Subject-Level Signal Quality Limits EEG Stimulus Classification: A Cross-Dataset Study of Classifiability in Oddball EEG

by

Seyed Mohammad Amin Miri

A thesis submitted to the
Department of Computer Science
in conformity with the requirements for
the degree of Master of Science

Bishop's University
Canada
August 2026

Abstract

Electroencephalography (EEG) provides high temporal resolution for studying stimulus-related brain activity, but EEG decoding remains difficult because scalp-recorded signals are noisy, non-stationary, and highly variable across subjects. This thesis examines EEG stimulus classification as both a model-selection problem and a subject-classifiability problem. The central objective is to determine how strongly decoding performance is associated with subject-level signal characteristics, relative to the choice of classifier.

The study used four OpenNeuro oddball or event-related potential (ERP) datasets: `ds003061`, `ds005863`, `ds006466`, and the pooled `ds006480` condition. The final analysis included 103 subject-recordings. Linear Discriminant Analysis (LDA), Logistic Regression, and EEGNet were evaluated using grouped within-subject and leave-one-subject-out (LOSO) designs. Performance was measured using the area under the receiver operating characteristic curve (ROC-AUC). In grouped within-subject evaluation, the cross-dataset mean AUC values were 0.7871 for LDA, 0.7842 for Logistic Regression, and 0.7404 for EEGNet. In LOSO evaluation, the corresponding values were 0.7285, 0.7488, and 0.7947. Paired subject-level comparisons showed that LDA and Logistic Regression did not differ significantly in the pooled within-subject analysis, whereas both linear models exceeded EEGNet. In LOSO evaluation, EEGNet exceeded both linear models.

The main contribution is a retrospective subject-level susceptibility analysis relating decoding performance to raw EEG, ERP, and spectral signal characteristics. Raw/ERP/power features produced 33 false-discovery-rate (FDR)-corrected model–feature effects in the within-subject analysis and 43 source-specific FDR-corrected effects in the LOSO analysis at $q \leq 0.05$. After AUC values were standardized within dataset, the standard deviation of subject-level mean AUC was approximately 2.8 times the standard deviation of classifier-level mean AUC. Subject-quality features also explained more incremental variance in AUC than classifier identity after accounting for dataset.

An additional analysis compared EEGNet with the better-performing linear model for subjects in the upper and lower quartiles of the Subject Quality Index. EEGNet did not provide a systematic gain in grouped within-subject evaluation, but it produced positive average LOSO gains in both groups. The larger average

gain observed among difficult subjects should be interpreted cautiously because the continuous association between subject quality and EEGNet gain was not statistically significant.

A training-only Gaussian-noise robustness ablation was also conducted for EEGNet under LOSO evaluation. Gaussian noise with $\sigma = 0.01$, selected through an inner-validation pilot, increased the pooled subject-level mean AUC from 0.7835 to 0.7846. The mean increase of 0.0011 was not statistically significant ($p = 0.072$). Augmentation slightly reduced the validation–test gap, suggesting modest regularization, but did not materially resolve held-out-subject performance variability.

Overall, the results indicate that classifier choice matters, particularly for cross-subject generalization, but decoding performance remains strongly associated with subject-level ERP separability and signal quality.

Acknowledgments

I would like to thank my supervisor, Dr. Russell Butler, for his guidance, feedback, and support throughout this thesis project. His comments helped shape the work from an initially broad source-separation proposal into a focused empirical study of EEG decoding, subject variability, and signal-quality limits. I would also like to thank Dr. Rachid Hedjam and the Department of Computer Science at Bishop's University for their support during my graduate studies.

I am grateful to my family and friends for their encouragement and patience throughout this work. Their support made it possible to continue through the many technical challenges, revisions, and changes in direction that shaped this thesis.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background and Motivation

Electroencephalography (EEG) is widely used in neuroscience, clinical research, and brain–computer interface (BCI) systems because it measures scalp electrical activity non-invasively with millisecond-scale temporal resolution [7, 12]. This temporal resolution makes EEG particularly suitable for studying event-related potentials (ERPs), which are voltage changes that are time-locked to sensory, cognitive, or motor events [20, 26]. In oddball paradigms, participants encounter frequent standard stimuli and less frequent or task-relevant target stimuli. Differences between the resulting target and standard responses can provide information for stimulus classification and BCI applications.

Reliable EEG decoding nevertheless remains difficult. Scalp recordings contain low-amplitude neural signals together with ocular activity, muscle activity, electrode noise, environmental interference, and other sources of variation [6, 29]. EEG responses also vary considerably across subjects and recording sessions. This variability may reflect differences in neurophysiology, attention, task engagement, anatomy, electrode contact, artifact burden, trial count, and experimental protocol [15, 37, 40]. Consequently, subjects exposed to the same experimental task may produce signals with substantially different levels of target–standard separability.

Many EEG classification studies are organized primarily as model comparisons. Several classifiers are evaluated, and the model with the highest average performance is identified as the preferred method [18, 24, 25]. Such comparisons are important, but an average model ranking does not explain why performance varies among subjects. A classifier may achieve high performance for a subject whose EEG contains a clear and stable ERP difference, while several classifiers may struggle when the corresponding response is weak, temporally variable, or contaminated by noise.

This thesis therefore treats subject-level classifiability as a separate object of analysis. The study examines not only how LDA, Logistic Regression, and EEGNet compare, but also whether subjects who are easy or difficult for one classifier tend

to remain easy or difficult for the others. It further examines whether interpretable raw EEG and ERP-derived features are associated with this variation. The purpose is not to replace classifier evaluation, but to determine how classifier identity and subject-level signal characteristics jointly contribute to decoding performance.

1.2 Problem Statement

EEG stimulus-classification performance is often summarized primarily through average classifier performance. This approach can obscure substantial variation among individual subjects. In oddball and ERP-style tasks, the magnitude, latency, spatial distribution, and trial-to-trial consistency of target–standard responses can differ across participants, datasets, recording systems, and task conditions [6, 20, 35]. As a result, an observed performance difference may reflect classifier properties, subject-level signal characteristics, dataset effects, or a combination of these factors.

This creates two related methodological problems. First, reporting only average AUC values does not reveal whether the same subjects remain easy or difficult across different classifiers. Second, without relating subject-level performance to measurable signal characteristics, it is difficult to determine whether poor performance is primarily associated with model choice or with weak and noisy task-related EEG responses.

The present study addresses these problems through a cross-dataset analysis of LDA, Logistic Regression, and EEGNet under grouped within-subject and LOSO evaluation. Subject-level AUC values are examined together with raw EEG, ERP, spectral, trial-structure, and exploratory source-quality features. The analysis asks whether subject-level signal characteristics are associated with classifiability, whether subject difficulty is consistent across classifiers, and whether subject-quality measures explain additional AUC variation after dataset identity has been taken into account.

1.3 Research Questions

This thesis addresses the following research questions:

1. To what extent are subject-level raw EEG, ERP-derived, and spectral features associated with grouped within-subject and LOSO classification performance?

2. After accounting for dataset identity, do subject-quality measures explain more variation in subject-level AUC than classifier identity?

3. Are subjects who are relatively easy or difficult to classify consistent across LDA, Logistic Regression, and EEGNet?

4. How do LDA, Logistic Regression, and EEGNet compare under grouped within-subject and leave-one-subject-out evaluation?

5. Does EEGNet provide greater benefit than the linear models for subjects with lower estimated signal quality, particularly in LOSO evaluation?

6. Are the observed decoding results supported by plausible post-stimulus ERP differences and time-resolved decoding patterns?

1.4 Main Contributions

This thesis makes the following contributions:

1. It develops and applies a systematically specified cross-dataset EEG stimulus-classification framework across four oddball or ERP-style datasets and 103 subject-recordings. The framework includes dataset-specific event mapping, consistent preprocessing, grouped within-subject evaluation, LOSO evaluation, and subject-level aggregation.

2. It compares two established linear classifiers, LDA and Logistic Regression, with the compact EEG-specific convolutional architecture EEGNet under two distinct generalization settings. This comparison shows that the relative model ranking depends on whether subject-specific calibration or unseen-subject generalization is evaluated.

3. It demonstrates that subject-level classification performance is highly consistent across classifiers. Subjects who are relatively easy or difficult for one model generally remain so for the other models.

4. It relates subject-level AUC to interpretable raw EEG, ERP, spectral, and trial-structure characteristics. The strongest corrected associations involve measures of ERP/P300 separability and signal variability, whereas the ICA/source-quality measures remain exploratory.

5. It quantifies the relative explanatory contributions of subject-level and classifier-level effects. After within-dataset standardization, the standard deviation of subject-level mean AUC was 2.85 times the standard deviation of classifier-level mean AUC in grouped within-subject evaluation and 2.83 times as large in LOSO evaluation. The retrospective Subject Quality Index also explained more incremental AUC variance than classifier identity after accounting for dataset.

6. It introduces an interpretable retrospective subject-quality index and evaluates its stability under alternative component and weighting choices. The index is used as an analysis tool for characterizing subject classifiability rather than as a validated prospective screening instrument.

7. It examines whether EEGNet provides gains over the better-performing linear classifier among relatively easy and difficult subjects. It also evaluates a training-only Gaussian-noise robustness ablation under LOSO. EEGNet's clearest advantage occurred in LOSO evaluation, while Gaussian-noise augmentation produced modest regularization without a statistically significant overall AUC improvement.

8. It provides ERP waveform and time-resolved decoding checks that connect the classification results to plausible post-stimulus target–standard differences while acknowledging that these analyses do not fully exclude non-neural contributions.

1.5 Thesis Organization

Chapter 2 reviews the relevant EEG, ERP, subject-variability, classifier, and cross-subject generalization literature and positions the thesis relative to prior work. Chapter 3 describes the datasets, preprocessing, classifier inputs, and subject-level features. Chapter 4 presents the evaluation framework, classifier implementations, training procedure, augmentation ablation, and statistical methods. Chapter 5 reports the benchmark, classifier-comparison, training-diagnostic, and augmentation results. Chapter 6 examines ERP plausibility, time-resolved decoding, and negative controls. Chapter 7 presents the Subject Quality Index, susceptibility, relative-importance, EEGNet-gain, and quality-stratified augmentation analyses. Chapter 8 discusses the findings, implications, limitations, future work, and conclusions.

Chapter 2

Background and Literature Review

2.1 EEG and ERP-Based Stimulus Classification

Electroencephalography records voltage fluctuations from electrodes placed on the scalp and is widely used for studying time-resolved neural responses [7, 12]. These signals reflect summed neural activity as well as non-neural artifacts and environmental noise. EEG is particularly useful for studying rapid cognitive processes because it provides temporal resolution on the order of milliseconds.

However, this advantage comes with challenges: the signal-to-noise ratio is low, spatial resolution is limited, and the measured channels contain mixtures of many underlying neural and non-neural sources [10, 27, 29]. These properties make EEG classification sensitive to preprocessing, evaluation design, and subject-level signal quality.

Stimulus classification is one way to evaluate whether EEG contains task-related information. In a typical target-versus-standard setting, a classifier is trained to distinguish EEG epochs following target stimuli from epochs following standard stimuli. Such classification can be useful in BCI systems, cognitive neuroscience, and clinical research. However, classification accuracy alone does not establish that the classifier is using physiologically meaningful information. Evaluation design, preprocessing choices, and subject-level signal quality all affect the result.

2.2 Oddball Paradigms and ERP/P300 Separability

Oddball paradigms are commonly used to elicit event-related responses to rare or task-relevant stimuli, including P300/P3-related responses [20, 26, 35]. In many oddball tasks, target stimuli evoke a larger late response than standard stimuli, often discussed in relation to the P300 or P3 component.

The P300 is not a single fixed waveform that appears identically across all tasks and subjects; its latency, amplitude, scalp distribution, and reliability can vary across

experimental conditions and participants [6, 26, 35]. For classification, the practical question is whether target and standard epochs are separable in the EEG signal. In this thesis, the term ERP/P300 separability is used broadly to refer to measurable post-stimulus target-standard differences in relevant time windows and regions of interest. This avoids overclaiming that every dataset shows identical textbook P300 morphology while still recognizing that late post-stimulus ERP structure can provide useful discriminative information.

2.3 BCI Performance Variability and Subject Dependence

Substantial variation among users is a persistent problem in EEG-based BCI research. Earlier literature often described users who did not achieve a predefined level of performance as exhibiting "BCI illiteracy" [1, 40]. More recent discussion commonly uses terms such as BCI inefficiency or BCI performance variability, which avoid treating the difficulty as a fixed property of the individual and recognize that performance can depend on the paradigm, recording conditions, preprocessing, feedback, and classifier [15, 37].

Much of the BCI-inefficiency literature has focused on motor-imagery paradigms and on physiological or behavioural markers that distinguish higher- and lower-performing users [1, 37, 40]. This work establishes that inter-subject variability is scientifically meaningful rather than merely random measurement error. It also shows that user performance may be influenced by factors outside the classifier itself.

The present thesis addresses a related but narrower question in oddball and ERP-style stimulus classification. Rather than defining a fixed category of successful and unsuccessful users, it treats classifiability as a continuous subject-level outcome measured by ROC-AUC. It then examines whether subject difficulty is consistent across three classifiers and whether interpretable EEG and ERP features are associated with this difficulty in both subject-specific and held-out-subject evaluation.

2.4 Classical EEG Classifiers

Classical linear classifiers remain important in EEG research because they are interpretable, computationally efficient, and often competitive in EEG-based brain–computer interface applications [24, 25]. Linear Discriminant Analysis has a long history in ERP classification, particularly when preprocessing or feature extraction produces approximately linear class separation, while Logistic Regression provides a regularized probabilistic alternative. In ERP-style tasks, discriminative information is often distributed across channels and time but remains sufficiently structured for linear models to perform strongly. Accordingly, LDA and Logistic Regression are treated in this thesis as established comparators rather than weak baselines.

2.5 Deep Learning and EEGNet

Deep neural networks can learn spatial and temporal representations directly from EEG epochs and have become increasingly common in EEG classification [2, 8, 14, 36]. EEGNet is a compact convolutional architecture designed specifically for EEG-based BCI applications, using temporal and depthwise spatial filtering while requiring substantially fewer parameters than many larger neural networks [21]. Related convolutional approaches have also demonstrated the value of learned temporal and spatial EEG representations [38].

In this thesis, EEGNet represents a compact EEG-specific nonlinear model rather than all deep-learning approaches. Its inclusion permits a controlled comparison with two established linear classifiers and tests whether learned representations alter subject-level classifiability, particularly under LOSO evaluation. Larger convolutional, recurrent, and Transformer-based architectures remain outside the scope of the final benchmark.

2.6 Cross-Subject Generalization

Within-subject and cross-subject evaluation address different generalization problems, and transfer to unseen subjects remains a major challenge in EEG-based BCI research [22, 37, 42]. Within-subject evaluation trains and tests on different observations from the same participant and is relevant when subject-specific calibration is available. Leave-one-subject-out evaluation instead trains on multiple participants and tests on an entirely held-out participant, requiring the classifier to accommodate inter-subject variation.

This distinction is central to the thesis because the classifier ranking differed across the two settings: the linear models were generally strongest in grouped within-subject evaluation, whereas EEGNet was strongest in LOSO. Classifier performance must therefore be interpreted in relation to the generalization setting rather than as a universal model ranking.

2.7 Signal Quality, Noise, and Classifiability

EEG signal quality is a broad and context-dependent concept. It can refer to electrode contact, artifact burden, background variability, trial count, temporal stability, or the reliability of a task-related response [6, 26, 29]. A recording may therefore be technically usable while still containing weak target–standard information, and a large averaged ERP may still coexist with substantial single-trial variability.

For target-versus-standard classification, two components are particularly relevant. The first is between-class separability: target and standard trials should differ in amplitude, temporal structure, spatial distribution, or another feature accessible

to the classifier. The second is within-class consistency: responses belonging to the same class should not vary so strongly that the class distributions overlap. Larger and more temporally consistent ERP differences can improve separability, whereas elevated baseline or post-stimulus variability can increase within-class dispersion and reduce classification performance.

The present study operationalizes these concepts using subject-level summaries. The features include baseline root-mean-square (RMS) amplitude, baseline standard deviation, post-stimulus variability, region-of-interest target–standard differences, P300-window amplitude and peak summaries, ERP signal-to-noise measures, single-trial effect sizes, spectral power features, and trial-count measures. These variables are used to study retrospective associations with subject-level AUC. They should not be interpreted as a complete or universally validated definition of EEG quality.

2.8 Source Separation and ICA as Exploratory Motivation

The project initially considered source separation and sparse autoencoders, while classical Independent Component Analysis remains widely used for EEG decomposition and artifact analysis [4, 10, 16, 27]. In the final thesis, ICA is retained only through exploratory source-quality features, including component kurtosis, spectral entropy, spatial roughness, and mixing-matrix summaries. These features were weaker than the raw EEG and ERP/P300 measures and are therefore not treated as the principal explanation of subject classifiability.

2.9 Positioning of the Present Work

The present study is related to several established areas of EEG and BCI research, but it does not duplicate any one of them. Research on BCI inefficiency investigates why some users achieve lower control or classification performance. EEG and ERP quality research examines the effects of artifacts, noise, preprocessing, and measurement reliability. Classifier-benchmarking studies compare algorithms under standardized evaluation procedures. Transfer-learning studies seek models that generalize across subjects, sessions, or datasets. These areas overlap, but they emphasize different objectives.

Table 2.1 summarizes how the present thesis is positioned relative to these research directions. The categories are not mutually exclusive; the table identifies their primary emphasis and the specific combination addressed here.

The originality of the thesis therefore does not rest on introducing a new classifier, identifying subject variability for the first time, or proposing a universal signal-quality metric. Its contribution lies in combining several complementary analyses within the same cross-dataset framework. Specifically, the study compares model performance under subject-specific and held-out-subject evaluation, quantifies the

Table 2.1: Positioning of the present thesis relative to related EEG and BCI research directions.

