

7. Set the parameters

7.1. Type the number of participants that the user intends to use for the estimation process in the "No of participants" text box.

7.2. BrainInvaders only: Specify the sampling rate of the dataset.

NOTE: BCI2000 files include the sample rate.

7.3. Choose a decimation value to downsample the dataset to approximately 20 Hz in order to improve classification performance⁴⁷. For example, if the sampling frequency is 256 Hz, then select a decimation value of 13.

7.4. Specify the time window for the classification in milliseconds.

NOTE: The recommended initial window size is specified, allowing the starting point to vary from 0 to 100 ms and the ending point from 700 to 800 ms. However, it is important to avoid making the window size excessively large to prevent overlapping with another P300 event.

7.5. Define the shift window for CBLE in milliseconds.

NOTE: The 'shift window' refers to the expanded time range that the CBLE method searches to find the P3 response. The classifier will be applied to the epoch starting at the first element of the shift window. The classifier is then sequentially applied to epochs starting one sample later, until the epoch would extend outside the shift window. Thus, the shift window must be larger than the original window; empirically, values less than 100 ms from each side perform well. In any case, the margin should be held to less than half the ISI.

- 7.6. BC2000 only: Enter the length of the subject ID indicated in the dataset files within the "ID length" field.

NOTE: The GUI expects the first sub-len characters of the filenames to encode the subject ID.

- 7.7. BCI2000 only: In the "Channel ID" field, indicate either the total number of channels or specify the specific channel numbers to be used for the analysis.

- 7.8. Click the "Set parameters" button to set all the parameters required for the analysis.

8. BrainInvaders only: Split the dataset into training and test set

- 8.1. Select a number of targets that represents the size of the training set. The remaining portion of the dataset will be considered as the test dataset.

NOTE: To ensure proper training of the model, it is essential to have a sufficiently large training sample. The recommended minimum training sample size is 20, though this may vary depending on the overall dataset size. If regression errors occur during training session, it is advisable to increase the training sample size.

- 8.2. Press the "Split the dataset" button to divide the dataset into the training and test sets.

NOTE: Each participant will have an equal amount of training data. However, the number of test data may not be equal for all participants due to the possibility of multiple attempts during the task. Consequently, the total number of targets or flashes presented may vary from person to person.

9. Train a model with training dataset

Note: Step 9.1 is applicable for “Brain Invaders” format and step 9.2 is applicable for “BCI2000” format.

9.1. BrainInvaders only: Click on the “Train a model” button to apply linear regression on the training dataset using equation 2 for training a classifier model.

9.2. BCI2000 only: Indicate training and testing filenames along with their data format (.dat) to distinguish the training and testing files from all files. Then, click on the “Train a model” button to apply linear regression on the training dataset.

10. Predict the accuracy of the test set

10.1. Click on Predict accuracy to apply the trained classifier model to the test feature set and predict accuracy using Equation 1.

11. Get X-target accuracies

11.1. . Select a maximum target number, X, to consider in the test set.

11.2. BCI2000 only: Select a test file number if user has multiple test files.

11.3. Press “Find X target accuracy.”

12. Calculate vCBLE

- 12.1. Click on the "Find vCBLE" button to get the vCBLE for all targets.
13. Calculate Root mean square error (RMSE) of BCI accuracy and vCBLE
 - 13.1. Click on the "Calculate RMSE" button to calculate the RMSE between both predictions based on vCBLE with BCI accuracy, and X-target accuracy with BCI accuracy.
14. Visualize the results of the analysis
 - 14.1. Click on the "Accuracy vs vCBLE" button to observe the relation between total accuracy and total vCBLE for all participants.
 - 14.2. Press on the "RMSE of BCI & vCBLE" button to show the RMSE curve of BCI accuracy and vCBLE.
15. Prediction of the performance of an individual participant
 - 15.1. To predict the accuracy of an individual participant, place the subject id in "Sub ID".

NOTE: Here, the dataset of all participants, excluding the test participant, will be used to train a linear regression model. The vCBLE scores of all other participants and their corresponding test accuracies will be utilized as predictors and labels, respectively, for the classifier.
 - 15.2. Select a target number, n. The prediction will be made based on the accuracy of n testing characters.
 - 15.3. Click the "Predict" button to get the predicted accuracy of the test participant.

Table 2.1: Standard parameters for BrainInvaders and Michigan datasets

Dataset name	Subject number	ID length	Channel ID	Sampling rate	Decimation value	Original window	CBLE window	Training sample no	Target number, X
Brain Invaders	64	N/A	[1:16]	512	26	[100, 600]	[0, 700]	20	10
Michigan	32	4	[1:16]	256	13	[4, 804]	[-100, 900]	N/A	20

2.4 Representative Results

The proposed protocol has been tested on two different datasets: “BrainInvaders” and the Michigan dataset. These datasets are already introduced briefly in the Introduction section. The parameters used for these two datasets are mentioned in Table 2.1. Fig. 2.2– 2.4 depict the findings obtained using the “BrainInvaders” dataset, whereas Fig. 2.5– 2.7 demonstrate the results achieved from the Michigan dataset.

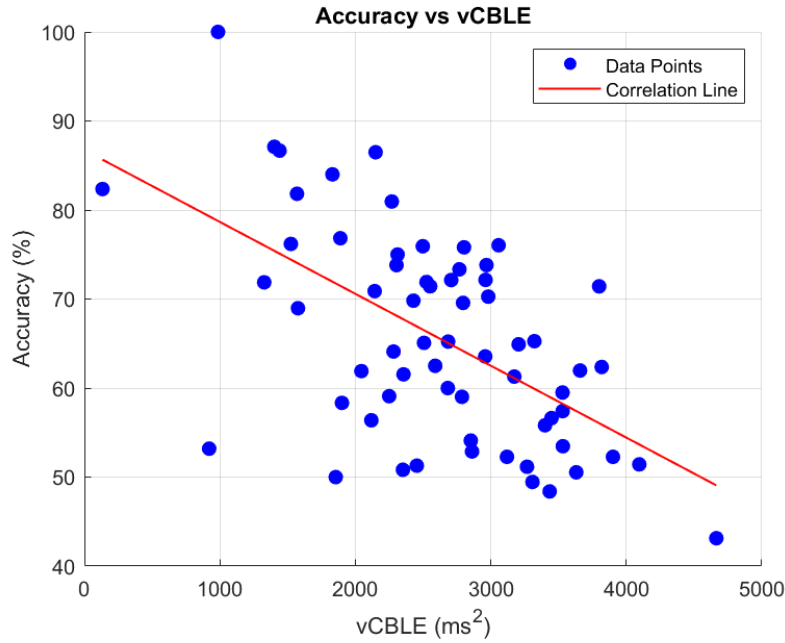


Figure 2.2: **Accuracy vs. vCBLE.** BCI accuracy plotted against vCBLE using “bi2014a” dataset. It shows a high negative correlation between accuracy and vCBLE.

The “BrainInvaders” dataset has 64 participants. Fig. 2.2 presents the relationship between BCI accuracy and vCBLE of all 64 participants. It shows that vCBLE is highly negatively correlated

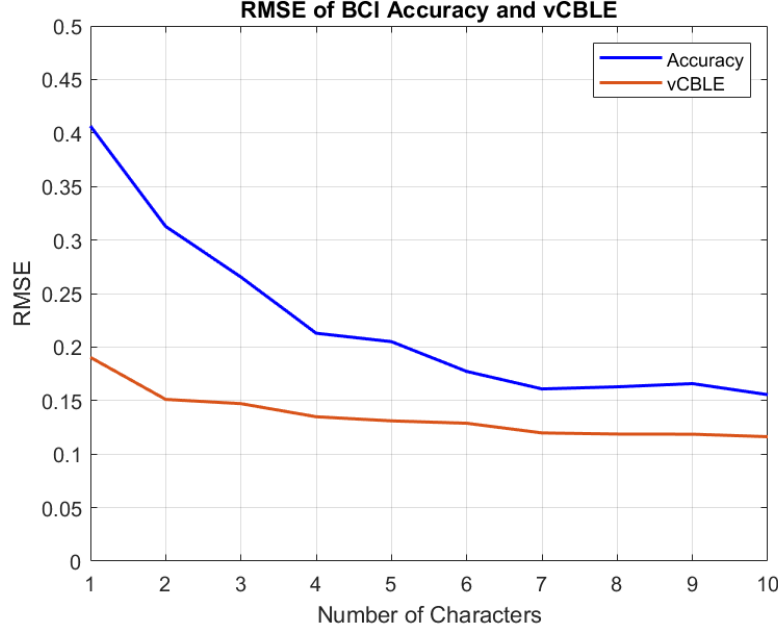


Figure 2.3: **RMSE of BCI accuracy and vCBLE.** The RMSE of vCBLE and accuracy were plotted against different test dataset sizes (1-10) using the “bi2014a” dataset. Overall, vCBLE performs better than BCI accuracy.

with BCI accuracy, although a few outliers are observed. Fig. 2.3 illustrates the RMSE of vCBLE and actual accuracy when the prediction was made based on the accuracy of testing characters. It shows the evidence that the RMSE for this prediction, based on the fit obtained by vCBLE, was lower than the accuracy based on any number of testing characters from 1 to 10. For the “BrainInvaders” dataset, vCBLE is capable of predicting BCI accuracy using only 7 characters. In Fig. 2.4, the prediction was made from the vCBLE of test sets with 2, 5, 7, and 10 characters, respectively. Here, a leave-one-participant-out approach was employed in the regression analysis to predict the accuracy of each individual participant. BCI accuracy and vCBLE were estimated over 100 repetitions. The lower and upper bounds are ± 2 standard deviations from the mean. All four conditions indicate that there is minimal variance observed when the number of participants in the training set exceeds 10. We conclude that about 10 individuals are required to build the regression model for the relationship between vCBLE and accuracy for a particular experimental paradigm.

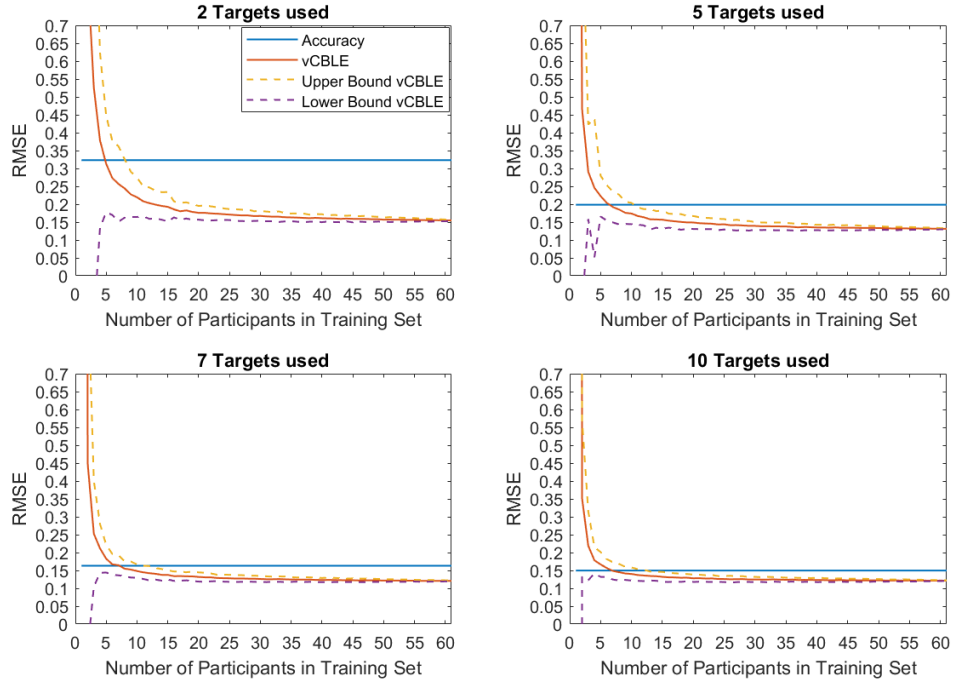


Figure 2.4: **Comparison of models using RMSE.** These models are built while predictions are performed from different sizes of test characters. Top left: 2 targets; top right: 5 targets; bottom left: 7 targets; bottom right: 10 targets.

In the second example, the Michigan dataset has 32 participants, in which all typed one training sentence on Day 1 and three testing sentences on Days 1, 2, and 3. The test sentences were 23 or 24 characters in length, and many participants made additional selections to correct errors made during online operation. In Fig. 2.5, it can be seen that the vCBLE model performed better when the training and testing datasets were collected on the same day. In fact, this prediction resulting from the fit provided by vCBLE resulted in a lower RMSE than the accuracy based for any number of testing characters from 1-20 when the training data and testing data were collected on the same day. Fig. 2.6 shows that overall, the vCBLE fit had a lower RMSE when the test included less than six characters. Additionally, it can be seen from Fig. 2.7 that the RMSE of the vCBLE accuracy estimation only decreases about 0.025 between three characters and the optimal number of characters used. This implies that there is not much benefit to collecting more than three

characters for the small test set.

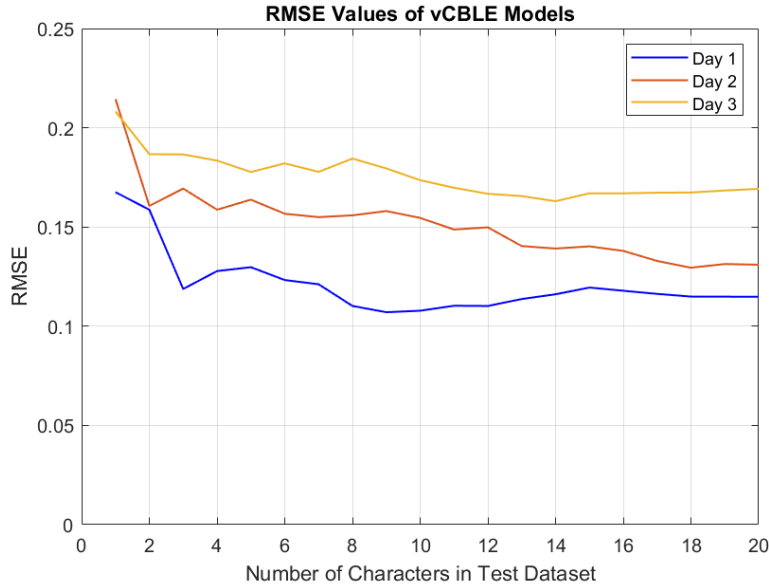


Figure 2.5: **RMSE Values of vCBLE Models.** A separate model was built to predict accuracy over three different days using the Michigan dataset. The RMSE values for models built using different test dataset sizes are shown.

2.5 Discussion

This article outlined a method for estimating BCI accuracy using a small P300 dataset. Here, the current protocol was developed based on the “bi2014a” dataset, although the efficacy of the protocol was confirmed on two different datasets. To successfully implement this technique, it is crucial to establish certain variables, such as the epoch window for the original data, the window for time shifting, the down-sampling ratio, and the size of both the training and testing datasets. These variables are determined by the characteristics of the dataset being used, including the number of targets or characters, the number of sequences, and the total number of participants.

The findings of the “bi2014a” dataset indicate that vCBLE’s prediction exhibits superior performance compared to character-level BCI accuracy for all test conditions (less than 10 characters),

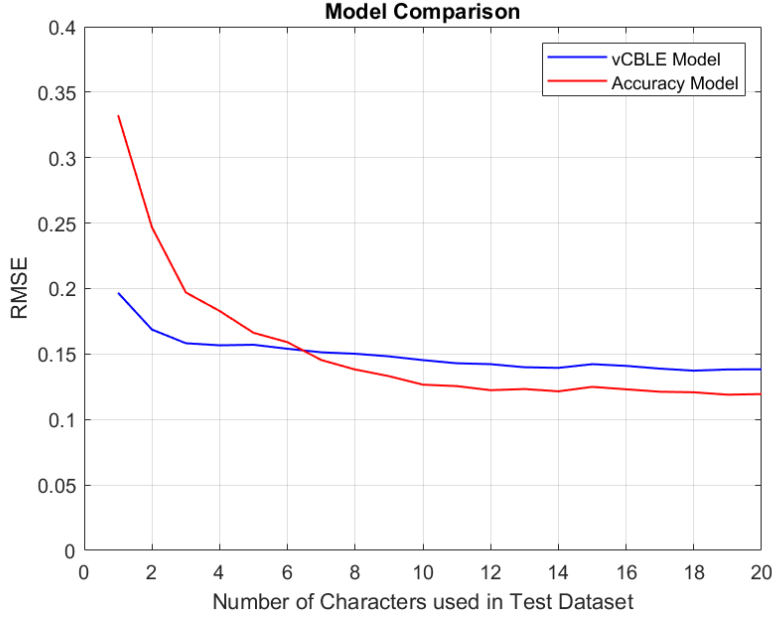


Figure 2.6: **Model comparison.** The mean of the RMSE over three days was computed for the vCBLE and the accuracy models using Michigan data.

which involve test datasets containing one to ten characters. However, when the test dataset comprises more than seven targets, the performance of vCBLE shows minimal variance. Results from the Michigan data suggest that using vCBLE to predict same-day performance will outperform the accuracy-based estimation if the test data set is less than six characters. Interestingly, increasing the amount of data used to build this model only improves marginally after the first few characters of data. Overall, this would imply that it is not necessary to collect large amounts of data to predict same-day accuracy.

According to the outcomes of the “bi2014a” dataset, it can be suggested that a minimum of 10 participants is necessary to construct a classifier model that can forecast an individual’s BCI accuracy. However, this also depends on the number of characters or the number of sequences used in both the training and testing phases. The ” bi2014a ” dataset includes several participants who had a relatively small number of total targets. It is worth mentioning that the vCBLE

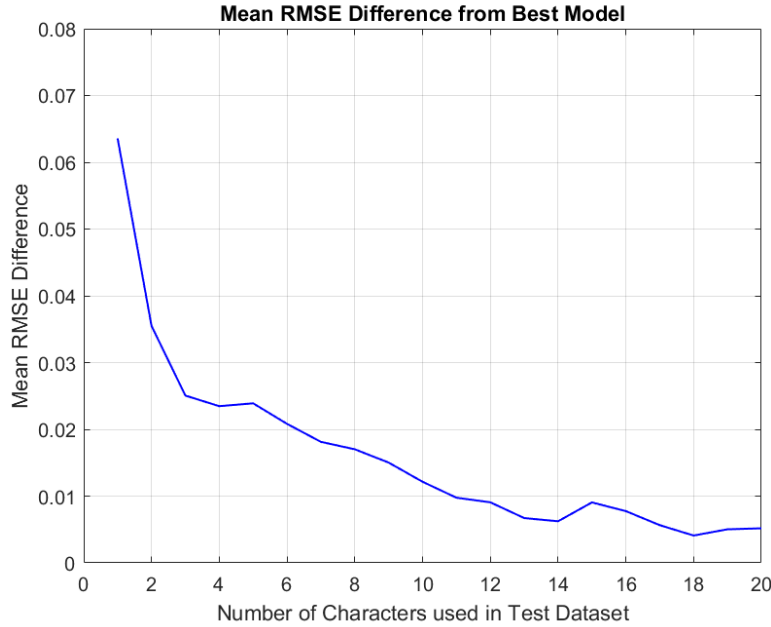


Figure 2.7: **Mean RMSE difference from the best model.** For each day, the minimum RMSE value was subtracted from each character’s RMSE value. The mean was calculated over the three days. This graph represents the average performance of a model using a certain data set size compared to the best model.

prediction method has already been successfully tested on small-size datasets consisting of 32 and 9 participants, respectively, and has demonstrated effective performance^{54:55}. However, these datasets have a relatively larger number of total targets, such as 19 characters in the training session and a minimum of 23 characters in the testing session.

There are a few limitations to be aware of when applying this method. From the analysis of the Michigan dataset, the vCBLE model seems to perform worse when the training and test data are collected on different days. Also, this method requires multiple participants to build a custom model for a given dataset. Moreover, the proposed method has been tested on four classifiers, including a least-squares classifier, stepwise linear discriminant analysis, support vector machine (SVM), and space autoencoder (SAE)^{54:55}. However, the protocol should be applicable to any time-sensitive classifier. Despite these limitations, the potential time savings to the research and

clinical communities warrant further investigation and application.

Acknowledgments

The data used for representative results were collected from the work supported by the National Institute of Child Health and Human Development (NICHD), the National Institutes of Health (NIH) under Grant R21HD054697, and the National Institute on Disability and Rehabilitation Research (NIDRR) in the Department of Education under Grant H133G090005 and Award Number H133P090008. The rest of the work was funded in part by the National Science Foundation (NSF) under award #1910526. Findings and opinions within this work do not necessarily reflect the positions of NICHD, NIH, NIDRR or NSF.

Chapter 3

A Comprehensive EEG Database for Human Emotion Analysis Induced by Multimodal Stimuli

Copyright notice: The content of this chapter is based on the manuscript "A Comprehensive EEG Database for Human Emotion Analysis Induced by Multimodal Stimuli," which is currently in the submission process for publication. Upon acceptance, copyright will be held by the publisher of the final published version.

In the previous chapter, the focus was on improving one important application of BCI, the P300 speller. This chapter explores another significant application of BCI, affective state estimation. To address the challenges of aBCI, the first approach involves developing a comprehensive multimodal EEG dataset that can effectively capture a wide range of emotional responses. The dataset integrates visual and auditory stimuli to elicit diverse affective states.

The chapter is organized into eight sections. Section [3.1](#) introduces the motivation and background for developing a comprehensive multimodal emotion database. Section [3.2](#) reviews related

work and existing EEG-based emotion datasets. Section 3.3 details the experimental design and protocol, including stimulus selection, hardware setup, and participant procedures. Section 3.4 describes the data repository. Section 3.5 analyzes subjective emotion ratings across continuous and discrete scales. Section 3.6 presents the procedure of affective state estimation, and Section 3.7 reports results from binary and discrete classification experiments and provides comparative analysis. Finally, Section 3.8 concludes the chapter with key findings and implications.

3.1 Introduction

Affect recognition involves accurately identifying and understanding the emotions expressed by an individual. In recent years, the estimation of affective states using physiological signals has gained considerable attention in the fields of neuroscience and psychology. Physiological signals offer important advantages for emotion recognition, as they can continuously monitor and provide real-time insights into a user's affective states. Prior research has explored a variety of physiological signals, including galvanic skin response (GSR)^{59–63}, blood pressure (BP)^{59;61}, eye gaze, skin temperature (ST)^{59;61;62}, heart rate⁶⁰, respiration pattern^{61–63}, electromyography (EMG)^{59–61;63}, electrocardiography (ECG)^{62–64}, electroencephalography (EEG)^{61;62;64}, and electrooculography (EOG)⁶¹, to assess emotional states. Among these, electroencephalography (EEG) has become particularly attractive for the estimation of the affective state. This is widely used because the brain is the primary source of electrical activity associated with emotional states, making neural signals some of the most reliable physiological indicators for evaluating affective states. An affective brain-computer interface (aBCI) is a sophisticated system designed to extract and analyze features related to affective states from these neurophysiological signals⁶⁵. Although several competing technologies exist within the domain of aBCIs, EEG stands out as a leading technique due to its non-invasive

nature, high temporal resolution, reliability, and relatively low cost²⁸. These attributes make EEG a preferred method for real-time and accurate affective state estimation.

Despite the advantages of EEG for emotion recognition, efficiently detecting an individual's emotional state from EEG signals remains a significant challenge. One primary issue is the inadequacy of existing EEG-based emotion detection databases, which often suffer from insufficient sample sizes. This limitation arises from a combination of factors, including the limited number of stimuli and participants, which complicates classification tasks. To improve model accuracy, it is essential to expand both the range of stimuli and participant numbers to ensure comprehensive coverage of emotional states. Creating a larger dataset is difficult, as it requires a substantial number of participants and stimuli per participant, making the process time-consuming and costly. Another significant factor is the type of stimulus used to evoke emotions in individuals. Common stimuli include pictures, facial expressions, sounds, and music. Pictures and images are frequently employed for emotion elicitation, while the connection between music and emotions has been studied for over a century. It is widely believed that music bridges cognition and emotion, influencing autonomic nervous system responses in both stimulating and soothing ways⁶⁶. Recent studies increasingly focus on combining auditory and visual modalities to enhance emotional elicitation, such as through the use of videos and films. By integrating multiple modalities, such as visual, auditory, and tactile stimuli, researchers can develop more robust emotion recognition models that offer a comprehensive assessment of emotional responses⁶⁷. Such multimodal integration can also provide a more comprehensive evaluation of stimulus effectiveness when tested with the same individual. While many databases have explored various stimuli to evoke emotions, most existing EEG-based emotion databases are restricted to single-modality stimuli. To our knowledge, no publicly available EEG-based emotion databases have been reported with multiple stimulus types

from the same individual within a single experiment.

