TEMPORAL DECODING IN NATURALISTIC FMRI WITH COMPUTATIONALLY EFFICIENT DEEP NEURAL NETWORKS

SARA ASADI

Bachelor of Science, Ferdowsi University of Mashhad, 2023

A thesis submitted in partial fulfillment of the requirement for the degree of

MASTER OF SCIENCE

In

NEUROSCIENCE

Department of Neuroscience

University of Lethbridge

LETHBRIDGE, ALBERTA, CANADA

TEMPORAL DECODING IN NATURALISTIC FMRI WITH COMPUTATIONALLY EFFICIENT DEEP NEURAL NETWORKS

SARA ASADI

Date of Defence: February 11, 2026

Dr. Chelsea Ekstrand	Associate Professor	Ph.D.
Dr. Hardeep Ryait	Assistant Professor	Ph.D.
Thesis Co-Supervisors		
Dr. David Euston	Associate Professor	Ph.D.
Dr. Artur Luczak	Professor	Ph.D.
Thesis Examination Committee Members		
Dr. Masami Tatsuno	Associate Professor	Ph.D.
Chair, Thesis Examination Committee		

DEDICATION

To my father and mother,
for your endless love, strength, encouragement, and sacrifices that made every step of this journey possible.

To my sisters, Sepideh and Simin,
for walking beside me through every step of my academic path, for your constant help, and the way you believe in me even when I doubt myself.

To my fiancé, Hojat,
for bringing peace into my days and joy into my life, for your patient, generous support, and for turning the future into something soft and hopeful.

I also dedicate this work to the people of Iran, who stand bravely in these difficult days, holding hope for a freer and brighter future.

With all my love and gratitude, always.

ABSTRACT

Deep learning has demonstrated strong potential for decoding cognitive states from functional magnetic resonance imaging (fMRI) data, but its substantial computational requirements and limited capacity to capture temporal dynamics often hinder practical application. This study introduces a computationally efficient deep learning framework designed for the classification of cognitive states (in this case, face processing) from single events in naturalistic fMRI, where stimuli are temporally rich and dynamically evolving. To facilitate efficient processing while preserving temporal complexity, fMRI related to face onset and face offset events from 86 participants from the Naturalistic Neuroimaging Database version 2.0 (Aliko et al., 2021) were transformed from their original 4D spatiotemporal format into 2D voxel-by-time matrices that explicitly incorporate temporal information. We developed a lightweight convolutional neural network (CNN) that extracts both spatial and temporal patterns in brain activity, enabling efficient analysis with minimal computational cost. The model achieved 82% test accuracy and 89% validation AUC while preserving temporal information. Attribution analysis using Integrated Gradients (via DeepExplain) highlighted activation patterns in expected face-processing regions including Fusiform Face Area (FFA), Occipital Face Area (OFA), and posterior superior temporal sulcus (pSTS), validating that the predictions were driven by relevant neural activity. Therefore, this framework provides a solution for moment-to-moment brain decoding in ecologically valid, naturalistic environments. This work is positioned as a proof-of-concept framework rather than as a performance benchmark.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisors, Dr. Chelsea Ekstrand and Dr. Hardeep Ryait, for giving me the opportunity to pursue my master's degree and explore the growing field of NeuroAI at the University of Lethbridge. Your continued support has shaped both my academic progress and personal growth. You taught me that no challenge is too big to overcome, and you created an environment that allowed my curiosity and passion for science to flourish.

I would also like to thank my committee members, Dr. David Euston and Dr. Artur Luczak, for your helpful feedback, guidance, and support throughout these years. I am deeply grateful for the time and thought you invested in my work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AFNI	Analysis of Functional NeuroImages
AUC	Area under the curve
AWS	Amazon Web Services
BCI	Brain-Computer Interface
BOLD	Blood-oxygen-level-dependent
CNN	Convolutional neural network
CV	Cross-validation
EPI	Echo-planar imaging
FFA	Fusiform face area
fMRI	Functional magnetic resonance imaging
GPU	Graphics processing unit
HRF	Hemodynamic response function
IG	Integrated Gradients
L2	L2-norm (weight decay)
LSTM	Long short-term memory
ML	Machine learning
MP-RAG	Magnetization-prepared rapid gradient echo
MRI	Magnetic resonance imaging
NNDb	Naturalistic Neuroimaging Database
OFA	Occipital face area
pSTS	Posterior superior temporal sulcus
ReLU	Rectified linear unit
ROC	Receiver operating characteristic
ROI	Region of interest
RNN	Recurrent neural network
SD	Standard deviation
SHAP	Shapley additive explanations
TE	Echo time
TR	Repetition time
XAI	Explainable AI

CHAPTER 1: INTRODUCTION

1.1 FUNCTIONAL MAGNETIC RESONANCE IMAGING AS A WINDOW INTO COGNITION

Functional magnetic resonance imaging (fMRI) has become one of the most widely used methods in neuroscience for studying human brain function (Glover, 2011). As a non-invasive neuroimaging technique, fMRI enables the visualization of large-scale neural activity with high spatial resolution while also capturing its temporal dynamics, allowing researchers to examine how brain activity changes over time. FMRI data is inherently four-dimensional, consisting of three spatial dimensions and a time dimension, which makes it possible to track dynamic changes in brain activity across the entire brain during perception, action, and cognition (Parry & Matthews, 2002).

FMRI measures fluctuations in the blood-oxygen-level-dependent (BOLD) signal, a contrast mechanism first described by Ogawa et al. (1990). The BOLD signal reflects changes in cerebral blood flow and blood oxygenation that occur in response to neuronal activity. Increased neural activity leads to greater delivery of oxygenated blood, resulting in higher BOLD signal intensity. Because these signal changes provide an indirect but reliable marker of neural activity, they can be repeatedly measured across the whole brain, making fMRI an important tool for understanding human cognition and brain function (Buxton et al., 1998). This vascular response unfolds over several seconds and thus the BOLD signal is delayed relative to the underlying neural activity, typically peaking approximately 4-6 seconds after stimulus onset. This hemodynamic delay limits the temporal precision of fMRI compared to electrophysiological methods, but the

temporal structure of the signal remains informative and can be modeled to infer task- or stimulus-related brain activity (Polimeni & Lewis, 2021).

A major strength of fMRI lies in its high spatial resolution, which enables the localization of cognitive functions to specific brain regions. Seminal studies demonstrated the ability of fMRI to map functional specialization in the human brain, including visual processing in occipital cortex (Kwong et al., 1992), language lateralization (Binder et al., 2009), and attentional modulation of sensory areas (Heinze et al., 1994). FMRI has also been instrumental in identifying category-selective regions, such as the fusiform face area involved in face perception (Kanwisher et al., 1997), and in revealing how memory encoding and retrieval engage medial temporal and prefrontal regions (Buckner et al., 1996).

Although temporal precision is limited by the delayed hemodynamic response, fMRI's combination of whole-brain coverage, spatial specificity, and repeatability has substantially advanced our understanding of human cognition (Buckner et al., 1996; Glover, 2011). By allowing researchers to compare BOLD responses across conditions, time points, and populations, fMRI has provided critical insights into how cognitive processes are organized in the brain, how they change with learning and development, and how they are altered in neurological and psychiatric disorders (Dale & Halgren, 2001; Heinze et al., 1994).

1.2 CHALLENGES IN HIGH-DIMENSIONAL FMRI ANALYSIS

Although fMRI has greatly advanced our understanding of human brain function, the high-dimensional nature of the data creates several analytical challenges. As previously stated, fMRI is four-dimensional, consisting of three spatial dimensions (x, y, z) sampled repeatedly over time (t), resulting in large datasets that capture both spatial structure and temporal variation in brain activity.

This richness provides a powerful window into cognition but also requires methodological choices that balance interpretability, statistical power, and preservation of meaningful information (Glover, 2011).

Most widely used fMRI analysis methods were developed to address specific experimental questions by simplifying the data structure. In task-based fMRI, the dominant analysis framework is the general linear model (GLM), which estimates voxel-wise activation by relating the BOLD signal to predefined experimental regressors convolved with a canonical hemodynamic response function (HRF), which models the typical delay and shape of the BOLD signal following a brief burst of neural activity. This approach has been highly successful in identifying brain regions associated with specific cognitive functions and remains the standard for functional localization and hypothesis-driven analysis (Soares et al., 2016).

For resting-state fMRI, which measures slow, spontaneous fluctuations in the BOLD signal while subjects lie inactive in the MRI scanner (Biswal et al., 1995), traditional analyses tend to focus on functional connectivity, commonly computed as temporal correlations between regional BOLD time series. These correlations are often summarized as a single connectivity matrix representing average coupling across the entire scan. Such approaches have provided important insights into large-scale brain organization and intrinsic functional architecture (Belyaeva et al., 2021; Janoos et al., 2011; Li et al., 2018; Specht, 2020).

However, a key limitation of these traditional approaches is their reliance on temporal averaging and stationarity assumptions. By summarizing activity across extended time periods or enforcing a fixed response shape, they reduce sensitivity to trial-by-trial variability and moment-to-moment changes in neural activity. As a result, temporally evolving cognitive processes may be obscured, particularly in complex or continuous tasks (Haynes & Rees, 2006).

More advanced analysis techniques, such as multivoxel pattern analysis (MVPA) and related machine learning approaches, analyze distributed patterns of activity across voxels rather than treating each voxel independently. These methods have been shown to detect information about perceptual and cognitive states that may not be apparent in univariate activation analyses, offering increased sensitivity to subtle representational differences (Haynes & Rees, 2006; Mahmoudi et al., 2016). Despite these advances, many multivariate approaches still rely on temporally summarized features, such as averaged activation within time windows or conditions. Consequently, although spatial information is better preserved, fine-grained temporal dynamics are often only partially captured.

More recently, dynamic fMRI methods have been introduced to address the temporal limitations of static analyses. Approaches such as sliding-window functional connectivity analyses aim to capture time-varying changes in connectivity by computing correlations within successive temporal windows. These methods have provided evidence that functional interactions fluctuate over relatively short timescales during perception, attention, and emotion (Janoos et al., 2011; Li et al., 2018).

However, dynamic approaches also face important constraints. Sliding-window methods compute statistics (e.g., correlation, covariance) within each window, and results can be sensitive to analytic choices such as window length, overlap, and noise (Shakil et al., 2016). As a result, rapid transitions and fine temporal structure may remain partially obscured, and interpretability can be challenging (Belyaeva et al., 2021).

Overall, while traditional and advanced fMRI analysis techniques have been highly effective for many research questions, they often rely on simplified assumptions that limit the use of the data's full temporal richness. These limitations are particularly salient in studies using

naturalistic stimuli or continuous tasks, where cognitive processes unfold dynamically over time. Naturalistic stimuli refer to complex, continuous inputs such as movies, spoken narratives, and real-world tasks that unfold over time and more closely resemble everyday experience than tightly controlled, discrete trials. Addressing these challenges requires analysis methods that preserve temporal information while remaining computationally feasible and interpretable.

1.3 IMPORTANCE OF PRESERVING TEMPORAL STRUCTURE & RELEVANCE OF NATURALISTIC STIMULI

The temporal dimension of fMRI data carries essential information about how neural activity evolves over time, making it critical for studying dynamic cognitive processes (Dale & Halgren, 2001; Janoos et al., 2011). Early work by Dale and Halgren (2001) demonstrated that even with the delayed hemodynamic response, fMRI contains meaningful temporal structure that can be leveraged to study the sequence and timing of neural events during cognition. Their findings showed that cognitive processes (such as perception, attention, language comprehension, and memory) are not instantaneous but evolve over time, and that ignoring temporal information can lead to incomplete or misleading interpretations of brain function.

Collapsing or averaging fMRI timepoints can also obscure transient fluctuations in neural activity and remove patterns that reflect rapid changes in perception or cognition. Further, averaging across trials assumes that neural responses are stable over time, an assumption that does not hold for tasks with continuous or time-varying input. Task-based studies using naturalistic stimuli, such as movie viewing, have shown that preserving the full BOLD time course retains meaningful information about how brain activity changes moment by moment during stimulus processing. For example, Hasson and colleagues showed that time-resolved fMRI responses during movie watching show reliable, stimulus-locked patterns across participants that are lost when

signals are averaged over time (Hasson et al., 2004). These findings indicate that preserving the temporal evolution of neural activity retains meaningful information about ongoing perceptual and cognitive dynamics that is lost through temporal averaging, particularly in tasks involving continuous or rapidly changing stimuli. Preserving temporal structure therefore allows models to capture meaningful moment-to-moment variations in the BOLD signal, including trial-specific differences in amplitude, latency, and duration (Serences, 2004).

Several methodological advances further highlight the value of preserving temporal information during fMRI analyses. For example, Janoos et al. (2011) used state-space modeling, which represents brain activity as transitions between latent functional states over time, to show that brain activity changes consistently over time, even within a single scanning session. Their work showed that cognitive and perceptual states can be inferred from the temporal evolution of the fMRI signal, supporting the idea that cognition is best characterized as a dynamic process rather than a static one.