Research direction	Primary emphasis in the cited literature	Distinction of the present thesis
BCI inefficiency and subject variability [1, 15, 37, 40]	Explaining why users differ in BCI performance, often in motor-imagery paradigms, using physiological, behavioural, or session-level factors.	Examines continuous subject-level AUC in oddball and ERP-style classification, tests whether difficulty is consistent across three classifiers, and compares subject-level variation with classifier-level variation.
EEG/ERP signal quality and methodological reliability [6, 26, 29]	Characterizing artifacts, noise, preprocessing choices, ERP reliability, and factors that affect physiological interpretation.	Relates interpretable noise and ERP-separability measures directly to subject-level classification performance across datasets and evaluation settings.
Classifier benchmarking [18, 24, 25]	Comparing algorithms and reporting average performance under standardized or reproducible evaluation procedures.	Uses classifier comparison as one component of the analysis, but additionally asks whether between-subject differences are larger and more consistent than between-classifier differences.
EEG-specific deep learning [8, 21, 38]	Developing neural architectures that learn temporal and spatial EEG representations and comparing them with classical methods.	Uses one compact EEG-specific neural model as a controlled comparator and studies whether its gains vary according to subject quality and evaluation setting.
Cross-subject transfer and generalization [22, 42]	Reducing distribution differences among subjects, sessions, or domains to improve performance on new users.	Evaluates held-out-subject generalization through LOSO but does not introduce a new transfer-learning algorithm. The analysis instead examines which held-out subjects remain difficult and how this difficulty relates to their signal characteristics.
Present thesis	Cross-dataset subject-classifiability analysis in oddball and ERP-style EEG.	Combines grouped within-subject and LOSO evaluation, three classifier families, subject-level susceptibility analysis, cross-classifier consistency, incremental explanatory comparisons, and EEGNet-gain analysis in one framework.

consistency of subject difficulty across models, relates that difficulty to interpretable EEG and ERP characteristics, and directly compares the additional AUC variation explained by subject quality and classifier identity.

This positioning also defines the limits of the contribution. The analysis concerns four oddball or ERP-style datasets, three classifiers, and retrospective subject-quality features derived from labelled data. It does not establish that the same relationships hold for every EEG paradigm, every classifier family, or prospective users without calibration data.

2.10 Summary and Thesis Gap

Existing research shows that EEG classification is affected by preprocessing, artifacts, evaluation design, classifier choice, and substantial inter-subject variability [6, 24, 37]. Related work has examined BCI inefficiency, algorithm benchmarking, EEG quality, and cross-subject transfer, but these perspectives are not always integrated within the same oddball or ERP-style decoding analysis.

This thesis addresses that gap by evaluating three representative classifiers under grouped within-subject and LOSO designs, measuring whether subject difficulty is consistent across models, relating subject-level AUC to interpretable EEG and ERP characteristics, and comparing the additional variation explained by subject quality and classifier identity after accounting for dataset. Its contribution is an analysis of the sources and consistency of decoding difficulty rather than the introduction of a new classification algorithm.

Chapter 3

Datasets, Task Definitions, and Preprocessing

3.1 Dataset Overview

The final analysis used four publicly available OpenNeuro EEG datasets containing oddball or event-related potential tasks represented as binary target-versus-standard classification problems [28]. The datasets followed the Brain Imaging Data Structure (BIDS) or compatible conventions for organizing electrophysiological recordings, events, channels, and metadata [3, 11, 34].

The benchmark included 103 subject-recordings: all 13 available subjects from `ds003061` and 30 subjects from each of `ds005863`, `ds006466`, and `ds006480`. The datasets differed in stimulus modality, participant population, event coding, channel count, trial count, and recording setup; the analysis therefore tested related cross-dataset patterns rather than assuming equivalent protocols. The active auditory-oddball condition was used for `ds006466`, while target and standard events were pooled across the included `ds006480` conditions and distractors were excluded.

Table 3.1 summarizes the analyzed tasks, sample sizes, retained channel counts, and dataset-specific conditions.

3.2 Dataset Sources

The `ds003061` dataset contains auditory oddball EEG recordings [9]; `ds005863` contains cognitive electrophysiology recordings across adulthood, including a visual oddball task [17]; and `ds006466` and `ds006480` contain corresponding older- and younger-adult auditory oddball recordings [30, 31]. Differences in participants, tasks, markers, channels, recording systems, and trial counts were treated as dataset effects rather than ignored, and the results should not be interpreted as generalization to every oddball paradigm.

Table 3.1: Datasets included in the final analysis. Channel counts refer to the EEG channels retained after channel-type selection.

Dataset	Analyzed task or condition	Subjects	EEG channels	Analysis note
ds003061	Auditory oddball/P300 task	13	64	All available subjects
ds005863	Visual oddball task	30	26	Target-versus-standard event mapping
ds006466	Active auditory oddball condition	30	65	Older-adult dataset
ds006480	Pooled auditory oddball conditions	30	65	Young-adult dataset; distractors excluded
Total	–	103	–	Subject-recordings

3.3 Task Definitions and Event Mapping

Each dataset was converted into a binary target-versus-standard classification problem. Standard trials were assigned label 0, and target trials were assigned label 1. Event definitions were specified separately for each dataset because the original annotation descriptions and numeric marker codes were not identical.

For ds003061, the annotation stimulus/standard defined the standard class. The target class was the union of stimulus/oddball and stimulus/oddball_with_response; the implementation also accepted the corresponding misspelled annotation variant when present. For ds005863, target codes were 11, 22, 33, 44, and 55, while the remaining included stimulus codes listed in Table 3.2 were treated as standards.

The final ds006466 analysis used the active auditory-oddball condition. Within that condition, code 113 represented a standard event and code 114 represented a target event. The event-loading utility also supported the corresponding passive and shock-condition codes, but these additional conditions were not part of the final ds006466 benchmark.

For the pooled ds006480 analysis, standard codes 106, 113, 125, and 132 and target codes 107, 114, 126, and 133 were combined across control/shock and active/passive conditions. Distractor codes were not included in either class. After event selection, target and standard epochs were returned to their original chronological order using their event sample indices.

The exact event definitions are consolidated in Appendix A.

Table 3.2: Summary of event definitions used in the final binary classification tasks.

Dataset	Standard definition	Target definition
ds003061	stimulus/standard	stimulus/oddball and stimulus/oddball_with_response
ds005863	12, 13, 14, 15, 21, 23, 24, 25, 31, 32, 34, 35, 41, 42, 43, 45, 51, 52, 53, 54	11, 22, 33, 44, 55
ds006466	113, active condition	114, active condition
ds006480 pooled	106, 113, 125, 132	107, 114, 126, 133

3.4 Channel Handling

Only channels identified as EEG were retained for classifier input and for the principal raw/ERP feature analyses. Auxiliary physiological channels, response channels, stimulus channels, and miscellaneous sensors were excluded. Dataset-specific channels such as EXG and GSR channels were reassigned to non-EEG channel types before EEG channel selection.

The retained EEG channels were rereferenced to their common average. When a recording did not already contain montage information, the standard 10–20 montage supplied by MNE-Python was assigned using channel-name matching, with unmatched channel locations left unspecified. This fallback was used to support topographic and region-of-interest analyses without requiring identical channel names across all datasets.

The resulting EEG channel counts were 64 for ds003061, 26 for ds005863, and 65 for both ds006466 and ds006480. Dataset-specific centroparietal regions of interest were used for ERP summaries because a single set of channel names was not available across all four datasets.

3.5 Filtering, Resampling, and Epoching

Continuous recordings were loaded and processed using MNE-Python and MNE-BIDS [3, 12]. After EEG channel selection and average rereferencing, the final classifier stream was band-pass filtered from 0.1 to 12 Hz. Because the implementation did not override MNE-Python's filtering method, its default zero-phase finite impulse response (FIR) implementation was used, with automatically determined filter length and transition bandwidth.

A 60 Hz notch-filter operation was also included in the standardized loading pipeline. In the final classifier stream, the preceding 12 Hz low-pass filter already strongly attenuated frequencies around the line frequency; the notch operation was therefore not expected to materially change the retained classification bandwidth.

Recordings were resampled to 128 Hz after filtering.

An EEG epoch was defined as a fixed-duration segment of the continuous recording aligned to one stimulus event. The stimulus-locked EEG epochs extended from −0.2 to 0.8 s relative to event onset. In later descriptions of neural-network optimization, a training epoch instead refers to one complete pass through the inner training set. Baseline correction was applied to every final dataset by subtracting, separately for each channel and epoch, the mean voltage over the −0.2 to 0 s prestimulus interval. This interval was available for all recordings included in the final benchmark. Epochs overlapping annotations marked as bad were excluded using MNE-Python's annotation-based rejection. No additional fixed peak-to-peak amplitude rejection threshold was imposed in the final loader.

The resulting epoched arrays had the form

$$X \in \mathbb{R}^{N \times C \times T}$$

where N is the number of retained stimulus trials, C is the dataset-specific number of EEG channels, and T is the number of samples. The classifiers used the post-stimulus crop from 0 to 0.8 s. At 128 Hz, and with both endpoints retained by the epoching convention, this crop contained 103 temporal samples per channel.

No Independent Component Analysis cleaning was applied to the principal final classifier benchmark. ICA-derived quantities were computed separately as exploratory subject-level source-quality features and were not used to replace the main 0.1–12 Hz classifier input.

3.6 Final Input Representation

The final benchmark used the same 0 to 0.8 s post-stimulus information for all three classifiers, but the temporal representation differed between the linear and neural models. Each retained EEG epoch remained one training or test observation for every model. Temporal bins were not treated as independent examples.

Time-binned representation for the linear models

For LDA and Logistic Regression, the preprocessed epoch was averaged within nominal 50 ms non-overlapping temporal bins. At the 128 Hz sampling rate, the implemented bin width was

$$n_{\text{bin}} = \text{round}(0.050 \times 128) = 6 \text{ samples}$$

corresponding to an effective duration of

$$\frac{6}{128} = 0.046875 \text{ s}$$

Table 3.3: Final classifier input dimensions. Each row corresponds to one EEG epoch and therefore one classifier observation.

Dataset	EEG channels	Temporal samples	Linear input	EEGNet input
`ds003061`	64	103	$64 \times 17 = 1088$	$1 \times 64 \times 103$
`ds005863`	26	103	$26 \times 17 = 442$	$1 \times 26 \times 103$
`ds006466`	65	103	$65 \times 17 = 1105$	$1 \times 65 \times 103$
`ds006480`	65	103	$65 \times 17 = 1105$	$1 \times 65 \times 103$

The 0 to 0.8 s crop contained 103 samples. The binning implementation retained 102 samples to form 17 complete six-sample bins; the remaining endpoint sample was omitted rather than padded. For epoch i, channel c, and temporal bin b, the binned feature was

$$\overline{x}_{i,c,b} = \frac{1}{6} \sum_{j=1}^{6} x_{i,c,6(b-1)+j}$$

The resulting channel-by-bin matrix was then flattened into one vector:

$$\mathbf{x}_i^{\text{linear}} \in \mathbb{R}^{C \times 17}$$

Consequently, the linear input dimensionality was $17C$. This operation reduced temporal dimensionality while retaining a coarse representation of the post-stimulus ERP time course.

Unbinned representation for EEGNet

EEGNet received the same preprocessed 0 to 0.8 s epochs without temporal averaging. A singleton convolutional input dimension was added, giving one input tensor per epoch of shape

$$\mathbf{X}_i^{\text{EEGNet}} \in \mathbb{R}^{1 \times C \times 103}$$

The first dimension is the convolutional input-map dimension, C is the number of retained EEG channels, and 103 is the number of temporal samples. This representation allowed EEGNet to learn temporal and spatial filters directly from the preprocessed epoch [21].

Table 3.3 compares the resulting input dimensions for all four datasets and both model families.

The time-resolved decoding analysis in Chapter 6 was a separate diagnostic procedure. In that analysis, classifiers were fitted using one temporal window at a time to determine when discriminative information appeared. It should not be interpreted as passing each temporal bin independently to the final benchmark classifiers or as increasing the number of training observations for the linear models.

3.7 Subject-Level Feature Extraction

Subject-level features were computed separately from the trial-level inputs used to train LDA, Logistic Regression, and EEGNet. They did not enter the classifier feature vectors. Instead, they summarized each subject's recording and were later related to that subject's aggregated classification AUC.

A separate broadband preprocessing stream from 0.1 to 45 Hz was used for the raw/ERP/power susceptibility analysis. This wider bandwidth preserved spectral information that would not be available after the 12 Hz classifier low-pass filter. The broadband feature stream otherwise used the same 128 Hz sampling rate, -0.2 to 0.8 s epoch interval, and -0.2 to 0 s baseline correction.

The raw and ERP-derived feature families included:

- baseline root-mean-square amplitude and standard deviation;
- post-stimulus root-mean-square amplitude and variability;
- target-minus-standard ERP differences;
- centroparietal region-of-interest summaries;
- P300-window mean, peak, and root-mean-square differences;
- ERP signal-to-noise summaries;
- single-trial target–standard effect sizes;
- spectral power and band-ratio measures; and
- numbers and proportions of target and standard trials.

The principal P300 analysis window was 0.25 to 0.50 s after stimulus onset. Region-of-interest channels were selected using dataset-compatible centroparietal channel candidates rather than requiring identical channel labels across datasets. ICA-derived component kurtosis, spectral entropy, spatial roughness, and mixing-matrix summaries were computed as a separate exploratory source-quality family.

These subject-level features were calculated retrospectively from the available labelled trials. In particular, target-minus-standard and single-trial separability measures require knowledge of the event labels. They therefore characterize subject classifiability within the analyzed datasets and should not be interpreted as a prospectively validated quality estimate for a new user without labelled calibration data.

The next chapter describes the training, validation, and independent test partitions, the controls against temporal and subject-level leakage, and the implementation of the three classifiers.

Chapter 4

Evaluation Framework and Classifiers

4.1 Why Evaluation Design Matters

EEG classification performance can be strongly influenced by how training and test observations are constructed. Random trial-level partitioning may distribute temporally adjacent or otherwise related trials across both sets, potentially producing an optimistic estimate of generalization [6, 18, 24]. This concern is particularly relevant when epochs originate from continuous recordings and may share slow signal drift, task-block structure, recording conditions, or neighbouring artifacts.

The evaluation framework was therefore designed around complete chronological groups rather than independently sampled trials. All feature extraction, model fitting, normalization, validation-based model selection, and testing followed the resulting data hierarchy. Two outer evaluation settings were used:

1. grouped within-subject evaluation, which measured subject-specific decodability using different chronological groups from the same subject; and
2. leave-one-subject-out (LOSO) evaluation, which measured generalization to a subject who was entirely excluded from model fitting.

The term *outer test set* refers to the partition used for final performance estimation. For EEGNet, a separate *inner validation set* was created only from the corresponding outer training data. The outer test observations were not used for normalization, early stopping, checkpoint selection, or hyperparameter selection.

4.2 Grouped Within-Subject Outer Evaluation

For each subject, target and standard epochs were returned to their original chronological order using the event sample indices. The ordered trials were then assigned

Table 4.1: Grouped within-subject outer-fold design. The nominal group size is the preferred number of chronological trials per group. One complete neighbouring group was purged on each available side of the test block.

Dataset	Subjects	Nominal group size	Requested folds	Valid folds	Purged groups
ds003061	13	80	65	65	1 per side
ds005863	30	20	150	150	1 per side
ds006466	30	40	150	146	1 per side
ds006480	30	40	150	149	1 per side
Total	103	–	515	510	–

to contiguous groups. The preferred number of trials per group depended on the dataset:

- 80 trials for ds003061;
- 20 trials for ds005863;
- 40 trials for ds006466; and
- 40 trials for ds006480.

When a nominal group boundary was reached before the group contained a target trial, the group was allowed to extend beyond its preferred size until at least one target was included, when possible. This safeguard reduced the number of grouped folds that lacked one of the classes while preserving chronological ordering.

The ordered sequence of groups was divided into up to five contiguous outer test blocks. For each outer fold, the groups in one block formed the test partition. The immediately adjacent group on each side of the test block was excluded from the training partition. Thus, the final purge width was one complete neighbouring group before and one complete neighbouring group after the test block, except at the beginning or end of a recording where only one adjacent side existed.

A fold was retained only when:

1. both the training and test partitions contained observations;
2. both partitions contained both classes; and
3. both partitions contained at least one target trial.

Consequently, the analysis requested five folds per subject but did not force invalid folds to remain in the benchmark. Table 4.1 reports the resulting number of valid outer folds.

The final grouped within-subject analysis therefore contained 510 distinct outer training–test partitions. Different folds from the same subject were used to estimate performance across different chronological portions of that subject's recording; they were not treated as independent subjects in the subsequent statistical comparisons.

4.3 Leave-One-Subject-Out Outer Evaluation

LOSO evaluation used an entire subject as the outer test set. For each dataset, one subject was removed, all remaining subjects formed the outer development pool, and the procedure was repeated until every subject had served as the held-out test subject once.

For `ds003061`, each LOSO model was trained using 12 subjects and evaluated on the remaining subject. For each of the three 30-subject datasets, the corresponding model was trained using 29 subjects and evaluated on one held-out subject. This produced 13 + 30 + 30 + 30 = 103 distinct LOSO outer partitions.

Group identifiers were offset before the training subjects were concatenated so that groups originating from different subjects remained distinct. For EEGNet, the validation partition was then selected from complete groups belonging only to the training subjects. The held-out subject was not used to estimate normalization statistics, select an epoch, tune a model, or define early stopping.

LOSO should therefore be interpreted as unseen-subject generalization within each dataset. It is not equivalent to cross-dataset transfer because the training and held-out subjects came from the same OpenNeuro dataset.