Another important criterion for automated emotion recognition is the conceptual framework used to define emotional states. Two predominant models are employed: continuous and discrete. The continuous model, developed by J.A. Russell, utilizes a circumplex model where emotions are characterized along two dimensions: valence and arousal³⁴. Valence measures the pleasantness of the stimulus, while arousal reflects the level of activation. An additional dimension, dominance, can also be incorporated to represent emotional control, thereby creating a three-dimensional space⁶⁸. In contrast, the discrete model, introduced by Paul Ekman, categorizes emotions into six basic types: happiness, sadness, surprise, fear, anger, and disgust³². While there have been many attempts by researchers to map valence and arousal into discrete emotional categories, this approach remains distinct from discrete models that classify emotions based on specific target emotions⁶⁹. Emotions within the same discrete category can vary significantly in their levels of valence and arousal⁷⁰. Obtaining ratings for a stimulus in both continuous and discrete models would be advantageous for comparing and detecting emotions across these frameworks. However, most EEG-based emotion databases tend to focus on either continuous or discrete models exclusively.

To address the current limitations of EEG-based emotion datasets mentioned above, we aimed to introduce a unique **Database for EEG-based Emotion Analysis using Multimodal Stimuli (DEEAMS)**. EEG data were collected using a 64-channel system at a sampling rate of 500 Hz. The signals were recorded from 30 participants as they were exposed to 240 different stimuli. These stimuli included three distinct types of multimodal inputs: emotional pictures, facial expression images, and music. Each stimulus was evaluated using two types of responses: ratings on the continuous model (assessing valence, arousal, and dominance) and ratings on the discrete model (based on six basic emotions: happiness, sadness, fear, disgust, surprise, and anger). We conducted

multiple classification experiments to recognize emotions in both continuous and discrete formats in order to set a baseline result. These experiments included modality-specific analyses for each type of stimulus (images, facial expressions, and music) to evaluate their individual classification performance. Additionally, we implemented an integrative classification approach, which combined all modalities into a single task. This approach aimed to assess the model’s ability to process and classify multiple types of stimuli simultaneously. Based on Table 3.1, this dataset has a relatively large number of participants (30) and nearly three times the number of stimuli (240) compared to existing EEG-based emotion databases. Additionally, it uniquely integrates three types of stimuli within a single experiment to elicit emotional responses.

3.2 Related Work

Extensive research has been conducted on EEG-based emotion recognition, with the creation of affective databases playing a crucial role in advancing this field. While many studies used their own collected data for emotion recognition, only a limited number of databases are publicly available. The recent advancements in this area have been discussed in several review papers^{69;84}. We also provided a summary of the latest physiological databases in Table 3.1. In previous studies, a range of stimuli, such as pictures, music, sounds, and videos, have been used to elicit emotional responses in participants. Two of the most widely utilized resources are the International Affective Picture System (IAPS)⁸⁵ for images and the International Affective Digital Sounds (IADS)⁸⁶ for auditory stimuli, both of which have been extensively applied in emotion detection research. For instance, Bos⁷¹ used stimuli from both IAPS and IADS to induce emotions and recorded EEG signals. However, the dataset included only 5 participants. Each was exposed to 36 emotion-eliciting stimuli, along with 3 neutral practice stimuli. In recent years, videos have emerged as a prominent

Table 3.1: The summary of EEG based emotion databases

No	Database	Year	Physiological Signal	Stimuli Modality	Stimuli Number	Subject No	Emotional Model	Publicly Availability
1	⁷¹	2006	EEG	IAPS, IADS	36	5	Continuous	N
2	⁷²	2012	EEG	Music	75	9	Likings only	N
3	DEAP ⁶¹	2012	EEG, EOG, EMG, RSP, BP, ST, GSR	Video	40	32	Continuous	Y
4	MANHOB-HCI ⁶²	2012	EEG, ECG, RSP, GSR, ST, eye gaze	Exp1: Video; Exp2: Images and videos	Exp 1: 20; Exp 2: 28+14	27	Continuous	Y
5	SEED ⁷³	2015	EEG	Video	15	15	Discrete	Y
6	⁷⁴	2017	EEG	Video	16	30	Continuous, discrete	Y
7	DREAMER ⁶⁴	2018	EEG, ECG	Video	18	23	Continuous	Y
8	RCLS ⁷⁵	2018	EEG	Video	15	14	Discrete	Y
9	⁷⁶	2018	EEG, blood volume pulse	Video	9	20	Discrete	N
10	SEED-IV ⁶⁷	2019	EEG, eye tracking	Video	72	15	Discrete	Y
11	MPED ⁶⁹	2019	EEG, ECG, TSP, GSR	Video	28	23	Discrete	Y
12	HR-EEG4EMO ⁷⁷	2020	EEG, GSR, ECG, RSP, SpO2, pulse rate	Video	20	27	Discrete	Y
13	AMIGOS ⁷⁸	2021	EEG, ECG, GSR	Video	16	40	Continuous, discrete	Y
14	SEED-V ⁷⁹	2021	EEG, eye tracking	Video	45	16	Discrete	Y
15	WeDea ⁸⁰	2022	EEG	Video	79	30	Discrete	N
16	SEED-FRA ⁸¹	2022	EEG, eye tracking	Video	21	8	Discrete	Y
17	SEED-GER ⁸²	2022	EEG, EYE TRACKING	Video	20	8	Discrete	Y
18	THU-EP ⁸³	2023	EEG	Video	28	80	Continuous, discrete	Y

stimulus in the field of emotion detection. They are frequently utilized in various emotion detection databases to elicit emotional responses^{64;67;69;73–76;78;79;83}. The DEAP database⁶¹, one of the most widely utilized resources in emotion research, relied solely on video stimuli. It collected various physiological signals (EEG, EOG, EMG, RSP, GSR, skin temperature) from 40 participants, with 40 videos per participant. The participants reported their emotional responses using a dimensional model that assessed emotions in terms of valence, arousal, dominance, and liking. Despite the widespread use of these established stimuli, only a limited number of studies have explored the integration of multiple types of traditional stimuli (e.g., pictures, sounds, and music) within a single experiment. The MANHOB.HCI⁶² dataset employed both images and videos to elicit emotional responses from 27 participants, with continuous-scale ratings provided for the emotional experiences.

Other prominent emotion datasets, including SEED, SEED-IV, SEED-V, SEED-FRA, and SEED-GER, also employed video stimuli to evoke emotional responses^{73;73;79;81;82}. These datasets, however, reported participants' responses using a discrete model. Despite the valuable insights they offer, these datasets are considered small-scale due to their relatively limited participant pool, typically ranging from 8 to 16 individuals, and a relatively small number of trials per participant. SEED-IV is an exception with 72 trials per subject. Several studies have utilized a large number of trials per participant, but many of these datasets are not publicly accessible. For instance, Hadjidimitriou⁷² conducted an experiment with 9 participants using 75 different pieces of music to evoke emotions, with participants providing emotional responses based solely on their preferences. Naksia et al.⁷⁶ conducted a study where multiple physiological signals, including EEG, were recorded from 20 participants as they viewed 9 different videos. The participants' emotional responses were reported solely using a discrete scale. Katsigiannis and Ramzan⁶⁴ developed the

DREAMER dataset, which includes EEG and ECG recordings from 23 participants. Emotions were assessed using a continuous scale focusing on valence, arousal, and dominance. In contrast, in the RCLS database⁷⁵, they collected EEG data from 14 participants while they viewed videos and rated their emotional responses using a discrete scale (positive, negative, neutral). Kim et al. introduced the WeDea database⁷⁸, which provides a comprehensive collection of EEG data from 30 participants. This database is notable for its extensive coverage, as it includes a substantial number of stimuli per participant, specifically 79.

Some widely used databases in affective state analysis are MPED⁶⁹, HR-EEG4EMO⁷⁷, and AMIGOS⁷⁸. These databases feature a variety of physiological signals beyond EEG, such as ECG, galvanic skin response (GSR), respiration, peripheral oxygen saturation (SpO2), pulse rate, and skin temperature. MPED, which includes data from 23 participants, reports emotions using a discrete scale with multiple emotional categories (joy, funny, anger, fear, disgust, sadness, and neutral). Very few databases have reported participants' responses in both scales (continuous and discrete)^{78;83}. AMIGOS stands out for its larger participant pool (40) and its provision of both continuous and discrete ratings (valence, arousal, control, familiarity, liking, and basic emotions). The most recent EEG-based emotion recognition database, THU-EP⁸³, published by Shen et al., involves the largest participant group among existing databases, with 80 participants. Although the number of video stimuli is relatively modest at 28, participants provided ratings on 12 emotional items, including anger, disgust, fear, sadness, amusement, joy, inspiration, tenderness, as well as arousal, valence, familiarity, and liking. Many existing databases provide high-quality data using various stimuli, involve large numbers of subjects or samples, and utilize one or both emotional models for stimulus assessment. However, a notable gap exists for databases that combine multimodal stimuli to evoke emotions, feature a large dataset, and include participants' self-assessments using both continuous

and discrete scales. This highlights a need for a multidimensional database to advance research in EEG-based emotion recognition.

3.3 Experimental Design and Protocol

3.3.1 Stimuli Selection

Our dataset is distinctive due to its use of three different modalities: pictures, facial images, and music. Specifically, this study included a total of 240 stimuli, consisting of 80 images from the International Affective Picture System (IAPS) [30]⁸⁵, 80 facial expression images from the Pictures of Facial Affect (POFA) database³¹, and 80 music tracks from the MediaEval Database for Emotional Analysis of Music (DEAM) dataset⁸⁷. This combination of modalities provides a comprehensive approach to studying emotional responses.

Pictures

80 pictures were chosen from a total of 1,182 images in IAPS, based on their original ratings provided in the IAPS dataset. Two well-established methods from cognitive theory were used to represent emotions in an n-dimensional space: the 2D valence-arousal (VA) model and the 3D valence-arousal-dominance (VAD) model^{34;88}. The valence-arousal model categorizes emotions into four quadrants: low arousal/low valence (LALV), low arousal/high valence (LAHV), high arousal/low valence (HALV), and high arousal/high valence (HAHV). 80 images were selected manually in such a way that they would be able to represent each quadrant, ensuring a diverse emotional range. The ratings of the dominance axis were also considered during the image selection process. The original rating scale for the IAPS pictures ranges from 1 to 9. However, since our study employed a 5-point scale, we converted the IAPS ratings to a corresponding 1 to

5 scale. The mean valence and arousal ratings of the selected images, as reported in the IAPS manual, are illustrated in Fig. 3.1.

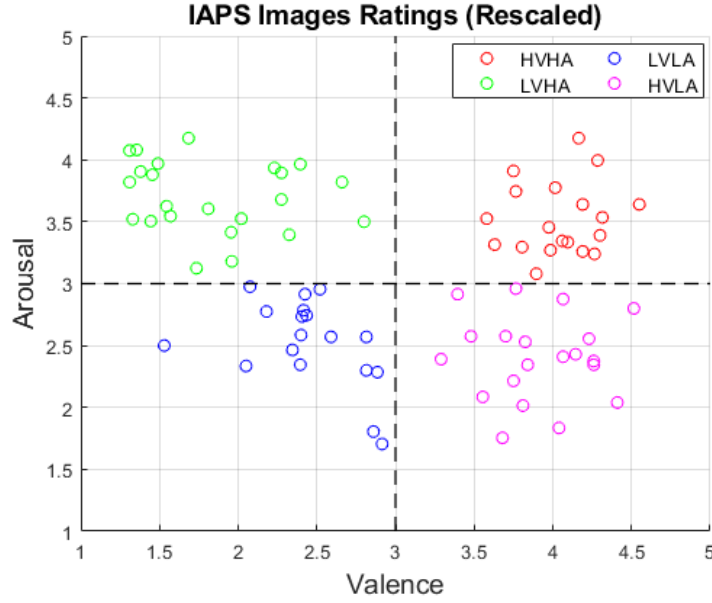


Figure 3.1: Mean ratings of selected IAPS images

Facial Images

The POFA dataset comprises images depicting various facial expressions associated with fundamental emotions, including happiness, sadness, fear, anger, surprise, disgust, and neutrality³¹. Each image in the POFA database is labeled according to one of these emotions. To achieve a comprehensive representation of these emotions, we selected 80 images from a total of 110 POFA images. Our selection process aimed to ensure that a substantial number of images were included for each emotion, thereby capturing a wide range of emotional expressions. The specific number of images chosen for each emotion is mentioned in Table 3.2.

Table 3.2: Number of POFA images

Emotion name	Number of images
Happy	12
Sad	12
Fear	11
Disgust	11
Surprise	13
Anger	12
Neutral	9

Music

In a preliminary phase, 80 music tracks were manually selected from a total of 1,802 tracks available in the DEAM database⁸⁷. This selection was based on their valence-arousal (V-A) ratings provided in the DEAM manual. The selection procedure followed the same protocol used for the picture selection from the IAPS database, based on the 2D valence-arousal model. Fig. 3.2 displays the average valence and arousal ratings for the selected music tracks, according to the DEAM database. The DEAM database comprises 58 full-length songs and 1,744 excerpts, each with a duration of 45 seconds. For the final analysis, each of the 80 chosen tracks was shortened to include only the first 20 seconds, ensuring consistency and focus in the auditory stimuli presented.

3.3.2 Hardware Setup

EEG data were acquired using the Cognionics Mobile-72 system, a high-density wireless EEG device with 64 channels. Signals were recorded at a sampling frequency of 500 Hz to ensure high temporal resolution. Active Ag/AgCl electrodes were positioned according to the international 10–20 system shown in Fig. 3.3, with reference and ground electrodes placed on the right and left mastoids, respectively. The entire experimental setup was controlled using the BCI2000 software⁸⁹, which was employed to present the stimuli to the participants. Participants were instructed to adjust

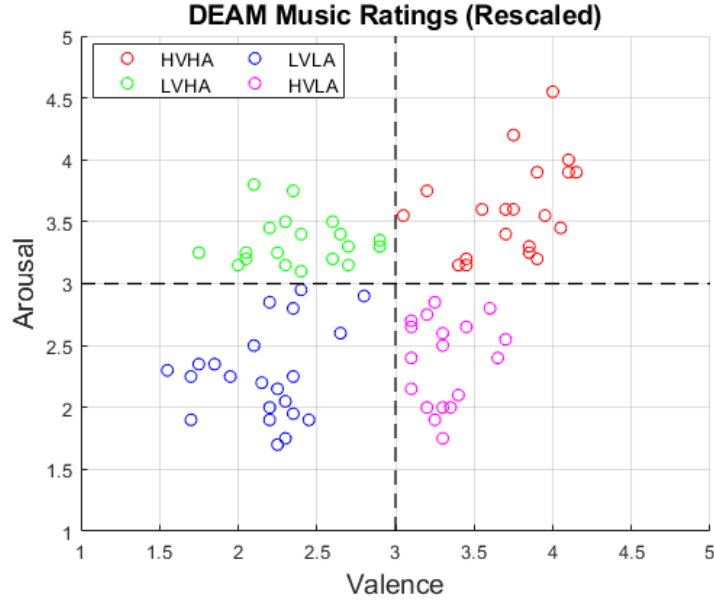


Figure 3.2: Mean ratings of selected musics

both the screen position and volume levels to ensure their comfort, and data were collected in an acoustically isolated room to minimize external noise and distraction.

3.3.3 Experimental Design

The experimental design and data collection were performed under the oversight of the Kansas State University Institutional Review Board (IRB), under protocol #8328. Before the experiment began, each participant was required to sign an informed consent form and complete a detailed questionnaire. This questionnaire gathered comprehensive information, including the participant's age, gender, ethnicity, and hand dominance. Additionally, it inquired about any previous health issues or ongoing medications that might influence the experimental outcomes. The questionnaire also assessed whether participants had any previous experience with brain-computer interface (BCI) experiments, which is crucial for understanding their baseline knowledge and potential adaptability to the experimental setup. This preliminary data collection was essential for ensuring

participant safety, personalizing the experimental conditions, and providing context for interpreting the recorded data.



Figure 3.3: Electrode placements for the International 10-20 system

The experimenter provided detailed instructions to the participant about the entire experimental protocol. In addition, participants were instructed on the self-assessment process, which involved categorizing emotions and using rating scales. Upon confirming that the participants fully understood the instructions, they were asked to initiate the experimental procedures. To prevent participant fatigue associated with long experimental sessions, the study was organized into 20 distinct blocks distributed over a period of 5 days. Each block contained 12 stimuli—comprising 4 images from IAPS, 4 images from POFA database, and 4 music clips from the DEAM dataset. The presentation sequence within each block was structured as follows: one IAPS image, one facial expression image, followed by one music clip. This arrangement was intended to ensure consistent participant engagement and maintain high data quality throughout the study.

The detailed protocol is shown in Fig. 3.4. The experiment for each block started with a 4-

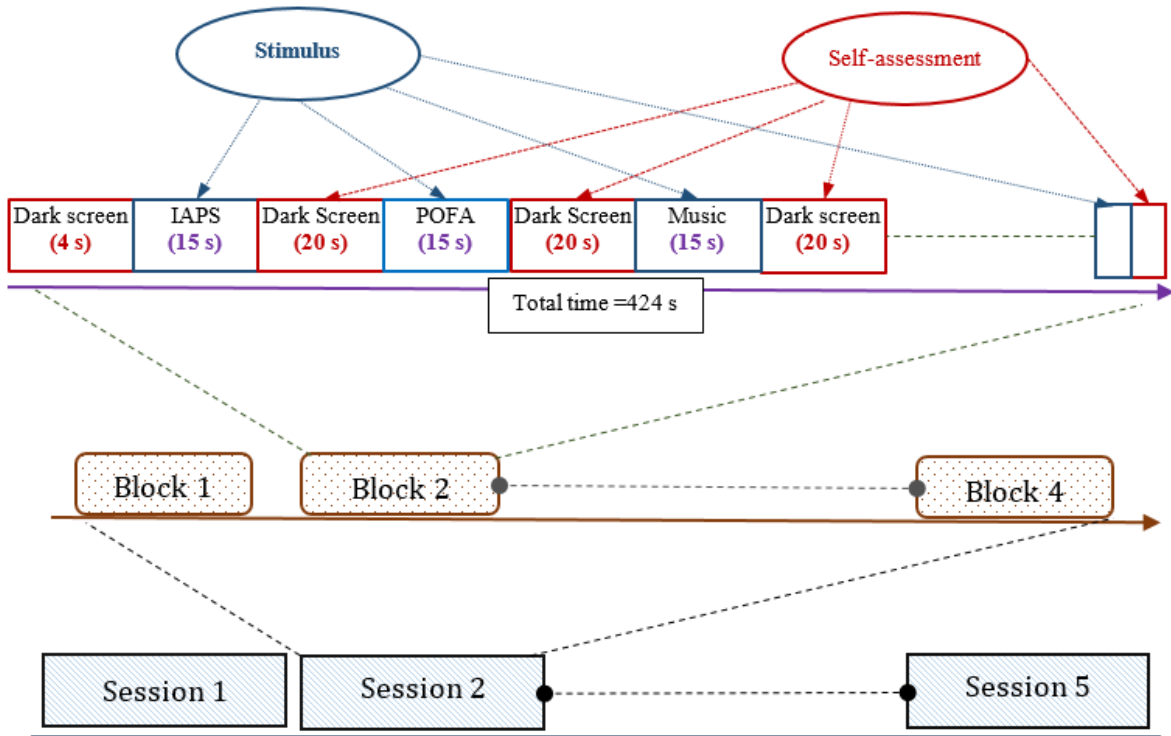


Figure 3.4: Flow chart of experimental protocol

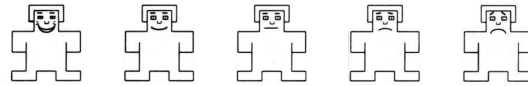
second black screen to prepare the participants for the upcoming stimulus. This was followed by the presentation of a single stimulus on the display, and then another black screen was shown to provide time for the participants to perform a self-assessment. Each stimulus was displayed for 15 seconds, with a 20-second interval between stimuli. Following each block, participants were asked to take a short break, during which they could consume snacks or beverages. The experimenter then verified the electrode placements and signal quality before proceeding to the next block. To complete the study, participants were required to attend five sessions, each comprising four blocks. At the end of each stimulus, participants were asked to rate each stimulus using the Self-Assessment Manikin (SAM)⁸¹, a widely utilized tool for emotional assessment. The self-assessment required evaluations on both discrete and continuous dimensions of emotional experience, as depicted in Fig. 3.5. For the discrete axes, participants selected one of six basic emotion categories: happiness,

Please fill out how what you just saw made you feel. **There is no wrong answer** – please just answer honestly, even if you think your response is unusual or strange.

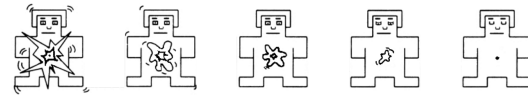
Select the closest emotion (circle just one):



How **good** (left) or **bad** (right) did the cue make you feel?



How **excited** (left) or **calm** (right) did the cue make you feel?



How **controlled** (left) or **in control** (right) did the cue make you feel?

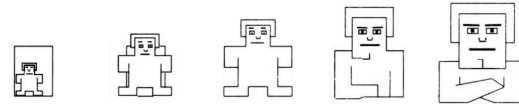


Figure 3.5: Self-assessment using SAMS Manikin

sadness, fear, disgust, surprise, or anger. On the continuous axes, participants assessed the stimulus across three distinct emotional dimensions: valence, arousal, and dominance (see Table 3.3).

Table 3.3: Self-assessment scale scoring

Scale	Valence	Arousal	Dominance
5	Very happy/pleasant	Excited	Dependent/controlled
4	Happy	Wide-awake	Powerlessness
3	Neutral	Neutral	Neutral
2	Sad	Dull	Powerful
1	Very Sad/unpleasant	Calm	Independent/In control

3.3.4 Participants

The data collection phase involved 34 participants from diverse cultural and professional backgrounds, ranging from undergraduate students to employed professionals. Among these partic-

ipants, 30 completed the study in its entirety, while the remaining withdrew due to personal circumstances. Therefore, the final dataset was comprised of responses from these 30 individuals. Within this subset, 17 participants were male, and 13 were female. The participants' ages ranged from 19 to 42 years, with a mean age of 26.8 years and a standard deviation of 5.9 years. The summary of participants' demographic data is provided in the Table 3.4. This demographic diversity was integral to ensuring the robustness and generalizability of the study's findings.

Table 3.4: Summary of demographic data

	No of participants	Age limit	Mean	Median	SD
All	30	19 – 42	26.83	27.5	5.93
Male	17	19 – 35	25.77	26	5.46
Female	13	19 – 42	27.65	28	6.3

3.4 Data Record

The DEEAMS dataset is available to reviewers via a private link for the initial submission. The repository follows a standardized layout; a folder tree is shown in 3.6. It contains:

1. *DEEAMS/* (root).

The root directory includes three subfolders—*metafile/*, *PreProcessed_Data/*, and *Raw EEG Data/*, and a *README* file, which contains the description of the dataset.

2. *metafile/* (participant metadata and ratings).

This folder contains one demographics workbook and 30 ratings workbooks.

- (a) *Demographics.xlsx*: a workbook summarizing demographic information for all 30 participants (one sheet per participant).

(b) *Emotion_rID.xlsx*: one workbook per participant (e.g., *Emotion_r205.xlsx*), each containing 20 sheets (one per block). Every sheet lists the 12 trials and includes both SAM scale ratings (e.g., valence/arousal/dominance) and discrete emotion labels for each trial.

3. *Raw EEG Data/* (original recordings).

This folder contains the raw EEG recordings for all 30 participants, organized as one sub-directory per participant named *Emotion_ID* (e.g., *Emotion_205*). Data were acquired using BCI2000 and saved in .dat format; no additional preprocessing or compression was applied at this level. Each participant folder includes 20 .dat files, corresponding to 5 sessions $\times$ 4 runs (blocks) per session. All files follow the same naming convention:

Emotion_<ID>S<session>R<run>.dat

where *<ID>* is the participant identifier, *S<session>* is the session index (001–005), and *R<run>* is the run/block index (01–04). For example, *Emotion_205S001R01.dat* denotes participant 205, session 1, run 1. This organization is consistent across all participant folders.