Similarly, Kriegeskorte & Douglas (2018) argued that understanding cognition requires models that explain how neural activity changes as information is processed over time. Rather than viewing brain activity as a static response to a stimulus, they described cognition as a sequence of computational steps, in which neural representations are gradually transformed, combined with prior knowledge, and updated as a task progresses. These intermediate representational states are reflected in the evolving patterns of activity across neural populations. From this perspective, cognitive information is carried by the temporal progression of multivariate activity patterns, and analyses based on single time points or averaged responses may miss important aspects of how information is processed in the brain.

These considerations become especially important in experimental paradigms where cognitive processes unfold continuously over time rather than in isolated trials. Evidence for the importance of temporal dynamics is particularly strong in studies using naturalistic stimuli. In language comprehension, for example, Brennan et al. (2016) demonstrated that neural responses depend strongly on sequential context, with brain activity reflecting ongoing integration of linguistic information across time. Their findings showed that models incorporating temporal structure better explain cortical responses during continuous speech compared to models that treat stimuli as independent events.

Most traditional fMRI paradigms are built around block-based or event-related designs, in which stimuli are presented as discrete, well-separated trials and neural responses are analyzed relative to predefined onsets. While these approaches have been successful for isolating specific cognitive functions, they impose assumptions about temporal structure and often rely on averaging across repeated events. As discussed above, such assumptions limit sensitivity to temporally continuous and context-dependent cognitive processes.

In this context, naturalistic stimuli provide a powerful framework for studying cognition evolving over time. Unlike traditional block or event-related designs, naturalistic paradigms engage multiple cognitive processes simultaneously and dynamically, including perception, attention, memory, and social cognition (Sonkusare et al., 2019). These stimuli evoke reliable, time-locked neural responses across individuals, particularly in perceptual, language, and social brain networks (Hasson et al., 2004; Lee & Chen, 2022). Importantly, naturalistic stimuli have also helped reveal hierarchical temporal organization in the brain, ranging from fast sensory responses to slower integrative processes that support narrative comprehension and meaning construction (Baldassano et al., 2017).

However, despite this rich information across time, naturalistic fMRI datasets are often analyzed using the same techniques developed for traditional task-based or resting state designs, which rely heavily on temporal averaging, fixed response models, or dimensionality reduction. These approaches tend to suppress trial-specific timing differences, transient event boundaries, and context-dependent changes in neural activity, features that are central to cognition in naturalistic settings (Nastase et al., 2020). As a result, valuable information about how cognitive processes evolve and interact over time may be attenuated or lost.

Preserving temporal information is therefore essential for capturing event-specific and context-dependent neural responses in realistic environments, where cognitive states are shaped by ongoing sensory input and prior context rather than isolated stimulus onsets. Analytical approaches that collapse temporal structure are thus poorly matched to the nature of naturalistic data, motivating the use of models that explicitly retain and exploit temporal information.

1.4 DEEP LEARNING AS A TOOL FOR HIGH-DIMENSIONAL FMRI

Recent advances in artificial intelligence (AI) have provided new ways to analyze complex and high-dimensional fMRI data (Deniz et al., 2019; Huth et al., 2016). Deep learning, a subset of AI, has become an influential framework for studying high-dimensional fMRI data because it can capture complex, nonlinear relationships that traditional statistical models often miss. Unlike conventional GLM or MVPA approaches, deep neural networks learn hierarchical feature representations directly from the data, allowing them to model the spatial, temporal, and functional complexity present in whole-brain measurements (Avberšek & Repovš, 2022; St-Yves et al., 2023). These methods are particularly effective for large and complex datasets, making them increasingly common for decoding cognitive states, predicting behaviour, and uncovering latent neural patterns embedded in fMRI signals (Thomas et al., 2019).

For example, seminal work by Huth et al. (2016) demonstrated that deep neural network models trained on natural images and movies could predict voxel-wise responses across much of visual cortex. Importantly, this work showed that representations learned by deep models align with hierarchical processing in the human visual system, with early layers corresponding to low-level visual features and deeper layers mapping onto higher-order visual and semantic areas. These findings provided strong evidence that deep learning models can capture representational structure that mirrors known functional organization in the brain.

Building on this representational alignment, Horikawa and Kamitani (2017) demonstrated that hierarchical visual features derived from deep neural networks can be decoded from fMRI activity to identify object categories beyond those used during training. Rather than training classifiers on a fixed set of categories, their approach mapped brain activity onto intermediate feature representations learned by vision models, enabling generalization to novel and even imagined objects. This work showed that fMRI signals contain rich, abstract visual information that can be expressed in a high-dimensional feature space, further supporting the use of deep learning frameworks for flexible and generalizable neural decoding.

Extending these ideas to naturalistic settings, Deniz et al. (2019) used deep neural networks to characterize category-selective responses during naturalistic movie viewing, revealing distributed and overlapping representations of objects, actions, and social information across cortical regions. Rather than isolating single regions, these studies showed that cognitive content is encoded in distributed patterns spanning large portions of cortex, supporting a shift away from localized activation-based interpretations toward population-level representations. Deep learning approaches have also advanced understanding of temporally evolving neural processes. Dvornek et al. (2017) demonstrated that models incorporating temporal relationships across consecutive

fMRI time points improve decoding of cognitive states compared to static representations, highlighting the importance of sequential information for interpreting brain activity.

Many deep learning approaches applied to fMRI rely on convolutional neural networks (CNNs; LeCun et al., 2015), which are particularly well suited for modeling spatially structured data such as voxel-wise brain measurements. In neuroimaging, CNNs are typically used as computational tools that operate on spatial or spatiotemporal representations derived from fMRI data, rather than as direct models of the neural mechanisms that generate the measured signals. This architectural bias toward locality and hierarchy aligns naturally with the distributed and hierarchical organization observed in cortical representations (Eickenberg et al., 2017).

Together, these studies illustrate that deep learning methods are particularly well suited for fMRI analyses that aim to preserve spatially distributed and temporally structured information. Rather than reducing data to static summaries, deep learning models have enabled researchers to probe how cognitive representations are organized across the brain and how they evolve during continuous, real-world tasks. As a result, deep learning has become a valuable tool for studying high-dimensional, temporally rich fMRI data, offering insights into neural representations that are difficult to obtain with traditional analysis techniques.

1.5 GAPS IN THE LITERATURE

Despite major progress in neuroimaging and application of deep learning to fMRI data, several important gaps remain in how dynamic cognitive processes are modeled. These gaps are particularly evident in studies that aim to capture cognition under naturalistic conditions. First, many decoding studies rely on tightly controlled experimental designs such as isolated images, short trials, or simple representations. For example, Horikawa and Kamitani (2017) showed that

hierarchical visual features derived from deep models can be decoded from fMRI responses to static images, revealing how visual information is organized across cortical areas. While this work provided important insight into visual representations in the brain, it was based on discrete, time-collapsed stimuli and did not address how neural responses evolve over continuous input. As a result, such designs improve experimental control but do not reflect the richness of real-world experience (Atteveldt et al., 2018; Hay et al., 2022).

Second, many CNN-based approaches treat fMRI data as static input, either by analyzing single 3D volumes or time-averaged features (Eickenberg et al., 2017; Güçü & van Gerven, 2015; Horikawa & Kamitani, 2017; VanRullen & Reddy, 2019). This design overlooks the fact that neural responses are dynamic, especially during naturalistic tasks, and that the temporal structure itself can contain important information. Although recurrent or hybrid models offer ways to integrate time, they often require long sequences or large training datasets, limiting their accessibility and interpretability. As a result, many naturalistic fMRI studies still rely on representations that collapse time, making trial-by-trial or event-level decoding difficult or impossible (Bottenhorn et al., 2018; Simony & Chang, 2020).

At the same time, there is a growing interest in decoding neural responses at the level of individual events or moments in time, rather than relying on averaged responses. Preserving temporal structure is essential for capturing such trial-by-trial variability, which has been shown to carry meaningful information about perception, memory, and attention. However, standard analysis pipelines are not well suited for this goal when temporal information is reduced early in the modeling process.

Taken together, these gaps highlight the need for methods that (1) preserve temporal information, (2) remain computationally efficient for high-dimensional fMRI, and (3) generalize

across individuals and diverse naturalistic stimuli. This thesis addresses these challenges by introducing a voxel-by-time decoding framework that captures event-level neural dynamics during movie viewing using a lightweight convolutional neural network.

1.6 FACE PERCEPTION AS A BENCHMARK SYSTEM

Face perception is one of the most well-studied domains in cognitive neuroscience and provides a strong benchmark for evaluating models of human visual processing. The brain contains specialized mechanisms for detecting, recognizing, and interpreting faces, reflecting the social and biological importance of facial information (Kanwisher & Yovel, 2006). These mechanisms support core processes such as identity recognition, emotion decoding, gaze interpretation, and social judgment, all of which are essential for everyday interaction (Haxby et al., 2000). Because face perception involves well-defined neural pathways and highly reproducible behavioural signatures, it serves as a reliable system for testing and validating computational models.

A substantial body of work has characterized the distributed network that supports face processing, which operates in a hierarchical and partially modular manner (Haxby et al., 2000; Kanwisher & Yovel, 2006). The core face-processing system includes three primary regions in occipitotemporal cortex (Figure 1). The occipital face area (OFA) supports early feature detection and part-based analysis. The fusiform face area (FFA) is specialized for holistic and identity-related processing. The posterior superior temporal sulcus (pSTS) is sensitive to changeable facial cues such as expression, eye gaze, and motion (Kanwisher et al., 1997; Haxby et al., 2000).

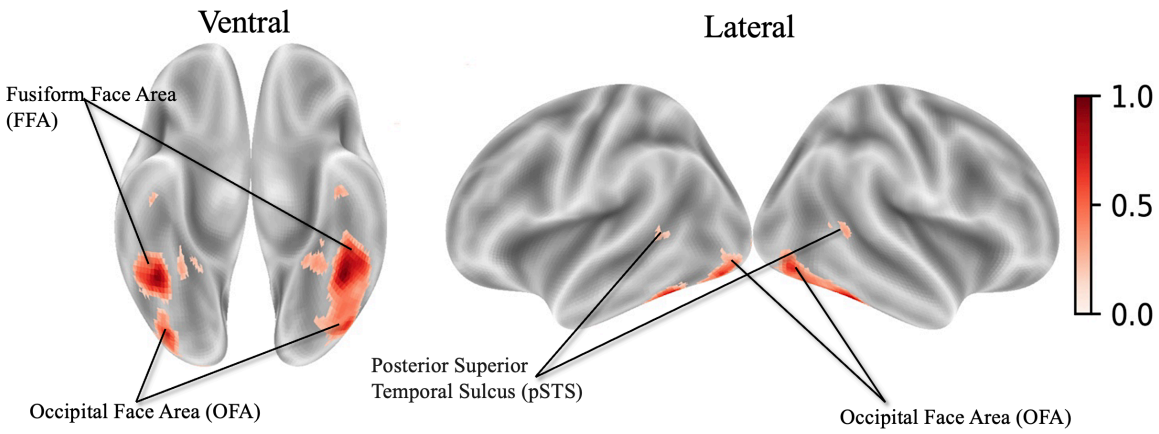

Figure 1: Face-selective cortical regions derived from meta-analytic fMRI data

Temporally, information flows from early visual cortex in the occipital lobes (e.g., V1 and V2) to the OFA for initial feature extraction, then diverges into two pathways: a ventral stream through the FFA for view-invariant identity recognition and a dorsal stream through the pSTS for dynamic social cues (Calder & Young, 2005). This core system interacts with an extended network that includes the amygdala (emotional valence), anterior temporal regions (semantic and biographical knowledge), and prefrontal cortex (higher-order social reasoning; Gobbini & Haxby, 2007). Together, these areas provide a comprehensive and well-characterized framework for evaluating neural models.

Face perception is also a powerful test case for models that incorporate temporal information. Naturalistic face stimuli, such as shifting gaze, changing expressions, or rapid social interactions, evoke dynamic neural responses that unfold over time within face-selective regions and broader social-cognitive networks (Furl et al., 2011; Pitcher & Ungerleider, 2021). These dynamics allow researchers to examine whether computational models can capture not only spatial activation patterns but also the temporal evolution of neural responses. As such, face perception offers an ideal benchmark for assessing deep learning methods designed to model fMRI signals.

In this thesis, face perception serves as the benchmark domain for evaluating our proposed voxel-by-time CNN. By focusing on face onset and face offset events during naturalistic movie viewing, the study tests whether the proposed model can capture the dynamic neural processes that unfold in real time within the constraints of fMRI.