4.4 Outer and Inner Data Hierarchy

Across both evaluation settings, outer test observations remained untouched until final evaluation. In grouped within-subject evaluation, EEGNet's inner training and validation groups were selected only from the outer training groups, while the chronological outer test block remained excluded from normalization, checkpoint selection, and model fitting. In LOSO evaluation, the held-out subject was excluded from every development step, and EEGNet's inner split used only the remaining subjects. LDA and Logistic Regression used each fixed outer training partition directly. Together, the two designs produced 510 + 103 = 613 distinct outer training–test partitions.

4.5 Inner Validation for EEGNet

EEGNet required an inner validation set because its weights were optimized iteratively and its training epoch was selected using early stopping. Approximately 15% of the complete groups in each outer training pool were assigned to validation.

Group membership was preserved, so trials from the same group were not divided between inner training and validation.

The split routine performed up to 200 candidate group selections. Candidate splits were required to contain both classes in both the inner training and validation partitions. Among valid candidates, the split whose class proportions most closely matched the target proportion of the complete outer training pool was retained.

Because group sizes and class composition differed, the validation partition was approximately 15% of the outer training *groups*, rather than exactly 15% of the outer training trials. The remaining groups formed the inner training set.

Channel-wise normalization statistics were estimated only from the inner EEGNet training set. For channel c, the mean and standard deviation were calculated across all training epochs and temporal samples:

$$\begin{aligned} \mu_c &= \frac{1}{N_{\text{tr}} T} \sum_{i=1}^{N_{\text{tr}}} \sum_{t=1}^{T} X_{i,c,t} \\ \sigma_c &= \sqrt{\frac{1}{N_{\text{tr}} T} \sum_{i=1}^{N_{\text{tr}}} \sum_{t=1}^{T} \left(X_{i,c,t} - \mu_c\right)^2} \end{aligned}$$

The same training-derived values were then applied to the inner training, validation, and outer test partitions:

$$\widetilde{X}_{i,c,t} = \frac{X_{i,c,t} - \mu_c}{\sigma_c + 10^{-8}}$$

This prevented information from the validation or outer test data from influencing normalization.

LDA and Logistic Regression did not use an inner validation set because their configurations were fixed before the final benchmark and they did not require epoch-wise checkpoint selection. Their scaling and classifier parameters were fitted inside each outer training partition and then applied to the corresponding outer test partition.

4.6 Classifier Selection and Scope

The final benchmark was intentionally restricted to LDA, Logistic Regression, and EEGNet. LDA represents an established discriminant approach for ERP and BCI classification, Logistic Regression provides a regularized probabilistic linear alternative, and EEGNet provides a compact EEG-specific convolutional model capable of learning temporal and spatial filters directly from multichannel epochs [21, 24, 25]. This controlled model set allowed the same subject-level analyses to be completed across four datasets and two evaluation settings.

Support Vector Machines, Random Forests, gradient-boosting methods, larger convolutional networks, and Transformer-based architectures were not included

Table 4.2: Effective linear decision-function dimensionality. Counts include one coefficient per input feature and one intercept. Fitted scaling statistics and LDA covariance estimates are described in the text but are not counted as independent decision coefficients.

Dataset	Input features p	LDA decision coefficients	Logistic Regression coefficients
`ds003061`	1088	1089	1089
`ds005863`	442	443	443
`ds006466`	1105	1106	1106
`ds006480`	1105	1106	1106

because their fair evaluation would require a substantially broader hyperparameter-selection study. The conclusions are therefore restricted to the three evaluated classifiers and do not establish that subject quality would explain more variability than every possible model family.

4.7 Linear Classifiers

LDA and Logistic Regression were implemented using scikit-learn [33]. Both models were fitted as pipelines beginning with a `StandardScaler` estimated only from the corresponding outer training partition.

Linear Discriminant Analysis

The LDA estimator used the least-squares solver with automatic covariance shrinkage. It modeled class-specific means and a shared regularized within-class covariance matrix. For an input vector $\mathbf{x} \in \mathbb{R}^p$, its binary decision function was

$$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$$

where $\mathbf{w} \in \mathbb{R}^p$ and b are the discriminant coefficients and intercept. The effective decision function therefore contained $p + 1$ coefficients; the fitted class means, priors, covariance estimate, and scaling statistics were additional stored quantities rather than independently optimized decision coefficients.

Logistic Regression

The Logistic Regression estimator used the LBFGS solver, an L_2 penalty with $C = 1.0$, an intercept, balanced class weighting, and a maximum of 2000 iterations. Its target-class probability was

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + \exp[-(\mathbf{w}^\top \mathbf{x} + b)]}$$

Table 4.3: Implemented EEGNet architecture for a post-stimulus epoch with C channels and 103 temporal samples. N denotes batch size.

Stage	Operation and hyperparameters	Output shape
Input	Preprocessed, unbinned EEG epoch	$N \times 1 \times C \times 103$
Temporal filtering	8 convolutions; kernel 1×32; no bias; batch normalization	$N \times 8 \times C \times 104$
Spatial filtering	Depthwise convolution across C channels; depth multiplier $D = 2$; no bias	$N \times 16 \times 1 \times 104$
First reduction	Batch normalization; ELU; average pooling 1×4; dropout 0.25	$N \times 16 \times 1 \times 26$
Separable temporal filtering	Depthwise kernel 1×16, followed by 16 pointwise 1×1 outputs	$N \times 16 \times 1 \times 27$
Second reduction	Batch normalization; ELU; average pooling 1×8; dropout 0.25	$N \times 16 \times 1 \times 3$
Classifier	Flatten 48 values; fully connected layer with two outputs	$N \times 2$

Its effective prediction function likewise contained p feature coefficients and one intercept, together with the training-derived scaling statistics.

Table 4.2 reports the effective dimensionality of both linear decision functions.

4.8 EEGNet Architecture

EEGNet was implemented using the PyTorch deep-learning framework [21, 32] and received input tensors of shape $(N, 1, C, 103)$, where N is the batch size, C is the dataset-specific channel count, and 103 is the number of post-stimulus samples [21]. The implemented architecture used the following hyperparameters:

$$F_1 = 8, \qquad D = 2, \qquad F_2 = 16$$

with temporal kernel lengths of 32 and 16 samples and dropout probability 0.25. Each temporal convolution used zero padding on both temporal sides: 16 samples per side for the 32-sample kernel and 8 samples per side for the 16-sample kernel. Because the kernels had even lengths, this padding increased the temporal dimension by one sample: the first convolution changed the input length from 103 to 104 samples, and the second temporal convolution changed its input length from 26 to 27 samples, as reported in Table 4.3.

The first block applied eight temporal convolutional filters with kernel size 1×32, followed by batch normalization. A depthwise spatial convolution then applied two spatial filters per temporal filter across the complete EEG channel dimension, producing $F_1 D = 16$ feature maps. This operation was followed by batch

Table 4.4: Number of trainable EEGNet parameters by dataset.

Dataset	EEG channels C	Trainable parameters
`ds003061`	64	1970
`ds005863`	26	1362
`ds006466`	65	1986
`ds006480`	65	1986

Table 4.5: EEGNet optimization and regularization settings.

Setting	Value
Optimizer	Adam
Learning rate	1×10^{-3}
Weight decay	1×10^{-4}
Batch size	64
Maximum training epochs	60
Early-stopping patience	10 epochs
Checkpoint criterion	Validation ROC-AUC
Minimum improvement	1×10^{-4}
Dropout probability	0.25
Gradient-clipping norm	1.0
Loss	Class-weighted cross-entropy
Validation groups	Approximately 15% of outer training groups
Validation split attempts	Up to 200
Random seeds	42, 43, and 44

normalization, an exponential linear unit (ELU), average pooling with width four, and dropout.

The second block used a depthwise temporal convolution with kernel size 1×16, followed by a 1×1 pointwise convolution producing 16 output feature maps. Batch normalization, ELU activation, average pooling with width eight, and dropout followed. The output was flattened into 48 values and passed to a two-output linear classification layer.

The number of trainable EEGNet parameters depended on the EEG channel count because the depthwise spatial convolution used a kernel spanning all C channels. The total number of trainable parameters was $946 + 16C$. The dataset-specific counts are reported in Table 4.4.

4.9 EEGNet Training and Regularization

EEGNet was trained using the Adam optimizer. Table 4.5 summarizes the optimization and regularization configuration.

The target class was less frequent than the standard class. The cross-entropy weights were calculated from the inner training labels as

$$w_0 = 1, \qquad w_1 = \frac{n_0}{n_1}$$

where n_0 and n_1 are the numbers of standard and target trials in the inner training partition.

Training batches were shuffled, whereas validation and test batches retained fixed order. At every training epoch, the implementation recorded training loss, validation loss, validation ROC-AUC, validation precision–recall AUC, and validation balanced accuracy.

A checkpoint was retained whenever validation ROC-AUC improved by more than 10^{-4}. The early-stopping counter was reset after an improvement. Training ended when validation ROC-AUC failed to improve for ten consecutive epochs or when 60 epochs were completed. The parameter state corresponding to the highest validation ROC-AUC was restored before the untouched outer test data were evaluated.

Gradients were clipped to a maximum norm of 1.0. Dropout and weight decay provided additional regularization. The model was repeated using random seeds 42, 43, and 44. Each seed initialized Python, NumPy, CPU and GPU PyTorch random-number generators; deterministic cuDNN behavior was enabled and cuDNN benchmarking was disabled.

Across the 613 distinct outer partitions, the three seeds produced $613 \times 3 = 1839$ EEGNet training runs. These repetitions used the same outer test definitions but different neural initializations and training-order randomness. They therefore measured training stability rather than creating additional independent test subjects or test folds.

4.10 Gaussian-Noise Augmentation Ablation

The primary classifier benchmark used the non-augmented EEGNet procedure described in the preceding section. A separate robustness ablation examined whether conservative training-input perturbation improved EEGNet generalization to unseen subjects.

The complete augmentation comparison was restricted to LOSO evaluation, where EEGNet had shown its strongest relative performance. All four datasets and all 103 held-out subject-recordings were included.

Augmentation transformation

Additive zero-mean Gaussian noise was selected as a conservative, label-preserving input perturbation. It slightly changes signal amplitude without shifting event

timing, removing channels, altering trial counts, changing class proportions, or synthesizing a new ERP morphology. In the normalized input space, it was intended to represent small nonspecific measurement variability rather than a complete physiological model of EEG artifacts or neural noise.

Gaussian noise was applied after channel normalization using statistics estimated exclusively from the inner training partition. For a normalized EEG training epoch $X \in \mathbb{R}^{C \times T}$, the augmented epoch was

$$\widetilde{X} = X + \epsilon, \qquad \epsilon_{c,t} \sim \mathcal{N}(0, \sigma^2)$$

Because each channel had already been normalized using inner-training data, σ is expressed in channel-standard-deviation units.

The perturbation was generated dynamically for each inner-training batch. Inner-validation epochs and held-out-subject outer-test epochs were not augmented. The transformation did not alter the event labels, number of trials, temporal alignment, channel organization, or class proportions.

A separate PyTorch random-number generator was used for augmentation noise. For a given held-out subject and neural seed, the perturbation draws therefore did not alter the random stream used for model initialization or shuffled training-batch order.

Validation-only selection of noise magnitude

The prespecified nonzero candidates were

$$\sigma \in \{0.01,\ 0.025,\ 0.05\}$$

They were compared in a validation-only pilot containing 36 paired partitions. For each dataset, the pilot included three selected subjects and two grouped within-subject folds per selected subject, together with three selected held-out subjects under LOSO evaluation.

The pilot sampled both evaluation settings so that the selected magnitude would not be tied exclusively to one split regime. The complete statistical comparison nevertheless remained restricted to LOSO, which was the evaluation setting of primary interest for unseen-subject generalization.

For each partition, the baseline and the three nonzero conditions used the same inner split and neural seed. Candidate selection used only the paired inner-validation ROC-AUC. Outer-test AUC was not calculated or used to select the noise magnitude.

All three nonzero candidates produced validation performance close to the non-augmented baseline. The selected value was

$$\sigma = 0.01,$$

because it produced the highest mean validation AUC among the nonzero candidates and the smallest mean decrease from the baseline.

Complete LOSO comparison

The complete analysis contained

$$103 \text{ held-out subject-recordings} \times 3 \text{ neural seeds} = 309$$

augmented EEGNet fits.

The seeds were

$$42, \quad 43, \quad 44$$

For every dataset, held-out subject, and seed, the augmented result was paired with the exact corresponding non-augmented EEGNet result from the principal LOSO benchmark. The paired conditions therefore shared:

- the same held-out subject;
- the same training subjects;
- the same inner-validation procedure;
- the same preprocessing and normalization;
- the same EEGNet architecture;
- the same optimizer and regularization settings; and
- the same neural random seed.

Only the augmented condition received Gaussian perturbation during inner-training batches.

Augmentation statistics

The overall subject-level outer-test AUC comparison was the primary augmentation endpoint. The quality-stratified comparisons and validation–test-gap analysis were secondary diagnostics.

Seed-level values were averaged within subject and condition before primary statistical inference. The resulting analysis contained one baseline and one augmented AUC for each of the 103 held-out subject-recordings.

The augmentation effect was defined as

$$\Delta_s^{\text{aug}} = \text{AUC}_{s,\text{augmented}} - \text{AUC}_{s,\text{baseline}}$$

The overall effect was evaluated using a two-sided paired Wilcoxon signed-rank test. The matched rank-biserial correlation was reported as an effect size. A 95% confidence interval for the mean subject-level change was calculated using 10,000 bootstrap resamples of the 103 subjects.

To examine possible differences according to subject quality, lower- and upper-quartile Subject Quality Index groups were defined separately within each dataset. Paired tests were performed within the hard and easy groups, with Holm correction across the two group-specific tests. A Mann–Whitney U test compared augmentation gains between the groups. A continuous Spearman correlation between Subject Quality Index and augmentation gain was also calculated.

A secondary analysis compared the validation–test gap

$$G_s = \mathrm{AUC}_{s,\text{validation}} - \mathrm{AUC}_{s,\text{outer test}}$$

between the baseline and augmented conditions. Neural seeds were averaged within subject before this paired test.

4.11 Negative Controls

The label-permutation and prestimulus procedures described in Section 6.5 were negative controls rather than augmentation methods.

In the label-permutation control, labels were shuffled independently within each subject while preserving that subject's original class counts. This destroyed the relationship between an EEG epoch and its stimulus class while retaining the number of target and standard trials.

In the prestimulus-only control, classifier input was restricted to the interval from -0.2 to 0 s relative to stimulus onset. This tested whether above-chance decoding could be obtained from prestimulus, chronological, or recording-related structure rather than post-stimulus target–standard information.

These controls were performed during the development-stage `ds003061` analysis and were not repeated over the complete 103-subject benchmark. They are therefore reported as supporting robustness checks rather than as complete cross-dataset controls.

Gaussian-noise augmentation served a different purpose. It retained the correct labels and modified only inner-training EEG inputs to evaluate regularization and held-out-subject generalization.

4.12 Performance Metrics and Aggregation

The primary performance measure was the area under the receiver operating characteristic curve (ROC-AUC). ROC-AUC estimates the probability that a randomly selected target trial receives a higher score than a randomly selected standard trial. It was selected because oddball tasks contain substantially more standard than target trials and because it does not depend on selecting one classification threshold.

Precision–recall AUC, accuracy, balanced accuracy, and F_1 score were also recorded for the neural training and diagnostic outputs, but the principal cross-model conclusions were based on ROC-AUC.

For grouped within-subject evaluation, fold-level AUC values were averaged to obtain one AUC value for each subject, model, and seed. Neural results were then averaged across the three seeds. For LOSO evaluation, each held-out subject produced one test AUC per model and seed, followed by averaging across seeds.

Cross-dataset headline values were computed as unweighted macro-averages of the four dataset-specific model means. They should not be interpreted as subject-count-weighted averages over all 103 subject-recordings.

4.13 Statistical Comparison of Classifiers

Classifier comparisons were performed using paired subject-level AUC values rather than treating outer folds or random seeds as independent observations. Separate analyses were conducted for grouped within-subject and LOSO evaluation.

For each setting, a Friedman test was used as the omnibus non-parametric comparison across LDA, Logistic Regression, and EEGNet. When the omnibus result indicated a classifier difference, paired Wilcoxon signed-rank tests were used for:

1. LDA versus Logistic Regression;
2. LDA versus EEGNet; and
3. Logistic Regression versus EEGNet.

The three pairwise p-values were adjusted using Holm's sequential procedure. Paired median AUC differences and matched rank-biserial effect sizes were reported alongside the significance tests. Statistical conclusions used a two-sided significance threshold of $\alpha = 0.05$.

The detailed classifier-comparison results are presented in Chapter 5. Subject-level susceptibility correlations, false-discovery-rate correction, the subject-quality index, and incremental R^2 analyses are described in Chapter 7 and Appendix B.

4.14 Chapter Summary

The evaluation framework contained 510 valid grouped within-subject partitions and 103 LOSO partitions. Outer test observations remained separate from fitting and normalization, while EEGNet used group-aware inner validation for checkpoint selection and early stopping. The three classifiers were evaluated using fixed implementations, and a separate LOSO Gaussian-noise ablation used validation-selected perturbation strength while preserving the original held-out subjects, seeds, architecture, and optimization settings. Chapter 5 reports the resulting benchmark, statistical comparisons, training diagnostics, and augmentation results.

Chapter 5

Benchmark Classification Results

5.1 Final Benchmark Overview

The final benchmark compared LDA, Logistic Regression, and EEGNet across 103 subject-recordings from four datasets. LDA and Logistic Regression received the time-binned feature vectors described in Section 3.6, whereas EEGNet received preprocessed, unbinned EEG epochs over the same 0 to 0.8 s post-stimulus interval.

Two evaluation settings were analyzed. Grouped within-subject evaluation measured subject-specific decoding using chronological training and test groups from the same recording. Leave-one-subject-out (LOSO) evaluation measured generalization to a subject who was entirely excluded from model fitting.