4. *PreProcessed Data/* (trial EEG + labels).

This folder contains the pre-processed EEG data for all participants. There are 30 preprocessed files in .mat format, with one MATLAB file per participant named *Emotion_<ID>.mat* (e.g., *Emotion_205.mat*). Each file is derived from the raw BCI2000 recordings using a minimal, standardized pipeline consisting of zero-phase band-pass filtering (0.5–48 Hz, 4th-order IIR). The data were then segmented into 15-second trials with epoching relative to stimulus onset (7,500 samples at 500 Hz) and extraction of the first 4-second baseline for every block (2,000 samples). All 70 channels are retained, with

channel labels provided, including corrections where channel 61 corresponds to Oz→PO10 and channel 64 corresponds to TP9→PO9. During data collection, several challenges were encountered, including technical equipment malfunctions, unexpected interruptions, and instances where participants missed tasks. These missing segments and their corresponding iterations were thoroughly documented in the README file. During preprocessing, *corrupted* or incomplete trials were excluded. Subject-specific exclusions were applied during preprocessing. Subject 2 had one .dat file excluded (12 trials missing), Subject 11 had one .dat file excluded (12 trials missing), and Subject 24 had four .dat files excluded from blocks 1–4 (48 trials missing). Each .mat file includes:

- RawSignal: ($N_{trials} \times 7500 \times 70$) double; band-passed, zero-phase filtered trial data
- baseline_signal: ($N_{blocks} \times 2000 \times 70$) double, first 4 s of each block (filtered)
- labels: A table with ratings of 240 trials
 - Valence, Arousal, Dominance (1–5 scale)
 - *Emotion Name*, *Emotion Number* — the discrete emotion category for each trial (e.g., 1=Happy, 2=Sad, 3=Fear, 4=Disgust, 5=Surprise, 6=Anger)
- ChannelName: 70×1 cell array of electrode names (with the PO10/PO9 corrections)
- SamplingRate: scalar (500 Hz)
- participant_id: participant identifier string (e.g., Emotion_205)

This structure ensures consistent, programmatic access to synchronized EEG signals, baselines, and labels for each participant.

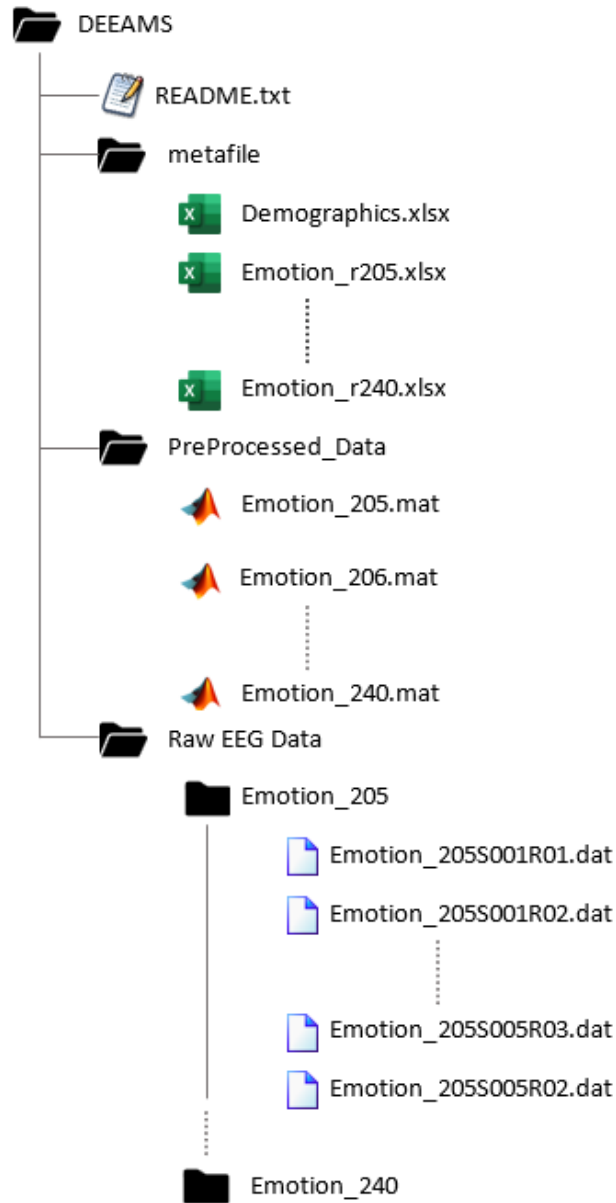


Figure 3.6: Structure of the DEEAMS dataset repository. The directory tree shows the overall organization of the dataset, including the metadata files (e.g., demographics and emotion annotation spreadsheets), preprocessed data files in .mat format, and raw EEG recordings stored as .dat files.

3.5 Subjective Rating Analysis

In this section, the ratings provided by the participants were thoroughly analyzed to identify any anomalies and assess the overall effectiveness of the stimuli by comparing them with the original ratings provided in their respective databases. Given that participants rated the stimuli using both the SAM scale and discrete labels, the analysis is divided into two parts: one focusing on the SAM ratings and the other on the discrete ratings. This approach allows for a comprehensive evaluation of the stimuli's performance and effectiveness.

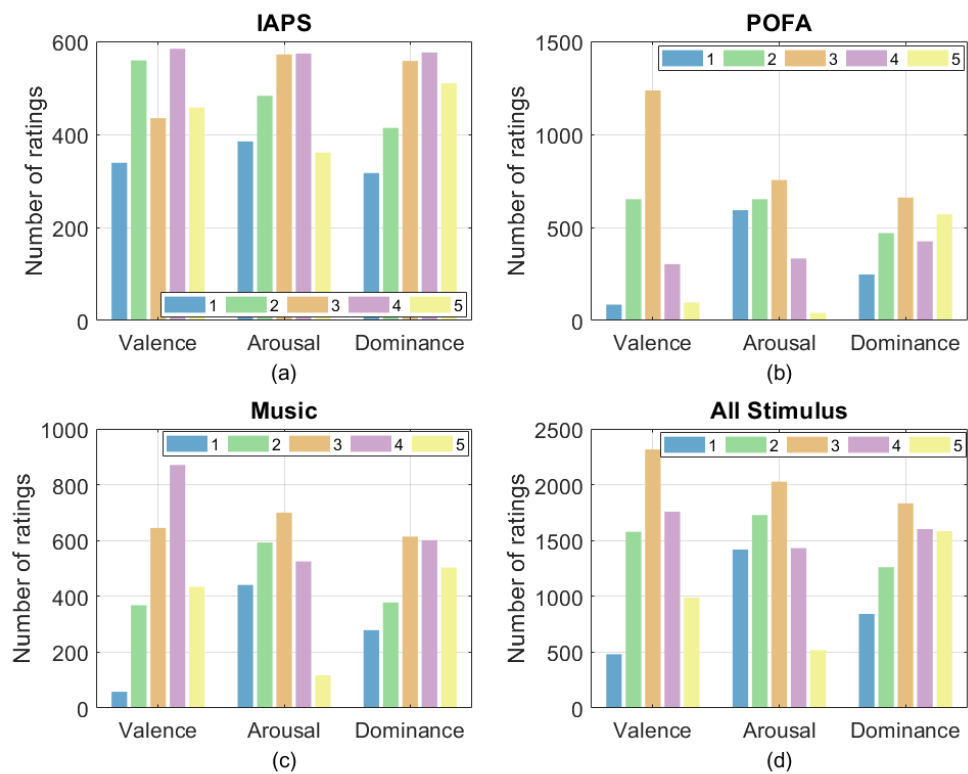


Figure 3.7: Distribution of all participants' ratings across 5-point scale of the valence, arousal, and dominance axes for different stimulus types: (a) IAPS, (b) POFA, (c) DEAM, (d) combined stimuli.

3.5.1 SAM Ratings

The distribution of all participants' ratings are presented in Fig. 3.7. Initially, we conducted an analysis of participant ratings by categorizing them based on the three types of stimuli. As shown in Fig. 3.7a, the ratings for the IAPS stimuli indicate a balanced distribution across the valence, arousal, and dominance axes, reflecting a well-represented dataset. In contrast, Fig. 3.7b reveals that the POFA stimuli were predominantly rated as neutral on the valence axis. Notably, there is an underrepresentation of ratings for the highest valence and arousal (specifically at a rating of 5). This is somewhat expected, as eliciting high arousal from facial expression images alone is inherently challenging. For the music stimuli, as shown in Fig. 3.7c, the highest number of ratings is observed for happy (at a rating of 4) on the valence axis, while the arousal and dominance axes show a nearly balanced distribution of ratings. Finally, Fig. 3.7d presents the combined ratings across all three types of stimuli, offering a comprehensive overview of the participants' responses.

The stimuli were categorized into four distinct quadrants within the Valence-Arousal (V-A) space, based on their respective valence and arousal values, as illustrated in Fig. 3.8. Based on the data presented in Fig. 3.8a, it can be observed that the participant-reported emotions for the IAPS pictures cover a broad range of the V-A space, indicating a diverse emotional response. In contrast, the distribution of ratings for the POFA images, as shown in Fig. 3.8b, is concentrated primarily in the lower arousal region, suggesting a more limited emotional impact. For the music stimuli, illustrated in Fig. 3.8c, the participant ratings predominantly occupy the high valence area, indicating that the music evoked more positive emotional responses.

Furthermore, we conducted a comparison between the ratings assigned to each stimulus by each participant and the original ratings provided for those stimuli. This comparison was designed to assess the effectiveness of the stimuli in eliciting the intended emotional responses. To evaluate

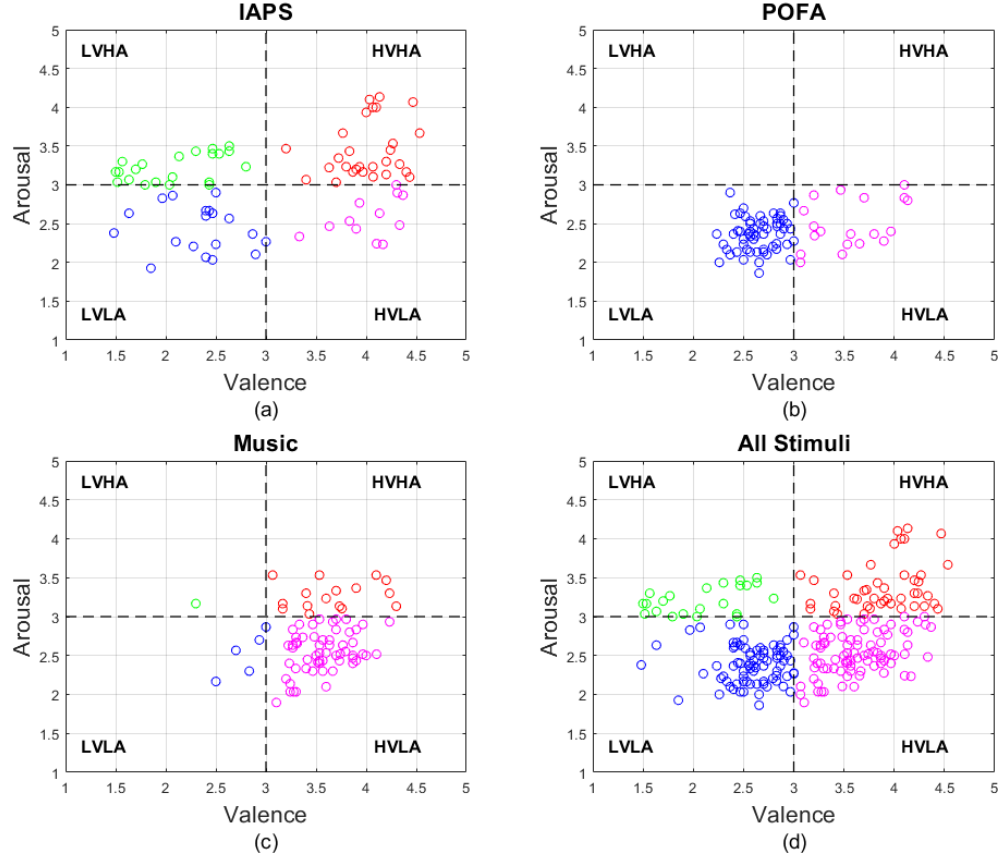


Figure 3.8: Mean rating distribution of all participants into four distinct quadrants within the Valence-Arousal (V-A) space: (a) IAPS, (b) POFA, (c) DEAM, (d) combined stimuli.

the effectiveness of the stimuli in eliciting the intended emotions, we computed the Pearson correlation coefficient between the ratings provided by participants and the original ratings for each stimulus. The average correlation coefficients across all 30 participants are presented in Table 3.5 for both IAPS and DEAM stimuli. This comparison was conducted across all three emotional dimensions—valence, arousal, and dominance—for the IAPS stimuli. For the DEAM music stimuli, the comparison was limited to the valence and arousal dimensions, as the DEAM database only includes ratings for these two axes. The analysis of the correlation coefficients revealed that IAPS pictures demonstrated a strong correlation in the valence axis, indicating effective emotion elicitation for this dimension. In terms of arousal, while there was some emotion elicitation, the correlation coefficient was relatively low at 0.29591. Unfortunately, the dominance axis showed a

negative and very low correlation coefficient of -0.06416, suggesting poor effectiveness in eliciting emotions related to dominance. For the music stimuli from the DEAM database, the correlation coefficients were positive for both the valence and arousal axes, with values of 0.1413 and 0.257, respectively. This indicates that the music stimuli were moderately successful in generating the target emotions in these dimensions.

Table 3.5: Mean correlation coefficient between participants' rating and original ratings of the stimuli

	Valence	Arousal	Dominance	Discrete label
Pictures (IAPS)	0.77249	0.29591	-0.06416	–
Music (DEAM)	0.1413	0.257	–	–
Facial expressions (POFA)	–	–	–	0.6865

3.5.2 Discrete Ratings

During the self-assessment phase, participants were instructed to classify each stimulus into one of six discrete emotional categories: happiness, sadness, fear, disgust, surprise, and anger. The distribution of these categorical ratings across all stimulus types is depicted in Fig. 3.9. For IAPS images (Fig. 3.9a), the predominant classification was "Happy," indicating that these stimuli were most frequently associated with positive emotional responses. In contrast, Fig. 3.9b reveals a more even distribution of emotional ratings for the POFA images, with a relatively balanced representation across all six emotional categories. It is noteworthy that the "Fear" and "Anger" categories received relatively few ratings for music stimuli (Fig. 3.9c), suggesting a lower elicitation of these emotions. Fig. 3.9d provides a comprehensive overview of the rating distribution when combining all three types of stimuli. This aggregate data shows that the "Happy" category consistently gained the highest number of ratings across all stimulus types, highlighting a dominant tendency towards

positive emotional responses in the overall dataset. Additionally, we performed a comparative

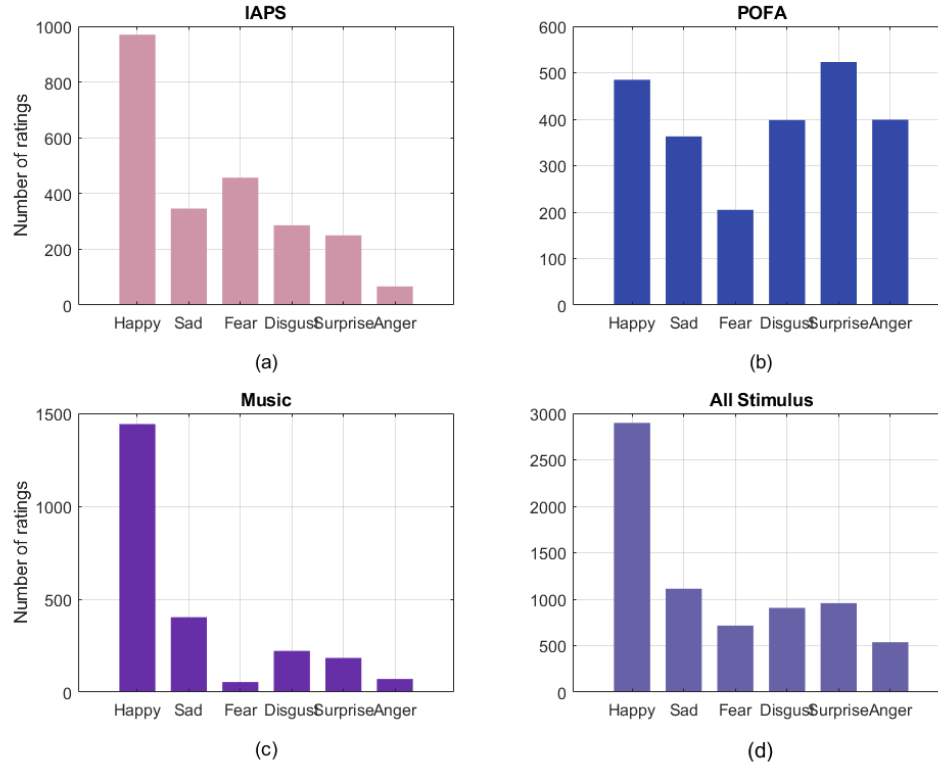


Figure 3.9: The distribution of discrete emotion categorical ratings across all stimulus types: (a) IAPS, (b) POFA, (c) DEAM, (d) combined stimuli.

analysis between participants' ratings and the original ratings of the POFA images. This comparison was focused solely on the POFA images, as the IAPS and DEAM databases did not provide discrete emotional labels for their stimuli. It is important to note that the POFA database includes images labeled with six basic emotions and a neutral category. However, our study only considered the six basic emotions, so we excluded the neutral images to ensure a valid comparison. We then calculated the Pearson correlation coefficient between participants' ratings and the original ratings for each POFA image. The average Pearson correlation coefficient across all 30 participants was 0.6865 (Table 3.5), indicating a strong correlation. This suggests that the POFA images effectively convey the intended emotions, validating their use in the study.

3.6 Affective State Estimation

In this section, we described the methodology for emotion classification using our dataset to provide a baseline result that demonstrates the reliability and validity of the proposed dataset. We conducted four separate classification analyses: the first three were modality-specific, addressing each type of stimulus individually, while the fourth analysis integrated all modalities into a single classification task. Additionally, since participants rated the stimuli using both SAM ratings and discrete labels, we performed separate classifications for each labeling method across all types of stimuli. All the classification methods employed in this study—including preprocessing, feature extraction, classifier selection, and performance metrics—were consistently applied across all types of classification tasks mentioned above.

3.6.1 Data Preprocessing

During initial preprocessing, raw signals were filtered using a 4th-order Butterworth band-pass filter (0.5–48 Hz) and then segmented into trials. As outlined in the experiment protocol, each block was designed to include 12 distinct stimuli. These blocks were segmented into 12 trials, each with a duration of 15 seconds. This resulted in a total of 240 trials for the entire recorded EEG dataset. Finally, z-score normalization was applied to each segmented trial.

$$\text{normalized signal} = \frac{x(t) - \mu}{\sigma} \quad (3.1)$$

Here, $x(t)$ is the signal, μ is the mean of the whole segmented signal, and σ is the standard deviation of the signal.

3.6.2 Feature Extraction

Power Spectral Density (PSD) is a key metric that quantifies the distribution of signal power across different frequencies, providing a detailed representation of the frequency content of EEG signals. It is a well-established method for analyzing EEG signals within defined frequency bands, making it particularly valuable in neuroscience and affective computing. PSD is especially effective in identifying correlations between EEG signals and an individual's affective state, as demonstrated in numerous studies⁹⁰⁻⁹². Previous studies on emotion recognition have consistently shown that five specific frequency bands in EEG data—delta (1-4 Hz), theta (4-8 Hz), alpha (8-14 Hz), beta (14-30 Hz), and gamma (30-42 Hz)—are strongly associated with human emotional states^{90;91;93}. In this study, we concentrated on four key frequency bands: Theta, Alpha, Beta, and Gamma, as these bands have shown a strong relationship with emotional processing in the brain. The EEG signals were decomposed into these frequency bands to isolate the relevant frequency-specific information. PSD was estimated with Welch's method using 2-s windows with 50% overlap. PSD was derived from each of the 64 EEG channels, resulting in 64 features per frequency band (see Fig. 3.10). In this study, only the PSD features were utilized for machine learning classification, without applying any further processing or dimensionality reduction techniques.

3.6.3 Classifier

In this study, participants provided emotional ratings for each stimulus modality (pictures, facial expressions, and music) using the Self-Assessment Manikin (SAM) on a 5-point scale across three emotional dimensions: valence, arousal, and dominance. To perform binary classification, these SAM ratings were converted into binary categories by applying a threshold of 3 on the SAM rating scale, resulting in "high" and "low" labels for each dimension. Four classification tasks were

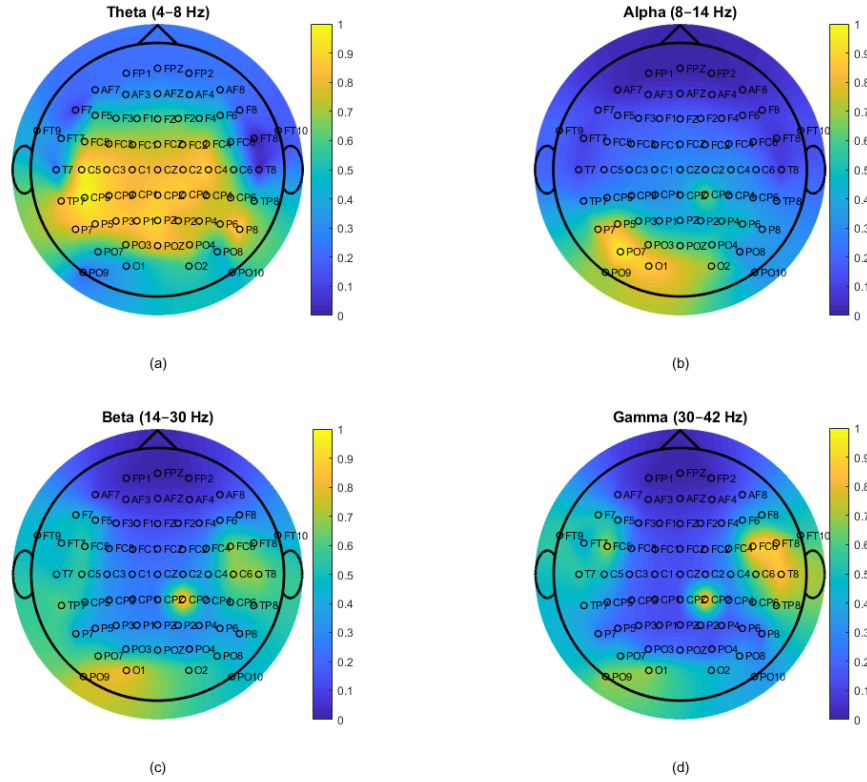


Figure 3.10: EEG data power spectral analysis. Characteristic power spectral density (PSD) is shown for each frequency band: Theta (4–8 Hz), Alpha (8–14 Hz), Beta (14–30 Hz), and Gamma (30–42 Hz).

conducted: one for each stimulus type separately, plus one combining all three types together. Binary classification is a widely employed method in emotion recognition, with various algorithms having been explored in prior research, including decision trees, logistic regression, random forest, Naive Bayes, k-nearest neighbor (kNN), support vector machine (SVM), and neural networks (NN). Here, a k-nearest neighbors (kNN)⁹⁴ classifier was used to classify binary emotional states across the three key dimensions. The classifier used Euclidean distance to measure similarity, k=5 nearest neighbors, and MATLAB's 'fitcknn' function. A subject-dependent analysis was carried out to evaluate the model's performance. For each participant individually, a 5-fold cross-validation approach was employed, where that participant's trials were partitioned into five subsets. The

model was trained on four of these subsets and tested on the remaining subset. This process was repeated five times for each participant. The final accuracy reported is the mean accuracy averaged across all participants, where each participant's accuracy was computed as the mean of their five-fold accuracies. The same kNN classifier was also utilized for classification tasks involving discrete labels, ensuring consistency across both labeling approaches.