1.7 PROPOSED DEEP LEARNING FRAMEWORK

This thesis introduces a deep learning framework designed to decode event-level neural activity from naturalistic fMRI while preserving both spatial and temporal information. The central idea is to reorganize the fMRI signal in a way that maintains short-term temporal dynamics but remains computationally efficient for high-dimensional data. To achieve this, we selected 10-second-long face onset and face offset events from naturalistic fMRI data and converted each 4D fMRI segment into a 2D voxel-by-time matrix. This representation preserves the evolving BOLD response across the 10-second window while avoiding the complexity of full 3D or 4D convolutions. It also ensures that the temporal dimension, which carries essential information about dynamic cognition, is not removed or averaged. A lightweight CNN is then applied to these matrices. Because the input structure is simplified and consistent across participants, the model can learn distributed patterns in both space and time without the large parameter demands normally associated with full spatiotemporal models. This design allows efficient training, reduces the risk of overfitting, and supports interpretability.

To validate the biological relevance of the model's decisions, attribution analysis is used to compute voxel-level attribution maps for each trial. These attributions are projected back into brain space to test whether the model relies on established face-selective regions such as the FFA, OFA, and pSTS. This step ensures that the decoding framework is grounded in known neural mechanisms rather than noise or dataset-specific artifacts. The contribution of this framework is

conceptual rather than architectural. Instead of introducing a highly complex model, the goal is to demonstrate that preserving trial-level temporal structure enables efficient and interpretable decoding of neural events in naturalistic settings. This approach supports the broader aim of modeling how cognitive states evolve in real time and provides a practical method for studying dynamic brain activity with deep learning.

CHAPTER 2: DECODING FACE PROCESSING IN FMRI WITH CNN AND XAI

2.1 BRIEF INTRODUCTION TO THE STUDY

Naturalistic movie-watching provides a rich way to study how the brain responds to visual events as they unfold over time. In this thesis, we focus on one well-defined visual event, moments when a face appears or disappears on the screen during audiovisual movies and use these moments to test a custom deep learning approach for classifying event-level brain activity in fMRI. Building on the motivation outlined in Chapter 1, the goal here is to evaluate whether preserving short temporal windows of fMRI (rather than averaging across time) allows a lightweight convolutional neural network (CNN) to learn meaningful patterns linked to face perception. Figure 2 provides an overview of the proposed decoding framework.

To accomplish this, 4D fMRI segments associated with face onset and face offset events in audiovisual movies were converted into 2D voxel-by-time matrices that retain the 10-second temporal structure of the BOLD signal. These representations were then used to train a CNN to classify each event type. To assess biological validity, attribution analysis was applied to check whether the model relied on known face-selective regions. We hypothesized that our proposed fMRI representations would support reliable decoding of face onset versus face offset events, with attribution maps highlighting well-established face-selective regions, including the OFA, FFA, and pSTS. This chapter describes the full experimental process, from data preparation and model design to classification results and neural interpretability and provides a proof-of-concept demonstration that temporally preserved representations can support efficient and biologically plausible decoding of dynamic fMRI signals.

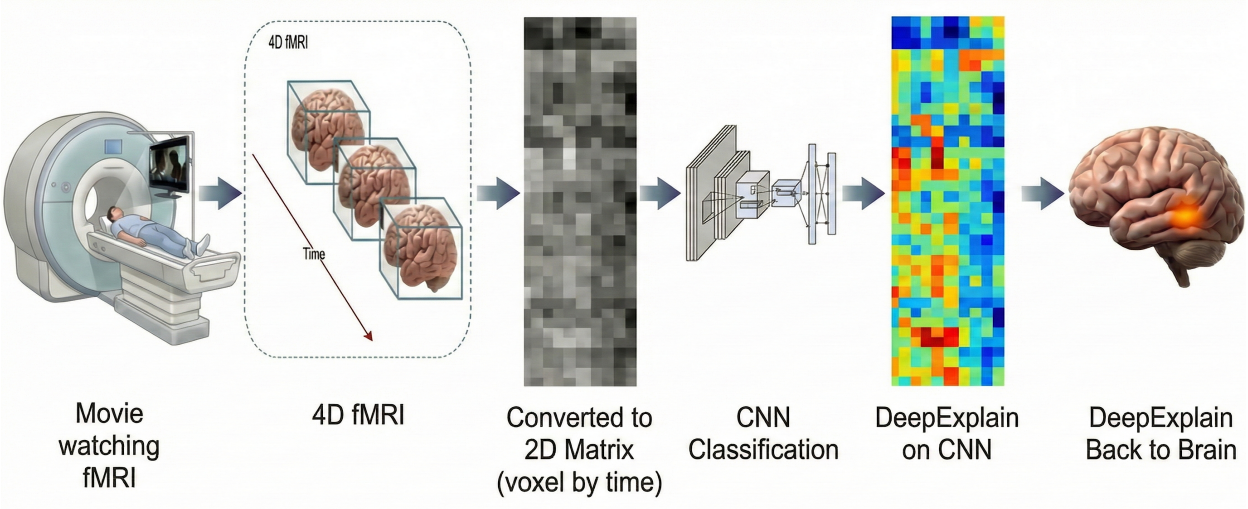

Figure 2: Overview of the fMRI decoding framework

2.2 METHODS

2.2.1 PARTICIPANTS AND FMRI DATA ACQUISITION

This study used preprocessed data from version 2.0 of the Naturalistic Neuroimaging Database (NNDb, Aliko et al., 2020) which includes fMRI data from 86 healthy participants (42 females, ages 18-58, M = 26.81, SD = 10.09 years). All participants were native English speakers with no reported psychiatric or neurological disorders and had unimpaired hearing and vision. Each participant viewed one of ten full-length audiovisual movies during fMRI. These films were chosen for their diverse emotional and social content, providing rich, naturalistic stimuli that elicit dynamic brain responses that approximate real-life processing. Collection of the original data was approved by the ethics committee of University College London and participants provided written informed consent to participate in the study and share anonymized data. All methods were carried out in accordance with the Code of Ethics of the World Medical Association for experiments involving humans.

Detailed methods information is available in Aliko et al. (2021). Briefly, each participant completed up to six fMRI runs during the movie-watching task, along with a T1-weighted high-resolution anatomical scan and over one hour of cognitive testing using the NIH toolbox. Scanning was performed on a Siemens Prisma 3T scanner with a 32-channel head coil. Functional images were collected using a multiband echo-planar imaging (EPI) sequence (repetition time (TR) = 1000 ms, echo time (TE) = 37 ms, voxel size = 2.4 mm^3, 60 slices, multiband factor = 6). T1-weighted scans used an MP-RAGE sequence (voxel size = 0.8 mm^3, TR = 2400 ms, TE = 2.22 ms). The authors of the database used the afni_proc.py pipeline from AFNI (Cox & Hyde, 1997) for preprocessing, which included slice-time correction, despiking, motion correction, spatially aligning the data to an MNI template with a resampling size of 3x3x3mm, and correcting for timing to align the fMRI time series and the film. Temporal filtering was performed using '3dTproject' to detrend the time series, which included the removal of low-frequency drifts and motion-related artifacts.

In addition to the neuroimaging data, the NNDb contains comprehensive annotations of face-presence events, enabling accurate labeling of time intervals for classification purposes (Aliko et al., 2020). These annotations were generated with automated tools including AWS Rekognition and Google Cloud Video Intelligence APIs (Aliko et al., 2020). These tools identified the presence and duration of faces in the movies with high accuracy, ensuring a strong link between the visual input and corresponding neural activity. Each annotation entry specifies the start time and length of a face event. For this study, we focused on 10-second segments after face onset and 10-seconds after face offset. Only events where the face remained on or off the screen for the entire 10 second window were included.

A subject-level data split was used to support both model training and evaluation. Six participants were completely excluded from model training and validation. Each of these participants viewed a different movie. These six participants were reserved only for continuous whole-time-course decoding analyses and were never seen by the model during optimization. Data from the remaining 80 participants was then used to build the model. All preprocessing and data preparation steps were applied in the same way to all participants before this split. This design allowed a clear separation between data used for training the model and data used to test generalization to new participants and movies.

2.2.2 MOVIE STIMULI

Participants in the NNDb each watched one of 10 feature length films, spanning drama, comedy, romance, documentary, and thriller genres (see **Table 1** for movie names and number of participants). To reduce familiarity effects, each participant viewed a movie that they had not previously seen. Film durations ranged from 92 to 148 minutes and were presented in as few continuous segments as possible. These extended, uninterrupted narratives produce rich, temporally structured neural responses that are well suited for studying dynamic cognitive processes such as face perception.

Table 1

Number of Participants for Each Film

Movie	Number of Participants
500 Days of Summer	20
Citizen Four	18
The Usual Suspects	6
Pulp Fiction	6
The Shawshank Redemption	6
The Prestige	6

Back to the Future	6
Split	6
Little Miss Sunshine	6
12 Years a Slave	6

2.2.3 FACE ANNOTATIONS AND EVENT EXTRACTION

The NNDb also contains annotations about when faces appear and disappear throughout each movie. These annotations specify the onset time and duration of each face event and were used in this study to define face onset and face offset intervals for classification (Aliko et al., 2020). As described by the dataset authors, after post-processing (merging overlapping detections and excluding low-confidence intervals), the final annotations were distributed as plain-text files with two columns: onset time and duration of each face event.

Using these annotations, we categorized time intervals into two event types that served as labels for classification: face onset and face offset. Face onset events were defined as intervals in which a detected face remained continuously visible on the screen for at least 10 seconds. Only events meeting this duration criterion were retained, and for each such event a 10-second fMRI segment was later extracted following a hemodynamic lag (see below).

Face offset events were defined using the transitions between consecutive face detections. Specifically, an offset event was identified when the time between the end of one face event and the start of the next was longer than 10 seconds. In these cases, the interval beginning when the face disappeared was treated as a possible face offset interval. Only intervals with at least 10 consecutive seconds without any detected faces were kept as valid face offset events.

For each movie, we first counted how many valid face offsets intervals were available (i.e., how many $g_i > 10$ s). We then took all valid face-onset events ($d_i \geq 10$ s), randomly shuffled

them, and selected an equal number to match the count of face offset intervals. This procedure produced a balanced set of face onset and face offset events for each movie.

Table 2 summarizes the number of valid face onset and face offset intervals identified across the ten NNDb films. These counts reflect the natural variation in how frequently faces appear and disappear in different narratives. All movies contained more face-onset intervals (faces on screen for ≥10 s) than offset gaps (≥10 s with no faces), but the magnitude of this difference varied across films. For example, The Prestige and 12 Years a Slave included many prolonged face-onset periods, whereas films such as Little Miss Sunshine and The Usual Suspects had fewer long face offset gaps. Because this imbalance reflects the structure of the films rather than the annotation procedure, we used the number of offset events in each movie as the limiting factor and selected an equal number of onset events by randomly sampling from all available onset intervals.

Within each movie, all selected onset and offset events were shuffled before segment extraction, ensuring that the training set did not preserve the original narrative order. This was done to better reflect the variety of naturalistic viewing, in which cognitive events do not occur in a fixed or repeated sequence, and to reduce the possibility that the CNN could rely on movie-specific ordering or recurring narrative structure. By randomizing event order separately within each film, the model was encouraged to learn event-related neural patterns rather than similarities tied to the temporal progression of a particular movie.

Table 2

Counts of Face Onset and Face Offset Events per Movie

Movie	Face Onset ≥ 10s	Face Offset gaps ≥ 10s
12 Years a Slave	147	32

500 Days of Summer	104	15
Back to the Future	121	19
Citizenfour	103	32
Little Miss Sunshine	122	9
Pulp Fiction	144	29
Split	105	28
The Prestige	154	6
The Shawshank Redemption	139	26
The Usual Suspects	104	16

Note. Counts reflect events ≥10 seconds in duration; onset counts were chosen to match offset events within each film.

The BOLD signal reflects neural activity but lags behind the underlying neural response by several seconds due to neurovascular coupling (Hillman, 2014). To account for this delay, we applied a fixed hemodynamic lag of 4 seconds (4 TRs, given TR = 1 s in this dataset). For both face onset and face offset events, fMRI segments were extracted starting 4 seconds after the event time.

From each lag-adjusted start point, a 10-second window (10 TRs) was extracted from the preprocessed fMRI time series. Segments that would have extended beyond the end of a run were excluded. Each remaining segment was labeled as either face onset or face offset based on the event that defined it. This procedure preserved the temporal structure of the BOLD signal while aligning each window with the expected peak hemodynamic response to the visual event.

2.2.4 GROUP-LEVEL BRAIN MASK

To reduce dimensionality and ensure anatomical consistency across participants, we constructed a group-level brain mask from individual T1-derived masks. For each participant, NNDb provides a binarized brain mask (sub-*_T1w_mask.nii.gz) in MNI space, where voxels inside the brain are coded as 1 and non-brain voxels as 0. These individual masks were averaged

voxel-wise across the 86 participants, and the resulting average mask was thresholded at 0.99. Voxels with values greater than 0.99 were retained, meaning they were present as part of the brain mask for at least 99% of participants. This produced a conservative group mask that excluded most boundary and non-brain voxels. The final mask contained 41,489 voxels. The thresholded group mask was then flattened into a 1D array, and the indices of nonzero entries were stored. These indices were later used to select the same set of reliable voxels from every lag-corrected 10-second segment.