For every dataset and classifier, Table 5.1 and Table 5.2 report the mean subject-level area under the receiver operating characteristic curve (ROC-AUC) together with the standard deviation (SD) across subjects. EEGNet AUC values were first averaged across the three random seeds for each subject. The cross-dataset headline values reported in the text are unweighted macro-averages of the four dataset-level means; they are not weighted by the number of subjects in each dataset.

Figures 5.1 and 5.2 show 95% confidence intervals for each dataset-level mean. For a dataset containing n subjects, the interval was calculated as

$$\overline{x} \pm t_{0.975,n-1} \frac{s}{\sqrt{n}}$$

where $\overline{x}$ is the subject-level mean AUC, s is the subject-level SD, and $t_{0.975,n-1}$ is the corresponding critical value from the Student t-distribution. The SD values in the tables describe between-subject variability, whereas the confidence intervals in the figures describe uncertainty in the estimated dataset-level mean.

Table 5.1: Grouped within-subject ROC-AUC by dataset and classifier. Values are subject-level mean ± standard deviation after averaging EEGNet results across the three random seeds. Bold values indicate the highest descriptive mean within each dataset.

Dataset	N	LDA	Logistic Regression	EEGNet
ds003061	13	**0.9061** ± 0.0825	0.8884 ± 0.0922	0.8900 ± 0.0912
ds005863	30	**0.7535** ± 0.0941	0.7481 ± 0.0947	0.6342 ± 0.0702
ds006466	30	0.8058 ± 0.0966	**0.8183** ± 0.0859	0.7837 ± 0.0760
ds006480 pooled	30	**0.6829** ± 0.0785	0.6819 ± 0.0731	0.6539 ± 0.0755

Table 5.2: Leave-one-subject-out ROC-AUC by dataset and classifier. Values are subject-level mean ± standard deviation after averaging EEGNet results across the three random seeds. Bold values indicate the highest descriptive mean within each dataset.

Dataset	N	LDA	Logistic Regression	EEGNet
ds003061	13	0.7949 ± 0.1204	0.7912 ± 0.1063	**0.8626** ± 0.0929
ds005863	30	0.7762 ± 0.0938	0.7608 ± 0.0925	**0.8017** ± 0.0946
ds006466	30	0.7006 ± 0.1112	0.7801 ± 0.1079	**0.8255** ± 0.1114
ds006480 pooled	30	0.6421 ± 0.0711	0.6632 ± 0.0733	**0.6890** ± 0.0769

5.2 Grouped Within-Subject Results

Table 5.1 reports the grouped within-subject results. The highest dataset-level mean was obtained by LDA for ds003061, ds005863, and pooled ds006480. Logistic Regression had the highest mean for ds006466. EEGNet remained close to the linear models for ds003061 and ds006466, but showed a larger reduction for ds005863.

The unweighted macro-average of the four dataset means was 0.7871 for LDA, 0.7842 for Logistic Regression, and 0.7404 for EEGNet. These averages indicate that the two linear models were generally stronger in the subject-specific setting. However, the difference between LDA and Logistic Regression was small and requires paired statistical testing before either model can be described as superior.

5.3 Leave-One-Subject-Out Results

Table 5.2 reports performance when each subject was evaluated using models trained only on the other subjects from the same dataset. EEGNet produced the highest descriptive mean for all four datasets.

The unweighted macro-average of the dataset means was 0.7947 for EEGNet, 0.7488 for Logistic Regression, and 0.7285 for LDA. The LOSO ranking therefore

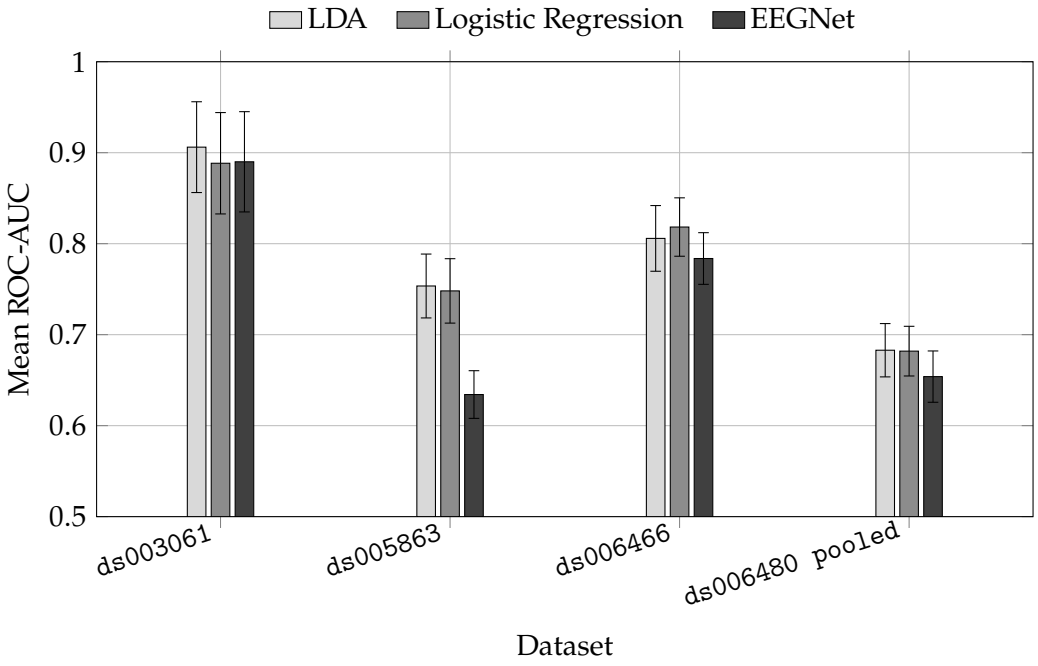

Figure 5.1: Grouped within-subject ROC-AUC by dataset and classifier. Bars show dataset-level means, and error bars show 95% confidence intervals calculated from subject-level AUC values. EEGNet results were averaged across the three random seeds before the subject-level summaries were computed.

differed from the grouped within-subject ranking. EEGNet was strongest when the test subject was entirely absent from model fitting, whereas the linear models were generally stronger when subject-specific training data were available.

The between-subject SD values remained substantial for every classifier. For example, the LOSO SD ranged from approximately 0.071 to 0.120 across datasets and models. Thus, the higher average performance of EEGNet did not eliminate variation among held-out subjects.

5.4 Paired Statistical Comparison of Classifiers

Classifier comparisons used one aggregated AUC value per dataset, subject, and classifier. Fold-level values were averaged within subject, and EEGNet values were additionally averaged across the three neural seeds. This produced 103 matched subject-recordings per classifier in each evaluation setting and avoided treating outer folds or neural runs as independent observations.

A Friedman test was first applied across the three classifiers, followed by two-sided paired Wilcoxon signed-rank tests with Holm correction. The matched

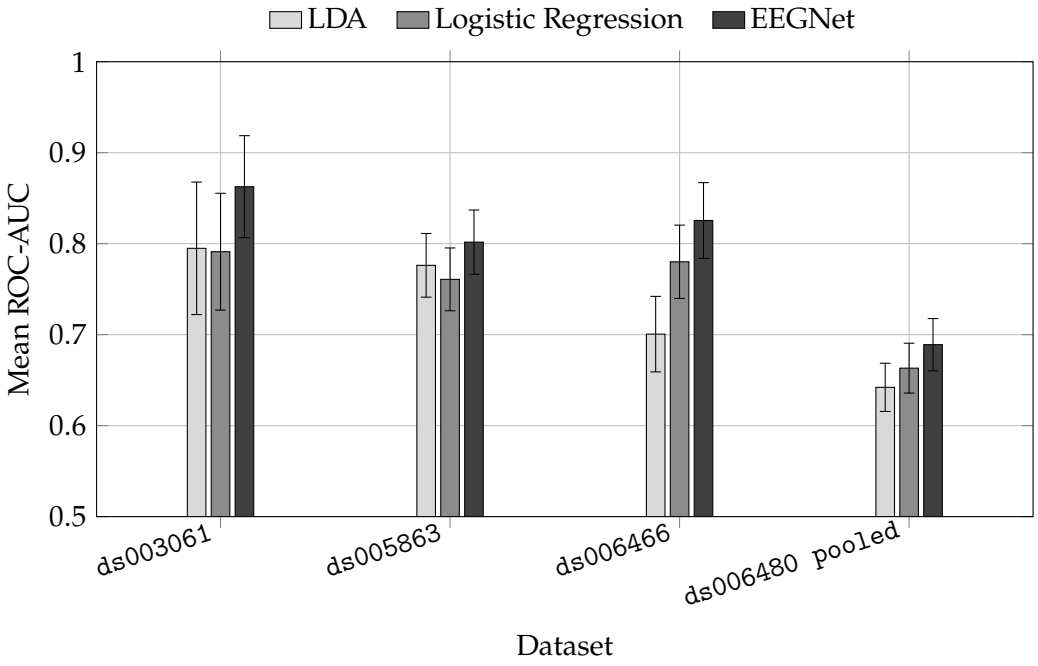

Figure 5.2: Leave-one-subject-out ROC-AUC by dataset and classifier. Values are subject-level mean ± standard deviation after averaging EEGNet results across the three random seeds. Bold values indicate the highest descriptive mean within each dataset.

rank-biserial correlation r_{rb} was reported as the pairwise effect size; positive values indicate higher AUC for the first classifier in the stated comparison.

The grouped within-subject omnibus result was significant,

$$\chi^2_F(2) = 58.39, \qquad p = 2.09 \times 10^{-13}$$

as was the LOSO result,

$$\chi^2_F(2) = 103.71, \qquad p = 3.02 \times 10^{-23}$$

Table 5.3 reports the corrected pairwise comparisons and matched effect sizes.

In grouped within-subject evaluation, LDA and Logistic Regression did not differ significantly, whereas both linear classifiers had significantly higher subject-level AUC than EEGNet. In LOSO evaluation, EEGNet significantly exceeded both linear models, and Logistic Regression had a smaller but statistically significant advantage over LDA. Negative LOSO effect-size signs indicate that the second classifier in the stated comparison performed better.

These conclusions concern the pooled set of 103 subject-recordings. They should not be interpreted as demonstrating that every pairwise difference is significant within each individual dataset.

Table 5.3: Paired classifier comparisons. Median differences are calculated as the first classifier minus the second classifier. r_{rb} is the matched rank-biserial effect size, and p_{Holm} is the two-sided Wilcoxon signed-rank p-value after Holm correction.

Evaluation	Comparison	Median AUC difference	r_{rb}	p_{Holm}
Within-subject	LDA – Logistic Regression	0.0025	0.044	0.698
Within-subject	LDA – EEGNet	0.0372	0.776	2.44×10^{-11}
Within-subject	Logistic Regression – EEGNet	0.0362	0.761	4.03×10^{-11}
LOSO	LDA – Logistic Regression	−0.0070	−0.341	0.00267
LOSO	LDA – EEGNet	−0.0549	−0.927	6.40×10^{-16}
LOSO	Logistic Regression – EEGNet	−0.0372	−0.940	3.78×10^{-16}

5.5 EEGNet Training Dynamics and Generalization

The saved EEGNet histories included 1530 grouped within-subject runs and 309 LOSO runs, corresponding to three neural seeds for every valid outer partition. At every training epoch, the logs recorded class-weighted training loss, unweighted validation loss, validation ROC-AUC, validation precision–recall AUC, and validation balanced accuracy. Training-set ROC-AUC was not recorded.

Early stopping generally selected checkpoints before the 60-epoch maximum. Within subject, the median selected checkpoint was epoch 16, the median completed training length was 26 epochs, 94.8% of runs stopped before epoch 60, and only 0.7% selected epoch 60 as the best checkpoint. In LOSO, the corresponding values were epoch 38, 48 completed epochs, 73.5%, and 3.2%. These distributions do not suggest that improved validation performance generally required unrestricted training to the maximum epoch.

Table 5.4 reports run-level performance at the selected checkpoints. These values differ from the headline subject-level macro-averages because every outer run contributes equally, whereas Tables 5.1 and 5.2 first aggregate within subject and then macro-average dataset means.

To make the optimization behaviour explicit, Figure 5.3 shows a representative LOSO training history from the final benchmark. The example was selected from the 309 LOSO EEGNet runs without using outer-test performance. Runs were ranked first by proximity to the overall median selected checkpoint of epoch 38, then by proximity to the median completed training length of 48 epochs, and finally by proximity to the median best validation AUC. The resulting example was `ds005863` subject `021`, seed 44; it selected epoch 38 and completed 48 training epochs.

For the representative run, validation ROC-AUC reached its recorded maximum at epoch 38, matching the selected checkpoint shown in Figure 5.3. Validation loss reached lower values at some other epochs, but checkpoint selection was based on validation ROC-AUC rather than validation loss, as specified in Section 4.9.

Table 5.4: EEGNet training, validation, and outer-test diagnostics. Training loss is class-weighted cross-entropy, whereas validation loss is unweighted cross-entropy, both reported at the selected checkpoint. Validation and test columns report run-level ROC-AUC averages. The validation–test gap is validation AUC minus outer-test AUC.

Setting	Training loss	Validation loss	Validation AUC	Test AUC	AUC gap	Seed SD
Within-subject	0.461	0.540	0.811	0.716	0.095	0.060
LOSO	0.486	0.478	0.819	0.784	0.035	0.015

Because training loss was class-weighted whereas validation loss was unweighted, and because dropout was active during training but disabled during validation, the absolute training and validation loss values are not directly comparable under identical conditions. Their trajectories are therefore used as optimization diagnostics rather than as a quantitative train–validation generalization gap. Validation and outer-test ROC-AUC provide the more relevant generalization comparison.

The mean validation–test AUC gap was 0.095 within subject and 0.035 in LOSO. The within-subject dataset-level run averages were 0.021 for `ds003061`, 0.135 for `ds005863`, 0.079 for `ds006466`, and 0.103 for `ds006480`. The mean seed-to-seed test-AUC SD was also lower in LOSO than within subject, at 0.015 versus 0.060, consistent with the larger multi-subject training pool used in LOSO.

The validation–test gap is not a pure estimate of overfitting because the validation and outer-test partitions differed in group composition and, in LOSO, represented different subjects. Lower outer-test AUC can therefore reflect both validation-selection optimism and genuine partition difficulty or cross-subject distribution shift. The results show that validation-based checkpointing and regularization controlled training while meaningful generalization variability remained; they do not establish that overfitting was entirely absent.

5.6 Gaussian-Noise Augmentation Results

The Gaussian-noise ablation examined whether conservative train-only perturbation improved EEGNet generalization under LOSO evaluation. The principal classifier benchmark remained unchanged. The analysis paired the original non-augmented LOSO results with augmented results using identical held-out subjects and neural seeds.

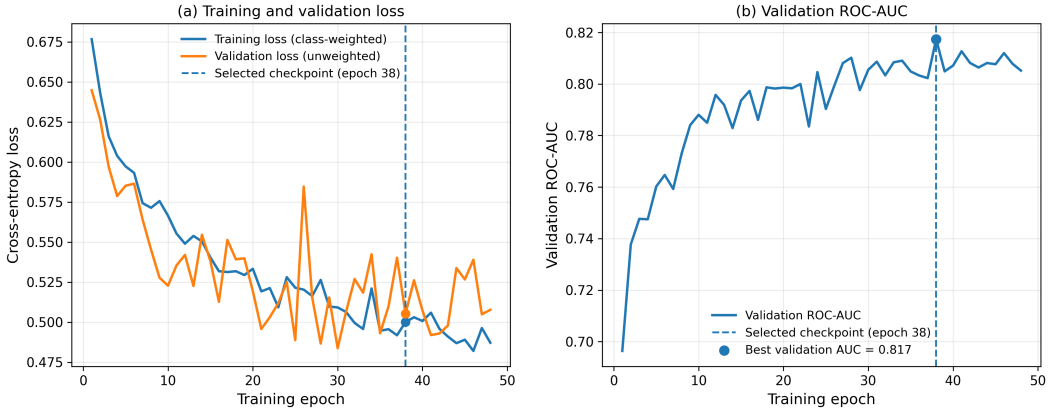

Figure 5.3: Representative EEGNet LOSO training diagnostics for `ds005863` subject `021`, seed 44. Panel (a) shows the class-weighted training-loss trajectory and the unweighted validation-loss trajectory. Panel (b) shows validation ROC-AUC, which was the checkpoint-selection criterion. The dashed vertical line marks the selected checkpoint at epoch 38. The example was selected according to typical LOSO training dynamics and validation performance; outer-test AUC was not used to choose the illustrated run.

Validation-only selection

Table 5.5 reports the validation-only candidate comparison. The baseline mean validation AUC across the 36 pilot partitions was 0.81681. Every nonzero candidate produced a mean validation AUC close to, but slightly below, the baseline. Among the nonzero candidates, $\sigma = 0.01$ produced the highest mean validation AUC and the smallest mean change and was therefore retained for the complete analysis.

The pilot should not be interpreted as demonstrating a validation improvement. It identified $\sigma = 0.01$ as the least disruptive of the three prespecified nonzero candidates.

Table 5.5: Validation-only selection of Gaussian-noise magnitude. Changes are candidate validation AUC minus paired non-augmented validation AUC. Outer-test results were not used during candidate selection.

σ	Paired partitions	Mean validation AUC	Mean change	Proportion improved
0.010	36	0.81677	−0.00004	0.306
0.025	36	0.81643	−0.00038	0.333
0.050	36	0.81561	−0.00120	0.222

Table 5.6: Overall subject-level effect of Gaussian-noise augmentation on EEGNet LOSO performance.

Measure	Value
Held-out subject-recordings	103
Baseline mean ROC-AUC	0.78350 ± 0.11317
Augmented mean ROC-AUC	0.78459 ± 0.11271
Mean augmented-minus-baseline change	0.00109 ± 0.00631
Median change	0.00024
Bootstrap 95% CI for mean change	[−0.00009, 0.00233]
Subjects improved	58 of 103 (56.3%)
Subjects unchanged	2 of 103 (1.9%)
Subjects decreased	43 of 103 (41.7%)
Paired Wilcoxon p-value	0.0720
Matched rank-biserial correlation	0.206

Overall LOSO effect

The complete comparison included all 103 held-out subject-recordings and all three neural seeds. Seed-level results were averaged before subject-level inference.