3.6.4 Performance Metrics

Here, three performance metrics were considered: regular accuracy, balanced accuracy, and the F1 score. These metrics were calculated from the confusion matrix parameters using true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

Regular Accuracy: This metric quantifies the proportion of correct predictions made by the classifier. It is calculated by dividing the number of correct by the total number of predictions mentioned in the below equation:

$$Regular\ accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.2)$$

Balanced Accuracy: In the context of imbalanced datasets, such as those used in emotion recognition, standard accuracy may not accurately reflect the classifier's performance. Standard accuracy can overestimate performance compared to balanced accuracy⁹⁵. Balanced accuracy provides a more accurate measure of classification effectiveness across both classes⁹⁵. Balanced accuracy is computed by averaging the individual class accuracy for all classes⁹⁶. For binary

classification, it is defined as:

$$\text{Balanced accuracy} = \frac{1}{2} (Acc_1 + Acc_2) \quad (3.3)$$

where Acc represents the accuracy of each class.

F1 Score: The F1 score measures a model's performance by integrating both precision (the proportion of true positive predictions among all positive predictions) and recall (the proportion of actual positives correctly identified). In this study, we utilized the macro-averaging F1 score, where precision and recall are calculated for each class individually, and the F1 scores are then averaged across all classes. The macro-average F1 score is the mean of the F1 scores for each class.

$$F1 \text{ score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (3.4)$$

3.7 Results

3.7.1 Binary Classification

In the classification experiments, binary classifications were performed for each emotional dimension: high versus low valence, high versus low arousal, and high versus low dominance. Table 3.6 presents the average regular accuracy of all participants obtained across the four experiments. The table also includes the classification accuracy for each of the four frequency bands individually, as well as the combined accuracy across all bands for each modality type. As shown in Table 3.6, the highest average accuracy was 81.3% for POFA stimuli on the valence dimension in the beta band, 79.6% for POFA stimuli on the arousal dimension in the beta band, and 80.3% for POFA stimuli on the dominance dimension in the beta band.

Table 3.6: Regular accuracy of binary classification

Valence					
Stimulus Type	Theta	Alpha	Gamma	Beta	All band
All	58.25	60.61	59.32	59.03	58.36
IAPS	55.53	55.13	55.98	53.75	53.72
POFA	80.24	78.27	80.08	81.31	79.87
Music	62.36	64.45	65.02	63.54	62.55
Arousal					
Stimulus Type	Theta	Alpha	Gamma	Beta	All band
All	67.17	67.06	67.93	67.80	67.93
IAPS	58.36	59.13	59.61	58.01	59.30
POFA	77.67	77.65	78.11	79.61	78.33
Music	68.16	65.50	65.73	67.29	67.51
Dominance					
Stimulus Type	Theta	Alpha	Gamma	Beta	All band
All	70.38	69.76	71.48	70.94	70.37
IAPS	63.04	64.49	61.58	62.67	63.61
POFA	77.72	78.48	77.90	80.28	78.89
Music	71.31	71.36	72.61	72.75	72.18

Given that a 5-point rating scale was employed, the decision threshold was positioned at the midpoint of the scale. This choice led to imbalanced class distributions for certain participants and rating scales, where the frequency of one class (e.g., "high" or "low" state) was significantly different from the other. To provide a more comprehensive evaluation of the classification performance under these conditions, we calculated and reported the balanced accuracy and F1 score in Table 3.7. Comparing the results in Table 3.7 with those in Table 3.6, we observed that the mean balanced accuracy decreased across all cases due to class imbalances. Regarding the F1 score, the best results were 52.9% for valence with the alpha band, considering all stimuli, 52.7% for the arousal dimension with the theta band for music stimuli, and 53.9% for the dominance dimension with the beta band for POFA stimuli.

Table 3.7: Balanced accuracy and macro F1 score of binary classification

Valence										
Stimulus Type	Balanced Accuracy					F1 Score				
	Theta	Alpha	Gamma	Beta	All band	Theta	Alpha	Gamma	Beta	All band
All	51.71	53.54	52.39	51.80	51.32	51.26	52.89	51.86	51.06	50.63
IAPS	51.57	51.31	52.32	50.05	49.87	50.72	50.54	51.74	49.45	49.08
POFA	51.52	48.59	50.91	52.04	50.51	49.80	46.23	49.04	50.39	48.34
Music	49.54	51.08	51.48	49.79	49.42	48.22	49.50	49.77	47.89	47.82
Arousal										
Stimulus Type	Balanced Accuracy					F1 Score				
	Theta	Alpha	Gamma	Beta	All band	Theta	Alpha	Gamma	Beta	All band
All	50.56	51.16	51.69	51.50	51.84	49.25	50.14	50.71	50.40	50.89
IAPS	49.68	50.74	51.51	49.72	51.21	48.17	49.65	50.75	48.29	49.95
POFA	50.14	49.80	50.28	51.77	50.37	48.55	47.98	48.54	50.18	48.61
Music	53.74	51.15	51.22	53.21	53.34	52.65	49.77	50.06	52.00	52.50
Dominance										
Stimulus Type	Balanced Accuracy					F1 Score				
	Theta	Alpha	Gamma	Beta	All band	Theta	Alpha	Gamma	Beta	All band
All	52.22	52.26	52.39	52.87	52.35	51.39	51.85	52.76	52.05	51.65
IAPS	51.52	53.08	49.71	51.04	52.47	50.51	51.94	48.85	49.96	51.64
POFA	51.32	52.18	52.45	55.24	52.17	49.56	50.80	50.73	53.88	50.50
Music	51.67	51.34	52.36	52.42	52.31	50.45	49.87	50.92	50.93	51.14

3.7.2 Discrete Classification

In the discrete label classification task, participants categorized all stimuli into one of six basic emotions: happy, sad, fear, disgust, surprise, and anger. This classification was performed using a subject-dependent approach, where each emotion was evaluated in a one-vs-all manner, meaning each emotion was classified against all other emotions combined. Table 3.8 provides a detailed summary of the mean regular accuracy, mean balanced accuracy, and Table 3.9 provides the mean F1 score for each of the six emotion labels across all four frequency bands. These metrics were calculated based on the classification results from all 30 participants, providing a comprehensive evaluation of performance across different emotional categories and frequency bands. Here, we only reported the classification accuracy when considering trials from all types of stimuli together. In particular, for the "happy" class, the highest regular accuracy achieved was 61.5% in the alpha

band, where balanced accuracy is 54.6%.

Additionally, the F1 score for the "happy" class is above chance levels for all bands, indicating that the classification performance for this emotion was better than random guessing in these frequency bands. In contrast, for the remaining five emotions (sad, fear, disgust, surprise, and anger), there was a significant disparity between the regular and balanced accuracies. This discrepancy is illustrated in Fig. 3.9d, where it is evident that most stimuli were predominantly categorized as "happy" by the participants. Consequently, the other emotion classes received substantially fewer ratings, leading to a notable imbalance that affected the accuracy measures.

Table 3.8: Regular accuracy of discrete classification

Stimulus Type	Regular Accuracy					Balanced Accuracy				
	Theta	Alpha	Gamma	Beta	All band	Theta	Alpha	Gamma	Beta	All band
Case 1: Happy	58.67	61.45	59.67	58.33	58.97	52.01	54.59	53.16	51.90	52.20
Case 2: Sad	82.22	82.68	82.52	82.39	82.78	50.29	51.09	50.42	50.42	50.68
Case 3: Fear	88.87	88.38	89.07	89.16	88.93	50.08	50.30	50.25	50.36	50.31
Case 4: Disgust	85.86	85.81	85.90	86.06	85.93	49.89	50.28	50.09	49.90	49.98
Case 5: Surprise	85.01	84.72	84.96	84.73	84.79	50.41	50.36	50.09	49.90	50.30
Case 6: Anger	91.85	92.09	92.05	91.87	92.31	49.75	50.07	50.20	49.98	50.48

Table 3.9: F1 score for discrete classification

Stimulus Type	Theta	Alpha	Gamma	Beta	All band
Case1: Happy	51.21	53.95	52.60	51.34	51.44
Case2: Sad	48.27	49.31	48.38	48.39	48.65
Case3: Fear	48.31	48.83	48.56	48.70	48.70
Case4: Disgust	47.57	48.38	48.01	47.55	47.68
Case5: Surprise	48.51	48.62	47.94	47.78	48.49
Case6: Anger	48.05	48.50	48.73	48.46	49.12

3.7.3 Comparative Analysis

A comparative analysis was conducted between our proposed dataset and existing datasets for binary emotion classification using EEG data, utilizing baseline results from the DEAP, DREAMER, and MANHOB-HCI databases, as shown in Table 3.10. It is important to note that these referenced databases reported only regular accuracy, rather than balanced accuracy. Consequently, we used regular accuracy from our dataset to ensure a consistent basis for comparison. The table indicates that our dataset achieved higher accuracies on the valence (58.4%) and dominance (70.4%) axes compared to the other datasets, though having class biases of 71.2% and 70.6%, respectively. For the arousal axis, our dataset's mean accuracy is 67.9%, which is higher than those of DEAP, MANHOB-HCI, and DREAMER datasets, with a lower-class bias of 56.3% compared to 59% in the DEAP dataset. Overall, the analysis indicates that the baseline performance of our proposed dataset is consistent and comparable with that of the existing databases. Additionally, the analysis addressed class imbalances observed across all three emotional axes in the datasets, providing insight into the distribution and impact of class biases on classification accuracy.

Table 3.10: Comparison of accuracies between the proposed DEEAMS dataset and prior EEG emotion datasets

Dataset	Regular Accuracy			Balanced Accuracy		
	Valence	Arousal	Dominance	Valence	Arousal	Dominance
DEAP	57.60%	62%	N/A	N/A	N/A	N/A
Manhob-HCI	57%	52.40%	N/A	N/A	N/A	N/A
DREAMER	62.49%	62.17%	61.84%	N/A	N/A	N/A
BBS-dataset	58.36%	67.93%	70.37%	51.32%	51.84%	52.35%
Class ratio of BBS Dataset (majority class)	71.19%	56.29%	70.55%			
Class ratio of DEAP Dataset (majority class)	57.00%	59.00%				

Note: Class-ratio = number of trials in majority class / total number of trials

3.8 Conclusion

In this study, we introduced a novel and comprehensive dataset specifically developed for the analysis of emotion elicitation across different modalities using electroencephalography (EEG) signals. The dataset comprises EEG recordings from 30 participants using a 64-channel EEG system. Participants were exposed to a total of 240 stimuli, which were selected from three distinct modalities: pictures from the IAPS, facial expression images from POFA, and music tracks from DEAM. This diverse range of stimuli was chosen to evoke a broad spectrum of emotional responses. Each participant provided ratings for all stimuli on both the continuous model—corresponding to dimensions such as valence, arousal, and dominance—and the discrete model, reflecting categorical emotions such as happiness, sadness, fear, anger, disgust, and surprise.

To the best of our knowledge, this dataset is the first to integrate these three modalities of stimuli within a single experiment designed to elicit emotional responses, making it a unique resource for the research community. Notably, each modality includes 80 stimuli, culminating in a total of 240 stimuli per participant, which significantly exceeds the trial counts typically seen in EEG-based emotion studies. To our knowledge, no other emotion-related EEG database includes such an extensive number of trials per participant, making this dataset particularly valuable for robust statistical analysis and model training. Furthermore, the dataset includes ratings based on both the dimensional model of emotion (assessing valence, arousal, and dominance) and the discrete model (categorizing emotions as happy, sad, fear, anger, disgust, and surprise). This dual-rating approach provides a more nuanced understanding of emotional responses, allowing for cross-comparison between these two prominent models of emotion. The inclusion of 30 participants also represents a significant sample size, which is comparable only to the DEAP, AMIGOS and THU-EP databases. This large participant pool enhances the dataset’s generalizability and its potential

for use in developing and validating emotion recognition systems. A comprehensive analysis was performed to evaluate the efficacy of various stimuli, both individually—focusing on stimuli from the IAPS, POFA, and music categories—and in combination. The analysis involved comparing participant ratings for each stimulus with the original ratings reported in their respective databases to determine the effectiveness of these stimuli in emotion elicitation. In the dimensional model, IAPS pictures demonstrated a strong capability in eliciting emotions along the valence axis, as evidenced by a high positive correlation coefficient of 0.77249. Similarly, POFA images were found to be effective within the discrete label model, achieving a notable correlation coefficient of 0.6865. This analysis underscores the distinct emotional impact of different types of stimuli and their varying effectiveness depending on the model employed for evaluation.

Finally, subject-dependent classification was performed on both SAM and discrete scales to establish baseline results for the proposed dataset. Results were evaluated using three performance metrics: regular accuracy, balanced accuracy, and macro F1 score. This approach provides a more comprehensive view of classification performance compared to other databases that typically report only regular accuracy and F1 score focused on major classes, potentially overlooking dataset imbalances. While the overall balanced accuracy of the proposed dataset was not as high as expected, regular accuracy aligned with baseline results from other datasets. This indicates that there is significant potential for improving the classifier and feature extraction processes to achieve higher accuracy in future work. The dataset has been made publicly available. This opens avenues for future research to explore the relationship between EEG signals and emotional states using both continuous and discrete models from the same stimuli. Researchers are also encouraged to apply and test their own classification methods with this diverse dataset.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. #1910526. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This work is also supported by the Cognitive and Neurobiological Approaches to Plasticity (CNAP) Center through the NIH NIGMS COBRE Phase II award P20GM113109. Graduate research assistantship and administrative support during parts of the data analysis were provided through CNAP, the CNAP SRO, and the CN Core at Kansas State University.

Chapter 4

The Role of Review Process Failures in Affective State Estimation: An Empirical Investigation of the DEAP Dataset

Copyright notice: The content of this chapter has been submitted for publication. A preprint version is publicly available on arXiv⁹⁷. Upon acceptance, the publisher will hold the copyright for the published version.

In the previous chapter, a comprehensive multimodal emotion database was developed to facilitate affective state estimation. However, creating a high-quality database alone is not sufficient. It requires robust and reliable estimation methods to be truly effective. While the literature presents numerous models claiming high accuracy on their own or limited datasets, their reliability in real-world applications remains questionable. Rather than proposing yet another high-accuracy model to an already saturated field, it is more critical to identify and address the shortcomings of existing approaches. To accomplish this, a systematic review of 101 studies using the DEAP dataset for emotion recognition was conducted to identify common methodological issues and evaluate their

impact on reported performance. This chapter also performed an experiment to assess the amount of performance inflation through experimental analysis.

The content of this chapter is organized into four sections. Section 4.1 provides an introduction and motivation for research. Section 4.2 presents the materials and methods, including the review strategy, dataset descriptions, and experimental analysis procedures. Section 4.3 details the results of the systematic review and the experimental validation. Concluding statements, concerns and recommendations are given in Section 4.4.

4.1 Introduction

Emotional human–computer interaction could be advanced by the development of affective brain–computer interfaces (aBCIs). These technologies enable the development of adaptive and automated user interfaces, as well as integration into larger systems as specialized subcomponents, enhancing user experience and interaction⁹⁸. To enhance the generalizability of emotion recognition algorithms across diverse subjects, researchers are increasingly focusing on developing various algorithmic approaches. These efforts include exploring different preprocessing techniques, feature extraction methods, feature selection strategies, and classifier combinations, alongside the integration of machine learning algorithms to optimize performance and accuracy. Machine learning (ML) and artificial intelligence (AI) have become integral components of modern healthcare, transitioning from experimental research to practical deployment. This shift is evident in the increasing number of AI/ML-powered medical devices receiving regulatory approval from the U.S. Food and Drug Administration (FDA)⁹⁹.

Although there have been significant advancements in the field of affective state estimation with the application of machine learning^{100;101}, we contend that the validity of these approaches

is compromised by common errors and pitfalls frequently observed in aBCI experiments. The reliability of aBCI systems depends on the systematic execution of each stage within the machine learning pipeline, which includes data acquisition, noise removal, signal processing, feature extraction and selection, model selection, and performance evaluation. Common factors observed in affective state estimation using EEG that can affect performance include the mixing of training and test datasets, data leakage due to inappropriate feature selection methods, incorrect hyperparameter optimization, classifier selection, and a flawed validation process. These methodological errors vary significantly in their impact, with some causing minimal degradation while others severely compromise estimation accuracy.

Segmentation, or windowing, is a widely used practice in EEG analysis, typically employed to increase the effective sample size by dividing long trials into multiple shorter segments^{102–105}. Segmenting a single trial into multiple segments is not inherently problematic. However, it is crucial to ensure that all segments from the same trial are assigned to the same data partition (training, testing, or validation). Distributing segments from a single trial across multiple partitions introduces data leakage, which can lead to significantly inflated performance estimates and compromise the generalizability of the model^{106;107}. Another common pitfall is applying data preprocessing using global statistical properties computed from the whole dataset^{108;109}. For example, baseline correction, a common preprocessing step in EEG-based machine learning, can inadvertently cause data leakage if not applied correctly. Baseline correction often involves computing the mean or median EEG activity before a stimulus and subtracting it from the entire trial. If this correction is performed before splitting the data into train and test sets, it may introduce dependencies between training and test sets. Commonly utilized techniques in machine learning model development include feature selection, hyperparameter optimization, and the selection of both classifiers and

feature sets. However, if these processes are performed using information from the test set or the entire dataset, it can result in data leakage¹⁰⁸. Besides that, bias in emotion recognition datasets can lead to performance overestimation in machine learning models¹¹⁰. When certain emotions are overrepresented in the data, the model may favor these emotions. Class imbalance can mislead performance metrics by favoring the majority class^{95;110;111}. All these issues, which are present in published studies, lead to overfitting, reduced generalizability, and misleading evaluation metrics that do not accurately reflect the real scenario of aBCI model.

Notably, the methodological challenges discussed in this context are not unique and centered only on aBCI applications¹¹². They are also commonly observed in other machine learning applications, including healthcare, medical diagnostics, neuroimaging, data analysis, cyber security, Internet of Things (IoT), and data mining^{108;113–121}. Schroeder et al. systematically compared four cross-validation strategies to assess their impact on classifier performance estimation in passive BCI experiments¹²². Kapoor et al. identified issues of data leakage and the reproducibility crisis in machine learning-based research across 17 scientific disciplines¹²¹. Similarly, Balendran et al. explored the impact and prevalence of various perturbations in machine learning for healthcare, including data imbalance, feature extraction and selection techniques, training strategies, and hyperparameter tuning¹¹³. These factors significantly influence model robustness, generalization, and performance in medical applications. A recent review provides a comprehensive overview of how machine learning can be intentionally or unintentionally misused in brain imaging¹⁰⁹. Other previous studies have explored the impact of data bias on medical diagnostics and healthcare outcomes^{112;123–125}. The presence of temporal bias, such as data collected at specific times or specific conditions, in current AI and robotic systems is addressed in¹²⁶.

Despite the promising advancements in aBCI systems and their high classification performance

in controlled research environments, their transition to real-world applications remains challenging. A comprehensive analysis of aBCI research published between 2019 and 2023 revealed that only 4.58% of studies explicitly focused on online emotion classification¹¹⁰. Another survey based on aBCI claimed that 90% of the reviewed research utilized offline classification methods, while only 8% implemented online classification, which are more applicable to real-time scenarios¹²⁷. This limitation is not exclusive to aBCI but reflects a broader trend in health research. The research and development community in BCI has not sufficiently prioritized the evaluation methods necessary for transitioning these systems into practical applications¹. Alarmingly, these pitfalls persist despite being widely recognized, raising concerns about the rigor of the peer-review process^{128–130}. The persistent approval and publication of studies that overlook these critical issues indicate a lack of awareness among reviewers.

In this study, we hypothesized that these common pitfalls occur frequently and consistently in the field of aBCI-based classification algorithms and are often inadequately addressed by reviewers during the evaluation process. Moreover, this limitation not only hinders the practical implementation of aBCI technologies but also results in inefficient use of research efforts and funding. Research inefficiency in biomedical fields represents a significant concern, with a study indicating that approximately 85% of biomedical research is largely inefficient and wasted^{131–133}. Additionally, multiple studies highlighted that these methodological flaws, lack of reproducibility, and limited applicability in real-world settings lead to the inefficient use and misallocation of healthcare research funding, ultimately restricting the contribution to scientific progress^{93;134;135}.

Therefore, we aimed to investigate the methodological flaws in affective state estimation as presented in the literature and to provide recommendations for avoiding these issues to enhance their reliability in real-world scenarios. To achieve this, we performed a comprehensive review of

101 emotion recognition papers using EEG, specifically focusing on the DEAP dataset⁶¹, which is a widely cited publicly available dataset in the field of aBCI. Subsequently, we expanded our investigation to validate the occurrence of observed pitfalls identified in the reviewed papers by conducting experimental analyses. For this purpose, we utilized both the DEAP dataset and data from EEG sensors applied to a watermelon collected in our own laboratory, inspired by the approach in¹³⁶. As part of this analysis, we also examined the extent of performance inflation caused by these methodological errors. Finally, we compare this performance inflation to a standardized fair analysis. This could be used as a reference for best practices to avoid these pitfalls.

4.2 Materials and Methods

4.2.1 Review Paper Strategy

Search strategy and selection criteria: In this study, we used the Scopus database to identify relevant studies for our review. Our search criteria were designed to be selective, focusing on core research papers central to our topic as of February 29, 2024. We specifically targeted affective state estimation studies that used the DEAP dataset, as DEAP is one of the most widely cited databases in emotion recognition⁸⁴. Therefore, we focused on "DEAP" as a keyword in our search. Using keyword combinations such as "EEG," "emotion," "recognition," "using," and "DEAP," we conducted an initial search (see Fig. 4.1). We limited the publication years to 2017 through 2023. Afterward, we applied filters to restrict the language to "English," document type to "article," and source type to "journal". These filtering criteria were implemented to focus solely on journal publications, which typically provide more comprehensive methodological details and novel applications compared to other publication formats. The collected publications were then organized

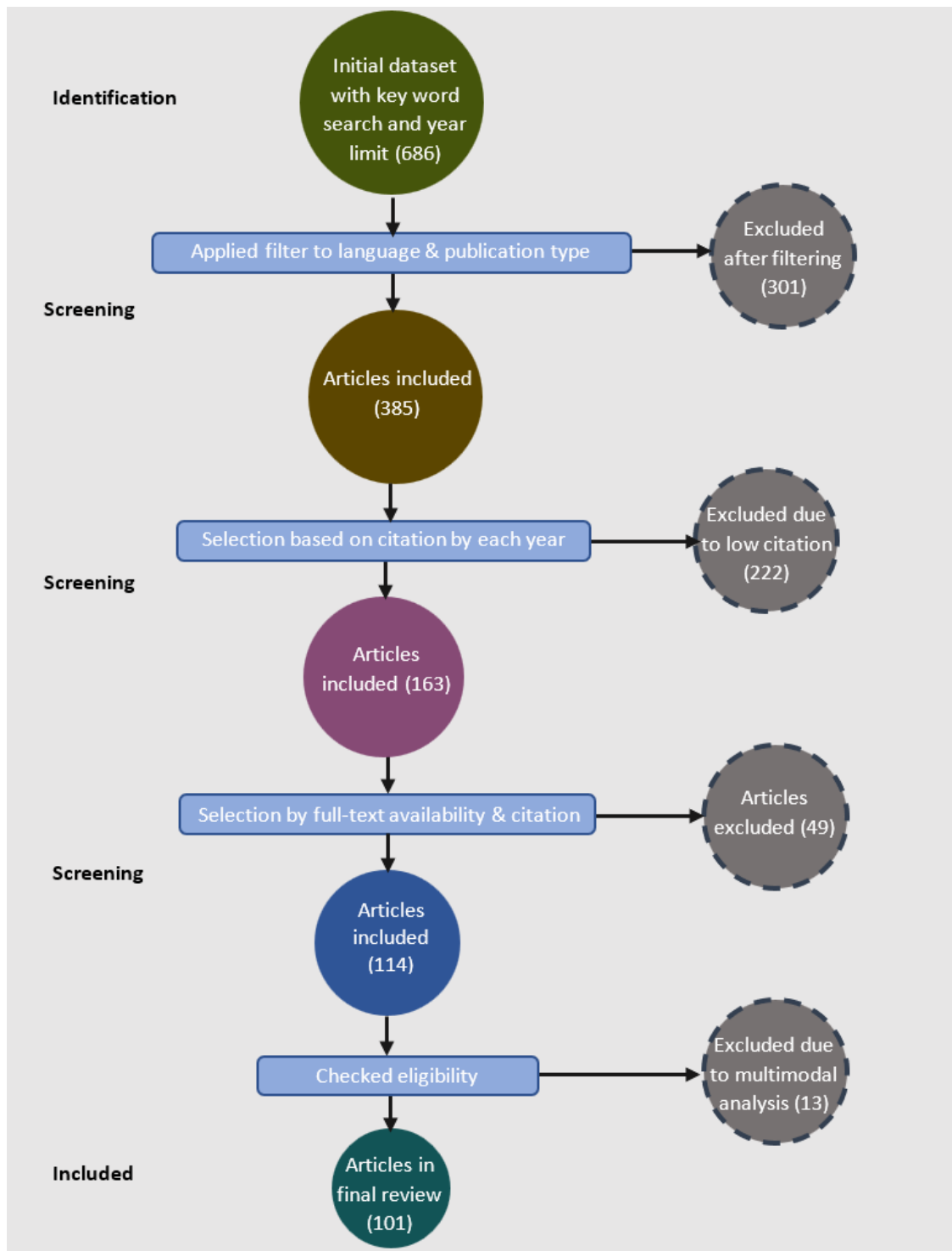


Figure 4.1: Systematic review selection process for EEG-based emotion recognition studies using the DEAP dataset. This flowchart shows the sequential screening and selection methodology, including keyword filtering, citation-based selection, full-text availability assessment, and final inclusion criteria.

for each year, by ranking them from highest to lowest citations. To examine the methodological developments over time, we initially aimed to select the 30 most highly cited papers from each year between 2023 and 2020, while including all papers from 2019, 2018, and 2017 due to their limited numbers. Then, we attempted to select an equal number of 20 papers from each year by considering the full article availability and citation metrics. Finally, we manually excluded the studies that applied multimodal emotion recognition to restrict the focus to EEG-based affective state estimation.