2.2.5 TRANSFORMATION TO VOXEL-BY-TIME MATRICES AND FINAL DATASET ORGANIZATION

Each 10-second 4D segment $X_e(x, y, z, t)$ was first reshaped by collapsing the three spatial dimensions (x, y, z) into a single voxel dimension, while keeping the time dimension unchanged. Before masking, this flattening step converts a volume of size $X \times Y \times Z \times 10$ into a 2D array of size

$$V_{\text{full}} \times 10,$$

where $V_{\text{full}} = X \times Y \times Z$ is the total number of voxels in the 3D grid. We then applied the group-level mask by selecting only the 41,489 voxels indicated by the mask indices. This yielded a masked voxel-by-time matrix of size

$$V \times 10, \qquad V = 41{,}489,$$

in which each row corresponds to a reliable voxel shared across participants and each column corresponds to one second (one TR) within the 10-second window. This transformation retained the temporal evolution of the BOLD signal while greatly reducing the dimensionality of the spatial component and enforcing a common voxel space across subjects and movies.

The full preprocessing pipeline, event definition, hemodynamic lag correction, 10-second segment extraction, group-level masking, and voxel-by-time transformation, was applied to all participants and all movies in NNDb v2.0. Each resulting 41,489 × 10 matrix was saved as an individual file, organized by subject and condition ("face onset" vs. "face offset"). After applying all inclusion criteria, the final dataset comprised 3,606 voxel-by-time matrices (1,803 face onset and 1,803 face offset), each representing a temporally preserved pattern of whole-brain activity. These matrices served as the input to the CNN described in the next chapter.

Once the data had been converted into 2D voxel-by-time matrices, a custom deep learning model was developed to classify each 10-second fMRI window as either a face onset or a face offset event. All modeling was implemented in Python using TensorFlow/Keras 2.10, and training was carried out on an NVIDIA GeForce RTX 5090 GPU (32 GB VRAM) with CUDA 13.0. Because no pre-existing architecture was designed for the highly sparse, short-duration, and high-dimensional nature of these matrices, the network was trained entirely from scratch. This allowed the model design to be matched precisely to the structure of the fMRI inputs, which differ substantially from typical computer vision inputs in both their dimensionality and their statistical properties.

Each sample entered the network as a voxel-by-time matrix of size $V \times T$, where $V = 41{,}489$ voxels and $T = 10$ timepoints (one per second). An explicit channel dimension was added, producing an input shape of $41{,}489 \times 10 \times 1$. This representation preserved the short temporal evolution of the BOLD signal around each event while keeping a consistent spatial frame shared across all participants and movies. The entire dataset consisted of 3,606 such matrices, 1,803 face-onset segments and 1,803 face-offset segments, and was stored as a large, serialized object for efficient loading during training.

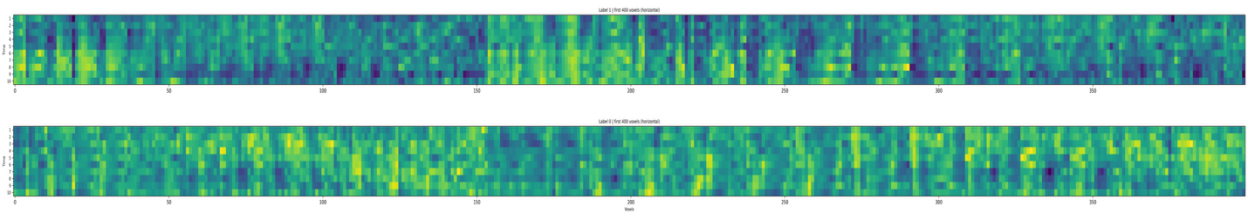

Figure 3: Example 2D Voxel-by-Time Matrices

2.2.6 DETAILED LAYER-BY-LAYER ARCHITECTURE

The final architecture consisted of three convolutional blocks followed by a fully connected classifier. Each convolutional layer used He-normal kernel initialization and L2 regularization (weight decay of 0.001), which helped stabilize training and limit the risk of overfitting. The first convolutional block contained 64 filters with a kernel size of 5×3. This relatively large temporal kernel allowed the model to detect intermediate-scale transitions such as rising or falling BOLD trends across several seconds. The activation function for all convolutional layers was Leaky ReLU (alpha = 0.1). This choice was based on preliminary tests with ReLU, ELU, Tanh, swish, and sigmoid, all of which led to either vanishing gradients or dead neurons in deeper layers. The sparse nature of fMRI data means that many input values are close to zero, and Leaky ReLU helped maintain propagating gradients even when activations were negative (Parhi & Nowak, 2020).

Each convolution was followed by batch normalization to reduce internal covariate shift and improve convergence. Max pooling was then used to down-sample the feature maps. In the first block, a pooling window of 2×1 reduced the voxel dimension by half while keeping all timepoints intact. This design preserved the short temporal resolution of the 10-second window, which would have been lost if early pooling reduced the time dimension too aggressively. A dropout rate of 0.3 at the end of the block provided additional regularization.

The second convolutional block used 32 filters with the same 5×3 kernel size. It followed the same structure, Leaky ReLU, batch normalization, max pooling with a 2×2 window, and dropout of 0.35. Reducing both the voxel and time dimensions at this stage helped compress the representation and encouraged the model to learn more abstract features rather than memorizing local noise. The third block also used 32 filters but with a smaller 3×3 kernel, allowing the model to extract fine-grained patterns after the earlier compression steps. Pooling again used a 2×2 window, and dropout was set to 0.3.

Following the convolutional layers, the output was flattened into a one-dimensional vector and passed into a fully connected dense layer with 128 units. As with the convolutional layers, this dense layer used He-normal initialization, L2 regularization, Leaky ReLU activation, batch normalization, and dropout (rate = 0.4). L2 regularization was chosen instead of L1 because fMRI data often involve distributed activity across many voxels, and L1 would encourage weights to become zero, eliminating informative small-scale patterns. By contrast, L2 keeps weights small without enforcing sparsity and is more suitable for brain-wide patterns. The final layer of the network was a single neuron with a sigmoid activation function, which produced a probability between 0 and 1 indicating whether the segment corresponded to a face onset (1) or face offset (0). Figure 2 depicts the architecture of the CNN model implemented in this study.

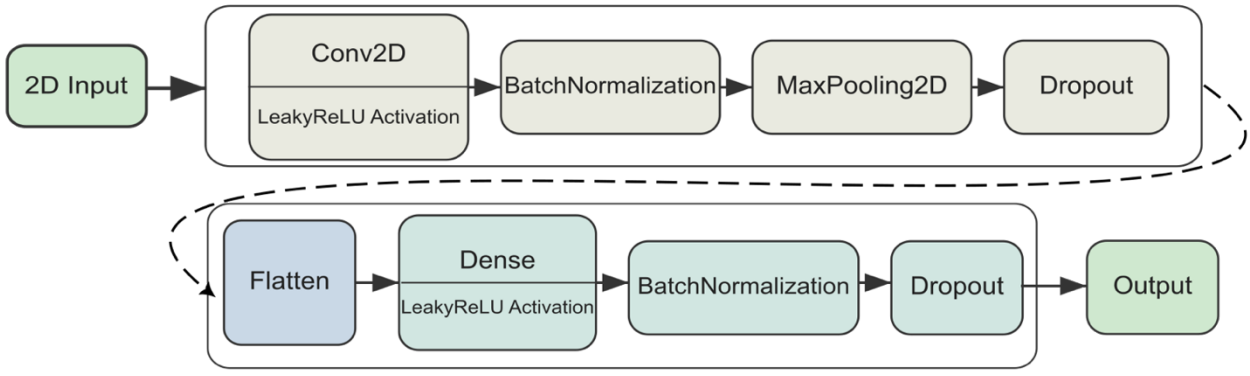

Figure 4: Convolutional Neural Network Architecture

2.2.7 REGULARIZATION STRATEGY, CROSS-VALIDATION, OPTIMIZATION AND TRAINING

Given the high dimensionality of the data and the relatively small number of samples, careful regularization was essential. The combination of L2 weight decay, dropout at multiple depths, and batch normalization formed a multi-layered defense against overfitting. L2 weight decay prevented the convolutional filters from growing excessively large, ensuring that learned features remained general rather than specific to subjects or movies (Laarhoven, 2017). Dropout encouraged the network to rely on distributed patterns rather than individual voxel-time locations, which is especially important for biological data where the informative signal is spread across distributed networks. Batch normalization provided training stability by reducing the effect of shifts in layer activations, and the max-pooling layers served as an implicit regularizer by reducing the effective feature map size. Early iterations of the architecture without these regularization components showed unstable validation curves and rapid overfitting, reinforcing the need for the multi-step strategy.

To obtain reliable and generalizable performance estimates, the model was trained using 4-fold stratified cross-validation. From the full dataset, 80 participants were used for model development, while six participants were completely excluded from training and validation. These six participants were reserved for continuous whole-time-course decoding and were never seen by the model during optimization. Each held-out participant viewed a different movie, ensuring that both subject identity and movie content were unseen during evaluation. Stratification ensured that both face-onset and face-offset samples were proportionally represented across training and validation sets in each fold. For each fold, class weights were computed to correct for minor imbalances in the training labels. The model was trained independently on each fold, with weights

randomly reinitialized each time to avoid information leakage. Each fold was trained for up to 150 epochs with a batch size of 20. After training, the best model for each fold was saved based on validation accuracy, allowing the selection of a single best-performing fold for final evaluation and visualization.

The model was trained using the Adam optimizer (Kingma & Ba, 2017) with a learning rate of 0.0002. Binary cross-entropy was used as the loss function, and accuracy was tracked as the primary metric. Training included a learning-rate scheduler (ReduceLROnPlateau) that monitored validation loss and reduced the learning rate by a factor of 0.5 when no improvement occurred for 15 consecutive epochs. The learning rate was allowed to decay to a minimum of 1×10^{-11}. This adaptive learning rate schedule stabilized training by allowing larger steps early in training while supporting finer parameter updates later. Model checkpointing ensured that only the best-performing model for each fold, based on validation accuracy, was saved. Although GPU computation introduces small amounts of nondeterminism, global random seeds were set to achieve consistent results across runs.

After all cross-validation folds were trained, the single best model (the one with the highest validation accuracy across folds) was selected as the final model. This model was saved in to enable later use in evaluation and interpretability analyses. The corresponding validation set was retained for generating performance curves and attribution analyses.

Model performance was assessed using the held-out validation data from the best-performing fold. Accuracy and loss curves across the 150 epochs were plotted to examine convergence behavior, and these curves showed smooth and stable learning dynamics with no signs of severe overfitting. The model's discriminative ability was summarized using the area under the ROC curve (AUC), computed from predicted probabilities. A confusion matrix was

produced using a probability threshold of 0.5, and a classification report was generated to summarize precision, recall, and F1-scores for both classes. To assess calibration, a reliability diagram was plotted by comparing predicted probabilities with observed event frequencies. These combined metrics provided a comprehensive summary of model performance and helped verify that the network learned meaningful and generalizable patterns associated with face onset and face offset events.

2.2.8 INTERPRETABILITY ANALYSIS WITH DEEPEXPLAIN

To understand how the CNN arrived at its predictions and to validate that the model relied on meaningful neurobiological patterns rather than noise or spurious correlations, we performed attribution-based interpretability analysis using the DeepExplain framework (Ancona et al., 2018). Interpretability is especially important in neuroimaging due to the high dimensionality of fMRI data, the distributed nature of neural representations, and the need to ensure that machine learning models reflect plausible brain processes (Bedel & Çukur, 2023). In naturalistic fMRI, where the data include complex, continuous stimuli and whole-brain responses, interpretability becomes even more essential for assessing whether the model has learned stimulus-driven or physiologically relevant structure rather than overfitting to idiosyncratic patterns (Simony & Chang, 2020).

DeepExplain is an open-source framework that provides a unified interface for several attribution methods, including Integrated Gradients (Sundararajan et al., 2017), Gradient × Input (Shrikumar et al., 2019), DeepLIFT (Shrikumar, Greenside, & Kundaje, 2017), and classical saliency maps. In this study, Integrated Gradients was used because it provides stable and theoretically grounded attributions that are less sensitive to gradient saturation and noise (Sundararajan et al., 2017). The method computes feature importance by integrating the gradients

along a straight-line path from a baseline input to the input, allowing relevant contributions to be recovered even local gradients are near zero.

$$IG_i(x) = (x_i - x_i') \int_0^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} \, d\alpha$$

This quantity reflects how much each voxel–time feature x_i contributes to the model output relative to the baseline x'.