The pooled subject-level mean AUC increased by approximately 0.0011. The confidence interval contained zero, and the paired Wilcoxon result did not meet the $\alpha = 0.05$ threshold. Gaussian-noise augmentation therefore did not produce a statistically supported overall increase in outer-test AUC.

The baseline mean of 0.7835 in Table 5.6 is the subject-weighted mean over all 103 recordings. It differs from the headline EEGNet LOSO value of 0.7947 because the latter is an unweighted macro-average of the four dataset means.

Dataset-level effects

The mean augmented-minus-baseline AUC change was +0.00165 for `ds003061`, −0.00015 for `ds005863`, +0.00279 for `ds006466`, and +0.00038 for pooled `ds006480`. The corresponding proportions of improved subjects were 69.2%, 40.0%, 73.3%, and 50.0%, respectively. The descriptive change was therefore positive in three datasets and slightly negative in `ds005863`, but all changes were small relative to between-subject variability.

Effects according to subject quality

The lower-quality quartile contained 28 subjects and showed a mean AUC increase from 0.75246 to 0.75491 ($\Delta = 0.00245$). The upper-quality quartile also contained 28 subjects and showed a mean decrease from 0.82579 to 0.82513 ($\Delta = -0.00066$). The lower-quality-group result was not significant after Holm correction ($p_{\text{Holm}} = 0.1477$),

Table 5.7: Subject-level validation–test gap before and after Gaussian-noise augmentation. A positive value indicates that inner-validation AUC exceeded held-out-subject outer-test AUC.

Measure	Value
Baseline mean validation–test gap	0.03505
Augmented mean validation–test gap	0.03377
Mean augmented-minus-baseline gap change	−0.00128
Bootstrap 95% CI	[−0.00260, −0.00004]
Subject-level paired Wilcoxon p-value	0.0344
Matched rank-biserial correlation	−0.240

and the upper-quality-group result was also not significant ($p_{\text{Holm}} = 0.3869$).

The direct hard-versus-easy comparison was inconclusive ($U = 508,\ p = 0.0584$), and the continuous relationship between Subject Quality Index and augmentation gain was not significant ($\rho = -0.157,\ p = 0.113$). The larger descriptive gain among lower-quality subjects therefore does not establish that Gaussian-noise augmentation reliably resolved difficult-subject performance.

Validation–test gap and training behaviour

The validation–test gap was averaged across the three neural seeds before subject-level inference.

Table 5.7 compares the subject-level validation–test gap before and after augmentation.

Gaussian noise reduced the mean validation–test gap from 0.03505 to 0.03377, a mean change of −0.00128. The paired result was statistically significant ($p = 0.0344$), but the change was very small and was not accompanied by a significant increase in outer-test AUC. Training behaviour was nearly unchanged: the median selected checkpoint remained epoch 38, the median completed training length remained 48 epochs, and approximately 73.5% of runs stopped before epoch 60 under both conditions.

The validation–test gap is not a pure overfitting measure because inner-validation observations came from the training-subject population, whereas outer-test observations came from an unseen subject. It therefore reflects both validation-selection optimism and genuine cross-subject distribution shift. The result supports modest regularization but does not identify severe overfitting as EEGNet's principal limitation.

5.7 Interpretation of the Benchmark

The benchmark did not identify one universally strongest classifier. LDA and Logistic Regression performed similarly in grouped within-subject evaluation and both exceeded EEGNet in the pooled paired analysis, whereas EEGNet significantly exceeded both linear models in LOSO. The model ranking therefore depended on whether subject-specific calibration or unseen-subject generalization was evaluated, consistent with broader EEG benchmarking evidence that performance depends on task, representation, sample size, and evaluation design [18, 24].

EEGNet's LOSO advantage suggests that its learned temporal and spatial representation transferred more effectively among subjects within each dataset, although the evaluated models do not establish its superiority over all cross-subject or transfer-learning methods [21, 42]. Substantial between-subject variability nevertheless remained for every classifier. Gaussian-noise augmentation slightly reduced the validation–test gap but produced only a small and statistically inconclusive change in outer-test AUC, so it did not materially alter the benchmark interpretation.

Chapter 6 next examines whether the discriminative information is temporally and physiologically plausible through ERP waveforms, time-resolved decoding, and negative controls. Chapter 7 then relates subject-level AUC to signal characteristics and compares the explanatory contributions of subject quality and classifier identity.

Chapter 6

ERP Sanity Checks and Time-Resolved Decoding

6.1 Purpose and Scope of the Physiological Checks

Classification performance alone does not establish that a model used physiologically plausible information. EEG classifiers may exploit task-related neural responses, but they can also be influenced by artifacts, temporal drift, block structure, recording instability, and other non-neural signal components [6, 29].

This chapter therefore evaluates dataset-level target–standard ERP waveforms, time-resolved grouped decoding, and development-stage label-permutation and prestimulus controls. These analyses are supporting plausibility checks rather than proof of a unique P300 generator or complete removal of non-neural contributions. Because the datasets differed in modality, task, population, channels, reference, and recording system, identical waveform morphology, polarity, and amplitude were not expected [20, 26, 35]. The relevant expectation was that discriminative structure should occur in plausible post-stimulus intervals and should weaken when labels or post-stimulus information were removed.

6.2 Dataset-Level Target–Standard ERP Differences

Dataset-specific centroparietal regions of interest were used because channel layouts differed across datasets. Target and standard trials were averaged within subject, and the standard waveform was subtracted from the target waveform. The principal analysis interval was 0.25–0.50 s after stimulus onset, selected to capture late target-related activity while allowing variation in component latency across subjects and paradigms [20, 26, 35]. For each subject, the latency of the largest absolute target–standard difference in this interval was identified, and Table 6.1 reports the dataset-level mean.

Table 6.1: Mean latency of the largest absolute target–standard ERP difference in the 0.25–0.50 s analysis interval. Values are averages of subject-specific peak latencies.

Dataset	Mean peak latency (s)
`ds003061`	0.430
`ds005863`	0.428
`ds006466`	0.398
`ds006480` pooled	0.414

Mean peak latencies ranged from approximately 0.40 to 0.43 s, consistent with a late target-related difference but not necessarily an identical P3 subtype or polarity across datasets.

Late target–standard differences varied in magnitude, shape, and polarity. The clearest positive differences occurred in `ds003061` and `ds005863`; `ds006466` showed smaller differences, and pooled `ds006480` showed a predominantly negative late difference in the selected region. Polarity and morphology can vary with electrode selection, reference, task, population, overlapping components, and source orientation [6, 26]. The analysis therefore emphasized the existence and timing of separation rather than requiring a uniformly positive textbook waveform.

6.3 Time-Resolved Decoding Method

Time-resolved decoding tested when target–standard information became accessible to a classifier [7]. The analysis used LDA with the scaling and automatic covariance-shrinkage configuration described in Chapter 4 and retained the same grouped within-subject outer folds.

A nominal 50 ms window was advanced in nominal 25 ms steps. At 128 Hz, these settings produced

$$n_{\text{window}} = \text{round}(0.050 \times 128) = 6, \qquad n_{\text{step}} = \text{round}(0.025 \times 128) = 3$$

samples, corresponding to effective durations of

$$\frac{6}{128} = 46.875 \text{ ms}, \qquad \frac{3}{128} = 23.4375 \text{ ms}$$

For each window, the channel-by-time segment had shape

$$X^{(a:b)} \in \mathbb{R}^{N \times C \times 6}$$

and was flattened into one $6C$-dimensional vector per EEG epoch. A separate LDA model was fitted at each window, but every EEG epoch remained one observation; the six temporal samples were features rather than independent examples.

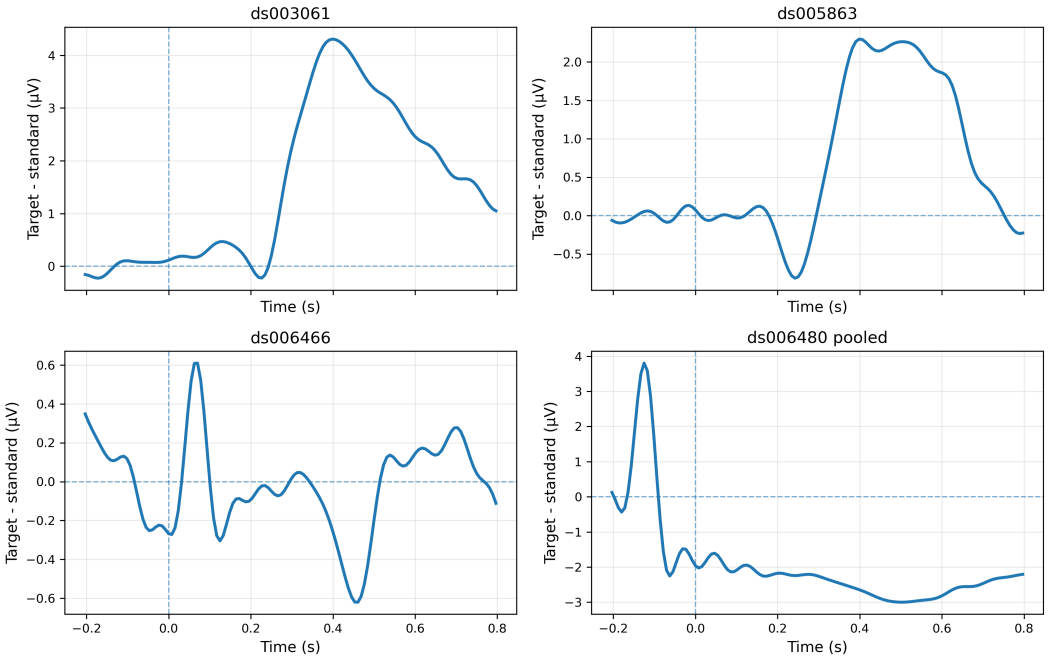

Figure 6.1: Dataset-level target-minus-standard ERP difference waveforms. The dashed vertical line marks stimulus onset, and the dashed horizontal line marks zero difference. The datasets show late post-stimulus differences with varying magnitude, shape, and polarity. Vertical scales differ across panels because the objective is to display the temporal structure within each dataset rather than compare absolute voltage directly across different recording systems and channel configurations.

ROC-AUC was calculated for every subject, outer fold, and temporal window, then averaged across folds within subject. Dataset-level curves summarized these subject-level profiles. Window locations were assigned using the mean time of the six included samples, so the reported peak times are window-centre values.

6.4 Time-Resolved Decoding Results

Table 6.2 reports the maximum dataset-level mean AUC and corresponding window centre.

Peak decoding occurred between approximately 0.44 and 0.49 s in all four datasets, overlapping the late target–standard differences shown in Figure 6.1.

Figure 6.2 shows the complete dataset-level decoding profiles across the analyzed post-stimulus interval.

Decoding was close to chance in the earliest post-stimulus windows and increased as the late event-related response developed. Its temporal peaks approximately

Table 6.2: Peak time-resolved LDA performance by dataset. Peak time is the centre of the six-sample sliding window with the highest dataset-level mean subject AUC.

Dataset	Peak window centre (s)	Peak ROC-AUC
ds003061	0.441	0.7910
ds005863	0.488	0.6837
ds006466	0.441	0.7385
ds006480 pooled	0.488	0.6293

Table 6.3: Development-stage negative-control results from the 10-subject ds003061 benchmark. These values were not produced by the complete final 103-subject benchmark and should not be compared directly with the cross-dataset means in Chapter 5.

Classifier	Permuted-label AUC	Prestimulus-only AUC
LDA	0.5011	0.5144
Logistic Regression	0.4972	0.5151
EEGNet	0.5096	0.5030

matched the waveform peak-latency interval. Because adjacent windows overlapped and could contain several neural and non-neural processes, this analysis identifies when discriminative information was available to LDA rather than a uniquely causal physiological component.

6.5 Development-Stage Negative Controls

Two controls were recorded during an earlier 10-subject ds003061 development-stage benchmark and were not repeated across the final 103-subject analysis. In the label-permutation control, labels were shuffled while preserving the trial data and grouped evaluation procedure. In the prestimulus control, classifier input was restricted to −0.2 to 0 s. These controls tested whether the earlier pipeline retained performance after removal of the true class relationship or post-stimulus information.

Table 6.3 reports the resulting near-chance control performance.

Permutation and prestimulus means were close to the chance value of 0.50. The slightly higher linear-model prestimulus values should not be interpreted as reliable decoding without confidence intervals and formal testing. These controls provide limited evidence that the earlier pipeline depended on the true labels and primarily post-stimulus information, but they do not prove that the final benchmark was free of every confound or that all post-stimulus information was neural. They were negative controls rather than augmentation procedures.

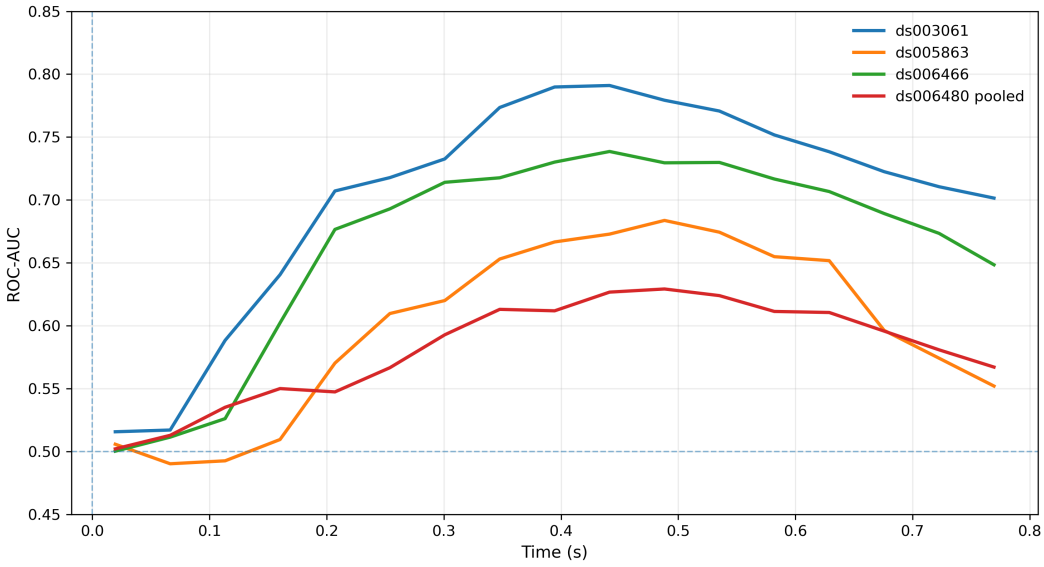

Figure 6.2: Time-resolved grouped within-subject LDA decoding. Each point represents the dataset-level mean of subject-level ROC-AUC values obtained from a six-sample window advanced in three-sample steps. The dashed horizontal line indicates chance-level ROC-AUC of 0.50, and the dashed vertical line marks stimulus onset. Performance increased primarily after stimulus onset and reached its maximum between approximately 0.44 and 0.49 s.

6.6 Physiological Interpretation

The ERP and time-resolved results provide several plausible explanations for the association between signal characteristics and classification performance. A larger target–standard mean difference increases between-class separation, while lower trial-to-trial variability reduces overlap between class distributions. Latency variability can weaken both averaged ERPs and single-trial decoding by temporally smearing the response, and a stable spatial distribution can provide a more consistent pattern for linear classifiers and learned spatial convolutions.

These relationships are compatible with established ERP principles but remain observational. Attention, fatigue, task engagement, age, trial count, target probability, electrode contact, eye movements, muscle activity, and other recording conditions could influence both ERP separability and AUC. A stronger ERP feature should therefore not be interpreted as the sole or necessarily causal explanation of higher classification performance.

The 0.1–12 Hz classifier filter reduced high-frequency muscle activity but could not eliminate slow ocular activity, electrode drift, movement transients, or other low-frequency artifacts [29]. Because the primary benchmark did not use ICA

cleaning, the physiological interpretation is limited to scalp-level target–standard separability.

Overall, benchmark performance was temporally aligned with late post-stimulus differences that are plausible in oddball and ERP-style tasks. The analyses do not uniquely identify the neural generator of the discriminative information or establish a causal relationship between ERP quality and classification performance.

6.7 Chapter Summary

Across datasets, the largest target–standard ERP difference occurred on average near 0.40–0.43 s, while six-sample time-resolved LDA windows reached their highest mean AUC near 0.44–0.49 s. Development-stage permutation and prestimulus controls remained close to chance. Together, these results support a cautious interpretation that useful benchmark information was concentrated in a plausible late post-stimulus interval, while not excluding non-neural contributions or establishing causality. Chapter 7 next examines whether ERP separability, signal variability, spectral characteristics, and trial structure are associated with subject-level decoding performance.

Chapter 7

Subject Classifiability and Signal-Quality Limits

7.1 Motivation and Analysis Scope

The benchmark showed that classifier ranking depended on the evaluation setting, while Chapter 6 located useful discriminative information in plausible late post-stimulus intervals. This chapter therefore examines why some subjects remained relatively easy or difficult across models. The term *subject classifiability* denotes the observed ability of the evaluated classifiers to distinguish a subject's target and standard trials; it is an empirical outcome rather than a fixed physiological trait [1, 15, 37, 40].

The analysis relates subject-level AUC to raw EEG, ERP, spectral, trial-structure, and exploratory source-quality features; constructs and tests a retrospective Subject Quality Index; compares subject-level and classifier-level performance differences; and examines whether EEGNet or training-only Gaussian-noise augmentation produced larger gains for lower-quality subjects.

All analyses are retrospective. Several features require labelled target and standard trials, so the results characterize the collected recordings and do not provide a prospectively validated screening method for a new user without calibration data.