Review process and data extraction: The paper review process was conducted in four stages to ensure rigor and transparency. In the first stage, the initial paper search and selection were performed by one author based on the criteria mentioned above. In the second stage, each selected paper was independently assessed by two authors to evaluate its methodological quality and pitfalls. In the third stage, any discrepancies or conflicts in the assessments were resolved through group discussions involving all authors. Finally, in the fourth stage, the reported outcomes were systematically re-evaluated by all authors to ensure clarity and consistency.

Methodological quality assessment criteria: Based on the extracted information from the revision, the papers were then categorized based on four common methodological pitfalls: data leakage due to segmentation, data leakage in hyperparameter optimization, data leakage in feature selection methods, and model selection using grid search. The defined concepts for each of these issues are detailed in Table 4.1. For each identified methodological pitfall, papers were systematically classified into three categories based on the clarity and appropriateness of their reported procedures:

- **Valid:** Papers that clearly described proper methodological procedures, with explicit indications of appropriate data partitioning and no indication of data leakage or improper practices.

- Invalid: Papers where the described methodology indicated violations of best practices, data leakage, or procedures that would compromise model generalizability.
- Undetermined: Papers that mentioned relevant procedures but provided insufficient detail to determine whether the approach followed a valid evaluation process, often due to vague or incomplete reporting.

Table 4.1: Definitions and examples of common methodological pitfalls identified in EEG-based emotion recognition studies

Pitfalls	Definitions	Examples
Data leakage due to segmentation	Presence of samples that are derived from the same trial after segmentation or windowing, in both the training and testing sets.	A 1-second segmentation was applied to each 60-second trial, resulting in 2400 segments from 40 trials. A k-fold cross-validation was then performed on these 2400 segments, revealing the presence of samples from the same trial in both the training and testing sets ¹³⁷ .
Bias in hyperparameter optimization	Hyperparameter optimization based on the accuracy of both the train and test sets.	Selected configuration and settings of the classifier based on the highest classifier performance ²⁸ .
Data leakage due to feature selection	Dimensionality or feature reduction algorithm was performed using the entire dataset before splitting into training and test sets, which can lead to data leakage.	Dimension reduction algorithm applied on the whole dataset before train/test set partition ¹³⁸ .
Model selection using grid search	Feature or model selection was conducted using grid search procedures where the selection criteria were based on performance metrics evaluated on the test set.	Selected classifier based on highest test set accuracy ¹³⁹ .

Additional reporting quality indicators: We evaluated additional quality indicators, including: (1) class imbalance acknowledgment, (2) statistical significance relative to chance performance, and (3) code availability for reproducibility. For class imbalance assessment, papers were classified as “reported” if they included any of the following metrics that account for class distribution: F1 scores, precision, recall, sensitivity, or specificity. Papers reporting only accuracy or similar aggregate metrics were marked as “not reported”. For statistical significance evaluation, we identified studies reporting confidence intervals or standard deviation bounds and determined whether these overlapped with theoretical chance levels. Theoretical chance levels were calculated based on the number of classes in each classification task (50% for binary classification, 25% for four-class classification). Studies were flagged when the chance classification level was within two of the reported standard deviations, indicating results potentially indistinguishable from random performance. For code availability assessment, papers were marked as “available” only if they provided accessible links to code repositories or pretrained models. Studies that mentioned code availability but lacked any accessible links or provided only pseudocode or implementation summaries within the paper were not counted.

Literature comparison bias analysis: To assess selective reporting bias, we investigated the extent to which published studies accurately report prior models with superior performance. To ensure fair comparison, we restricted our analysis to papers that performed participant-dependent binary classification using a threshold of 5 on the valence or arousal dimension. This resulted in a subset of 31 papers from our original collection. Of these, 21 papers included explicit comparison tables listing prior models. For each of these 21 papers, we extracted: (1) the reported accuracy of each paper’s proposed model, and (2) the highest accuracy reported in their comparison table

from prior studies. Pearson correlation coefficients were computed to assess the relationship between proposed and literature-reported accuracies for valence and arousal tasks separately. To further quantify underreporting, we cross-referenced our review database to identify how many high-performing models had been published prior to each paper’s submission date.

4.2.2 DEAP Emotion Dataset

The DEAP dataset⁶¹ is a widely utilized benchmark for emotion recognition using EEG signals. It consists of EEG recordings from 32 healthy participants (50% female, aged 19–37; mean age: 26.9 ± 4.45) while they watched 40 music videos designed to elicit emotional responses. The database provides two datasets: raw data and preprocessed data. The raw data includes 32 BioSemi .bdf files, each containing 48 recorded channels: 32 EEG channels, 12 peripheral channels (including electrooculography (EOG), electromyography (EMG), galvanic skin response (GSR), and respiration), three unused channels, and one status channel. The EEG signals were originally recorded at a sampling rate of 512 Hz. To facilitate analysis, a preprocessed version of the dataset was provided, incorporating several modifications. The data was downsampled to 128 Hz, EOG artifacts were removed, and a 4–45 Hz band-pass filter was applied to retain relevant EEG frequency components. Additionally, the signals were re-referenced using a common average reference, and the data were segmented into 60-second trials, with a 3-second pre-trial baseline. Each participant has 40 trials. Following each video stimulus, participants provided self-assessment ratings on a nine-point scale for four emotional dimensions: valence, arousal, dominance, and liking. Emotional states in the DEAP are often interpreted using Russell’s circumplex model, where valence and arousal define a two-dimensional emotional space.

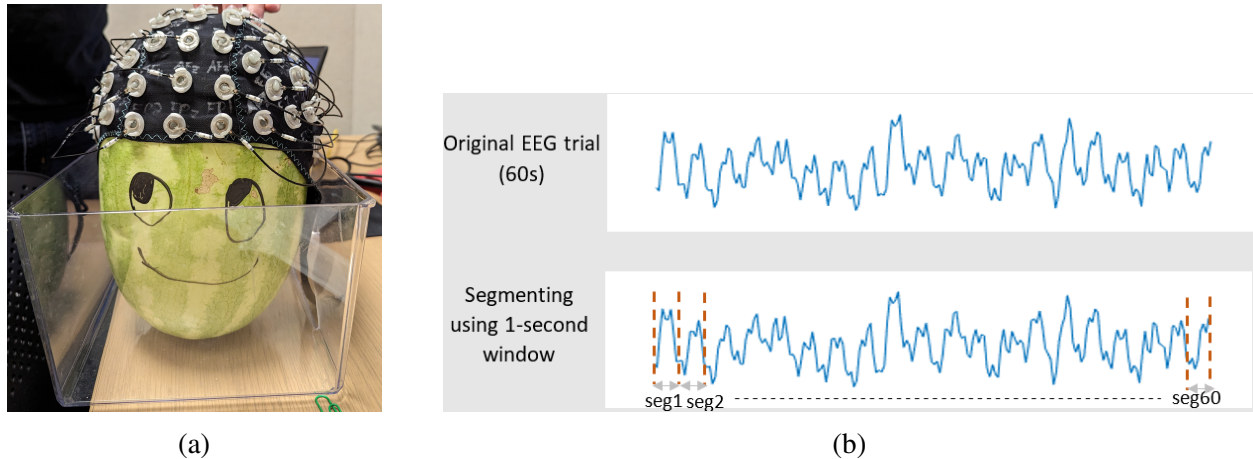


Figure 4.2: EEG data collection and segmentation. (A) Demonstration of EEG data collection using a watermelon phantom. (B) EEG trial segmentation process, illustrating a 60-second EEG trial divided into 1-second non-overlapping windows, resulting in 60 segments per trial. This approach is commonly used to augment sample size but may introduce data leakage if segments from the same trial are distributed across training and test sets. Segments are illustrative and not shown to scale.

4.2.3 Watermelon Dataset

The EEG collection method using watermelon was previously referred to as phantom EEG in earlier studies^{136;140;141}. The watermelon EEG data focuses solely on temporal correlation features without capturing any stimulus-driven neural responses¹³⁶. Here, watermelon EEG data acquisition was performed using the Cognionics Mobile-72 EEG system, a high-density mobile EEG device with 64 channels (Fig. 4.2a). We collected EEG data from the watermelon for more than one hour to ensure sufficient data for the classification task. The recordings were conducted at a sampling frequency of 500 Hz. The EEG cap, equipped with standard active Ag/AgCl electrodes, was positioned according to the international 10–20 system, a widely recognized standard in EEG studies. Reference and ground electrodes were placed on either side of the watermelon to mimic standard mastoid placements and ensure signal quality. The watermelon EEG data were then reorganized to follow the same structure as the DEAP dataset. From the first one-hour recording,

40 trials were extracted, each lasting 60 seconds. A 30-second interval was maintained between consecutive trials to avoid unwanted influences between trials. These 40 trials were then randomly assigned binary labels, either 1 or 0, for the final classification task.

4.2.4 Procedure of Experimental Analysis

In a typical EEG-based emotion recognition classification framework, the process generally follows a standardized pipeline that includes data preprocessing, feature extraction, feature selection, train-test partitioning, classifier selection, and performance evaluation using metrics such as accuracy and F1-score. In this study, we implemented all these steps; however, to systematically assess the significance and impact of specific methodological pitfalls, we prioritized one pitfall per experiment. Accordingly, we conducted separate experiments focusing on segmentation, feature selection, and hyperparameter optimization. In all cases, the same experimental protocol was applied consistently to both the DEAP and watermelon datasets.

Segmentation analysis: To investigate the improper use of EEG trial segmentation, we performed subject-dependent binary classification using the raw DEAP dataset. Each trial in the DEAP dataset originally lasted 60 seconds (Fig. 4.2b). To increase the sample size, we segmented each trial into t -second intervals, resulting in a total of $N=60/t$ segments per trial. For instance, segmenting into 1-second intervals resulted in a total of 2400 trials ($40 \text{ trials} \times 60 \text{ seconds}$). By varying the segment length (t), we assessed the effect of segmentation on classification performance. A threshold of 5 was applied, where ratings less than or equal to 5 (≤ 5) were considered low states, and ratings greater than 5 (> 5) were considered high states. For each emotional axis, there are two classes: high valence vs. low valence, high arousal vs. low arousal, and high dominance vs. low dominance.

In emotion recognition studies, EEG data are often analyzed across five frequency bands: delta (1-4 Hz), theta (4-8 Hz), alpha (8-14 Hz), beta (14-30 Hz), and gamma (30-42 Hz). The power spectral density (PSD), one of the most used features in aBCI classification, was computed using MATLAB built-in 'pwelch' function for those frequency bands as the feature for classification.

Various classification algorithms, including decision trees, logistic regression, Naive Bayes, k-nearest neighbors (kNN), support vector machine (SVM), CNN, and neural networks (NN) have been employed for binary classification in prior studies. In our study, we selected kNN as the primary classifier for categorizing binary emotional states across all three dimensions. The kNN classifier was implemented using MATLAB's built-in 'fitcknn' function with default parameters, with the number of nearest neighbors (k) specifically set to 5. To rigorously evaluate the classifier's performance, we performed two distinct validation methodologies: k-fold cross-validation and leave-one-trial-out validation. For the former approach, 5-fold cross-validation was applied to all 2400 samples from each subject. In the latter approach, we implemented a leave-one-trial-out validation method across the original 40 trials, ensuring that all segmented samples derived from a single trial remained within the same fold. This methodological choice removes the potential data leakage between training and testing sets that could occur due to segmentation. Specifically, the training data comprised segmented samples from 39 trials, while the testing data was exclusively derived from the remaining trial. This process was iteratively repeated until each trial had served as a test set. Here, a well-known performance metric, balanced accuracy, was also applied to evaluate the performance of both validation methods. Balanced accuracy, which accounts for the potential class imbalance in the emotional state classification task, is calculated by averaging the individual class accuracy for all classes⁹⁵. The same preprocessing protocols, feature extraction techniques, and classification methodology were subsequently applied to the watermelon dataset

for comparative analysis.

Feature selection analysis: Here, we employed the DEAP preprocessed dataset, which is also widely used for affective state estimation research. We analyzed all trials of 60-second experimental signals after removing the 3-second baseline, and applied the same thresholding with a value of 5 to divide the 40 trials into binary classes. As the DEAP preprocessed data was already filtered through a 4-45 Hz band-pass filter, we calculated the PSD across four frequency bands: theta (4-8 Hz), alpha (8-14 Hz), beta (14-30 Hz), and gamma (30-42 Hz). For feature selection, we implemented a statistical t-test to identify the most relevant features while reducing redundancy. Classification was performed using the kNN classifier with default parameters. However, we conducted subject-dependent experiments using 70% of the trials as the train set and 30% of the trials as the test set.

We systematically performed two feature selection methods and compared these strategies. The first is Global Selection, where t-test feature selection was applied to the entire dataset before splitting into train and test sets. This is a common but flawed practice that can introduce data leakage by allowing test data to influence feature selection, potentially inflating performance metrics. Second, we used Local Selection, where feature selection was performed only on the training set. At first, t-test feature selection was applied to the entire training set to rank features, ensuring that the test set remained completely unseen during feature selection. Using that fixed ranking, we ran 5-fold cross-validation on the training set with top-k features ($k=1-50$), and averaged the accuracies across the 5 folds. The k with the highest average accuracy was chosen, then we retrained on the full training set with those features and evaluated on the held-out test set. The latter approach maintains strict separation between training and test data, preventing information leakage and

yielding more realistic performance estimates despite potentially introducing variability in selected features across folds. The effectiveness of both approaches was assessed using both accuracy and balanced accuracy.

Hyperparameter optimization: In this analysis, we also employed the preprocessed DEAP dataset and maintained identical pre-processing and feature extraction methodologies as outlined in the preceding section. We also applied the same validation approach, a subject-dependent experiment using 70% of the data for training, and 30% for testing. To investigate potential data leakage during hyperparameter optimization, we implemented two distinct approaches with a kNN classifier. In the first approach, we repeatedly adjusted the classifier’s hyperparameters (number of neighbors, standardization setting, and distance metric), selecting the combination that yielded the highest classification accuracy on the test data. This method, while yielding optimal performance metrics, risks overfitting to the test set. In contrast, the second strategy followed proper machine learning practice by isolating the test set entirely during the tuning process. We employed a 5-fold cross-validation procedure within the training set to optimize the combination of hyperparameters, including the number of neighbors, standardization (enabled or disabled), and type of distance metric. For each hyperparameter setting, the training data was partitioned into five folds. In each fold, the model was trained on 4 folds and validated on the remaining one. The accuracy was computed on the validation fold and averaged across all five folds. The combination of hyperparameters that achieved the highest average balanced accuracy was selected as optimal. The final model was then retrained on the full training set using these optimal settings and subsequently evaluated on the untouched test set. This approach ensured that no test data was used during model tuning, thus avoiding data leakage and providing a more reliable estimate

of generalization performance. Both regular accuracy and balanced accuracy were chosen as the performance metrics to compare both approaches, providing insights into how optimization methodology impacts classification results.

4.3 Results

4.3.1 Study Selection

Our systematic literature search, using keywords related to emotion recognition and EEG (e.g., “DEAP”, “EEG emotion classification”), yielded 686 papers published between 2017 and 2023. Fig. 4.1 illustrates a summarized overview of the selection process. After applying filters, the dataset was refined to 385 publications. At this stage, we identified 146 papers from 2023, 98 from 2022, 51 from 2021, 47 from 2020, 22 from 2019, 10 from 2018, and 11 from 2017. We implemented a two-stage selection process. In the initial stage, we selected 163 candidate papers, including the 30 papers with the highest citations from each year. For years with fewer available papers, we included all publications: 22 from 2019, 10 from 2018, and 11 from 2017. Subsequently, we prioritized full-text availability and citation metrics to select 114 papers for comprehensive review (20 papers from each year between 2023 and 2020, 18 from 2019, 8 from 2018, and 8 from 2017). During the detailed analysis phase, 13 papers were excluded due to their multimodal methodological approaches. Consequently, the final review included 101 papers for our systematic analysis and synthesis.

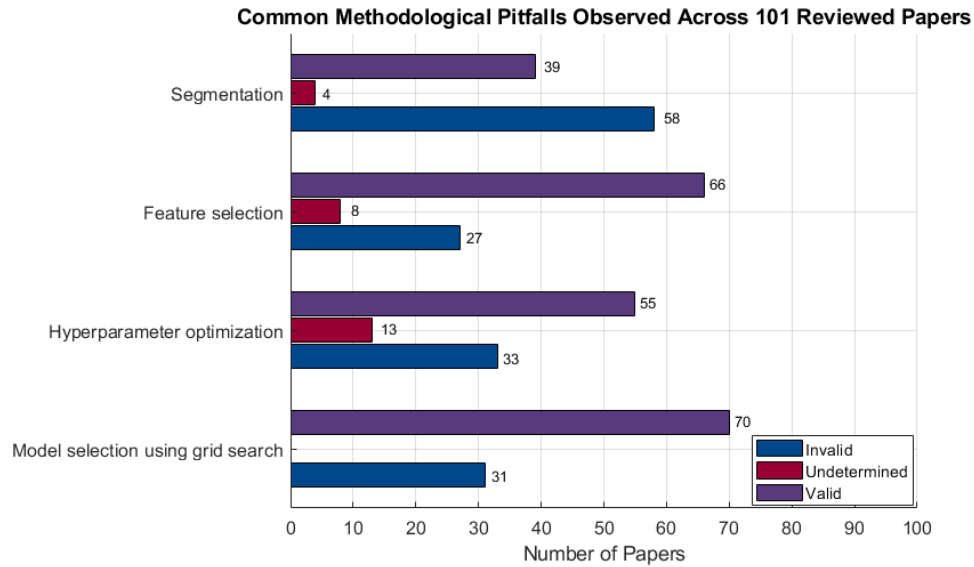
4.3.2 Prevalence of Methodological Pitfalls from Reviewed Papers

In our systematic review, we identified several prevalent methodological issues across the reviewed papers. Here, our analysis specifically focused on four critical methodological pitfalls that represent the most common sources of data leakage and performance inflation: segmentation strategy, feature selection method, hyperparameter optimization, and model selection using grid search (Table 4.1). For each category, papers were categorized as valid, invalid, or undetermined based on the methodological descriptions provided in the paper. For hyperparameter optimization and feature selection, papers were marked as invalid when the language in the text indicated that hyperparameter optimization was performed by evaluating model performance on the entire dataset or dimension reduction algorithm was applied on the whole dataset prior to train/test splitting. Some cases required judgement calls due to the lack of detailed reporting. However, the language used often led us, as secondary reviewers, to reasonably believe that the error occurred. Some papers mentioned performing hyperparameter search or feature reduction method. But the descriptions were so vague that it was unclear whether the process was conducted in a valid manner. These were marked as undetermined. Papers were marked as valid when there was no indication of methodological error in the text. For example, this included cases when reduction techniques were applied only to the training set, or when a dedicated validation set was used for optimization.

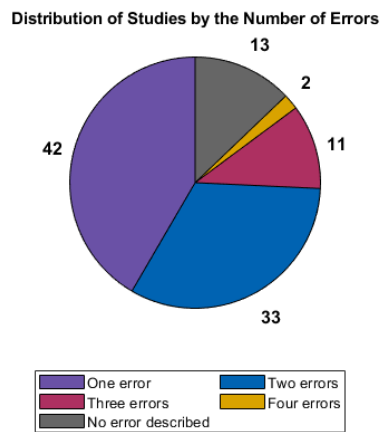
In the case of segmentation, papers were marked as valid when the text, figures, or reporting, including the order of steps described in the methods section or the structure of performance matrices, indicated that segmentation was performed after the train-test splitting. This also included cases where appropriate validation strategies, such as leave-one-trial-out or leave-one-subject-out, were explicitly mentioned. Papers were marked as invalid when segmentation appeared to have been conducted prior to the train-test split, based on their descriptions of using k-fold or percentage-

based cross-validation on already segmented data. Studies were marked as undetermined when segmentation was mentioned, but the available information was insufficient to determine whether proper evaluation procedures were followed. For model selection via grid search, papers were marked as invalid if they reported a table or figure where test set performance biased the feature or model selection process.

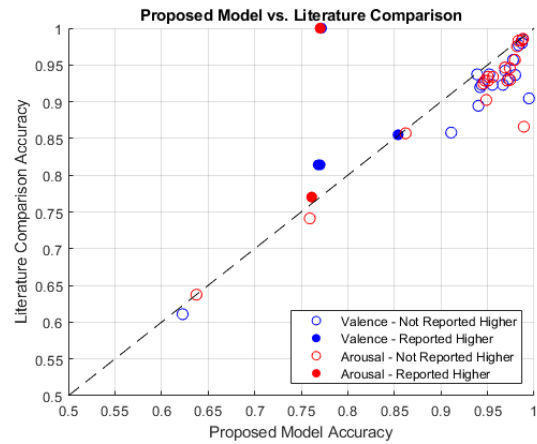
Fig. 4.3a presents the frequency of four common methodological pitfalls identified across 101 reviewed papers. The most frequent issue was observed in segmentation, with 58 papers marked as invalid and an additional 4 as undetermined. Hyperparameter optimization was the second prominent observed error, with 33 papers marked as invalid and 13 as undetermined. This observed error also indicates that nearly half of the reviewed studies either applied questionable optimization procedures or failed to report them clearly. Feature selection method was also a significant concern, with 27 papers marked as invalid and 8 as undetermined. Although feature selection methods showed 66 studies marked as valid in Fig. 4.3a, it is important to clarify that this high number is largely due to the fact that 64 out of 101 papers did not apply any feature or dimension reduction method at all. Among the 37 papers that did apply such algorithms, only 2 papers implemented them correctly by applying the procedure solely to the training set, and were therefore marked as valid. In contrast, 27 papers applied feature selection across the entire dataset, and 8 papers provided insufficient detail to assess their procedure. This highlights a critical issue: among studies that applied feature selection methods, the vast majority (35 out of 37) performed so improperly or unclearly- a concern masked by the aggregated valid count. Grid search-based model selection also showed a methodological issue, with 31 of the 101 papers exhibiting errors, indicating that this approach is not free from concern. To further examine the potential impact of this error, we analyzed the magnitude of inflation it could introduce. For papers that reported test



(a)



(b)



(c)

Figure 4.3: Prevalence and distribution of methodological errors in EEG-based emotion recognition studies. (A) Distribution of common methodological issues across 101 reviewed studies, focusing on segmentation, feature selection method, hyperparameter tuning, and model selection using grid search. Studies were categorized as valid (purple), undetermined (red), or invalid (blue). (B) Distribution of studies based on the number of methodological errors identified, showing the proportion of papers containing one error ($n=42$), two errors ($n=33$), three errors ($n=11$), four errors ($n=2$) and no errors described ($n=13$). (C) Comparison of proposed model accuracy versus the highest prior accuracy cited in 21 studies. Blue and red circles represent valence and arousal tasks, respectively; filled markers indicate papers that acknowledged higher-performing prior models. Correlations: $r = 0.78$ (valence) and $r = 0.79$ (arousal).

set performance under multiple configurations, we collected the highest and lowest reported test accuracies—reflecting the selection of models or features based on test set results. The difference in reported test accuracy ranged from 0.24 to 49.7 percentage points. While we do not claim that all papers experienced large inflation due to these issues, the fact that accuracy could vary by up to 49 percentage points demonstrates that the impact of such methodological flaws cannot be ignored.