In this study, Integrated Gradients was used to quantify the contribution of each voxel and time point in the 10-second window to the model's classification outcome. The baseline input for all analyses was an all-zero matrix of the same size as the input (41,489 voxels × 10 time points × 1 channel). A zero baseline was chosen because it corresponds to a conceptual "absence of signal" state and is recommended when no natural baseline exists for the domain. Integrated Gradients then estimates attributions by evaluating the gradients at many points between this baseline and the actual sample, averaging these gradients across 200 interpolation steps. Using a high number of steps helped ensure smooth and stable estimates, which is particularly important for high-dimensional fMRI representations. The numerical approximation for Integrated Gradients used in this study was:

$$IG_i(x) \approx (x_i - x_i') \frac{1}{m} \sum_{k=1}^{m} \frac{\partial F\left(x' + \frac{k}{m}(x - x')\right)}{\partial x_i}, m = 200.$$

The CNN used a sigmoid activation at the final layer, which compresses its output into the $[0, 1]$ probability range. Working directly with these post-activation probabilities during attribution can lead to saturation issues and reduced sensitivity. To avoid this, attributions were

calculated with respect to the pre-activation logit, the value computed by the final dense layer before applying the sigmoid. The logit corresponds to the linear transformation:

$$z = W^{\top}h + b,$$

where h is the activation vector from the final hidden layer. Because z is unbounded, it is more sensitive to meaningful feature differences than the sigmoid output. The sigmoid activation,

$$\sigma(z) = \frac{1}{1 + e^{-z}},$$

has a derivative

$$\frac{d\sigma}{dz} = \sigma(z)(1 - \sigma(z)),$$

which approaches zero when the model is confident, making gradient-based attribution unstable. By computing Integrated Gradients on the logit instead, we avoided this saturation problem. For samples belonging to the face onset class (class 1), Integrated Gradients was computed directly on z. For face-offset samples (class 0), we used the negative logit:

$$F_{\text{offset}}(x) = -z,$$

so that attribution values consistently reflected evidence supporting the true class. This avoided asymmetric interpretations where only features increasing the face probability would be highlighted.

Because DeepExplain was developed for TensorFlow 1.x, the analysis was performed in TF 1.x compatibility mode with eager execution disabled. The entire model was rebuilt in TensorFlow's graph mode using the exact same architecture and weights as the best-performing CNN from cross-validation. Batch normalization and dropout layers were automatically set to

inference mode using K.set_learning_phase (0) to ensure that random behavior was disabled and that attribution maps would be stable and reproducible. Each fMRI segment was loaded from the saved NumPy files as a 2D voxel-by-time array and reshaped into the expected 4D input format $(1, 41{,}489, 10, 1)$.

Integrated Gradients was then computed for each individual sample, yielding an attribution map with dimensions $(1, 41{,}489, 10, 1)$. Each entry in this matrix represented how much a specific voxel at a specific time point contributed to the final decision. Positive values indicated features that pushed the model toward classifying the segment as a face onset, while negative values indicated features that supported the face offset class. All computations were performed in an isolated inference session within DeepExplain to ensure consistent graph states across samples.

Compared to raw gradient maps, the Integrated Gradients results were markedly more stable and interpretable. Raw gradients tended to highlight scattered and noisy voxel–time patterns, while Integrated Gradients produced smoother, distributed attribution profiles more consistent with known properties of the BOLD response. The method revealed which brain regions and time points contributed most strongly to the model's decisions, enabling later reconstruction of these attributions back into 3D brain space for visualization and neuroscientific interpretation. These results formed the basis for the attribution maps and cortical projections presented in Chapter 3, where the spatial patterns of model attention are examined in detail for both face-onset and face-offset conditions.

2.2.9 RECONSTRUCTING NEURAL ATTRIBUTIONS IN BRAIN SPACE

The Integrated Gradients procedure produced attribution maps in the same format as the model inputs: a 2D voxel-by-time matrix with dimensions $41{,}489 \times 10$ which it does not directly

reflect the anatomical structure of the brain. To interpret the results in a neuroscientific context, it was necessary to reconstruct these voxel-level attributions back into a full 3D+time NIfTI image for each sample. This step allowed every Integrated Gradients attribution map to be visualized in native MNI space and compared across subjects.

The reconstruction process made use of the same group-level brain mask used during preprocessing. This mask contained 41,489 nonzero voxels, representing brain locations that were present in the brain masks for at least 99% of participants. During the CNN preprocessing stage, only these voxels were retained; therefore, each Integrated Gradients attribution vector corresponded exactly to this mask index set. To revert the attributions back into three-dimensional space, a zero-filled 3D grid matching the original mask shape was created, and the Integrated Gradients values were inserted into the appropriate voxel locations according to the mask indices. This process was repeated separately for each of the ten timepoints in the Integrated Gradients matrix, resulting in a 4D volume with dimensions $X \times Y \times Z \times 10$. The affine transformation from the original mask was preserved so that the reconstructed images remained anatomically aligned.

After reconstruction, attribution maps were averaged within each subject across all segments belonging to the same condition (face onset or face offset). Because all reconstructed maps shared the same spatial grid and alignment, this averaging could be performed directly without additional registration. This procedure yielded subject-level attribution maps that summarize consistent, model-relevant neural patterns over time.

Together, this reconstruction and averaging process transformed the model's attribution outputs into anatomically interpretable brain maps, allowing examination of which brain regions most strongly contributed to distinguishing face-onset and face-offset events. These maps form the basis for the individual- and group-level analyses presented in Chapter 3.

2.3 RESULTS

2.3.1 MODEL PERFORMANCE

The CNN model was evaluated using 4-fold stratified cross-validation on 80 participants with 3,462 voxel-by-time matrices (1,731 "face onset" and 1,731 "face offset" segments). Across folds, the model demonstrated robust performance in classifying face onset versus face offset segments from naturalistic fMRI data. The final model, selected based on the highest validation accuracy across folds, achieved an average validation accuracy of 82.2% (SD = 1.1%), with a validation loss of 0.829 and an area under the receiver operating characteristic curve (AUC) of 0.892. These metrics indicate strong discriminative ability, particularly given the complexity and variability of naturalistic stimuli. Detailed classification metrics from the final evaluation on the held-out test set are summarized in **Table 3**. Precision, recall, and F1-score were balanced across classes, with slightly higher precision for the "face onset" class (0.823) compared to "face offset" (0.819), suggesting the model was marginally more sensitive to detecting face-related neural activity. The macro-averaged F1-score of 0.82 reflects consistent performance across both conditions, accounting for any minor class imbalances in the test split.

Table 3

Classification results for face onset vs. offset on unseen data

Class	Precision	Recall	F1-Score	Accuracy	Support
0 (Face Offset)	0.819	0.824	0.821	0.821	433
1 (Face Onset)	0.823	0.817	0.820	0.820	433

Training convergence was stable, with loss decreasing steadily across epochs and no evidence of overfitting, as indicated by comparable training and validation curves. The use of class weights and regularization techniques contributed to this balance, ensuring the model generalized

well to unseen data from diverse movies and participants. Performance visualizations are provided in Figure 5, which includes (A) the confusion matrix, (B) ROC curve, and (C) the reliability diagram (calibration curve). The reliability diagram shows that the model's predicted probabilities are reasonably well-calibrated, closely following the ideal line, though with slight overconfidence in higher probability bins. The ROC curve confirms the high AUC of 0.89, demonstrating excellent separability between classes. The accuracy and loss curves illustrate steady improvement during training, with validation accuracy stabilizing around 0.82 and validation loss at 0.8, suggesting potential for further optimization, such as increased regularization to narrow the gap between training and validation loss. Together, these performance metrics indicate that the model reliably distinguishes neural states associated with the presence or absence of faces, motivating the subsequent interpretability analyses identifying which brain regions drive these decisions.

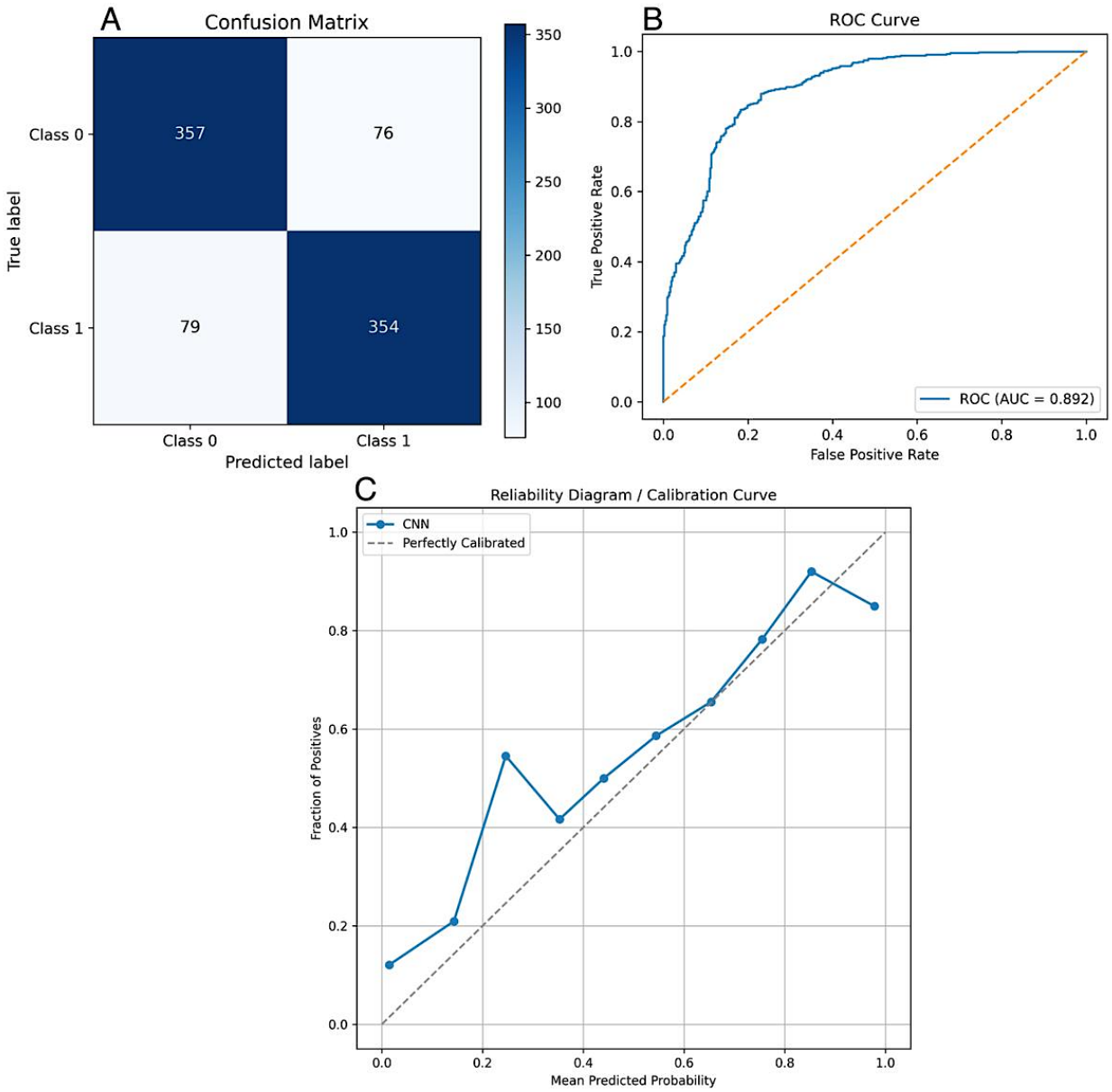

Figure 5: Model Performance Metrics for CNN-Based Face Classification

2.3.2 ATTRIBUTION PATTERNS

To understand how the model responds to individual neural events, the attribution patterns were examined for face-onset and face-offset segments. These single-trial attributions provide a direct view of how the convolutional network interprets the voxel-by-time representation moment by moment. Examining single trials is important because it shows whether the model relies on

consistent and meaningful neural signals rather than noise or artifacts. For face-onset events, single-trial attributions showed activity in known face-selective regions as the event progressed through the ten-second window. Even without any spatial smoothing or temporal averaging, many trials exhibited concentrated attribution in the right fusiform face area, suggesting that face appearance elicited reliable neural responses that the model leveraged for classification.