7.2 Subject-Level Susceptibility Analysis

For each evaluation setting, the subject-level classifier results were merged with the features described in Section 3.7. Each row represented one dataset, subject, classifier, aggregated ROC-AUC value, and the corresponding signal and trial-structure summaries. Within subject, the features and AUC described the same participant whose trials were divided across outer folds. In LOSO, the features described the completely held-out participant and therefore characterized the

recording to which models trained on other subjects attempted to generalize.

The analyzed families included baseline and post-stimulus variability, ROI ERP and P300-window differences, single-trial effect sizes, spectral power and band ratios, trial counts and proportions, and exploratory ICA/source-quality measures. Because datasets differed in scale and overall difficulty, features and AUC were standardized within dataset:

$$z_s = \frac{x_s - \mu_d}{\sigma_d}$$

where μ_d and σ_d are the dataset-specific mean and population standard deviation. Pooled associations were measured using Spearman rank correlation.

Multiple-comparison correction

The Benjamini–Hochberg false discovery rate procedure was applied to the tested model–feature combinations [5]. Effects with $q \leq 0.05$ were treated as significant and those with $q \leq 0.10$ as suggestive. Within subject, correction was applied separately within each classifier over the reportable raw/ERP/power features. In LOSO, raw/ERP/power and ICA/source-quality tests formed separate correction families. Reported counts therefore represent significant model–feature combinations rather than the same number of distinct physiological variables.

7.3 Construction of the Subject Quality Index

The Subject Quality Index was designed as a compact summary of task-response strength and signal variability. It was intentionally based on interpretable raw/ERP measures rather than ICA-derived quantities because the raw/ERP features produced the clearest corrected susceptibility results.

The index contained 20 components:

- 12 positively oriented task-response components; and
- 8 negatively oriented noise or variability components.

The exact components are listed in Tables 7.1 and 7.2.

Feature-selection rationale

The 20 components were selected from the available numeric raw/ERP feature table using fixed, interpretable naming and orientation rules. Identifier columns, classifier outcomes, latency variables, columns with fewer than ten numeric values, and nearly constant columns were excluded.

Columns describing P300-window differences, ERP signal-to-noise ratio, peaks, target–standard differences, Cohen's d, and task-window response magnitude were

Table 7.1: Task-response components of the primary Subject Quality Index. Positive orientation means that larger standardized values increased the index.

Component family	Component and implementation variable	Orientation o_k
Task response	ROI P300-window signed mean target–standard difference `roi_p300_mean_diff_uv`	+1
Task response	ROI P300-window absolute mean target–standard difference `roi_p300_abs_mean_diff_uv`	+1
Task response	ROI P300-window RMS target–standard difference `roi_p300_rms_diff_uv`	+1
Task response	ROI P300-window signed peak target–standard difference `roi_p300_peak_signed_uv`	+1
Task response	ROI P300-window absolute peak target–standard difference `roi_p300_peak_abs_uv`	+1
Task response	ROI ERP signal-to-noise ratio `roi_erp_snr_ratio`	+1
Task response	ROI ERP signal-to-noise ratio in decibels `roi_erp_snr_db`	+1
Task response	All-channel P300-window RMS target–standard difference `all_ch_p300_rms_diff_uv`	+1
Task response	P300-window epoch RMS amplitude `p300_epoch_rms_uv`	+1
Task response	ROI P300-window standard deviation `roi_p300_std_uv`	+1

assigned positive orientation. Columns describing general baseline or post-stimulus RMS, standard deviation, or noise were assigned negative orientation. Two task-window components, `p300_epoch_rms_uv` and `roi_p300_std_uv`, were assigned positive orientation because they were intended to summarize response magnitude within the P300 analysis window rather than general recording noise. Because larger values for these two components may also reflect increased variability, the leave-one-component-out, median-aggregation, and equal-family-weighting analyses in Section 7.4 were used to assess whether the main findings depended strongly on these orientation choices. Classifier AUC was not used to optimize component weights or to select a subset that maximized association with performance.

This procedure reduced direct outcome-based tuning, but the index remains retrospective. The component definitions were derived from labelled target and standard trials, and their interpretation was informed by the intended meaning of the feature names.

Table 7.2: Noise and variability components of the primary Subject Quality Index. Negative orientation means that the standardized value was multiplied by −1, so that larger original values decreased the index.

Component family	Component and implementation variable	Orientation o_k
Task response	Signed single-trial ROI P300 Cohen's d `trial_roi_p300_cohen_d`	+1
Task response	Absolute single-trial ROI P300 Cohen's d `trial_roi_p300_abs_cohen_d`	+1
Noise/variability	ROI baseline RMS target–standard difference `roi_baseline_rms_diff_uv`	−1
Noise/variability	All-channel baseline RMS target–standard difference `all_ch_baseline_rms_diff_uv`	−1
Noise/variability	Baseline RMS amplitude `baseline_rms_uv`	−1
Noise/variability	Post-stimulus RMS amplitude `poststim_rms_uv`	−1
Noise/variability	Baseline standard deviation `baseline_std_uv`	−1
Noise/variability	Post-stimulus standard deviation `poststim_std_uv`	−1
Noise/variability	ROI baseline standard deviation `roi_baseline_std_uv`	−1
Noise/variability	Median single-trial ROI baseline standard deviation `trial_roi_baseline_std_median_uv`	−1

Within-dataset normalization

Let $x_{s,k}$ denote component k for subject s, and let $d(s)$ identify the dataset containing that subject. Each component was standardized within its dataset:

$$z_{s,k} = \frac{x_{s,k} - \mu_{d(s),k}}{\sigma_{d(s),k}}$$

The population standard deviation was used in the implementation. No clipping or winsorization was applied.

An orientation coefficient $o_k \in \{-1, +1\}$ was then applied:

$$\widetilde{z}_{s,k} = o_k z_{s,k}$$

After orientation, a larger value consistently represented better expected classifiability according to the index definition.

Aggregation and weighting

For subject s, the primary index was the arithmetic mean of the available oriented components:

$$Q_s = \frac{1}{K_s} \sum_{k \in \mathcal{K}_s} \widetilde{z}_{s,k}$$

where $\mathcal{K}_s$ is the set of non-missing components for subject s, and $K_s = |\mathcal{K}_s|$.

All 103 subjects had all 20 components in the final index table, so $K_s = 20$ for every subject in the reported analysis.

Every component received equal weight. Because the index contained 12 task-response components and eight noise components, equal component weighting implicitly assigned: $\frac{12}{20} = 0.60$ of the aggregate weight to the task-response family and $\frac{8}{20} = 0.40$ to the noise family. This 60/40 family balance was a consequence of the number of selected components and was not separately optimized.

The resulting index is relative within the four analyzed datasets. A value of zero does not represent an absolute quality threshold. Positive and negative values indicate positions above or below the dataset-specific component means after aggregation.

7.4 Sensitivity of the Subject Quality Index

Three sensitivity analyses were used to test whether the main conclusions depended strongly on one component or one aggregation rule.

Leave-one-component-out analysis

The index was recomputed 20 times, omitting one component at a time. The Spearman correlation between each reduced index and the primary 20-component index ranged from

$$0.9847 \quad \text{to} \quad 0.9951.$$

Thus, no single component determined the overall subject ordering.

Median aggregation

A robust alternative index was computed as the median of the 20 oriented standardized components:

$$Q_s^{\text{median}} = \text{median}_k \left(\widetilde{z}_{s,k} \right).$$

Its Spearman correlation with the primary mean-based index was $\rho = 0.9173$.

Table 7.3: Sensitivity of the Subject Quality Index. Index correlations are Spearman correlations with the primary 20-component mean index. Incremental R^2 is the additional in-sample variance explained after dataset identity was included.

Index construction	Correlation with primary	Within-subject ΔR^2	LOSO ΔR^2
Primary component mean	1.0000	0.0704	0.1222
Component median	0.9173	0.0671	0.0983
Equal-family weighting	0.9497	0.0899	0.1307

Equal-family weighting

To remove the implicit 60/40 family weighting, the 12 positively oriented task-response components were averaged separately from the eight reversed noise components. The two family means then received equal weight:

$$Q_s^{\text{family}} = \frac{1}{2}\left[\frac{1}{12}\sum_{k\in\mathcal{P}}\widetilde{z}_{s,k} + \frac{1}{8}\sum_{k\in\mathcal{N}}\widetilde{z}_{s,k}\right]$$

where $\mathcal{P}$ and $\mathcal{N}$ denote the task-response and noise component sets.

The equal-family index had a Spearman correlation of $\rho = 0.9497$ with the primary index.

Table 7.3 shows how the alternative constructions affected the incremental R^2 results.

For comparison, classifier identity added 0.0439 after dataset in the within-subject analysis and 0.0577 in LOSO. Subject quality therefore remained more explanatory than classifier identity under both alternative index constructions.

These sensitivity analyses support the stability of the main result, but they do not establish that the selected index is uniquely optimal. Other feature definitions, physiological measurements, weighting methods, or prospective quality measures could produce different results.

7.5 Susceptibility Results

In grouped within-subject evaluation, the dataset-standardized raw/ERP/power analysis produced 33 model–feature effects at $q \leq 0.05$ and 64 at $q \leq 0.10$. In LOSO, source-specific correction produced 43 effects at $q \leq 0.05$ and 57 at $q \leq 0.10$.

Across both settings, the clearest associations involved ERP/P300 separability and signal variability. Larger target–standard differences, stronger single-trial effects, and higher ERP signal-to-noise summaries were generally associated with higher AUC, whereas greater baseline or post-stimulus variability was generally associated with

Table 7.4: Dataset-standardized Spearman associations between the primary Subject Quality Index and subject-level AUC. The reported p-values are two-sided and unadjusted; these six composite-index correlations are presented separately from the mass-univariate FDR analysis.

	LDA		Logistic Regression		EEGNet	
Setting	ρ	p	ρ	p	ρ	p
Within-subject	0.3305	0.0007	0.3451	0.0004	0.3158	0.0012
LOSO	0.3529	0.0003	0.3800	0.0001	0.3547	0.0002

lower AUC. These directions are consistent with stronger between-class separation and lower within-class overlap [6, 26].

ICA/source-quality effects were substantially weaker. None survived correction at $q \leq 0.05$; six were suggestive within subject and five in LOSO at $q \leq 0.10$. They are therefore treated as exploratory rather than as the principal explanation of classifiability.

The LOSO associations remain observational. Held-out subjects may differ from the training population in age, attention, anatomy, trial count, response latency, electrode quality, artifacts, or recording conditions. The analysis shows that measurable properties of a held-out recording were associated with cross-subject generalization difficulty, but it does not isolate one causal mechanism [22, 42].

7.6 Association Between the Subject Quality Index and AUC

The primary Subject Quality Index was positively associated with AUC for all three classifiers in both evaluation settings after the index and AUC were standardized within dataset. Table 7.4 reports the model-specific Spearman correlations.

The moderate correlations indicate that the index captured meaningful subject-level variation without fully explaining physiological, behavioural, and dataset-specific influences.

Figure 7.1 visualizes these relationships after within-dataset standardization.

7.7 Subject-Level Versus Classifier-Level Differences

Three complementary analyses were used to compare subject-level and classifier-level contributions:

1. the relative spread of subject and classifier mean AUC values;
2. the consistency of subject difficulty across classifiers; and

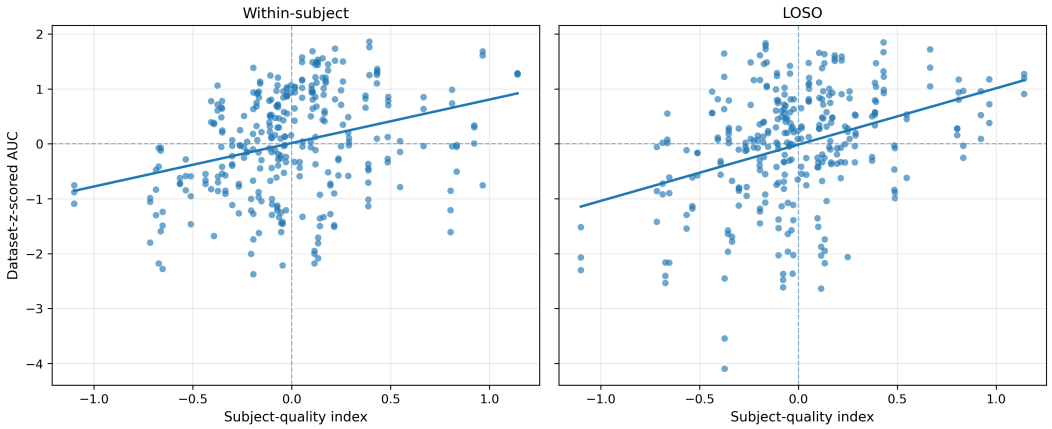

Figure 7.1: Relationship between the retrospective Subject Quality Index and subject-level AUC. Both variables are standardized within dataset. Each subject contributes one point for each classifier, producing three observations per subject in each evaluation setting. The fitted lines are descriptive; the model-specific Spearman analyses in Table 7.4 were calculated separately for LDA, Logistic Regression, and EEGNet.

Table 7.5: Subject-level and classifier-level spread in dataset-standardized AUC. The final column is the ratio of the subject-level SD to the classifier-level SD. It is a standard-deviation ratio, not a variance ratio.

Evaluation setting	Subject-mean AUC SD	Classifier-mean AUC SD	SD ratio
Within-subject	0.9058	0.3176	2.85
LOSO	0.9257	0.3269	2.83

3. incremental R^2 after accounting for dataset.

These analyses address different aspects of the same question and should not be interpreted as interchangeable causal estimates.

Standard-deviation comparison

AUC was first standardized within dataset. For each subject, the standardized AUC was averaged across the three classifiers. The standard deviation of these 103 subject means measured the spread of average subject performance.

For each classifier, standardized AUC was averaged across all 103 subjects. The standard deviation of the three classifier means measured the spread among the evaluated classifiers.

Table 7.5 compares the resulting subject-level and classifier-level spreads.

Table 7.6: Cross-classifier consistency of subject-level AUC rankings. Summary values are calculated across the 12 dataset-specific classifier-pair Spearman correlations in each evaluation setting.

Setting	Number of correlations	Mean ρ	Median ρ	Range
Within-subject	12	0.8363	0.9006	0.5488–0.9945
LOSO	12	0.8811	0.9430	0.6053–0.9835

Subject-level spread was approximately 2.8 times classifier-level spread in both settings. This is a standard-deviation comparison for the three evaluated classifiers and four datasets, not a universal measure of subject importance or a variance ratio.

Consistency of subject difficulty across classifiers

Subject-level AUC rankings were compared for the three classifier pairs within each of the four datasets, producing 12 Spearman correlations per evaluation setting.

Table 7.6 summarizes the cross-classifier rank consistency.

All correlations were positive, with mean values of 0.8363 within subject and 0.8811 in LOSO. Thus, model choice changed average performance but largely preserved the relative ordering of easy and difficult subjects.

Incremental R^2

A descriptive ordinary least-squares analysis compared the additional variance explained by classifier identity and the Subject Quality Index after dataset identity had been included.

Let: Y = subject-level AUC, D = dataset identity, C = classifier identity, and Q = Subject Quality Index. The classifier increment was

$$\Delta R^2_{\text{classifier}} = R^2(Y \sim D + C) - R^2(Y \sim D)$$

and the subject-quality increment was

$$\Delta R^2_{\text{quality}} = R^2(Y \sim D + Q) - R^2(Y \sim D).$$

The regression table contained 309 rows per evaluation setting: 103 subjects $\times$ 3 classifiers.

Table 7.7 compares the incremental in-sample contributions of classifier identity and subject quality.

The Subject Quality Index added more in-sample R^2 than classifier identity in both settings: approximately 1.6 times as much within subject and 2.1 times as much in LOSO. These descriptive increments are not causal estimates, and the subject-grouped cross-validation analysis below provides the more conservative generalization check.

Table 7.7: Incremental in-sample R^2 after accounting for dataset identity.

Evaluation setting	Classifier increment	Subject-quality increment	Quality/classifier ratio
Within-subject	0.0439	0.0704	1.60
LOSO	0.0577	0.1222	2.12

Table 7.8: Five-fold subject-grouped cross-validated R^2 for models using dataset, classifier identity, and the Subject Quality Index.

Evaluation setting	Dataset + classifier	Dataset + quality	Dataset + classifier + quality
Within-subject	0.4144	0.4394	0.4826
LOSO	0.2464	0.3175	0.3750

Subject-grouped cross-validation check

As a sensitivity analysis, five-fold cross-validation was performed with complete subject identifiers assigned to folds. The three classifier rows belonging to one subject therefore remained together in either the training or test partition.

Table 7.8 reports the subject-grouped cross-validated results.

Dataset plus subject quality produced higher cross-validated R^2 than dataset plus classifier identity in both settings, while combining classifier identity and subject quality produced the highest values. Subject quality therefore accounted for more variation than classifier identity alone, but classifier choice still contributed useful information.

7.8 Does EEGNet Benefit Difficult Subjects?

EEGNet gain for subject s was defined relative to the better linear classifier:

$$G_s = \mathrm{AUC}_{s,\mathrm{EEGNet}} - \max\left(\mathrm{AUC}_{s,\mathrm{LDA}}, \mathrm{AUC}_{s,\mathrm{LogReg}}\right)$$

Easy and hard groups were defined separately within dataset using the upper and lower Subject Quality Index quartiles. Pooling these quartiles produced 28 easy and 28 hard subject-recordings per evaluation setting; the middle 50% were excluded from the grouped comparison.

Table 7.9 compares EEGNet gain between the two subject-quality groups.

Within subject, EEGNet gain was negative in both quality groups, and the groups did not differ significantly. In LOSO, gain was positive in both groups and descriptively larger in the hard group; the exploratory Mann–Whitney comparison gave $p = 0.0413$.

Table 7.9: EEGNet gain relative to the better linear classifier. Values are mean ± subject-level SD. The p-value is from a two-sided Mann–Whitney U comparison between the easy and hard groups.