Fig. 4.3b illustrates the distribution of papers based on the number of methodological faults simultaneously present, focusing on four key areas. Our analysis revealed a concerning pattern: 42 papers contained one of these methodological errors, while 33 exhibited two errors. Also, 11 studies demonstrated three concurrent methodological flaws, and 2 studies presented with all four, potentially compromising their reported outcomes. Our review identified only 13 of the 101 studies that did not exhibit any of these four methodological errors based on their reported procedures and documentation.

4.3.3 Additional Reporting and Reproducibility Concerns

In addition to evaluating methodological procedures, we identified several additional qualities and reporting concerns. We assessed how class imbalance was addressed across the reviewed studies. Among the reviewed studies, only 38 papers reported F1 scores or complementary metrics such as precision, recall, sensitivity, and specificity, thus acknowledging potential class distribution concerns. Furthermore, we identified seven studies that reported results with confidence intervals or standard deviation bounds overlapping the theoretical chance level in at least one experimental task. Notably, these results were not acknowledged as statistically indistinguishable from chance in their respective discussions or limitations sections. This raises concerns about the rigor of peer review and highlights that studies with no statistically meaningful outcomes may still be

presented—and accepted—as valid contributions.

Another concern is comparison bias, where studies selectively omit prior models with superior performance in their comparison tables. Of the 31 eligible binary classification studies (based on threshold 5 and participant-dependent analysis), 21 included comparison tables with results from previous publications. Fig. 4.3c shows the relationship between the reported performance of each paper’s model and the highest accuracy listed in its own comparison table. Valence and arousal accuracies are shown separately, yielding correlations of 0.78 and 0.79 for valence and arousal tasks, respectively. Notably, only 4 out of 21 papers acknowledged literature results that exceeded their own accuracy, suggesting a tendency to underreport stronger prior models. To further quantify this bias, we identified how many higher-performing studies from our review had been published prior to each paper’s submission date. Our analysis revealed that thirteen of these papers (62%) had at least three or more previously published models that achieved superior performance but were not mentioned in their comparison tables. This is likely an underestimate because we only checked references among the 21 papers in our review. Even though these are highly cited, other high-performing papers outside our sample were probably overlooked.

As part of our review, we also assessed the reproducibility of the reported studies by checking the availability of code or pretrained models. Out of the 101 papers, only eight provided accessible links to their code repositories or shared models. However, most of these links lacked sufficient content to fully replicate the reported experiments. An additional three papers stated that their code would be made available, but did not include a link or any accessible resource at the time of our review.

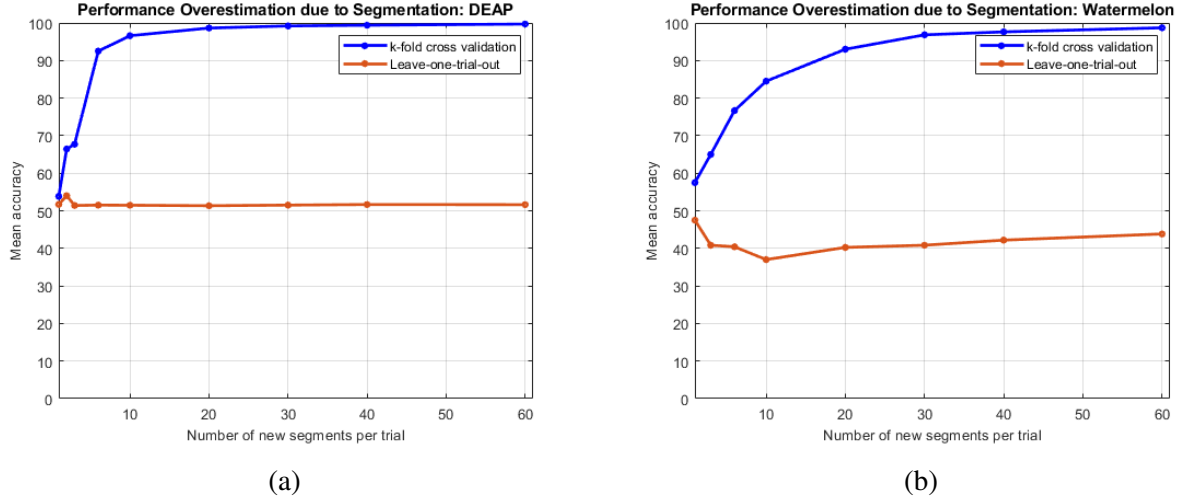


Figure 4.4: Performance overestimation due to windowing/segmenting trial in EEG-based binary classification. (A) Comparison of mean accuracy between k-fold cross validation and leave-one-trial-out validation approaches as segments per trial increase using DEAP dataset in the valence axis. (B) Binary classification accuracy trends for watermelon dataset under the same segmentation conditions. In both cases, k-fold cross validation (blue) results in inflated accuracy estimated compared to leave-one-trial-out validation (orange).

4.3.4 Experimental Analysis: Segmentation

In this experiment, we performed binary classification on the valence axis of the DEAP dataset, after segmenting each trial into multiple segments. Additionally, binary classification was conducted on the watermelon dataset with randomly generated binary labels. Fig. 4.4 shows how different validation strategies impact classification performance when segmentation or windowing is applied, using the DEAP dataset and watermelon dataset. The graph highlights a notable difference between two validation methods: k-fold cross-validation and leave-one-trial-out, as the number of segments per trial increases. In leave-one-trial out, all segments from an original trial are kept together, whereas in k-fold, segments from an original trial may end up in different folds. In Fig. 4.4a, the k-fold cross-validation method shows a significant improvement in performance, increasing from about 53% accuracy without segmentation to over 90% with just 6 segments per trial, and nearly 100% with 60 segments. On the other hand, the leave-one-trial-out validation method

remains relatively consistent at around 52% accuracy, regardless of the segmentation level. A comparable phenomenon was also observed in the binary classification experiments using the watermelon dataset, where similar patterns of performance inflation occurred despite the absence of meaningful neurophysiological signals (Fig. 4.4b). These results conclusively demonstrate that conventional k-fold cross-validation after segmentation can cause inflated performance estimates.

Table 4.2: Comparison of feature selection strategies in binary classification across DEAP and watermelon datasets

		Regular accuracy		Balanced accuracy	
		Global	Local	Global	Local
DEAP	Valence	77.41	61.98	75.31	59.46
	Arousal	76.77	62.63	70.39	54.74
	Dominance	77.22	61.85	71.63	54.72
Watermelon		65.17	50.33	–	–

4.3.5 Experimental Analysis: Feature Selection Method

To assess the performance inflation caused by improperly implemented feature reduction procedures, we conducted an experimental comparison of valid and invalid reduction strategies using the DEAP and Watermelon datasets (Table 4.2). To perform binary classification, we tested two validation approaches. In the Global Selection approach, feature selection was applied on the whole dataset, while Local Selection exclusively used training data for feature selection. Two performance metrics are reported: regular accuracy, which does not account for class imbalance, and balanced accuracy, which accounts for class bias. For the DEAP dataset in the valence axis, Global Selection inflated conventional accuracy by 15.5 percentage points (77.4% versus 61.9%) and balanced accuracy by 15.9 percentage points (75.3% versus 59.4%). Similar patterns were observed for arousal (inflation of 14.1 and 15.6 percentage points) and dominance (inflation of 15.3 and 16.9

percentage points) dimensions. Notably, the watermelon dataset, which contains no neurophysiological signals, showed a 14.8 percentage point inflation in regular accuracy. Since this dataset contains balanced class labels, regular accuracy and balanced accuracy are almost equivalent by definition; thus, only regular accuracy is reported. The Global Selection approach consistently yielded inflated accuracy compared to the Local Selection approach. These findings quantitatively demonstrate that improper implementation of feature reduction algorithms can substantially overestimate classification performance, with inflation magnitudes ranging from approximately 14-17 percentage points across all conditions.

Table 4.3: Comparison of hyperparameter optimization strategies in binary classification across DEAP and watermelon datasets

		Regular accuracy		Balanced accuracy	
		Incorrect	Correct	Incorrect	Correct
DEAP	Valence	84.42	62.52	83.16	58.91
	Arousal	82.26	60.69	78.95	53.79
	Dominance	82.96	61.56	78.01	53.04
Watermelon		77.66	50.50	–	–

4.3.6 Experimental Analysis: Hyperparameter Optimization

In our hyperparameter optimization experiments, we observed consistent performance disparities between compromised and methodologically correct approaches (Table 4.3). Two experimental approaches were employed: hyperparameter optimization applied to both the training and test sets, and optimization restricted to the training set only. In the valence axis, the regular accuracy increased by 21.9 percentage points when hyperparameter tuning was improperly performed using test data (84.4% vs. 62.5%). Similar inflations were observed for arousal (21.5 percentage points) and dominance (21.4 percentage points). The balanced accuracy metric, which compensates for

class imbalance, also showed substantial inflation: 24.2, 25.2, and 24.9 percentage points for valence, arousal, and dominance, respectively. Most strikingly, our watermelon dataset showed an increase in accuracy from 50.5% to 77.6% (a 27.1 percentage point inflation) when hyperparameters were improperly optimized using test data. These results highlight that improper inclusion of the test set in hyperparameter optimization leads to an overestimation of performance, with inflation magnitudes ranging from approximately 21 to 27 percentage points across all evaluated conditions.

4.4 Discussion

In this study, we investigated the prevalent methodological errors commonly encountered in aBCI experiments by analyzing the most highly cited publications utilizing the DEAP dataset from 2017 to 2023. Our findings indicate that while many studies report high-performance results in EEG-based emotion estimation, a significant portion of these results may be attributed to the methodological pitfalls. The high prevalence of data leakage during segmentation (58 publications) highlights a fundamental misunderstanding of proper validation protocols. Similarly, the misapplication of feature reduction algorithms (27 publications) and hyperparameter optimization (33 publications) across entire datasets rather than restricting them to the training phase indicates a critical methodological gap. Notably, 87% of the reviewed papers exhibited at least one or more of these four pitfalls. In addition to these explicit errors, we identified a substantial number of papers where the methodological description was insufficient to assess validity —specifically, 4 in segmentation, 8 in feature reduction, and 13 in hyperparameter optimization. Many of the papers in this review failed to document critical methodological details, such as thresholding criteria, data partitioning strategies, or whether analyses were subject-dependent. Although these cases were not labeled as definitively invalid, these undetermined categories also reflect a broader concern

with inadequate transparency. Vague or incomplete methodological descriptions obscure potential errors that should be identified during the peer review process.

This concern is further underscored by additional reporting issues we observed, including limited reporting of class-imbalance metrics, failure to acknowledge results statistically indistinguishable from chance, and a lack of acknowledgment of prior models that achieved higher accuracy. We found that even within the limited scope of this review, authors seem to be selectively choosing papers with lower performance than their own. Given that the primary contribution of these papers is often the model performance itself, failing to cite models with higher performance is a major issue. Together, these findings highlight not only pervasive implementation errors but also a lack of reporting rigor and review scrutiny.

Our experimental analysis demonstrated that high classification accuracy can be achieved through segmentation even when the signals themselves have no information relevant to the labels. In fact, we showed that if data leakage is allowed to occur through segmentation, this can cause an accuracy increase of up to 46 percentage points. Moreover, hyperparameter optimization or feature selection errors in an individual can inflate performance estimates by approximately 14-25 percentage points. More concerning is that multiple methodological errors can occur simultaneously, which could significantly inflate the classification accuracy. In this experimental study, we conducted an in-depth analysis of the segmentation, feature selection method, and hyperparameter optimization; however, the effects of bias and data imbalance were not extensively examined. For in-depth observation, we refer to our previous analysis, which demonstrated that neglecting class imbalance can also lead to misleading results⁹⁵.

Our findings also highlight that a lack of expertise in machine learning can lead to overestimation, raising concerns about the reliability of current aBCI research. Notably, our review was conducted

solely based on the documented methodologies and reported procedures in the publications. While only 8 out of 101 papers provided accessible code repositories, we examined available code when provided to verify methodological descriptions. However, the majority of our assessment relied on the information reported in the publications themselves, without access to training models or implementation details. The identification of these pitfalls through careful examination of published materials indicates that these issues should be detectable during peer review. Our ability to identify these methodological flaws using only the information available in publications suggests the potential gaps in the rigor of the review process, where critical methodological details may not receive sufficient scrutiny. This finding underscores the need for more thorough methodological assessment during peer review to ensure research validity and reproducibility.

Therefore, we propose some critical recommendations to mitigate the pitfalls observed in our review. A primary concern is preventing data leakage between training and testing sets, as such contamination invalidates the reported performance. To ensure a clear separation, all methodological procedures must remain independent of test data¹²¹. All methodological decisions, such as preprocessing techniques, feature selection approaches, and hyperparameter optimization, should be executed exclusively on training data, without incorporating information from the test set^{108;111}. Decisions regarding classifier selection and feature type, and frequency bands, should be derived solely from training data performance rather than prioritizing combinations that maximize test set outcomes. Choosing a suitable validation strategy is essential, especially for segmentation. Leave-one-trial-out or leave-one-subject-out are more appropriate. Our experimental analysis also demonstrates the correct application of these approaches, serving as a guideline to avoid common pitfalls. Additionally, appropriate performance metric selection is essential for addressing class imbalance issues; conventional metrics such as F1 score, precision, and sensitivity may

prove inadequate when calculated predominantly on majority class samples. Balanced accuracy represents a better alternative metric due to its inherent insensitivity to class imbalance⁹⁵. To ensure the model's robustness, it should be evaluated on multiple datasets and through online testing. Researchers must ensure that the test set is unbiased and representative of the target distribution. Furthermore, insufficient methodological details hinder model reproducibility. All model parameters and relevant information should be clearly reported. Additionally, all code, training models, and datasets should be accessible to all. Finally, ensuring that reviewers have comprehensive knowledge across relevant domains is crucial. Insufficient knowledge in any single aspect of BCIs or machine learning can lead to deficiencies in experimental design, statistical analyses, or result interpretation. Adding reviewers from multidisciplinary fields might be helpful to identify methodological shortcomings in both fields. Additionally, reviewers should critically evaluate model reproducibility and availability during the review process.

Our review methodology has certain limitations. We used only the Scopus database for our literature search, which may have restricted the scope of included studies compared to using multiple databases. The keyword search specifically targeted studies on emotion recognition using the DEAP dataset, which we selected due to its high citation frequency in the field. These limitations may have led to the exclusion of broader developments in the field. However, we believe that using a single dataset allows for a more consistent and comparable evaluation of methodological errors. Future reviews would benefit from incorporating multiple databases, expanding the study period, and including studies utilizing diverse emotional datasets to provide more comprehensive insights. Determining whether a method within a paper committed a methodological error based solely on the text is a difficult task, and inherently asks the reviewer to make many judgment calls. To reduce bias, multiple reviewers assessed each paper and resolved discrepancies through discussion. While

individual assessments may not be flawless, we believe this approach provides a valid evaluation of the literature and supports the reliability of our aggregate findings.

In conclusion, our comprehensive analysis of the affective BCI literature reveals widespread methodological shortcomings that significantly impact the field's scientific rigor. By identifying and replicating these issues, we emphasize the necessity for stricter validation strategies, transparent reporting, and unbiased evaluation metrics. Our findings are not merely a critique but a call to the scientific community to uphold higher standards in experimental design, peer review, and result interpretation. Failure to do so may compromise the integrity of the findings and continue to perpetuate biased and misleading studies. Addressing these concerns will enhance the reliability of aBCI research, fostering reproducible and impactful advancements in the field.

Acknowledgments

This work was supported by the National Science Foundation (NSF) under award #1910526 and by the Cognitive and Neurobiological Approaches to Plasticity (CNAP) Center through the NIH NIGMS COBRE Phase II award (P20GM113109). The authors also acknowledge support from the CN Core facilities at Kansas State University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Chapter 5

Conclusion and Future Work

Researchers worldwide are actively exploring ways to make brain-computer interfaces (BCIs) more effective for human communication and device control in real-world environments. This dissertation systematically addresses several critical challenges in BCI research: the impractical burden of lengthy performance validation procedures, the lack of large and standardized datasets for emotion recognition, and the widespread methodological flaws that compromise research reliability. The studies presented here work toward providing practical solutions to these issues and advancing BCIs toward more robust and reproducible real-world applications.

The summary of this dissertation and its contributions are discussed in Section [5.1](#). Section [5.2](#) and Section [5.3](#) discuss the limitations and suggestions for future work, respectively.

5.1 Contribution

5.1.1 P300-Based Brain-Computer Interface Speller Performance Estimation

In this published work³⁹, a protocol was introduced to predict a user’s P300 speller accuracy from a small P300 dataset. To implement this protocol, the Classifier-Based Latency Estimation (CBLE) approach with a least-squares classifier was applied. The method was evaluated using two datasets: the publicly available ‘Brain Invaders calibration-less P300-based BCI dataset using dry EEG electrodes’ dataset (bi2014a) and a P300 speller dataset collected at the University of Michigan. The results of this analysis showed that this protocol could estimate the BCI performance from a relatively small number of typed characters ($\sim 3-8$). The resulting confidence bounds are tighter than those produced by traditional methods. Additionally, a graphical user interface (GUI) titled “CBLE Performance Estimation” was developed to implement the proposed method by other researchers. The GUI was tested and validated on both datasets.

5.1.2 A Comprehensive EEG Database of Emotion Recognition

In addition to improving performance estimation for P300 spellers, this dissertation also addressed the need for comprehensive affective state databases. A database was developed for emotion analysis using EEG signals. This dataset represents one of the largest EEG-based emotion databases with 30 participants and 240 stimuli per participant. Participants rated their emotional responses in two ways: SAM ratings and discrete labels. Subject-dependent classification was performed using power spectral density (PSD) features with a k-Nearest Neighbor (kNN) classifier. The database performance is comparable to existing emotion recognition datasets. This work addresses critical

gaps in existing emotion databases: limited sample sizes, single-modality stimuli, and a lack of dual emotion rating formats. To our knowledge, it is the first EEG dataset to integrate all three stimulus modalities within a single, unified experiment per participant. This design enables stronger benchmarking and cross-modal analyses. The dataset will be made publicly available for the research community.

5.1.3 Methodological Assessment of Affective State Estimation Studies

While developing new databases is important, ensuring the reliability of research using these databases is equally critical. In this study, we reviewed 101 highly cited DEAP-based EEG emotion studies to identify common methodological flaws. Our study revealed that nearly 87% of studies have key issues such as data leakage from improper segmentation, biased feature selection, flawed hyperparameter tuning, class imbalance neglect, and inadequate reporting. To validate these concerns, we expanded our investigation by conducting experimental analyses using both the DEAP dataset and an EEG dataset from a watermelon. From experimental analysis, high classification accuracy on the DEAP dataset can be achieved (inflated by up to 46 percentage points) simply by segmenting single trials due to an improper validation approach. This high accuracy can also be observed even with non-human subjects such as watermelon. Additionally, errors in improper hyperparameter tuning and feature selection inflated results by 14–25 percentage points. These findings reveal how methodological oversights lead to inflated performance estimates, making the reliable implementation of affective state estimation in real-world applications challenging. The experimental analysis demonstrates correct application of these approaches and provides concrete guidelines to help researchers avoid common methodological pitfalls, ultimately improving the reliability and reproducibility of affective BCI research.

5.2 Limitations

While this dissertation addresses critical challenges in BCI implementation, several limitations should be acknowledged. Detailed limitations for each contribution are provided within the respective chapters. The main limitations across this dissertation are summarized here.

- The CBLE performance estimation method shows reduced effectiveness when training and testing data are collected on different days, and requires data from multiple participants to build reliable models. We also tested only a small set of time-sensitive classifiers.
- In the multimodal emotion database, the limitation of the self-assessment approach is the use of a 1-5 rating scale, which introduces participant-dependent bias. Individual differences in scale interpretation and usage patterns can affect ratings. Some participants may consistently use the full range, while others cluster around middle values. This variability complicates cross-participant comparisons and normalization.
- The methodological review focused exclusively on the DEAP dataset, which enabled consistent evaluation but may have limited the scope of findings. Additionally, the experimental analysis focused on three specific methodological issues, while other implementation steps (i.e. preprocessing) may also contribute to performance estimation inflation.

5.3 Future Work

Several directions for future research can extend and improve the work presented in this dissertation across all three contributions. As mentioned above, BCI speller performance-estimation protocol was validated on four classifiers. Future work should test the method with other time-sensitive

classifiers to establish its broader applicability across different classification approaches. The CBLE Performance Estimation GUI currently supports only Brain Invaders (bi2014a) and BCI2000 format. To increase accessibility and practical utility, the GUI should be expanded to accommodate other common dataset formats. This enhancement would allow researchers to apply the CBLE protocol to their own P300 datasets regardless of the acquisition system.

To address bias from the 1-5 rating scale, future work should explore alternative assessment methods or normalization techniques across participants, such as visual analog scales or standardized questionnaires. While this study provides a framework for evoking affective states in controlled laboratory settings, future data collection should incorporate more realistic, everyday scenarios to better reflect real-world BCI applications. The classification results provide baseline performance comparable to established datasets but suggest that traditional feature extraction methods have limited effectiveness for affective state estimation. Future work should explore advanced preprocessing techniques, alternative feature types, and different classifier combinations while carefully avoiding the methodological errors identified in this dissertation to prevent performance inflation.

Bibliography

- [1] He Pan, Peng Ding, Fan Wang, Tianwen Li, Lei Zhao, Wenya Nan, Yunfa Fu, and Anmin Gong. Comprehensive evaluation methods for translating bci into practical applications: usability, user satisfaction and usage of online bci systems. *Frontiers in Human Neuroscience*, 18, June 2024. ISSN 1662-5161. doi: 10.3389/fnhum.2024.1429130. URL <http://dx.doi.org/10.3389/fnhum.2024.1429130>.
- [2] Jinyi Yang. Challenges and trends in brain-computer interface technology. *Applied and Computational Engineering*, 152(1):56–62, May 2025. ISSN 2755-273X. doi: 10.54254/2755-2721/2025.23006. URL <http://dx.doi.org/10.54254/2755-2721/2025.23006>.
- [3] Giulia Liberati, Alessia Pizzimenti, Luca Simione, Angela Riccio, Francesca Schettini, Maurizio Inghilleri, Donatella Mattia, and Febo Cincotti. Developing brain-computer interfaces from a user-centered perspective: Assessing the needs of persons with amyotrophic lateral sclerosis, caregivers, and professionals. *Applied Ergonomics*, 50:139–146, September 2015. ISSN 0003-6870. doi: 10.1016/j.apergo.2015.03.012. URL <http://dx.doi.org/10.1016/j.apergo.2015.03.012>.
- [4] Jonathan R Wolpaw, Niels Birbaumer, Dennis J McFarland, Gert Pfurtscheller, and Theresa M Vaughan. Brain–computer interfaces for communication and control. *Clinical Neurophysiology*, 113(6):767–791, June 2002. ISSN 1388-2457. doi: 10.

1016/s1388-2457(02)00057-3. URL [http://dx.doi.org/10.1016/s1388-2457\(02\)00057-3](http://dx.doi.org/10.1016/s1388-2457(02)00057-3).