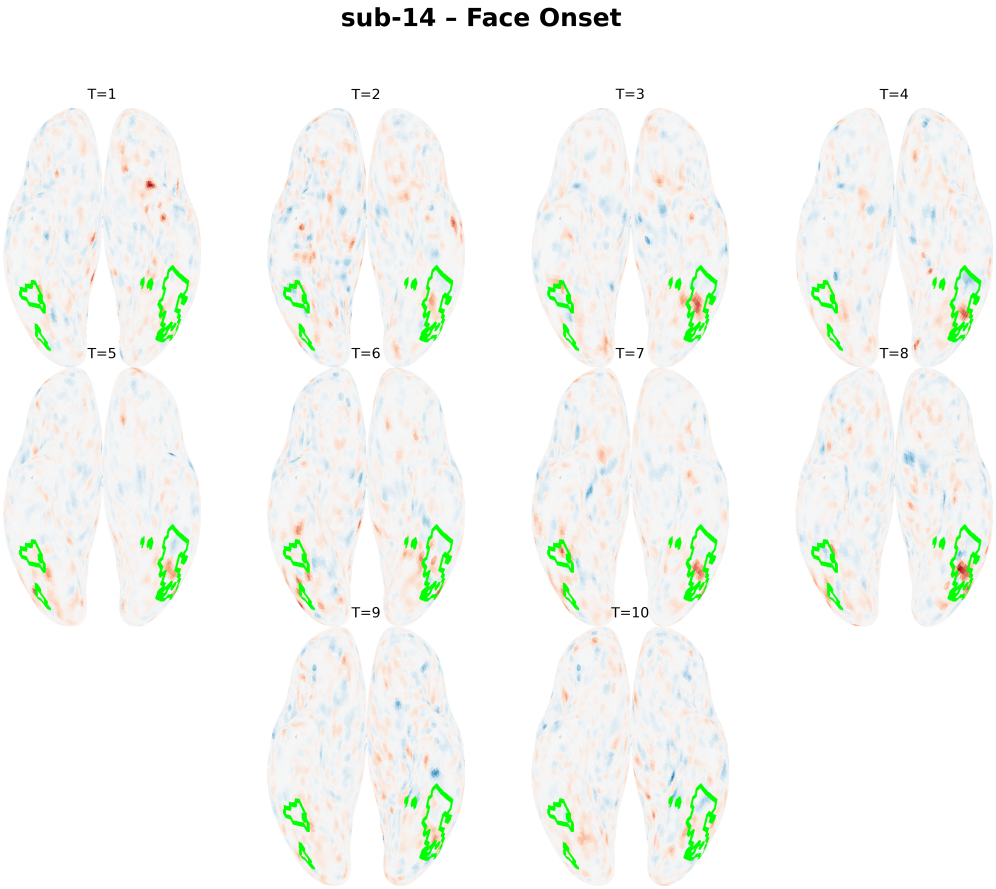

Figure 6: Single-Trial sample Subject 14 Attribution Patterns Face onset

In contrast, single trials of face-offset events showed reduced attribution values in the same regions, and in many cases, negative attribution values emerged in the fusiform and occipital face areas. These negative attributions indicated that decreased activity in these regions contributed evidence toward the offset prediction. Nevertheless, the pattern was still sufficiently distinct for the classifier to separate the two classes at the single-trial level.

Although single-trial maps naturally contained more variability than subject-averaged maps, the core structure of the face-processing network was clearly visible in the majority of them. Some trials exhibited small attribution clusters outside the canonical face-processing system, which is expected given the complexity of naturalistic scenes, variability in eye movement, and differences in visual context across events.

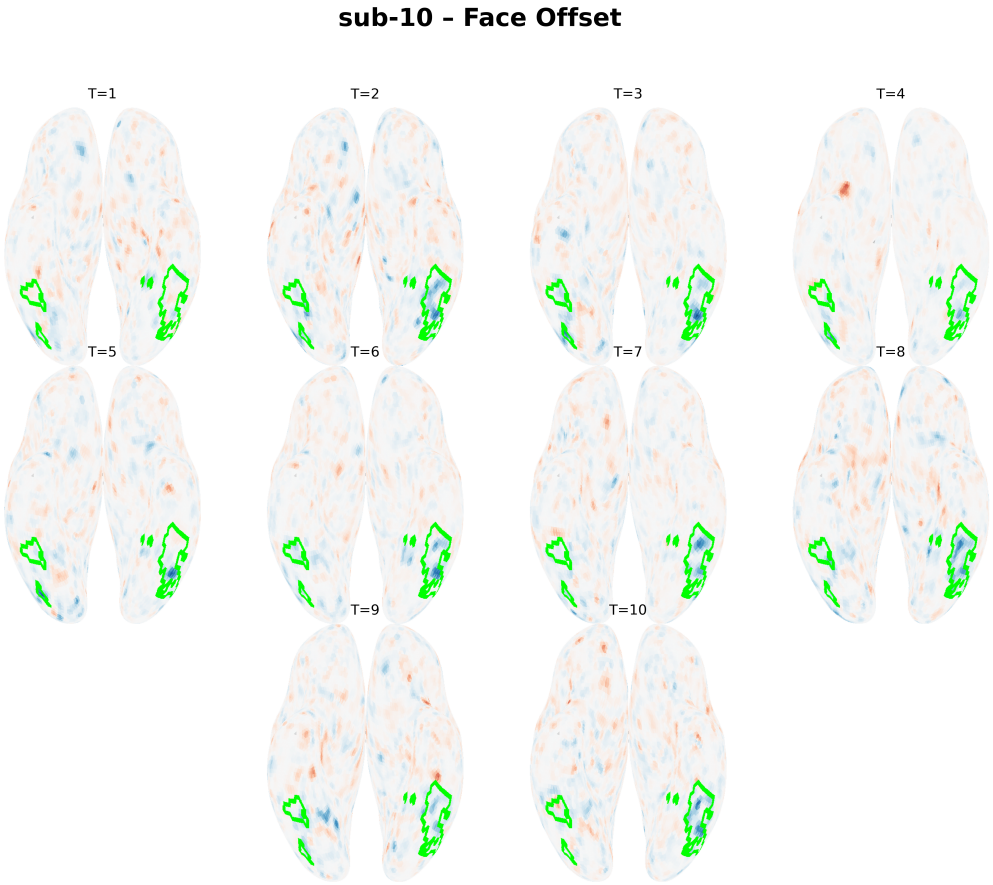

Figure 7: Single-Trial sample Subject 10 Attribution Patterns Face offset

Averaging face-onset and face-offset events within each participant reduced noise and revealed stable attribution patterns at the individual level. Across the 80 participants, attribution maps were highly consistent despite substantial variation in the viewed films, including lighting, motion, and scene transitions. During face-onset events, strong positive attribution values consistently appeared in the fusiform face area, occipital face area, and surrounding ventral

temporal cortex. These regions emerged as the dominant contributors to onset classification across subjects.

The ability to capture individual variability is a strength of this approach rather than a limitation, as it reflects sensitivity to meaningful differences in neural organization across participants. Accordingly, the spatial extent of these patterns varied between individuals, ranging from more focal to more distributed fusiform involvement, the same core face-selective regions were observed in nearly all participants. Temporal analysis further showed a stable progression of attributions over time. This consistent temporal and spatial structure across participants indicate that the model relied on meaningful neural differences between onset and offset conditions rather than stimulus-specific artifacts.

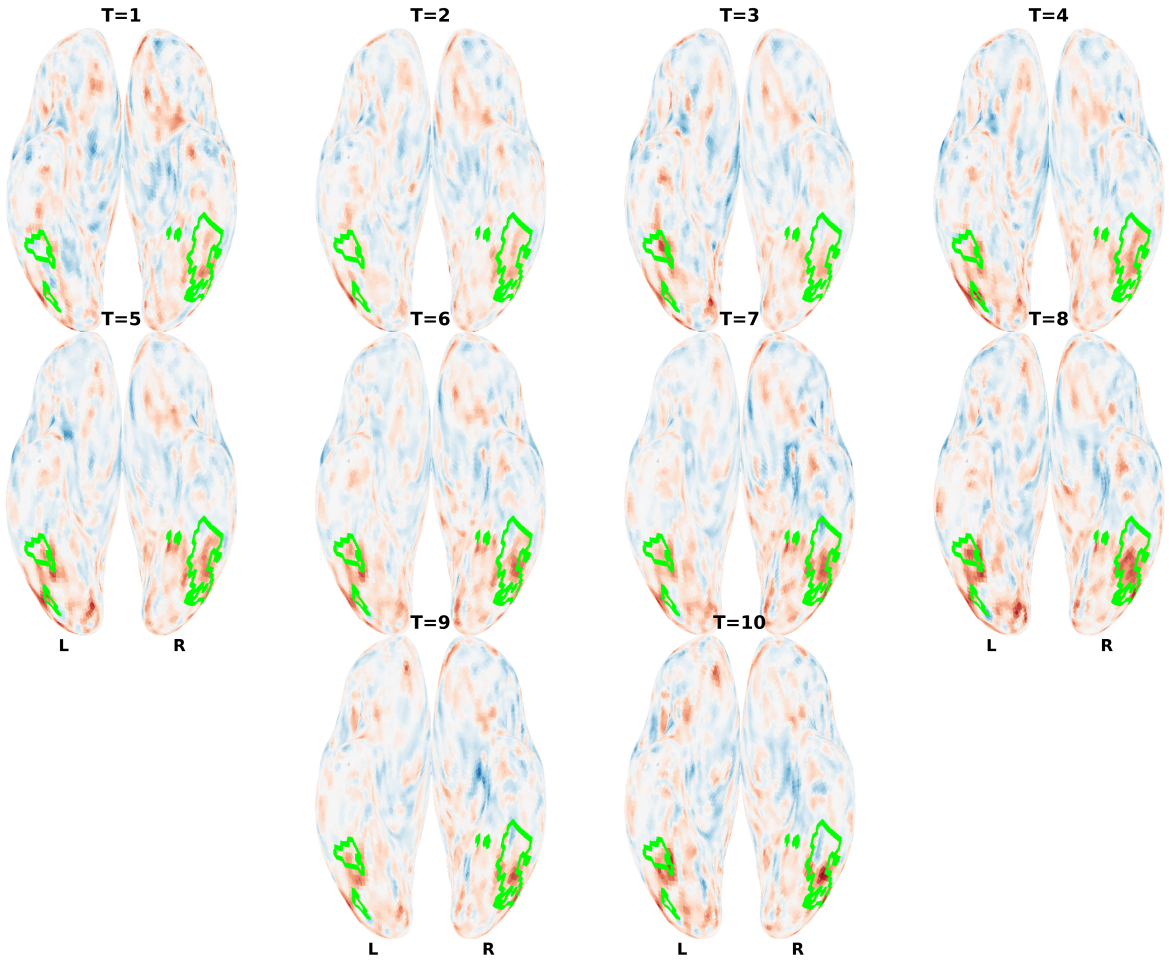

Figure 8: Individual-Level Attribution Maps sample subject 33 face onset

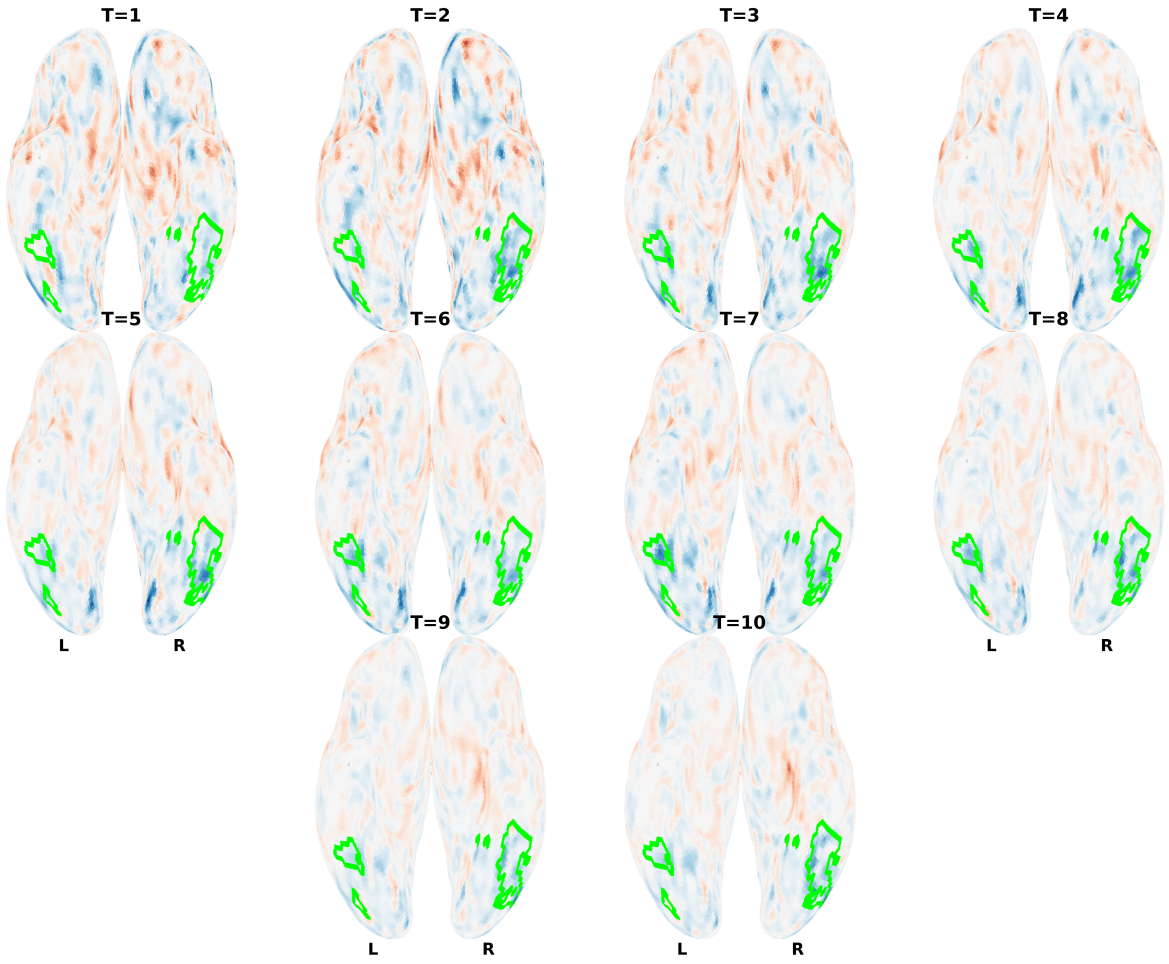

Figure 9: Individual-Level Attribution Maps sample subject 33 face offset

Group-level maps provide a population-level perspective on the neural signals that the classifier relied on. The group-averaged map for face-onset events showed a clear and highly concentrated activation pattern in the fusiform face area, with particularly strong lateralization toward the right hemisphere. The occipital face area and posterior superior temporal sulcus also displayed robust attribution peaks, forming a spatially continuous network that closely mirrors the canonical face-processing system described in prior fMRI studies. The concentration of attribution in these regions indicates that the classifier consistently relied on ventral visual stream activity to identify the presence of a face.