Setting	n per group	Easy gain	Hard gain	p
Within-subject	28	-0.0601 ± 0.0542	-0.0541 ± 0.0709	0.5718
LOSO	28	0.0302 ± 0.0293	0.0499 ± 0.0406	0.0413

This result should be interpreted cautiously because the groups were based on the retrospective, label-dependent index, the comparison was not adjusted for testing both evaluation settings, and quartile categorization discarded continuous information. The continuous quality–gain correlations were not significant within subject ($\rho = 0.052,\ p = 0.599$) or in LOSO ($\rho = -0.114,\ p = 0.250$). The supported conclusion is therefore that EEGNet improved LOSO performance on average, with a larger exploratory mean gain in the lower-quality quartile, rather than a demonstrated monotonic rescue of difficult subjects.

7.9 Can Training Augmentation Reduce Difficult-Subject Effects?

As reported in Section 5.6, the lower-quality quartile showed a larger descriptive augmentation gain than the upper-quality quartile (0.00245 versus −0.00066). However, the lower-quality-group result was not significant after Holm correction ($p_{\text{Holm}} = 0.1477$), the direct hard-versus-easy comparison did not reach the conventional significance threshold ($p = 0.0584$), and the continuous relationship between Subject Quality Index and augmentation gain was not significant ($\rho = -0.157,\ p = 0.113$). Gaussian-noise augmentation therefore did not reliably resolve difficult-subject performance.

7.10 Interpretation, Limitations, and Chapter Summary

The subject-level analyses produced a consistent pattern. Raw EEG and ERP/power features yielded 33 corrected model–feature effects within subject and 43 source-specific corrected effects in LOSO at $q \leq 0.05$, while ICA/source-quality effects remained exploratory. The 20-component Subject Quality Index was positively associated with AUC for all classifiers, and its main conclusions were stable under component removal and alternative aggregation and weighting rules.

Subject-level mean AUC showed approximately 2.8 times the standard deviation of classifier-level mean AUC, and subject rankings were highly consistent across

classifiers. The Subject Quality Index added more in-sample and grouped cross-validated explanatory value than classifier identity after dataset was considered, although combining both predictors produced the strongest cross-validated models. Classifier choice therefore remained important, particularly because EEGNet significantly improved LOSO performance.

The EEGNet-gain and augmentation analyses did not demonstrate that more sophisticated modeling reliably rescued lower-quality subjects. EEGNet produced positive average LOSO gains in both quartiles, with a larger exploratory gain in the hard group, but the continuous quality–gain association was not significant. Gaussian-noise augmentation likewise showed only a small and statistically inconclusive quality-dependent benefit.

These findings remain associative. The index is retrospective, label-dependent, and composed of correlated measurements rather than independent physiological mechanisms. Subject-level AUC and signal characteristics may both be influenced by attention, behaviour, trial count, electrode quality, artifacts, task design, participant population, or recording hardware. The conclusions are also restricted to four oddball or ERP-style datasets and three classifiers.

Overall, measurable subject-level signal characteristics accounted for substantial and consistent variation in decoding success, while classifier identity contributed additional information, especially under LOSO. Chapter 8 discusses the implications, dataset heterogeneity, limitations, practical relevance, and future validation requirements.

Chapter 8

Discussion and Conclusion

8.1 Summary and Interpretation of the Main Findings

This thesis examined EEG stimulus classification across four oddball or event-related potential datasets and 103 subject-recordings. Linear Discriminant Analysis, Logistic Regression, and EEGNet were evaluated using grouped within-subject and leave-one-subject-out (LOSO) designs.

The results support four principal conclusions. First, classifier ranking depended on the evaluation setting. LDA and Logistic Regression performed similarly within subject and both significantly exceeded EEGNet in the pooled paired comparison, whereas EEGNet significantly exceeded both linear models in LOSO. Classifier selection therefore remained important, particularly when generalization to an unseen subject was required.

Second, large differences among subjects remained under every classifier. After AUC was standardized within dataset, subject-level mean AUC had approximately 2.8 times the standard deviation of classifier-level mean AUC in both evaluation settings. Subject rankings were also highly consistent across classifiers, with mean cross-classifier Spearman correlations of 0.8363 within subject and 0.8811 in LOSO.

Third, retrospective raw EEG, ERP, spectral, and trial-structure features were associated with subject-level decoding performance. The strongest corrected effects involved target–standard separability, single-trial effect size, ERP signal-to-noise summaries, and signal variability. The Subject Quality Index was positively associated with AUC under all three classifiers and explained more additional variation than classifier identity after dataset was included. Its conclusions were stable under component-removal and alternative weighting analyses.

Fourth, the Gaussian-noise ablation produced only a modest regularization effect. The mean outer-test AUC change was 0.0011 and was not statistically significant ($p = 0.072$). The validation–test gap decreased slightly, but augmentation did not reliably resolve lower-quality-subject performance.

The central interpretation is therefore balanced: classifier identity affected

performance, especially under LOSO, but it did not fully account for the substantial and consistent variation among subjects.

8.2 Subject Classifiability, Classifier Choice, and Generalization

The main contribution of the thesis is not a new classification algorithm. It is an integrated cross-dataset analysis of subject classifiability that combines classifier benchmarking, interpretable signal characteristics, cross-classifier consistency, and relative explanatory comparisons.

Previous BCI research has established substantial variation among users and has related performance to physiological, behavioural, methodological, and recording factors [1, 15, 37, 40]. The present work extends this perspective to oddball and ERP-style stimulus classification by treating subject-level ROC-AUC as a continuous outcome and asking whether the same recordings remain relatively easy or difficult across models.

The high cross-classifier correlations show that classifier changes altered mean performance without completely changing the subject ranking. Stronger and more consistent target–standard differences were generally associated with higher AUC, whereas greater baseline and post-stimulus variability was generally associated with lower AUC. These findings suggest that the evaluated models encountered a shared difficulty associated with the information available in each recording.

This conclusion does not make classifier selection irrelevant. The linear models operated on temporally averaged ERP-oriented features and were effective when subject-specific calibration data were available, consistent with the continued competitiveness of linear EEG classifiers [24, 25]. EEGNet's learned temporal and spatial representation transferred more successfully among subjects within each dataset and produced the strongest LOSO performance [21].

The comparison was nevertheless restricted to three classifiers. It does not establish that EEGNet is superior to every cross-subject method or that subject quality would explain more variation than a broader collection of classical, convolutional, recurrent, adaptation-based, or Transformer models.

The augmentation analysis further indicates that the observed LOSO difficulty was not explained solely by uncontrolled late-epoch fitting. EEGNet used train-only normalization, dropout, weight decay, gradient clipping, validation-based checkpoint selection, and early stopping. Gaussian noise slightly reduced the validation–test gap but did not materially improve outer-test AUC. The gap therefore likely reflected a combination of validation-selection optimism, cross-subject distribution shift, and subject-level recording difficulty rather than overfitting alone.

The Subject Quality Index should also be interpreted within its intended scope. It is retrospective, depends partly on labelled target and standard trials, and

summarizes overlapping task-response and variability measures. It is an analytical description of the collected recordings, not a universal definition of EEG quality or a prospectively validated screening instrument.

8.3 Physiological Interpretation and Dataset Heterogeneity

The subject-level findings are physiologically plausible because classification depends on both between-class separation and within-class consistency. A larger and more repeatable target-related response increases separation between the target and standard distributions. Baseline noise, response-latency variation, trial-to-trial amplitude variation, and unstable spatial patterns increase overlap and may reduce performance even when the averaged ERP contains a visible difference.

The ERP and time-resolved analyses support this interpretation. The strongest time-resolved decoding occurred approximately 0.44–0.49 s after stimulus onset, overlapping the late target–standard waveform differences. This timing is compatible with task-related ERP/P3 processing in oddball paradigms [20, 26, 35].

The results should not be interpreted as showing identical canonical P300 morphology across datasets. Waveform shape, polarity, amplitude, and latency can vary with stimulus modality, task condition, reference, electrode location, participant population, and overlapping neural components. Attention, fatigue, age, trial count, electrode contact, ocular activity, muscle activity, and slow recording drift may also influence both measured signal characteristics and AUC.

The classifier stream was low-pass filtered at 12 Hz, which reduced high-frequency muscle contamination but did not remove all low-frequency ocular, movement, electrode, or muscle-related activity [29]. No ICA cleaning was used in the principal final benchmark. The negative controls provide supporting evidence that the development-stage pipeline depended mainly on the relationship between post-stimulus EEG and the labels, but they do not prove that every discriminative feature was neural.

ICA/source-quality features were retained because the project originated partly from a source-separation motivation and because ICA has an established role in EEG decomposition [4, 10, 16, 27]. However, no ICA effect survived the principal $q \leq 0.05$ correction. The available component summaries were therefore less informative than direct scalp-level ERP and variability measures and remain exploratory.

Using four datasets allowed replication across related EEG contexts but also introduced substantial heterogeneity. The datasets differed in modality, task conditions, participant populations, hardware, channel layout, marker conventions, trial counts, and overall difficulty. Within-dataset standardization, dataset covariates, dataset-specific reporting, and macro-averaged headline results reduced some of these differences, but they could not remove every interaction among protocol, population, acquisition, and signal morphology.

The study should consequently be described as cross-dataset replication rather

than cross-dataset transfer. LOSO models were trained and tested within the same dataset. Generalization to an independent acquisition site, recording system, population, or task remains untested.

8.4 Practical Implications for EEG and BCI Systems

The findings suggest that classifier selection should be considered together with acquisition quality and subject classifiability. When subject-specific calibration data are available, a linear model may provide strong performance with relatively low complexity. When transfer to an unseen subject is required, a learned cross-subject representation may be more appropriate.

Poor performance should not automatically lead to greater model complexity. Weak target–standard separation or high variability may instead motivate electrode checks, additional trials, shorter blocks, breaks, changes in task pacing or salience, individualized preprocessing, or greater uncertainty in the system output.

The current Subject Quality Index cannot yet guide such decisions prospectively because it requires labelled calibration data. Its appropriate present use is to motivate future quality-aware systems rather than to exclude difficult users. Quality analysis should identify where acquisition, calibration, preprocessing, task design, or model selection should adapt.

Table 8.1 summarizes practical responses that future systems could evaluate.

These responses require prospective evaluation. The thesis establishes the analytical motivation for quality-aware adaptation but does not evaluate a real-time adaptive BCI.

8.5 Limitations

Dataset scope and external validity

The analysis included four oddball or ERP-style datasets, which remains a limited sample of paradigms, populations, recording systems, and laboratories. There was no fifth independent dataset reserved exclusively for confirmation of the complete subject-quality framework. The findings therefore demonstrate replication within the analyzed collection rather than independent external validation or generalization to motor imagery, sleep, seizure detection, affective EEG, continuous decoding, or other EEG tasks.

LOSO evaluation measured unseen-subject generalization within each dataset, not transfer across acquisition sites or protocols. Dataset contributions were also unequal in subject count, trial count, channel count, and condition structure. Within-dataset standardization and dataset covariates reduced but did not eliminate these differences.

Table 8.1: Examples of practical responses to observed subject-level difficulty. These recommendations are implications for future system design and were not experimentally compared in the present thesis.

Observed issue	Possible system response	Purpose
High baseline variability	Check electrode contact, movement, and environmental noise; repeat affected blocks	Improve acquisition reliability
Weak target–standard separation	Collect additional labelled trials or modify task timing and salience	Improve task-related information
Unstable response latency	Use broader temporal representations or latency-adaptive features	Reduce sensitivity to temporal misalignment
Strong calibrated linear performance	Prefer a simpler subject-specific model when operational requirements permit	Reduce complexity and improve interpretability
Poor unseen-subject generalization	Use cross-subject pretraining, adaptation, or limited subject-specific fine-tuning	Reduce distribution mismatch
Low-quality or uncertain calibration	Report greater uncertainty and request additional data rather than producing overconfident output	Support safer decision-making

Retrospective analysis and possible confounding

The Subject Quality Index includes target–standard differences and single-trial effect sizes and therefore requires labelled trials. It was not prospectively tested as a screening, adaptive-calibration, or early-stopping measure. Its 20 components are correlated summaries of overlapping signal properties rather than independent physiological mechanisms.

The public datasets did not provide a common set of attention, fatigue, medication, sleep, impedance, behavioural, and artifact variables. Associations between signal characteristics and AUC may therefore reflect neural response strength, participant state, recording quality, trial structure, or combinations of these factors. The analyses are associative and do not establish that directly modifying one measured feature would cause AUC to improve.

Classifier and configuration scope

The benchmark included only LDA, Logistic Regression, and EEGNet. It did not include SVMs, ensemble methods, larger convolutional networks, recurrent models, domain-adaptation methods, Transformers, or pretrained EEG foundation models.

The relative subject-versus-classifier comparison depends on the spread among the three selected models and may change with a broader classifier set.

The final preprocessing and representation choices were informed by preliminary work on the same general dataset collection. Although outer tests remained separate from direct fitting and checkpoint selection, every preprocessing and architecture choice was not selected within a fully nested outer procedure using an independent development dataset.

Physiological and control limitations

ERP and time-resolved decoding supported plausible late post-stimulus discrimination but did not uniquely identify its neural source. No source localization, component labelling, electromyography, eye tracking, or simultaneous artifact measurement was available. The permutation and prestimulus controls were performed only during the earlier `ds003061` development stage and were not repeated across all four datasets.

Augmentation and difficult-subject analysis

The augmentation study examined one conservative Gaussian-noise transformation after validation-only selection among three nonzero magnitudes. Other perturbations could produce different results. The easy/hard comparisons used dataset-specific quartiles, excluded the middle 50%, and were secondary analyses. Their descriptive tendencies were not supported by significant continuous quality–gain associations and require independent replication.

8.6 Future Work

Prospective, external, and longitudinal validation

The most important next step is to estimate subject quality from an initial calibration segment and evaluate it against later untouched trials. A prospective study should compare the current label-dependent index with reduced label-independent recording-quality features and with behavioural, demographic, and acquisition metadata.

Complete datasets or acquisition sites should also be reserved for external testing. Such work would require channel harmonization, reference alignment, uncertainty estimation, and explicit treatment of distribution shift. Repeated-session data could further separate relatively stable subject differences from temporary effects of attention, fatigue, impedance, and environmental conditions.

Adaptive calibration and broader robustness methods

The practical responses in Section 8.4 should be evaluated experimentally. An early quality estimate could trigger additional trials, electrode adjustment, breaks, alternative task timing, individualized preprocessing, or different model families. The important outcome would be whether acting on the estimate improves later performance, not merely whether quality correlates retrospectively with AUC.

Future robustness studies should compare Gaussian noise with temporal jitter, channel or frequency masking, mixup, trial interpolation, and physiologically motivated perturbations. Candidate transformations should be selected inside nested training validation and evaluated on untouched subjects or independent datasets. Average AUC, between-subject variability, calibration, and lower-quality-subject performance should be analyzed separately.

Broader architectures and representation learning

Transformer and convolution–attention models may capture longer-range temporal and cross-channel relationships. EEG Conformer is one example combining local convolutional extraction with self-attention [39]. Pretrained models such as BIOT, LaBraM, and EEGPT seek transferable representations across heterogeneous biosignal or EEG collections [19, 43, 44].

Greater model scale does not guarantee elimination of subject variability, and specialist models may remain competitive under current EEG data regimes [23]. Future comparisons should therefore measure not only average AUC but also subject-ranking consistency, difficult-subject performance, calibration requirements, uncertainty, and external-dataset generalization.

The original source-separation motivation also remains relevant. Future ICA or learned representation methods should be assessed through independent outer-test classification, preservation of ERP structure, artifact robustness, stability across datasets, and improvement among recordings with weak baseline classifiability rather than through reconstruction quality alone.

8.7 Conclusion

This thesis examined EEG stimulus classification as both a classifier-selection and subject-classifiability problem. Across four oddball or ERP-style datasets, linear models were strongest in grouped within-subject evaluation, whereas EEGNet was strongest in LOSO. Classifier identity therefore influenced performance, especially for unseen-subject generalization.

Subject-level differences were nevertheless larger and more consistent than the mean differences among the three classifiers. Subjects maintained similar rankings across models, and retrospective ERP-separability and variability measures were

associated with AUC. The Subject Quality Index explained more additional variation than classifier identity after dataset was included, although both contributed to the strongest combined models.

These results do not establish that subject quality is fixed or unmodifiable. The measured differences may reflect neural responses, attention, task engagement, acquisition quality, trial count, protocol, and other factors. The index is retrospective and requires labelled data, and Gaussian-noise augmentation did not materially remove held-out-subject variability.

The central conclusion is therefore:

> Within the evaluated oddball and ERP-style datasets, classifier choice influenced performance, especially for unseen-subject generalization, but subject-level task-response separability and signal variability accounted for substantial and consistent differences in decoding success.

Improving EEG and BCI performance should consequently involve both better models and better characterization of the signal and recording conditions under which those models operate.

Bibliography

[1] M. Ahn, H. Cho, S. Ahn, and S. C. Jun, *High Theta and Low Alpha Powers May Be Indicative of BCI-Illiteracy in Motor Imagery*, PLOS ONE, 8 (2013), p. e80886.

[2] H. Altaheri, G. Muhammad, M. Alsulaiman, S. U. Amin, G. A. Altuwaijri, W. Abdul, M. A. Bencherif, and M. Faisal, *Deep Learning Techniques for Classification of Electroencephalogram Motor Imagery Signals: A Review*, Neural Computing and Applications, 35 (2023), pp. 14681–14722.

[3] S. Appelhoff, M. Sanderson, T. Brooks, M. v. Vliet, R. Quentin, C. Holdgraf, M. Chaumon, E. Mikulan, K. Tavabi, R. Höchenberger, D. Welke, C. Brunner, A. Rockhill, E. Larson, A. Gramfort, and M. Jas, *MNE-BIDS: Organizing Electrophysiological Data into the BIDS Format and Facilitating Their Analysis*, Journal of Open Source Software, 4 (2019), p. 1896.