- [5] Theresa M Vaughan and Jonathan R Wolpaw. The third international meeting on Brain-Computer interface technology: making a difference. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 14(2):126–127, June 2006.
- [6] B. Wilhelm, M. Jordan, and N. Birbaumer. Communication in locked-in syndrome: Effects of imagery on salivary ph. *Neurology*, 67(3):534–535, August 2006. ISSN 1526-632X. doi: 10.1212/01.wnl.0000228226.86382.5f. URL <http://dx.doi.org/10.1212/01.wnl.0000228226.86382.5f>.
- [7] Valentina Galiotta, Ilaria Quattrocioni, Mariagrazia D’Ippolito, Francesca Schettini, Pietro Aricò, Stefano Sdoia, Rita Formisano, Febo Cincotti, Donatella Mattia, and Angela Riccio. Eeg-based brain-computer interfaces for people with disorders of consciousness: Features and applications. a systematic review. *Frontiers in Human Neuroscience*, 16, December 2022. ISSN 1662-5161. doi: 10.3389/fnhum.2022.1040816. URL <http://dx.doi.org/10.3389/fnhum.2022.1040816>.
- [8] Olga Maslova, Yuliya Komarova, Natalia Shusharina, Alexander Kolsanov, Alexander Zakharov, Evgenia Garina, and Vasiliy Pyatin. Non-invasive eeg-based bci spellers from the beginning to today: a mini-review. *Frontiers in Human Neuroscience*, 17, August 2023. ISSN 1662-5161. doi: 10.3389/fnhum.2023.1216648. URL <http://dx.doi.org/10.3389/fnhum.2023.1216648>.
- [9] Erich E. Sutter. The brain response interface: communication through visually-induced electrical brain responses. *Journal of Microcomputer Applications*, 15(1):31–45, January

1992. ISSN 0745-7138. doi: 10.1016/0745-7138(92)90045-7. URL [http://dx.doi.org/10.1016/0745-7138\(92\)90045-7](http://dx.doi.org/10.1016/0745-7138(92)90045-7).
- [10] Andrea Kübler, Boris Kotchoubey, N. Ghanayim, Thilo Hinterberger, J. Perelmouter, Maggie Schauer, C. Fritsch, and N. Birbaumer. A thought translation device for brain computer communication. *Studia Psychologica*, 40:17–30, 01 1998.
- [11] Jonathan R. Wolpaw, Dennis J. McFarland, Gregory W. Neat, and Catherine A. Forneris. An eeg-based brain-computer interface for cursor control. *Electroencephalography and Clinical Neurophysiology*, 78(3):252–259, March 1991. ISSN 0013-4694. doi: 10.1016/0013-4694(91)90040-b. URL [http://dx.doi.org/10.1016/0013-4694\(91\)90040-b](http://dx.doi.org/10.1016/0013-4694(91)90040-b).
- [12] L.A. Farwell and E. Donchin. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology*, 70(6):510–523, December 1988. ISSN 0013-4694. doi: 10.1016/0013-4694(88)90149-6. URL [http://dx.doi.org/10.1016/0013-4694\(88\)90149-6](http://dx.doi.org/10.1016/0013-4694(88)90149-6).
- [13] Samuel Sutton, Patricia Tueting, Joseph Zubin, and E. R. John. Information delivery and the sensory evoked potential. *Science*, 155(3768):1436–1439, March 1967. ISSN 1095-9203. doi: 10.1126/science.155.3768.1436. URL <http://dx.doi.org/10.1126/science.155.3768.1436>.
- [14] Monica Fabiani, Gabriele Gratton, Demetrios Karis, and Emanuel Donchin. Definition, identification, and reliability of measurement of the p300 component of the event-related brain potential. *Adv Psychophysiol*, 2:1–78, 01 1987.
- [15] John Polich and Albert Kok. Cognitive and biological determinants of p300: an integrative

- review. *Biological Psychology*, 41(2):103–146, October 1995. ISSN 0301-0511. doi: 10.1016/0301-0511(95)05130-9. URL [http://dx.doi.org/10.1016/0301-0511\(95\)05130-9](http://dx.doi.org/10.1016/0301-0511(95)05130-9).
- [16] Reza Fazel-Rezai. *P300-Based Speller Brain-Computer Interface*. InTech, October 2009. ISBN 9789533070049. doi: 10.5772/7483. URL <http://dx.doi.org/10.5772/7483>.
- [17] Simone Fiuza Baptista de Lima Daiane Damaris de Lemos Menezes Pedro Figueiredo Frizzo Ana Cláudia Leite Romero, Ana Carla Regacone. Event-related potentials in clinical research: guidelines for eliciting, recording, and quantifying mismatch negativity, p300, and n400. *Audiology - Communication Research*, 2015. URL <https://www.redalyc.org/articulo.oa?id=391544062002>.
- [18] Kyungho Won, Moonyoung Kwon, Minkyu Ahn, and Sung Chan Jun. Eeg dataset for rsvp and p300 speller brain-computer interfaces. *Scientific Data*, 9(1), July 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01509-w. URL <http://dx.doi.org/10.1038/s41597-022-01509-w>.
- [19] Guanxiong Pei, Haiying Li, Yandi Lu, Yanlei Wang, Shizhen Hua, and Taihao Li. Affective computing: Recent advances, challenges, and future trends. *Intelligent Computing*, 3, January 2024. ISSN 2771-5892. doi: 10.34133/icomputing.0076. URL <http://dx.doi.org/10.34133/icomputing.0076>.
- [20] Xiao Sun and Man Lv. Facial expression recognition based on a hybrid model combining deep and shallow features. *Cognitive Computation*, 11(4):587–597, May 2019. ISSN 1866-9964. doi: 10.1007/s12559-019-09654-y. URL <http://dx.doi.org/10.1007/s12559-019-09654-y>.

- [21] Mihai Gavrilescu. Recognizing emotions from videos by studying facial expressions, body postures and hand gestures. In *2015 23rd Telecommunications Forum Telfor (TELFOR)*, pages 720–723, 2015. doi: 10.1109/TELFOR.2015.7377568.
- [22] Shaoling Jing, Xia Mao, and Lijiang Chen. Prominence features: Effective emotional features for speech emotion recognition. *Digital Signal Processing*, 72:216–231, January 2018. ISSN 1051-2004. doi: 10.1016/j.dsp.2017.10.016. URL <http://dx.doi.org/10.1016/j.dsp.2017.10.016>.
- [23] Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2016. doi: 10.1109/TAFFC.2015.2457417.
- [24] Manuel Oliva and Andrey Anikin. Pupil dilation reflects the time course of emotion recognition in human vocalizations. *Scientific Reports*, 8(1), March 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-23265-x. URL <http://dx.doi.org/10.1038/s41598-018-23265-x>.
- [25] S. Sharanyaa, Tini J Mercy, and Samyukthaa V.G. Emotion recognition using speech processing. In *2023 3rd International Conference on Intelligent Technologies (CONIT)*, pages 1–5, 2023. doi: 10.1109/CONIT59222.2023.10205935.
- [26] Donald Glowinski, Nele Dael, Antonio Camurri, Gualtiero Volpe, Marcello Mortillaro, and Klaus Scherer. Toward a minimal representation of affective gestures. *IEEE Transactions on Affective Computing*, 2(2):106–118, 2011. doi: 10.1109/T-AFFC.2011.7.

- [27] Tim Dalgleish. The emotional brain. *Nature Reviews Neuroscience*, 5(7):583–589, July 2004. ISSN 1471-0048. doi: 10.1038/nrn1432. URL <http://dx.doi.org/10.1038/nrn1432>.
- [28] Sofien Gannouni, Arwa Aledaily, Kais Belwafi, and Hatim Aboalsamh. Emotion detection using electroencephalography signals and a zero-time windowing-based epoch estimation and relevant electrode identification. *Scientific Reports*, 11(1), March 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-86345-5. URL <http://dx.doi.org/10.1038/s41598-021-86345-5>.
- [29] Edgar P. Torres, Edgar A. Torres, Myriam Hernández-Álvarez, and Sang Guun Yoo. Eeg-based bci emotion recognition: A survey. *Sensors*, 20(18), 2020. ISSN 1424-8220. doi: 10.3390/s20185083. URL <https://www.mdpi.com/1424-8220/20/18/5083>.
- [30] Abeer Al-Nafjan, Manar Hosny, Yousef Al-Ohali, and Areej Al-Wabil. Review and classification of emotion recognition based on eeg brain-computer interface system research: A systematic review. *Applied Sciences*, 7(12):1239, December 2017. ISSN 2076-3417. doi: 10.3390/app7121239. URL <http://dx.doi.org/10.3390/app7121239>.
- [31] Ekman P. Pictures of facial affect, 1976. URL <https://cir.nii.ac.jp/crid/1370567187526327572>.
- [32] Paul Ekman, Wallace V. Friesen, Maureen O’Sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E. Ricci-Bitti, Klaus Scherer, Masatoshi Tomita, and Athanase Tzavaras. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, 53(4):712–717, 1987. ISSN 0022-3514. doi: 10.1037/0022-3514.53.4.712. URL <http://dx.doi.org/10.1037/0022-3514.53.4.712>.

- [33] Paul Ekman and Wallace V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, 1971. ISSN 0022-3514. doi: 10.1037/h0030377. URL <http://dx.doi.org/10.1037/h0030377>.
- [34] James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, December 1980. ISSN 0022-3514. doi: 10.1037/h0077714. URL <http://dx.doi.org/10.1037/h0077714>.
- [35] Shangkai Gao, Yijun Wang, Xiaorong Gao, and Bo Hong. Visual and auditory brain–computer interfaces. *IEEE Transactions on Biomedical Engineering*, 61(5):1436–1447, May 2014. ISSN 1558-2531. doi: 10.1109/tbme.2014.2300164. URL <http://dx.doi.org/10.1109/TBME.2014.2300164>.
- [36] H. Serby, E. Yom-Tov, and G.F. Inbar. An improved p300-based brain-computer interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 13(1):89–98, 2005. doi: 10.1109/TNSRE.2004.841878.
- [37] Yisi Liu, Olga Sourina, and Minh Khoa Nguyen. Real-time eeg-based human emotion recognition and visualization. In *2010 International Conference on Cyberworlds*, pages 262–269, 2010. doi: 10.1109/CW.2010.37.
- [38] Xin Hu, Jingjing Chen, Fei Wang, and Dan Zhang. Ten challenges for eeg-based affective computing. *Brain Science Advances*, 5(1):1–20, March 2019. ISSN 2096-5958. doi: 10.1177/2096595819896200. URL <http://dx.doi.org/10.1177/2096595819896200>.
- [39] Nazmun N. Khan, Taylor Sweet, Chase A. Harvey, Seth Warschausky, Jane E. Huggins, and David E. Thompson. P300-based brain-computer interface speller perfor-

- mance estimation with classifier-based latency estimation. *Journal of Visualized Experiments*, (199), September 2023. ISSN 1940-087X. doi: 10.3791/64959. URL <http://dx.doi.org/10.3791/64959>.
- [40] Aya Rezeika, Mihaly Benda, Piotr Stawicki, Felix Gembler, Abdul Saboor, and Ivan Volosyak. Brain–computer interface spellers: A review. *Brain Sciences*, 8(4):57, March 2018. ISSN 2076-3425. doi: 10.3390/brainsci8040057. URL <http://dx.doi.org/10.3390/brainsci8040057>.
- [41] Janis J Daly and Jonathan R Wolpaw. Brain–computer interfaces in neurological rehabilitation. *The Lancet Neurology*, 7(11):1032–1043, November 2008. ISSN 1474-4422. doi: 10.1016/s1474-4422(08)70223-0. URL [http://dx.doi.org/10.1016/s1474-4422\(08\)70223-0](http://dx.doi.org/10.1016/s1474-4422(08)70223-0).
- [42] Niels Birbaumer. Breaking the silence: Brain–computer interfaces (bci) for communication and motor control. *Psychophysiology*, 43(6):517–532, September 2006. ISSN 1469-8986. doi: 10.1111/j.1469-8986.2006.00456.x. URL <http://dx.doi.org/10.1111/j.1469-8986.2006.00456.x>.
- [43] Angela Riccio, Luca Simione, Francesca Schettini, Alessia Pizzimenti, Maurizio Inghilleri, Marta Olivetti Belardinelli, Donatella Mattia, and Febo Cincotti. Attention and p300-based bci performance in people with amyotrophic lateral sclerosis. *Frontiers in Human Neuroscience*, 7, 2013. ISSN 1662-5161. doi: 10.3389/fnhum.2013.00732. URL <http://dx.doi.org/10.3389/fnhum.2013.00732>.
- [44] Andrea Finke, Alexander Lenhardt, and Helge Ritter. The mindgame: A p300-based brain–computer interface game. *Neural Networks*, 22(9):1329–1333, November 2009. ISSN

- 0893-6080. doi: 10.1016/j.neunet.2009.07.003. URL <http://dx.doi.org/10.1016/j.neunet.2009.07.003>.
- [45] Qi Li, Zhaohua Lu, Ning Gao, and Jingjing Yang. Optimizing the performance of the visual p300-speller through active mental tasks based on color distinction and modulation of task difficulty. *Frontiers in Human Neuroscience*, 13, April 2019. ISSN 1662-5161. doi: 10.3389/fnhum.2019.00130. URL <http://dx.doi.org/10.3389/fnhum.2019.00130>.
- [46] Dennis J. McFarland, William A. Sarnacki, George Townsend, Theresa Vaughan, and Jonathan R. Wolpaw. The p300-based brain–computer interface (bci): Effects of stimulus rate. *Clinical Neurophysiology*, 122(4):731–737, April 2011. ISSN 1388-2457. doi: 10.1016/j.clinph.2010.10.029. URL <http://dx.doi.org/10.1016/j.clinph.2010.10.029>.
- [47] Dean J Krusienski, Eric W Sellers, François Cabestaing, Sabri Bayoudh, Dennis J McFarland, Theresa M Vaughan, and Jonathan R Wolpaw. A comparison of classification techniques for the p300 speller. *Journal of Neural Engineering*, 3(4):299–305, October 2006. ISSN 1741-2552. doi: 10.1088/1741-2560/3/4/007. URL <http://dx.doi.org/10.1088/1741-2560/3/4/007>.
- [48] Eric W. Sellers and Emanuel Donchin. A p300-based brain–computer interface: Initial tests by als patients. *Clinical Neurophysiology*, 117(3):538–548, March 2006. ISSN 1388-2457. doi: 10.1016/j.clinph.2005.06.027. URL <http://dx.doi.org/10.1016/j.clinph.2005.06.027>.
- [49] E. Donchin, K.M. Spencer, and R. Wijesinghe. The mental prosthesis: assessing the speed of a p300-based brain-computer interface. *IEEE Transactions on Rehabilitation*

- Engineering*, 8(2):174–179, June 2000. ISSN 1063-6528. doi: 10.1109/86.847808. URL <http://dx.doi.org/10.1109/86.847808>.
- [50] Johannes Höhne. A novel 9-class auditory erp paradigm driving a predictive text entry system. *Frontiers in Neuroscience*, 5, 2011. ISSN 1662-453X. doi: 10.3389/fnins.2011.00099. URL <http://dx.doi.org/10.3389/fnins.2011.00099>.
- [51] Laura Acqualagna, Matthias Sebastian Treder, and Benjamin Blankertz. Chroma speller: Isotropic visual stimuli for truly gaze-independent spelling. In *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, page 1041–1044. IEEE, November 2013. doi: 10.1109/ner.2013.6696115. URL <http://dx.doi.org/10.1109/NER.2013.6696115>.
- [52] G. Townsend, B.K. LaPallo, C.B. Boulay, D.J. Krusienski, G.E. Frye, C.K. Hauser, N.E. Schwartz, T.M. Vaughan, J.R. Wolpaw, and E.W. Sellers. A novel p300-based brain–computer interface stimulus presentation paradigm: Moving beyond rows and columns. *Clinical Neurophysiology*, 121(7):1109–1120, July 2010. ISSN 1388-2457. doi: 10.1016/j.clinph.2010.01.030. URL <http://dx.doi.org/10.1016/j.clinph.2010.01.030>.
- [53] Christoph Guger, Shahab Daban, Eric Sellers, Clemens Holzner, Gunther Krausz, Roberta Carabalona, Furio Gramatica, and Guenter Edlinger. How many people are able to control a p300-based brain–computer interface (bci)? *Neuroscience Letters*, 462(1):94–98, September 2009. ISSN 0304-3940. doi: 10.1016/j.neulet.2009.06.045. URL <http://dx.doi.org/10.1016/j.neulet.2009.06.045>.
- [54] Md Rakibul Mowla, Jesus D. Gonzalez-Morales, Jacob Rico-Martinez, Daniel A. Ulichnie,

- and David E. Thompson. A comparison of classification techniques to predict brain-computer interfaces accuracy using classifier-based latency estimation. *Brain Sciences*, 10(10):734, October 2020. ISSN 2076-3425. doi: 10.3390/brainsci10100734. URL <http://dx.doi.org/10.3390/brainsci10100734>.
- [55] David E Thompson, Seth Warschausky, and Jane E Huggins. Classifier-based latency estimation: a novel way to estimate and predict bci accuracy. *Journal of Neural Engineering*, 10(1):016006, December 2012. ISSN 1741-2552. doi: 10.1088/1741-2560/10/1/016006. URL <http://dx.doi.org/10.1088/1741-2560/10/1/016006>.
- [56] David E. Thompson, Kirsten L. Gruis, and Jane E. Huggins. A plug-and-play brain-computer interface to operate commercial assistive technology. *Disability and Rehabilitation: Assistive Technology*, 9(2):144–150, April 2013. ISSN 1748-3115. doi: 10.3109/17483107.2013.785036. URL <http://dx.doi.org/10.3109/17483107.2013.785036>.
- [57] Louis Korczowski, Ekaterina Ostaschenko, Anton Andreev, Grégoire Cattan, Pedro Luis Coelho Rodrigues, Violette Gautheret, and Marco Congedo. Brain invaders calibration-less p300-based bci using dry eeg electrodes dataset (bi2014a), 2019. URL <https://zenodo.org/record/3266223>.
- [58] Md Rakibul Mowla, Jane E. Huggins, and David E. Thompson. Enhancing p300-bci performance using latency estimation. *Brain-Computer Interfaces*, 4(3):137–145, June 2017. ISSN 2326-2621. doi: 10.1080/2326263x.2017.1338010. URL <http://dx.doi.org/10.1080/2326263X.2017.1338010>.
- [59] MOHAMMAD SOLEYMANI, GUILLAUME CHANEL, JOEP J. M. KIERKELS, and THIERRY PUN. Affective characterization of movie scenes based on content analysis and

- physiological changes. *International Journal of Semantic Computing*, 03(02):235–254, June 2009. ISSN 1793-7108. doi: 10.1142/s1793351x09000744. URL <http://dx.doi.org/10.1142/S1793351X09000744>.
- [60] Julien Fleureau, Philippe Guillotel, and Quan Huynh-Thu. Physiological-based affect event detector for entertainment video applications. *IEEE Transactions on Affective Computing*, 3(3):379–385, July 2012. ISSN 1949-3045. doi: 10.1109/t-affc.2012.2. URL <http://dx.doi.org/10.1109/T-AFFC.2012.2>.
- [61] S. Koelstra, C. Muhl, M. Soleymani, Jong-Seok Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis;using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, January 2012. ISSN 1949-3045. doi: 10.1109/t-affc.2011.15. URL <http://dx.doi.org/10.1109/T-AFFC.2011.15>.
- [62] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, January 2012. ISSN 1949-3045. doi: 10.1109/t-affc.2011.25. URL <http://dx.doi.org/10.1109/T-AFFC.2011.25>.
- [63] Jonghwa Kim and Elisabeth Andre. Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12):2067–2083, December 2008. ISSN 2160-9292. doi: 10.1109/tpami.2008.26. URL <http://dx.doi.org/10.1109/TPAMI.2008.26>.
- [64] Stamos Katsigiannis and Naeem Ramzan. Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices. *IEEE Journal of*

- Biomedical and Health Informatics*, 22(1):98–107, January 2018. ISSN 2168-2208. doi: 10.1109/jbhi.2017.2688239. URL <http://dx.doi.org/10.1109/JBHI.2017.2688239>.
- [65] Steffen Steinert and Orsolya Friedrich. Wired emotions: Ethical issues of affective brain–computer interfaces. *Science and Engineering Ethics*, 26(1):351–367, March 2019. ISSN 1471-5546. doi: 10.1007/s11948-019-00087-2. URL <http://dx.doi.org/10.1007/s11948-019-00087-2>.
- [66] Carol L. Krumhansl. An exploratory study of musical emotions and psychophysiology. *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale*, 51(4):336–353, December 1997. ISSN 1196-1961. doi: 10.1037/1196-1961.51.4.336. URL <http://dx.doi.org/10.1037/1196-1961.51.4.336>.
- [67] Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE Transactions on Cybernetics*, 49(3):1110–1122, March 2019. ISSN 2168-2275. doi: 10.1109/TCYB.2018.2797176. URL <http://dx.doi.org/10.1109/TCYB.2018.2797176>.
- [68] Albert Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292, December 1996. ISSN 1936-4733. doi: 10.1007/bf02686918. URL <http://dx.doi.org/10.1007/BF02686918>.
- [69] Tengfei Song, Wenming Zheng, Cheng Lu, Yuan Zong, Xilei Zhang, and Zhen Cui. Mped: A multi-modal physiological emotion database for discrete emotion recognition. *IEEE Access*, 7:12177–12191, 2019. ISSN 2169-3536. doi: 10.1109/ACCESS.2019.2891579. URL <http://dx.doi.org/10.1109/ACCESS.2019.2891579>.

- [70] Stephan Hamann. Mapping discrete and dimensional emotions onto the brain: controversies and consensus. *Trends in Cognitive Sciences*, 16(9):458–466, September 2012. ISSN 1364-6613. doi: 10.1016/j.tics.2012.07.006. URL <http://dx.doi.org/10.1016/j.tics.2012.07.006>.
- [71] Danny Plass-Oude Bos. Eeg-based emotion recognition the influence of visual and auditory stimuli. 2007. URL <https://api.semanticscholar.org/CorpusID:16285681>.
- [72] Stelios K. Hadjidimitriou and Leontios J. Hadjileontiadis. Toward an eeg-based recognition of music liking using time-frequency analysis. *IEEE Transactions on Biomedical Engineering*, 59(12):3498–3510, December 2012. ISSN 1558-2531. doi: 10.1109/tbme.2012.2217495. URL <http://dx.doi.org/10.1109/TBME.2012.2217495>.
- [73] Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, September 2015. ISSN 1943-0612. doi: 10.1109/tamd.2015.2431497. URL <http://dx.doi.org/10.1109/TAMD.2015.2431497>.
- [74] Yong-Jin Liu, Minjing Yu, Guozhen Zhao, Jinjing Song, Yan Ge, and Yuanchun Shi. Real-time movie-induced discrete emotion recognition from eeg signals. *IEEE Transactions on Affective Computing*, 9(4):550–562, October 2018. ISSN 2371-9850. doi: 10.1109/taffc.2017.2660485. URL <http://dx.doi.org/10.1109/TAFFC.2017.2660485>.
- [75] Yang Li, Wenming Zheng, Zhen Cui, Yuan Zong, and Sheng Ge. Eeg emotion recognition based on graph regularized sparse linear regression. *Neural Processing Letters*, 49(2):555–571, April 2018. ISSN 1573-773X. doi: 10.1007/s11063-018-9829-1. URL <http://dx.doi.org/10.1007/s11063-018-9829-1>.