The group-level offset map showed the inverse structure of the onset map, with negative attributions appearing prominently in the fusiform and occipital face areas, demonstrating that reductions in activity contributed to the model's classification of face absence. Although the magnitude of the offset pattern was smaller than that of the onset pattern, the spatial structure remained clearly distinguishable. The close correspondence between attribution patterns and established face-processing regions supports the interpretability of the model. The consistent right-lateralization observed across subjects and in the group average aligns with well-documented hemispheric specialization for face processing, indicating that the convolutional neural network was sensitive to meaningful neural features rather than artifacts.

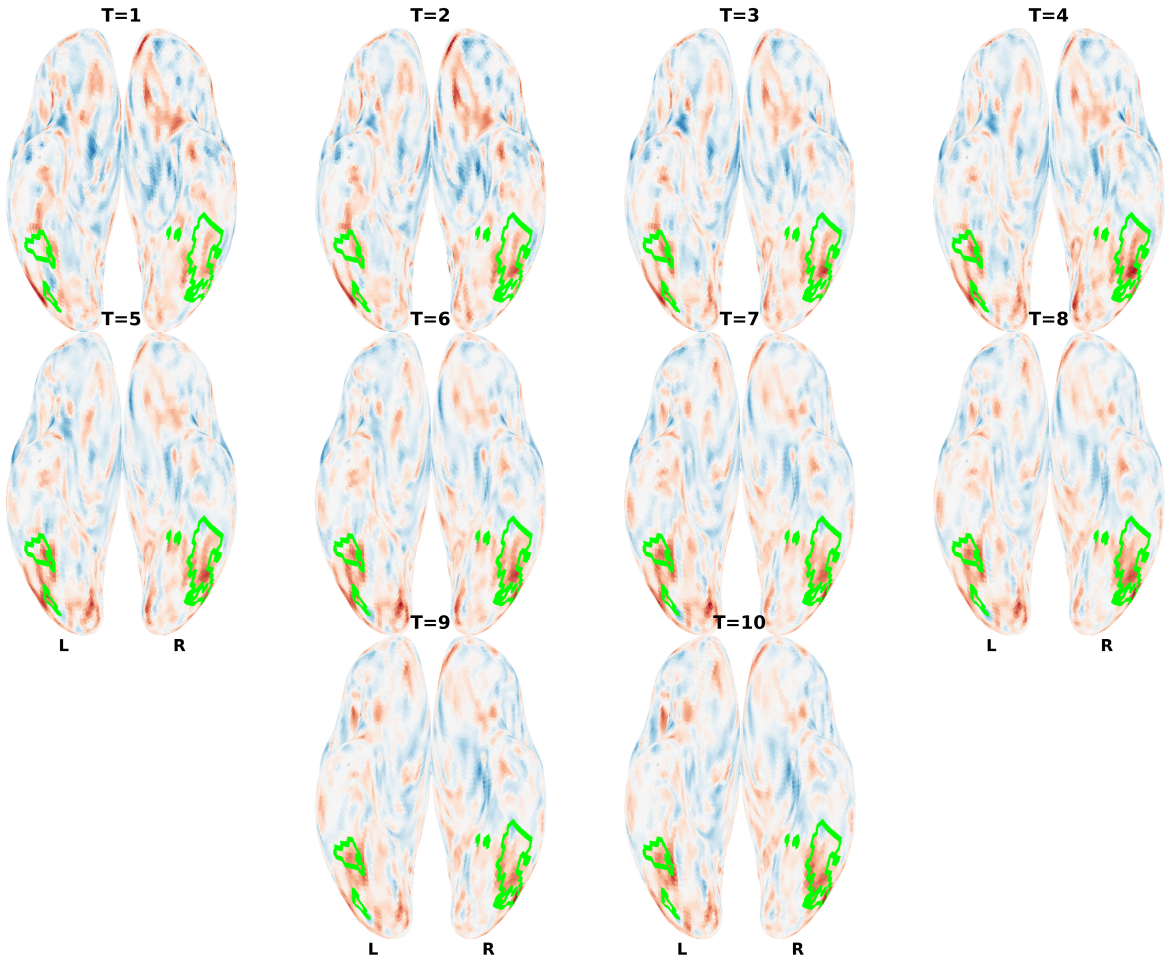

Figure 10: Group-averaged Integrated Gradients attributions after a face appears

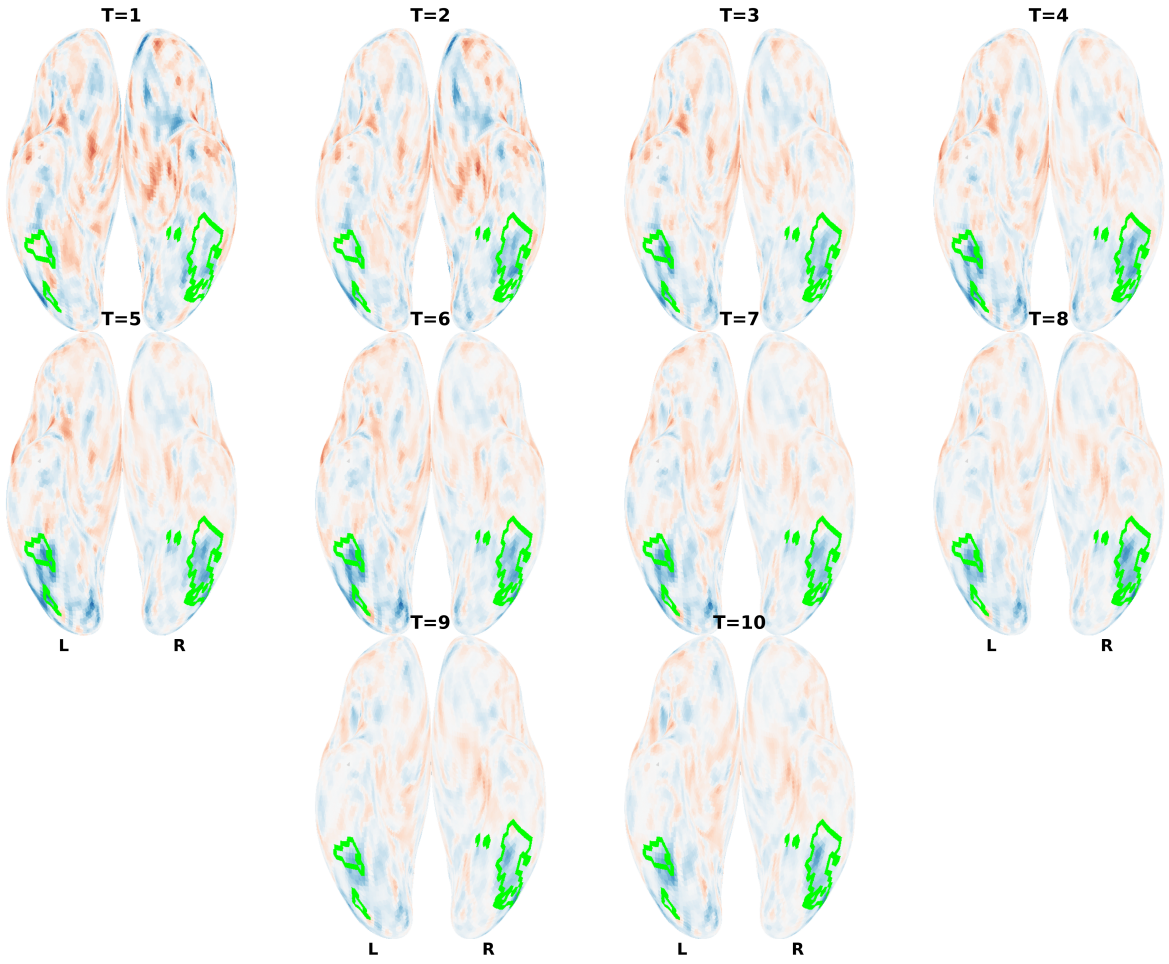

Figure 11: Group-averaged Integrated Gradients attributions after a face disappears

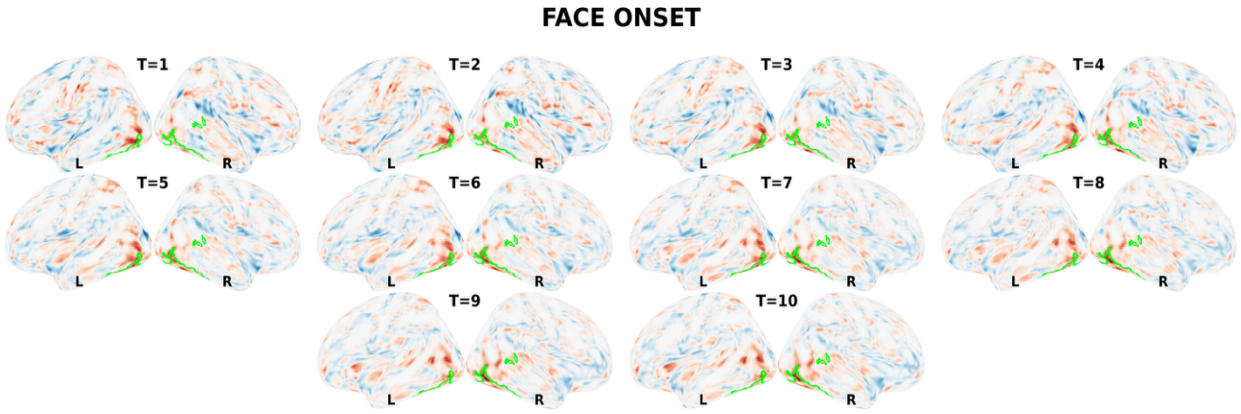

Figure 12: Lateral view of group-averaged attribution patterns following face onset

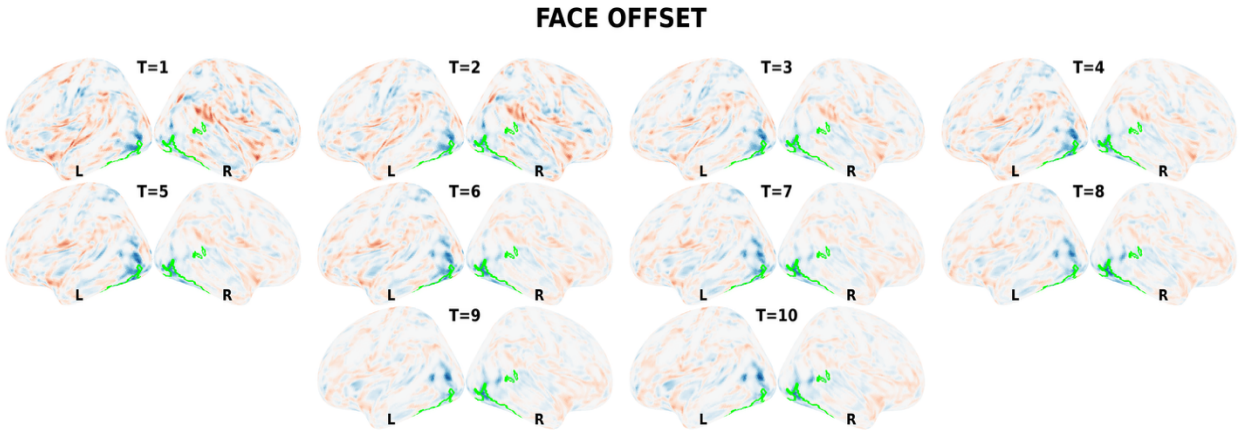

Figure 13: Lateral view of group-averaged attribution patterns following face offset

2.3.3 CONTINUOUS TIME-COURSE DECODING DURING NATURALISTIC MOVIE VIEWING

To evaluate whether the trained CNN could generalize beyond isolated face-onset and face-offset events, we performed continuous time-course decoding analysis using the complete fMRI sequence from six unseen participants each from different movies. This analysis was designed to

naturalistic viewing conditions, in which faces appear and disappear over extended periods and neural responses evolve continuously rather than as discrete trials. Unlike the event-based windows used during training, the model was required to generate predictions at every second across long segments of the movie, allowing evaluation of moment-by-moment sensitivity to face-related neural activity.