[4] A. J. Bell and T. J. Sejnowski, *An Information-Maximization Approach to Blind Separation and Blind Deconvolution*, Neural Computation, 7 (1995), pp. 1129–1159.

[5] Y. Benjamini and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing*, Journal of the Royal Statistical Society: Series B, 57 (1995), pp. 289–300.

[6] P. E. Clayson, S. A. Baldwin, H. A. Rocha, and M. J. Larson, *The Data-Processing Multiverse of Event-Related Potentials: A Roadmap for the Optimization and Standardization of ERP Processing and Reduction Pipelines*, NeuroImage, 245 (2021), p. 118712.

[7] M. X. Cohen, *Analyzing Neural Time Series Data: Theory and Practice*, MIT Press, Cambridge, MA, 2014.

[8] A. Craik, Y. He, and J. L. Contreras-Vidal, *Deep Learning for Electroencephalogram Classification Tasks: A Review*, Journal of Neural Engineering, 16 (2019), p. 031001.

[9] A. Delorme, *EEG Data from an Auditory Oddball Task*, 2022.

[10] A. Delorme and S. Makeig, *EEGLAB: An Open Source Toolbox for Analysis of Single-Trial EEG Dynamics Including Independent Component Analysis*, Journal of Neuroscience Methods, 134 (2004), pp. 9–21.

[11] K. J. Gorgolewski, T. Auer, V. D. Calhoun, R. C. Craddock, S. Das, E. P. Duff, G. Flandin, S. S. Ghosh, T. Glatard, Y. O. Halchenko, D. A. Handwerker, M. Hanke, D. Keator, X. Li, Z. Michael, C. Maumet, B. N. Nichols, T. E. Nichols, J. Pellman, J.-B. Poline, A. Rokem, G. Schaefer, V. Sochat, W. Triplett, J. A. Turner, G. Varoquaux, and R. A. Poldrack, *The Brain Imaging Data Structure, a Format for Organizing and Describing Outputs of Neuroimaging Experiments*, Scientific Data, 3 (2016), p. 160044.

[12] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. S. Hämäläinen, *MEG and EEG Data Analysis with MNE-Python*, Frontiers in Neuroscience, 7 (2013), p. 267.

[13] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, *Array Programming with NumPy*, Nature, 585 (2020), pp. 357–362.

[14] K. M. Hossain, M. R. Islam, S. Hossain, A. Nijholt, and M. A. R. Ahad, *Status of Deep Learning for EEG-Based Brain–Computer Interface Applications*, Frontiers in Computational Neuroscience, 16 (2023), p. 1006763.

[15] G. Huang, P. Xu, and D. Zhang, *Discrepancy Between Inter- and Intra-subject Variability in EEG-Based Motor Imagery Brain–Computer Interface*, Frontiers in Neuroscience, 17 (2023), p. 1122661.

[16] A. Hyvärinen and E. Oja, *Independent Component Analysis: Algorithms and Applications*, Neural Networks, 13 (2000), pp. 411–430.

[17] E. Isbell, A. N. Peters, D. M. Richardson, and N. E. R. De León, *Cognitive Electrophysiology in Socioeconomic Context in Adulthood*, 2024.

[18] V. Jayaram and A. Barachant, *MOABB: Trustworthy Algorithm Benchmarking for BCIs*, Journal of Neural Engineering, 15 (2018), p. 066011.

[19] W.-B. Jiang, L.-M. Zhao, and B.-L. Lu, *Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI*, in International Conference on Learning Representations, 2024.

[20] E. S. Kappenman, J. L. Farrens, W. Zhang, A. X. Stewart, and S. J. Luck, *ERP CORE: An Open Resource for Human Event-Related Potential Research*, NeuroImage, 225 (2021), p. 117465.

[21] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, *EEGNet: A Compact Convolutional Neural Network for EEG-Based Brain–Computer Interfaces*, Journal of Neural Engineering, 15 (2018), p. 056013.

[22] M. Li, F. He, B. Chen, and B.-L. Lu, *Transfer Learning in Motor Imagery Brain–Computer Interface*, Journal of Shanghai Jiao Tong University (Science), 29 (2024), pp. 85–101.

[23] D. Liu, Y. Chen, Z. Chen, Z. Cui, Y. Wen, J. An, J. Luo, and D. Wu, *EEG Foundation Models: Progresses, Benchmarking, and Open Problems*, arXiv preprint arXiv:2601.17883, (2026).

[24] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, *A Review of Classification Algorithms for EEG-Based Brain–Computer Interfaces: A 10 Year Update*, Journal of Neural Engineering, 15 (2018), p. 031005.

[25] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, *A Review of Classification Algorithms for EEG-Based Brain–Computer Interfaces*, Journal of Neural Engineering, 4 (2007), pp. R1–R13.

[26] S. J. Luck, *An Introduction to the Event-Related Potential Technique*, MIT Press, Cambridge, MA, 2 ed., 2014.

[27] S. Makeig, A. J. Bell, T.-P. Jung, and T. J. Sejnowski, *Independent Component Analysis of Electroencephalographic Data*, Advances in Neural Information Processing Systems, 8 (1996), pp. 145–151.

[28] C. J. Markiewicz, K. J. Gorgolewski, F. Feingold, R. Blair, Y. O. Halchenko, E. Miller, N. Hardcastle, J. Wexler, O. Esteban, M. Goncavles, A. Jwa, and R. A. Poldrack, *The OpenNeuro Resource for Sharing of Neuroscience Data*, eLife, 10 (2021), p. e71774.

[29] S. D. Muthukumaraswamy, *High-Frequency Brain Activity and Muscle Artifacts in MEG/EEG: A Review and Recommendations*, Frontiers in Human Neuroscience, 7 (2013), p. 138.

[30] OpenNeuro, *Older Adult Resting State and Auditory Oddball Task EEG Data*, 2025.

[31] ——, *Young Adult Resting State and Auditory Oddball Task EEG Data*, 2025.

[32] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, *PyTorch: An Imperative Style, High-Performance Deep Learning Library*, in Advances in Neural Information Processing Systems, vol. 32, 2019.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *Scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research, 12 (2011), pp. 2825–2830.

[34] C. R. Pernet, S. Appelhoff, K. J. Gorgolewski, G. Flandin, C. Phillips, A. Delorme, and R. Oostenveld, *EEG-BIDS, an Extension to the Brain Imaging Data Structure for Electroencephalography*, Scientific Data, 6 (2019), p. 103.

[35] J. Polich, *Updating P300: An Integrative Theory of P3a and P3b*, Clinical Neurophysiology, 118 (2007), pp. 2128–2148.

[36] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, *Deep Learning-Based Electroencephalography Analysis: A Systematic Review*, Journal of Neural Engineering, 16 (2019), p. 051001.

[37] S. Saha and M. Baumert, *Intra- and Inter-subject Variability in EEG-Based Sensorimotor Brain–Computer Interface: A Review*, Frontiers in Computational Neuroscience, 13 (2020), p. 87.

[38] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, *Deep Learning with Convolutional Neural Networks for EEG Decoding and Visualization*, Human Brain Mapping, 38 (2017), pp. 5391–5420.

[39] Y. Song, Q. Zheng, B. Liu, and X. Gao, *EEG Conformer: Convolutional Transformer for EEG Decoding and Visualization*, IEEE Transactions on Neural Systems and Rehabilitation Engineering, 31 (2023), pp. 710–719.

[40] C. Vidaurre and B. Blankertz, *Towards a Cure for BCI Illiteracy*, Brain Topography, 23 (2010), pp. 194–198.

[41] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, and P. van Mulbregt, *SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python*, Nature Methods, 17 (2020), pp. 261–272.

[42] D. Wu, Y. Xu, and B.-L. Lu, *Transfer Learning for EEG-Based Brain–Computer Interfaces: A Review of Progress Made Since 2016*, IEEE Transactions on Cognitive and Developmental Systems, 14 (2022), pp. 4–19.

[43] C. Yang, M. B. Westover, and J. Sun, *BIOT: Biosignal Transformer for Cross-Data Learning in the Wild*, in Advances in Neural Information Processing Systems, vol. 36, 2023.

[44] T. Yue, S. Xue, X. Gao, Y. Tang, L. Guo, J. Jiang, and J. Liu, *EEGPT: Unleashing the Potential of EEG Generalist Foundation Model by Autoregressive Pre-Training*, arXiv preprint arXiv:2410.19779, (2024).

Appendix A

Dataset Event Definitions

This appendix consolidates the dataset-specific definitions used to construct the binary target-versus-standard tasks. Standard trials were assigned label 0, target trials were assigned label 1, and retained epochs were returned to their original chronological order using their event sample indices. Events not listed in Table A.1, including distractors, were excluded.

Table A.1: Exact binary event definitions used in the final benchmark.

Dataset	Standard class	Target class	Final condition

`ds003061`	`stimulus/standard`	`stimulus/oddball`
		`stimulus/oddball_with_response`
		`stimulus/oddball_with_reponse`
			All 13 available subjects; target response variants pooled

`ds005863`	12, 13, 14, 15, 21 23, 24, 25, 31, 32 34, 35, 41, 42, 43 45, 51, 52, 53, 54	11, 22, 33, 44, 55	Visual oddball task; 30 subjects
`ds006466`	113	114	Active auditory-oddball condition; 30 subjects
`ds006480` pooled	106, 113, 125, 132	107, 114, 126, 133	Included active/passive and control/shock conditions; 30 subjects

For `ds006480`, distractor codes 108, 115, 127, and 134 were excluded. The general-purpose `ds006466` loader also recognized passive and shock-condition codes, but only codes 113 and 114 from the active condition were included in the final benchmark.

Appendix B

Implementation and Statistical Details

This appendix provides a compact replication summary. The rationale and complete procedures are presented in Chapters 3–7.

B.1 Classifier Inputs and Parameterization

Table B.1 consolidates the final classifier inputs and implementations.

The linear input contained 17 complete six-sample bins obtained from the 103-sample post-stimulus crop. EEGNet's trainable parameter count was $946 + 16C$. Table B.2 reports the resulting dataset-specific dimensions and parameter counts.

LDA additionally estimated class means, priors, and a shrinkage covariance matrix, while both linear pipelines stored training-derived scaling statistics. These quantities are not directly comparable with independently optimized neural-network weights.

Table B.1: Classifier inputs and final implementations. C is the dataset-specific number of EEG channels.

Classifier	Input per EEG epoch	Final configuration
LDA	$17C$-dimensional time-binned vector	`StandardScaler`; LSQR solver; automatic covariance shrinkage; one decision coefficient per input feature and one intercept
Logistic Regression	$17C$-dimensional time-binned vector	`StandardScaler`; LBFGS; L_2 penalty; $C = 1.0$; balanced class weights; intercept; maximum 2000 iterations
EEGNet	$1 \times C \times 103$ tensor	$F_1 = 8$, $D = 2$, $F_2 = 16$; temporal kernels 32 and 16; dropout 0.25; Adam; learning rate 10^{-3}; weight decay 10^{-4}; batch size 64; maximum 60 epochs; patience 10

Table B.2: Dataset-specific input dimensions and effective parameter counts. Linear counts include the decision coefficients and intercept.

Dataset	Channels C	Linear input	Linear coefficients	EEGNet parameters
`ds003061`	64	1088	1089	1970
`ds005863`	26	442	443	1362
`ds006466`	65	1105	1106	1986
`ds006480`	65	1105	1106	1986

Table B.3: Numbers of outer partitions and primary-benchmark classifier fits. The 309 augmented EEGNet fits are reported separately in Section B.4.

Evaluation	Outer partitions	LDA fits	Logistic Regression fits	EEGNet fits
Within-subject	510	510	510	1530
LOSO	103	103	103	309
Total	613	613	613	1839

B.2 Evaluation, Aggregation, and Statistical Analysis

The final analysis contained 510 valid grouped within-subject outer partitions and 103 LOSO outer partitions.

Table B.3 gives the corresponding primary-benchmark fit counts.

The EEGNet counts reflect three seeds, 42, 43, and 44, for every outer partition. These repetitions measured neural-training stability and did not create additional independent test observations.

Within subject, outer-fold AUC values were first averaged for each subject and classifier; EEGNet values were then averaged across seeds. In LOSO, EEGNet seed-level AUC values were averaged for each held-out subject. Cross-dataset headline values were unweighted macro-averages of the four dataset-level means:

$$\overline{\mathrm{AUC}}_{\mathrm{macro}} = \frac{1}{4}\sum_{d=1}^{4}\overline{\mathrm{AUC}}_{d}$$

Classifier comparisons used one matched observation per subject-recording and classifier. Friedman tests were followed by paired Wilcoxon signed-rank tests with Holm correction. Matched rank-biserial correlations were reported as pairwise effect sizes.

Mass-univariate susceptibility analyses used Spearman correlations. Benjamini–Hochberg false-discovery-rate correction was applied within the specified analysis families. Within subject, correction was performed separately within each classifier for the reportable raw/ERP/power features. In LOSO, raw/ERP/power and ICA/source-quality tests were corrected as separate source families. The primary threshold was $q \leq 0.05$, with $q \leq 0.10$ treated as suggestive.

B.3 Subject Quality Index

For subject s and component k, the original feature was standardized within dataset:

$$z_{s,k} = \frac{x_{s,k} - \mu_{d,k}}{\sigma_{d,k}}$$

Each component was oriented so that larger values represented better estimated quality:

$$\widetilde{z}_{s,k} = o_k z_{s,k}, \qquad o_k \in \{-1, +1\}$$

The primary Subject Quality Index was the unweighted mean of the 20 oriented components:

$$Q_s = \frac{1}{20} \sum_{k=1}^{20} \widetilde{z}_{s,k}$$

All 103 subject-recordings had complete values for these components. No winsorization or clipping was applied, and classifier AUC was not used to optimize component weights. The 12 response-related components and eight variability-related components therefore contributed 60% and 40% of the component-level weight, respectively. Leave-one-component-out, median, and equal-family alternatives were evaluated as sensitivity analyses.

Incremental explanatory values were defined as

$$\Delta R^2_{\text{classifier}} = R^2(Y \sim D + C) - R^2(Y \sim D)$$

and

$$\Delta R^2_{\text{quality}} = R^2(Y \sim D + Q) - R^2(Y \sim D)$$

where Y is AUC, D is dataset identity, C is classifier identity, and Q is the Subject Quality Index. A five-fold subject-grouped cross-validation analysis kept all classifier rows belonging to one subject in the same fold.

B.4 EEGNet Augmentation Replication Details

The Gaussian-noise ablation modified only normalized inner-training inputs:

$$\widetilde{X} = X + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2), \qquad \sigma = 0.01$$

Noise was generated independently for each training batch and was not applied to inner-validation or outer-test data. The magnitude was selected from $\{0.01,\ 0.025,\ 0.05\}$ using 36 paired validation partitions. Candidate selection used inner-validation ROC-AUC only; outer-test AUC was not used.

The complete LOSO comparison contained 103 held-out subject-recordings and three neural seeds, giving 309 augmented fits paired with the exact corresponding

non-augmented results. Conditions shared the held-out subject, training subjects, architecture, optimizer, normalization, inner split, early-stopping procedure, and neural seed. A separate random-number generator was used for Gaussian perturbations so that augmentation did not alter initialization or batch-order randomness.

Seed-level results were averaged before subject-level inference. The primary augmentation comparison used a paired Wilcoxon signed-rank test, matched rank-biserial effect size, and a 10,000-resample subject-level bootstrap confidence interval. Quality-stratified tests used lower and upper within-dataset Subject Quality Index quartiles and Holm correction across the two within-group comparisons.

Gaussian augmentation retained the original labels and altered only training inputs. By contrast, the development-stage label-permutation and prestimulus analyses removed the intended label relationship or post-stimulus information and were negative controls rather than augmentation procedures.

Appendix C

Computational Environment and Data Availability

Table C.1 records the software and hardware environment used for the final analysis.

EEG loading, filtering, referencing, resampling, epoching, and metadata handling used MNE-Python and MNE-BIDS [3, 12]. LDA and Logistic Regression used scikit-learn [33], and EEGNet used PyTorch [32]. Numerical and statistical processing used NumPy, pandas, and SciPy, and figures were generated using Matplotlib [13, 41].

Neural experiments used seeds 42, 43, and 44 for Python, NumPy, and PyTorch. CUDA generators were seeded when GPU training was available, deterministic cuDNN behavior was enabled, and cuDNN benchmarking was disabled. These settings reduce unintended variation but do not guarantee bitwise-identical results across operating systems, library versions, CUDA versions, or GPU architectures.

All four datasets are publicly available through OpenNeuro [28]: `ds003061` [9], `ds005863` [17], `ds006466` [30], and `ds006480` [31]. Raw EEG files are not redistributed with the thesis.

The thesis documents dataset accessions and event mappings, preprocessing

Table C.1: Recorded computational environment for the final analysis.

Component	Recorded specification
Processor	Intel Core i7-6700HQ CPU at 2.60 GHz
System memory	16 GB RAM
Graphics processor	NVIDIA GeForce GTX 960M with 4096 MiB memory
Operating system	64-bit Linux; kernel `7.0.0-28-generic`
Python	3.10.19
NumPy / pandas / SciPy	2.2.6 / 2.3.3 / 1.15.2
scikit-learn	1.7.2
PyTorch / CUDA build	2.7.1+`cu118` / CUDA 11.8
MNE-Python / MNE-BIDS	1.10.2 / 0.17.0
Matplotlib	3.10.7

intervals and parameters, channel handling, classifier inputs, architecture and parameter counts, outer and inner partitioning, training and early-stopping settings, random seeds, statistical aggregation, correction families, augmentation selection and run counts, and the computational environment. These details support methodological transparency and independent reconstruction, although exact numerical reproduction may depend on software, hardware, and implementation details not fully captured in the text.