- [76] Bahareh Nakisa, Mohammad Naim Rastgoo, Andry Rakotonirainy, Frederic Maire, and Vinod Chandran. Long short term memory hyperparameter optimization for a neural network based emotion recognition framework. *IEEE Access*, 6:49325–49338, 2018. ISSN 2169-3536. doi: 10.1109/access.2018.2868361. URL <http://dx.doi.org/10.1109/ACCESS.2018.2868361>.
- [77] Hanna Becker, Julien Fleureau, Philippe Guillotel, Fabrice Wendling, Isabelle Merlet, and Laurent Albera. Emotion recognition based on high-resolution eeg recordings and reconstructed brain sources. *IEEE Transactions on Affective Computing*, 11(2):244–257, April 2020. ISSN 2371-9850. doi: 10.1109/taffc.2017.2768030. URL <http://dx.doi.org/10.1109/TAFFC.2017.2768030>.
- [78] Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing*, 12(2):479–493, April 2021. ISSN 2371-9850. doi: 10.1109/taffc.2018.2884461. URL <http://dx.doi.org/10.1109/TAFFC.2018.2884461>.
- [79] Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):715–729, June 2022. ISSN 2379-8939. doi: 10.1109/tcds.2021.3071170. URL <http://dx.doi.org/10.1109/TCDS.2021.3071170>.
- [80] Sun-Hee Kim, Hyung-Jeong Yang, Ngoc Anh Thi Nguyen, Sunil Kumar Prabhakar, and Seong-Whan Lee. Wedea: A new eeg-based framework for emotion recognition. *IEEE*

- Journal of Biomedical and Health Informatics*, 26(1):264–275, January 2022. ISSN 2168-2208. doi: 10.1109/jbhi.2021.3091187. URL <http://dx.doi.org/10.1109/JBHI.2021.3091187>.
- [81] Wei Liu, Wei-Long Zheng, Ziyi Li, Si-Yuan Wu, Lu Gan, and Bao-Liang Lu. Identifying similarities and differences in emotion recognition with eeg and eye movements among chinese, german, and french people. *Journal of Neural Engineering*, 19(2):026012, March 2022. ISSN 1741-2552. doi: 10.1088/1741-2552/ac5c8d. URL <http://dx.doi.org/10.1088/1741-2552/ac5c8d>.
- [82] Alexandre Schaefer, Frédéric Nils, Xavier Sanchez, and Pierre Philippot. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition amp; Emotion*, 24(7):1153–1172, November 2010. ISSN 1464-0600. doi: 10.1080/02699930903274322. URL <http://dx.doi.org/10.1080/02699930903274322>.
- [83] Xinke Shen, Xianggen Liu, Xin Hu, Dan Zhang, and Sen Song. Contrastive learning of subject-invariant eeg representations for cross-subject emotion recognition. *IEEE Transactions on Affective Computing*, 14(3):2496–2511, July 2023. ISSN 2371-9850. doi: 10.1109/taffc.2022.3164516. URL <http://dx.doi.org/10.1109/TAFFC.2022.3164516>.
- [84] Hussein Ali Hamzah and Kasim K. Abdalla. Eeg-based emotion recognition datasets for virtual environments: A survey. *Applied Computational Intelligence and Soft Computing*, 2024(1), January 2024. ISSN 1687-9732. doi: 10.1155/2024/6091523. URL <http://dx.doi.org/10.1155/2024/6091523>.
- [85] Margaret M. Bradley and Peter J. Lang. *International Affective Picture System*, page 1–4. Springer International Publishing, 2017. ISBN 9783319280998. doi: 10.1007/

- 978-3-319-28099-8_42-1. URL http://dx.doi.org/10.1007/978-3-319-28099-8_42-1.
- [86] Ryan A. Stevenson and Thomas W. James. Affective auditory stimuli: Characterization of the international affective digitized sounds (iads) by discrete emotional categories. *Behavior Research Methods*, 40(1):315–321, February 2008. ISSN 1554-3528. doi: 10.3758/brm.40.1.315. URL <http://dx.doi.org/10.3758/BRM.40.1.315>.
- [87] Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani. Developing a benchmark for emotional analysis of music. *PLOS ONE*, 12(3):e0173392, March 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0173392. URL <http://dx.doi.org/10.1371/journal.pone.0173392>.
- [88] James A Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294, September 1977. ISSN 0092-6566. doi: 10.1016/0092-6566(77)90037-x. URL [http://dx.doi.org/10.1016/0092-6566\(77\)90037-X](http://dx.doi.org/10.1016/0092-6566(77)90037-X).
- [89] G. Schalk, D.J. McFarland, T. Hinterberger, N. Birbaumer, and J.R. Wolpaw. Bci2000: A general-purpose brain-computer interface (bci) system. *IEEE Transactions on Biomedical Engineering*, 51(6):1034–1043, June 2004. ISSN 0018-9294. doi: 10.1109/tbme.2004.827072. URL <http://dx.doi.org/10.1109/TBME.2004.827072>.
- [90] Kai Yang, Li Tong, Jun Shu, Ning Zhuang, Bin Yan, and Ying Zeng. High gamma band eeg closely related to emotion: Evidence from functional network. *Frontiers in Human Neuroscience*, 14, March 2020. ISSN 1662-5161. doi: 10.3389/fnhum.2020.00089. URL <http://dx.doi.org/10.3389/fnhum.2020.00089>.

- [91] Michela Balconi and Guido Mazza. Brain oscillations and bis/bas (behavioral inhibition/activation system) effects on processing masked emotional cues. *International Journal of Psychophysiology*, 74(2):158–165, November 2009. ISSN 0167-8760. doi: 10.1016/j.ijpsycho.2009.08.006. URL <http://dx.doi.org/10.1016/j.ijpsycho.2009.08.006>.
- [92] Harry W. Cole and William J. Ray. Eeg correlates of emotional tasks related to attentional demands. *International Journal of Psychophysiology*, 3(1):33–41, July 1985. ISSN 0167-8760. doi: 10.1016/0167-8760(85)90017-0. URL [http://dx.doi.org/10.1016/0167-8760\(85\)90017-0](http://dx.doi.org/10.1016/0167-8760(85)90017-0).
- [93] J. X. Chen, P. W. Zhang, Z. J. Mao, Y. F. Huang, D. M. Jiang, and Y. N. Zhang. Accurate eeg-based emotion recognition on combined features using deep convolutional neural networks. *IEEE Access*, 7:44317–44328, 2019. ISSN 2169-3536. doi: 10.1109/access.2019.2908285. URL <http://dx.doi.org/10.1109/ACCESS.2019.2908285>.
- [94] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, January 1967. ISSN 1557-9654. doi: 10.1109/tit.1967.1053964. URL <http://dx.doi.org/10.1109/TIT.1967.1053964>.
- [95] Md Rakibul Mowla, Rachael I. Cano, Katie J. Dhuyvetter, and David E. Thompson. Affective brain-computer interfaces: Choosing a meaningful performance measuring metric. *Computers in Biology and Medicine*, 126:104001, November 2020. ISSN 0010-4825. doi: 10.1016/j.combiomed.2020.104001. URL <http://dx.doi.org/10.1016/j.combiomed.2020.104001>.
- [96] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann.

- The balanced accuracy and its posterior distribution. In *2010 20th International Conference on Pattern Recognition*, page 3121–3124. IEEE, August 2010. doi: 10.1109/icpr.2010.764. URL <http://dx.doi.org/10.1109/ICPR.2010.764>.
- [97] Nazmun N Khan, Taylor Sweet, Chase A Harvey, Calder Knapp, Dean J. Krusienski, and David E Thompson. The role of review process failures in affective state estimation: An empirical investigation of deap dataset, 2025. URL <https://arxiv.org/abs/2508.02417>.
- [98] Naveed Ahmed, Zaher Al Aghbari, and Shini Girija. A systematic survey on multimodal emotion recognition using learning algorithms. *Intelligent Systems with Applications*, 17: 200171, February 2023. ISSN 2667-3053. doi: 10.1016/j.iswa.2022.200171. URL <http://dx.doi.org/10.1016/j.iswa.2022.200171>.
- [99] U.S. Food and Drug Administration. Artificial intelligence-enabled medical devices. <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices>, 2025. Accessed: 2025-10-14.
- [100] Haoran Liu, Ying Zhang, Yujun Li, and Xiangyi Kong. Review on emotion recognition based on electroencephalography. *Frontiers in Computational Neuroscience*, 15, October 2021. ISSN 1662-5188. doi: 10.3389/fncom.2021.758212. URL <http://dx.doi.org/10.3389/fncom.2021.758212>.
- [101] Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. Deep learning for electroencephalogram (eeg) classification tasks: a review. *Journal of Neural Engineering*,

- 16(3):031001, April 2019. ISSN 1741-2552. doi: 10.1088/1741-2552/ab0ab5. URL <http://dx.doi.org/10.1088/1741-2552/ab0ab5>.
- [102] Luis Alfredo Moctezuma, Takashi Abe, and Marta Molinas. Two-dimensional cnn-based distinction of human emotions from eeg channels selected by multi-objective evolutionary algorithm. *Scientific Reports*, 12(1), March 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-07517-5. URL <http://dx.doi.org/10.1038/s41598-022-07517-5>.
- [103] Theerawit Wilaiprasitporn, Apiwat Dittthapron, Karis Matchaparn, Tanaboon Tongbuasirilai, Nannapas Banluesombatkul, and Ekapol Chuangsuwanich. Affective eeg-based person identification using the deep learning approach. *IEEE Transactions on Cognitive and Developmental Systems*, 12(3):486–496, September 2020. ISSN 2379-8939. doi: 10.1109/tcds.2019.2924648. URL <http://dx.doi.org/10.1109/TCDS.2019.2924648>.
- [104] Chunmei Qing, Rui Qiao, Xiangmin Xu, and Yongqiang Cheng. Interpretable emotion recognition using eeg signals. *IEEE Access*, 7:94160–94170, 2019. ISSN 2169-3536. doi: 10.1109/access.2019.2928691. URL <http://dx.doi.org/10.1109/ACCESS.2019.2928691>.
- [105] Ning Zhuang, Ying Zeng, Kai Yang, Chi Zhang, Li Tong, and Bin Yan. Investigating patterns for self-induced emotion recognition from eeg signals. *Sensors*, 18(3):841, March 2018. ISSN 1424-8220. doi: 10.3390/s18030841. URL <http://dx.doi.org/10.3390/s18030841>.
- [106] Zhihui Zhang, Josep M. Fort, and Lluís Giménez Mateu. Mini review: Challenges in eeg emotion recognition. *Frontiers in Psychology*, 14, January 2024. ISSN 1664-1078. doi: 10.3389/fpsyg.2023.1289816. URL <http://dx.doi.org/10.3389/fpsyg.2023.1289816>.

- [107] Gabriel Ivucic, Saurav Pahuja, Felix Putze, Siqi Cai, Haizhou Li, and Tanja Schultz. The impact of cross-validation schemes for eeg-based auditory attention detection with deep neural networks. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, page 1–4. IEEE, July 2024. doi: 10.1109/embc53108.2024.10782636. URL <http://dx.doi.org/10.1109/EMBC53108.2024.10782636>.
- [108] Seshu Bhavani Mallampati and Seetha Hari. A comparative study on the impacts of data leakage during feature selection using the cic-iot 2023 intrusion detection dataset. In *2024 10th International Conference on Electrical Energy Systems (ICEES)*, page 1–6. IEEE, August 2024. doi: 10.1109/icees61253.2024.10776873. URL <http://dx.doi.org/10.1109/ICEES61253.2024.10776873>.
- [109] Steven Lemm, Benjamin Blankertz, Thorsten Dickhaus, and Klaus-Robert Müller. Introduction to machine learning for brain imaging. *NeuroImage*, 56(2):387–399, May 2011. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2010.11.004. URL <http://dx.doi.org/10.1016/j.neuroimage.2010.11.004>.
- [110] Huayu Chen, Junxiang Li, Huanhuan He, Jing Zhu, Shuting Sun, Xiaowei Li, and Bin Hu. Toward the construction of affective brain-computer interface: A systematic review. *ACM Computing Surveys*, 57(6):1–56, February 2025. ISSN 1557-7341. doi: 10.1145/3712259. URL <http://dx.doi.org/10.1145/3712259>.
- [111] Jaime A Riascos, Marta Molinas, and Fabien Lotte. Machine learning methods for bci: challenges, pitfalls and promises. In *ESANN 2024 proceedings*, ESANN 2024,

page 555–564. Ciaco - i6doc.com, 2024. doi: 10.14428/esann/2024.es2024-4. URL <http://dx.doi.org/10.14428/esann/2024.ES2024-4>.

- [112] Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R. Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, Alessandro Ruggiero, Anna Korhonen, Emily Jefferson, Emmanuel Ako, Georg Langs, Ghassem Gozaliasl, Guang Yang, Helmut Prosch, Jacobus Preller, Jan Stanczuk, Jing Tang, Johannes Hofmanninger, Judith Babar, Lorena Escudero Sánchez, Muhunthan Thillai, Paula Martin Gonzalez, Philip Teare, Xiaoxiang Zhu, Mishal Patel, Conor Cafolla, Hojjat Azadbakht, Joseph Jacob, Josh Lowe, Kang Zhang, Kyle Bradley, Marcel Wassin, Markus Holzer, Kangyu Ji, Maria Delgado Ortet, Tao Ai, Nicholas Walton, Pietro Lio, Samuel Stranks, Tolou Shadbahr, Weizhe Lin, Yunfei Zha, Zhangming Niu, James H. F. Rudd, Evis Sala, and Carola-Bibiane Schönlieb. Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. *Nature Machine Intelligence*, 3(3):199–217, March 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00307-0. URL <http://dx.doi.org/10.1038/s42256-021-00307-0>.
- [113] Alan Balendran, Céline Beji, Florie Bouvier, Ottavio Khalifa, Theodoros Evgeniou, Philippe Ravaud, and Raphaël Porcher. A scoping review of robustness concepts for machine learning in healthcare. *npj Digital Medicine*, 8(1), January 2025. ISSN 2398-6352. doi: 10.1038/s41746-024-01420-1. URL <http://dx.doi.org/10.1038/s41746-024-01420-1>.
- [114] Miseon Shim, Seung-Hwan Lee, and Han-Jeong Hwang. Inflated prediction accuracy of neuropsychiatric biomarkers caused by data leakage in feature selection. *Scientific Reports*,

- 11(1), April 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-87157-3. URL <http://dx.doi.org/10.1038/s41598-021-87157-3>.
- [115] Hyung-Tak Lee, Hye-Ran Cheon, Seung-Hwan Lee, Miseon Shim, and Han-Jeong Hwang. Risk of data leakage in estimating the diagnostic performance of a deep-learning-based computer-aided system for psychiatric disorders. *Sci. Rep.*, 13(1):16633, October 2023.
- [116] Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(1), February 2006. ISSN 1471-2105. doi: 10.1186/1471-2105-7-91. URL <http://dx.doi.org/10.1186/1471-2105-7-91>.
- [117] Christophe Ambroise and Geoffrey J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences*, 99(10):6562–6566, April 2002. ISSN 1091-6490. doi: 10.1073/pnas.102102699. URL <http://dx.doi.org/10.1073/pnas.102102699>.
- [118] Andy Stock, Edward J. Gregr, and Kai M. A. Chan. Data leakage jeopardizes ecological applications of machine learning. *Nature Ecology amp; Evolution*, 7(11):1743–1745, August 2023. ISSN 2397-334X. doi: 10.1038/s41559-023-02162-1. URL <http://dx.doi.org/10.1038/s41559-023-02162-1>.
- [119] Matthew Rosenblatt, Link Tejavibulya, Rongtao Jiang, Stephanie Noble, and Dustin Scheinost. Data leakage inflates prediction performance in connectome-based machine learning models. *Nature Communications*, 15(1), February 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-46150-w. URL <http://dx.doi.org/10.1038/s41467-024-46150-w>.
- [120] Daniel Arp, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke, Fabio Pierazzi,

- Christian Wressnegger, Lorenzo Cavallaro, and Konrad Rieck. Dos and don'ts of machine learning in computer security. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 3971–3988, Boston, MA, August 2022. USENIX Association. ISBN 978-1-939133-31-1. URL <https://www.usenix.org/conference/usenixsecurity22/presentation/arp>.
- [121] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, September 2023. ISSN 2666-3899. doi: 10.1016/j.patter.2023.100804. URL <http://dx.doi.org/10.1016/j.patter.2023.100804>.
- [122] Felix Schroeder, Stephen Fairclough, Frederic Dehais, and Matthew Richins. The impact of cross-validation choices on pbc1 classification metrics: lessons for transparent reporting. *Frontiers in Neuroergonomics*, 6, July 2025. ISSN 2673-6195. doi: 10.3389/fnrgo.2025.1582724. URL <http://dx.doi.org/10.3389/fnrgo.2025.1582724>.
- [123] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, July 2021. ISSN 1557-7341. doi: 10.1145/3457607. URL <http://dx.doi.org/10.1145/3457607>.
- [124] Arjun K. Manrai, Birgit H. Funke, Heidi L. Rehm, Morten S. Olesen, Bradley A. Maron, Peter Szolovits, David M. Margulies, Joseph Loscalzo, and Isaac S. Kohane. Genetic misdiagnoses and the potential for health disparities. *New England Journal of Medicine*, 375(7):655–665, August 2016. ISSN 1533-4406. doi: 10.1056/nejmsa1507092. URL <http://dx.doi.org/10.1056/NEJMSa1507092>.

- [125] *AMA Journal of Ethics*, 21(2):E167–179, February 2019. ISSN 2376-6980. doi: 10.1001/amajethics.2019.167. URL <http://dx.doi.org/10.1001/amajethics.2019.167>.
- [126] Ayanna Howard and Jason Borenstein. The ugly truth about ourselves and our robot creations: The problem of bias and social inequity. *Science and Engineering Ethics*, 24(5):1521–1536, September 2017. ISSN 1471-5546. doi: 10.1007/s11948-017-9975-2. URL <http://dx.doi.org/10.1007/s11948-017-9975-2>.
- [127] Soraia M. Alarcao and Manuel J. Fonseca. Emotions recognition using eeg signals: A survey. *IEEE Transactions on Affective Computing*, 10(3):374–393, July 2019. ISSN 2371-9850. doi: 10.1109/taffc.2017.2714671. URL <http://dx.doi.org/10.1109/TAFFC.2017.2714671>.
- [128] Hanan Khalil, Micah D.J. Peters, Patricia A. McInerney, Christina M. Godfrey, Lyndsay Alexander, Catrin Evans, Dawid Pieper, Erica B. Moraes, Andrea C. Tricco, Zachary Munn, and Danielle Pollock. The role of scoping reviews in reducing research waste. *Journal of Clinical Epidemiology*, 152:30–35, December 2022. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2022.09.012. URL <http://dx.doi.org/10.1016/j.jclinepi.2022.09.012>.
- [129] D G Altman. The scandal of poor medical research. *BMJ*, 308(6924):283–284, January 1994. ISSN 1468-5833. doi: 10.1136/bmj.308.6924.283. URL <http://dx.doi.org/10.1136/bmj.308.6924.283>.
- [130] C. Glenn Begley and John P.A. Ioannidis. Reproducibility in science: Improving the standard for basic and preclinical research. *Circulation Research*, 116(1):116–126, January 2015. ISSN 1524-4571. doi: 10.1161/circresaha.114.303819. URL <http://dx.doi.org/10.1161/CIRCRESAHA.114.303819>.

- [131] Iain Chalmers and Paul Glasziou. Avoidable waste in the production and reporting of research evidence. *The Lancet*, 374(9683):86–89, July 2009. ISSN 0140-6736. doi: 10.1016/S0140-6736(09)60329-9. URL [http://dx.doi.org/10.1016/S0140-6736\(09\)60329-9](http://dx.doi.org/10.1016/S0140-6736(09)60329-9).
- [132] Iain Chalmers, Michael B Bracken, Ben Djulbegovic, Silvio Garattini, Jonathan Grant, A Metin Gülmezoglu, David W Howells, John P A Ioannidis, and Sandy Oliver. How to increase value and reduce waste when research priorities are set. *The Lancet*, 383(9912):156–165, January 2014. ISSN 0140-6736. doi: 10.1016/S0140-6736(13)62229-1. URL [http://dx.doi.org/10.1016/S0140-6736\(13\)62229-1](http://dx.doi.org/10.1016/S0140-6736(13)62229-1).
- [133] Rustam Al-Shahi Salman, Elaine Beller, Jonathan Kagan, Elina Hemminki, Robert S Phillips, Julian Savulescu, Malcolm Macleod, Janet Wisely, and Iain Chalmers. Increasing value and reducing waste in biomedical research regulation and management. *The Lancet*, 383(9912):176–185, January 2014. ISSN 0140-6736. doi: 10.1016/S0140-6736(13)62297-7. URL [http://dx.doi.org/10.1016/S0140-6736\(13\)62297-7](http://dx.doi.org/10.1016/S0140-6736(13)62297-7).
- [134] Malcolm R Macleod, Susan Michie, Ian Roberts, Ulrich Dirnagl, Iain Chalmers, John P A Ioannidis, Rustam Al-Shahi Salman, An-Wen Chan, and Paul Glasziou. Biomedical research: increasing value, reducing waste. *The Lancet*, 383(9912):101–104, January 2014. ISSN 0140-6736. doi: 10.1016/S0140-6736(13)62329-6. URL [http://dx.doi.org/10.1016/S0140-6736\(13\)62329-6](http://dx.doi.org/10.1016/S0140-6736(13)62329-6).
- [135] Leonard P. Freedman, Iain M. Cockburn, and Timothy S. Simcoe. The economics of reproducibility in preclinical research. *PLOS Biology*, 13(6):e1002165, June 2015. ISSN

- 1545-7885. doi: 10.1371/journal.pbio.1002165. URL <http://dx.doi.org/10.1371/journal.pbio.1002165>.
- [136] Xiran Xu, Bo Wang, Boda Xiao, Yadong Niu, Yiwen Wang, Xihong Wu, and Jing Chen. Beware of overestimated decoding performance arising from temporal autocorrelations in electroencephalogram signals, 2024. URL <https://arxiv.org/abs/2405.17024>.
- [137] Yongqiang Zheng, Jie Ding, Feng Liu, and Dongqing Wang. Adaptive neural decision tree for eeg based emotion recognition. *Information Sciences*, 643:119160, September 2023. ISSN 0020-0255. doi: 10.1016/j.ins.2023.119160. URL <http://dx.doi.org/10.1016/j.ins.2023.119160>.
- [138] Arti Anuragi, Dilip Singh Sisodia, and Ram Bilas Pachori. Eeg-based cross-subject emotion recognition using fourier-bessel series expansion based empirical wavelet transform and nca feature selection method. *Information Sciences*, 610:508–524, September 2022. ISSN 0020-0255. doi: 10.1016/j.ins.2022.07.121. URL <http://dx.doi.org/10.1016/j.ins.2022.07.121>.
- [139] Guanhua Zhang, Minjing Yu, Yong-Jin Liu, Guozhen Zhao, Dan Zhang, and Wenming Zheng. Sparsedgcnn: Recognizing emotion from multichannel eeg signals. *IEEE Transactions on Affective Computing*, 14(1):537–548, January 2023. ISSN 2371-9850. doi: 10.1109/taffc.2021.3051332. URL <http://dx.doi.org/10.1109/TAFFC.2021.3051332>.
- [140] Johan N. van der Meer, Yke B. Eisma, Ronald Meester, Marc Jacobs, and Aart J. Nederveen. Effects of mobile phone electromagnetic fields on brain waves in healthy volunteers. *Scientific Reports*, 13(1), December 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-48561-z. URL <http://dx.doi.org/10.1038/s41598-023-48561-z>.

- [141] Dominik Freche, Jodie Naim-Feil, Avi Peled, Nava Levit-Binnun, and Elisha Moses. A quantitative physical model of the tms-induced discharge artifacts in eeg. *PLOS Computational Biology*, 14(7):e1006177, July 2018. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006177. URL <http://dx.doi.org/10.1371/journal.pcbi.1006177>.

Appendix A

Graphical User Interface Developed for CBLE Performance Estimation

MATLAB App

CBLE Performance Estimation

Select dataset format

BCI2000

Status: Off

BCI2000

Select input folder

Total data files:

Output folder

Run

No of participants

ID length

Channel ID

Decimation value

Original window

CBLE window

0

4

[1:16]

13

[0,800]

[-100,900]

Set Parameters

Training file name

*TRN.dat

Test file name

*TST.dat

Train a model

Predict accuracy for test set

Choose character number, X

10

Test file no

1

Find X target accuracy

Find vCBLE

Calculate RMSE

Check List

☐ Load input data

☐ Set the parameters

☐ Train a model

☐ Predict accuracy for test set

☐ Find X_target accuracy

☐ Calculate vCBLE

☐ Calculate RMSE

Prediction of the performance of an individual participant

Sub ID

0

Target number

Predict

Predicted accuracy

Plot the figures

Select an option

None

Accuracy vs vCBLE

RMSE of BCI & vCBLE

Appendix B

Reuse Permissions from Publishers

Re: Request for Permission to Include Published Article in Dissertation

From Anastasia Gomez <anastasia.gomez@jove.com>
Date Tue 9/16/2025 10:06 AM
To Nazmun Nahar Khan <nkhan1@ksu.edu>
Cc editorial-office@jove.com <editorial-office@jove.com>

This email originated from outside of K-State.

Hi,
Thank you for reaching out!

Consider this an explicit permission to use the requested JoVE content as described in your request.

Please ensure to cite JoVE accordingly.

Best,
Anastasia

Anastasia Gomez

Associate Director of Production | [JoVE](#)

E: anastasia.gomez@jove.com

Follow Us: [LinkedIn](#) | [X](#) | [Facebook](#)

[Publish with JoVE](#)