Across the unseen participants, the model achieved a mean classification accuracy of 83.2%. Performance was consistent across individual films, with accuracies of 85% for 500 Days of Summer, 85% for Pulp Fiction, 82% for The Usual Suspects, 84% for Little Miss Sunshine, and 80% for The Prestige. Notably, trained CNN used here was trained on a very limited amount of data, consisting of a small number of short, 10-second event-centered windows (e.g., ~20 segments per movie for each participant). Despite this, the trained model successfully generalized to continuous whole-movie decoding spanning approximately 8,000 seconds of fMRI data. This strong performance indicates that the model learned meaningful and stable neural representations rather than overfitting to short training segments.

For each subject, the entire BOLD time series was segmented into overlapping 10-second windows with a 1-second stride. A fixed 4-second hemodynamic lag was applied so that each voxel-by-time matrix was aligned with the expected peak of the BOLD response to visual input. These windows were passed to the pretrained CNN without additional fine-tuning, and produced a continuous probability estimate for the presence of a face at each time point. The resulting probability trace reflects the model's evolving confidence about whether a face was present on the screen throughout the movie.

The model produced clear and interpretable temporal transitions. Predicted probabilities increased following face appearance and decreased after face offset, with transition points

generally aligned with the annotated ground truth. During sustained face-present intervals, the model typically maintained elevated probabilities, although smaller fluctuations were observed. Such variability is expected in naturalistic data, because long face presentations contain fewer sharp changes in the signal than face onsets, especially for a model trained to detect onset–offset events. Importantly, performance did not degrade during these intervals, indicating preserved discriminability across the full-time course.

Because the full movie time course spans approximately 8,000 s and is difficult to interpret at full scale, a zoomed segment from t = 2400 s to t = 3100 s is shown for visualization (Figure 14). Figure 14 presents the thresholded binary version of the model's predictions, indicating when each moment is classified as "face" or "no face." The binary trace closely aligns with the annotated ground-truth intervals, highlighting the model's ability to detect face-related neural activity on a near second-by-second basis. Together, these results demonstrate that the CNN can track meaningful cognitive events continuously over time, despite being trained only on isolated 10-s onset and offset segments.

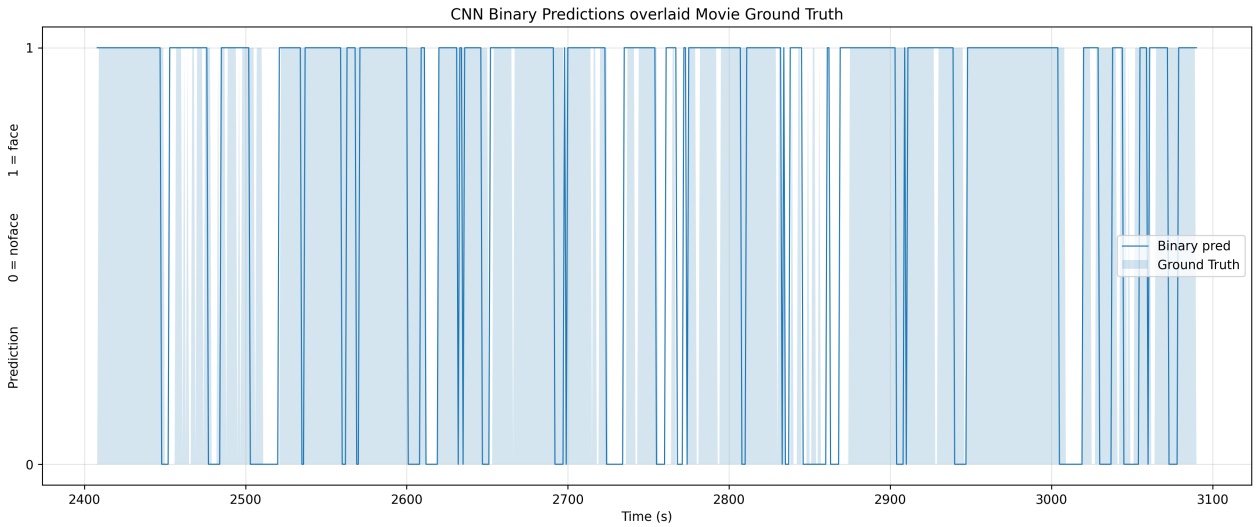

Figure 14: Binary Face/No-Face Prediction During Naturalistic Viewing

CHAPTER 3: GENERAL DISCUSSION

3.1 SUMMARY OF FINDINGS

This thesis investigated whether preserving temporal structure in fMRI representations enables efficient and biologically meaningful decoding of cognitive events during naturalistic movie watching. To do so, fMRI data were reorganized into voxel-by-time matrices that retained the temporal changes of the BOLD signal following face onset and face offset events. A lightweight convolutional neural network was trained on these representations preserving temporal dynamics to distinguish between face appearance and disappearance at the event level.

The proposed framework achieved reliable classification performance across participants and movies, demonstrating that proposed voxel-by-time matrices contain sufficient information to distinguish between different perceptual stimuli (i.e., face onset vs. face offset). Importantly, the model generalized beyond isolated event windows and successfully tracked face presence during continuous movie viewing in unseen participants, indicating sensitivity to moment-to-moment neural dynamics rather than reliance on static or stimulus-specific features.

Across participants and movies, face-related neural patterns were reliably identified, and interpretability analyses showed that these patterns were driven by activity in established face-processing regions, including the FFA and OFA. Attribution strength followed a temporally structured profile, suggesting that the preserved temporal information contains meaningful signals about the timing and organization of face perception in the brain.

3.2 INTERPRETATION OF FINDINGS

A central interpretation is that preserving temporal structure in proposed voxel-by-time fMRI representations enables reliable, moment-to-moment decoding of naturalistic cognitive events. This work is positioned as a proof-of-concept framework, showing feasibility and biological grounding through interpretable attributions, rather than as a benchmark optimized for maximum performance. Many fMRI decoding pipelines effectively compress time for example, by averaging over windows or using single volumes, which can hide meaningful dynamics that unfold after an event (Boynton et al., 1996; Hutchison et al., 2013).

In contrast, the voxel-by-time representation used here retains short-term temporal evolution in the BOLD signal, allowing the model to access how neural responses change across several seconds following. This design reflects the view that cognition is best characterized as a temporally evolving pattern of distributed activity (Dale & Halgren, 2001; Kriegeskorte & Douglas, 2018). By preserving within-window temporal structure, the CNN distinguished face-onset and face-offset events that engage overlapping spatial regions but differ in the direction and progression of their neural responses. Thus, discriminative information was available in the temporal evolution of activity across the 10-second window.

A key interpretive result is the close correspondence between decoded signals and known face-selective brain systems. Attribution analyses consistently highlighted canonical regions such as the FFA, OFA, and pSTS at both the individual trial and group level, and the contrast between face onset and face offset produced complementary spatial patterns. This mirrored structure serves as an internal consistency check, suggesting that the same neural systems contribute differently depending on whether face-related information is increasing or decreasing. Such alignment is especially important in high-dimensional naturalistic data, where decoding success can sometimes

arise from confounds unrelated to cognition, including motion, scanner drift, or run structure (Samek et al., 2017; Hosseini et al., 2020).

Finally, the model's ability to generalize from isolated event windows to continuous time-course decoding reinforces this interpretation. Although the CNN was trained only on short, event-centered segments, it successfully tracked face presence during extended naturalistic viewing in unseen participants and movies. This performance indicates that the learned voxel-by-time fMRI data representations are sufficient to support reliable decoding in naturalistic fMRI data. Together, these whole-movie results support the central claim of this framework: that a lightweight and computationally efficient CNN can decode high-level visual events from fMRI on a moment-to-moment basis in naturalistic settings.

Together, these findings indicate that temporal information in fMRI carries essential information about how cognitive states unfold over time. By preserving short-term temporal structure, the proposed framework enables efficient decoding of dynamic perceptual events while remaining aligned with known neural organization and physiological constraints, supporting its suitability for studying cognition in naturalistic settings.

3.3 RELATION TO PRIOR LITERATURE

The present study relates to three overlapping literatures: research on naturalistic fMRI, studies of face perception, and computational approaches for decoding brain activity. Each of these literatures has developed largely in parallel, and relatively few studies explicitly integrate all three. As a result, there remains an open methodological challenge: how to decode time-varying cognitive events from realistic stimuli while retaining interpretability and biological grounding.

This thesis addresses this gap by adapting deep learning tools to the specific constraints and opportunities of naturalistic fMRI data.

This work extends a growing literature emphasizing the value of naturalistic stimuli in revealing reliable and time-locked neural dynamics during continuous perception. Movie-based fMRI has been shown to drive rich, synchronized activity across visual, social, and narrative networks (Hasson et al., 2004), and is increasingly viewed as a powerful alternative to highly controlled experimental paradigms (Nastase et al., 2020; Sonkusare et al., 2019). However, many decoding approaches applied to naturalistic data rely on temporally averaged or collapsed inputs, limiting sensitivity to transient event-level dynamics (Simony & Chang, 2020; Bottenhorn et al., 2019).

Deep learning approaches have previously been applied to fMRI using convolutional architectures that preserve spatial structure, including 3D and 4D CNNs designed to model spatiotemporal data. Such models have been used for tasks ranging from visual decoding to disease classification and functional connectivity estimation (Vieira et al., 2017; Yin et al., 2022). While these approaches demonstrate the expressive power of CNNs, they often require large datasets and extensive computational resources, and they can be difficult to interpret due to architectural complexity.

In contrast, the approach adopted in this thesis emphasizes representational simplicity rather than architectural depth. By flattening spatial dimensions and retaining time explicitly, the model avoids the need for full spatiotemporal convolutions while still capturing event-related dynamics. This design choice aligns with recent work showing that careful feature organization can sometimes yield greater gains than increasing model complexity, particularly in high-dimensional, low-sample regimes such as fMRI (Alvarez-Gonzalez & Mendez-Vazquez, 2022).

Prior work applying deep learning to face perception in fMRI has largely focused on controlled experimental paradigms or static image presentations. CNN-derived features have been used to map hierarchical organization along the ventral visual stream (Güçlü & van Gerven, 2015; Eickenberg et al., 2017), and deep generative models have reconstructed faces from fMRI activity patterns (VanRullen & Reddy, 2019). While these studies demonstrate that face-related information is decodable from fMRI, they typically rely on isolated stimuli and temporally collapsed responses.

The present work differs in two keyways. First, it targets event-level decoding within continuous movie viewing rather than trial-based classification. Second, it explicitly models short temporal trajectories, allowing the network to learn how face-related signals unfold over time. By demonstrating that a lightweight CNN can recover canonical face-processing regions under naturalistic conditions, this thesis extends prior face-decoding work into a more ecologically valid domain while retaining interpretability.

3.4 STRENGTHS AND CONTRIBUTIONS

A major strength of this project is its use of naturalistic fMRI, which captures neural responses under more true-to-life audiovisual conditions. By analyzing brain activity during continuous movie viewing, the study captures neural responses as they unfold in settings that more closely resemble everyday experience, rather than relying on simplified or tightly controlled stimuli. This ecological validity strengthens the relevance of the findings for understanding how face-selective networks operate in real-world contexts.

A central contribution is designing a framework that balances biological validity, temporal fidelity, and computational efficiency for decoding fMRI signal. Representing each 4-dimensional

fMRI as a voxel-by-time matrix made it possible to examine how neural responses evolve following the appearance and disappearance of faces, rather than reducing activity to static or time-averaged summaries. This approach revealed that face-related information is carried not only by which brain regions are engaged, but also by how activity within these regions unfolds over several seconds. Temporally structured responses were consistently observed in canonical face-processing areas including the FFA, OFA, and pSTS, with timing aligned to known hemodynamic delays and the hierarchical organization of the ventral visual stream.

Importantly, these insights were evident even at the level of individual events, without requiring extensive temporal averaging. Single-trial analyses demonstrated that our representation with preserved temporal information can reveal stable and interpretable neural structure directly from high-dimensional fMRI data.

More broadly, this work contributes a framework for studying cognition as a temporally continuous process. Rather than focusing on isolated trials or static representations, the approach emphasizes how distributed patterns of brain activity evolve over time in response to meaningful events embedded within ongoing experience. By demonstrating that face-related neural dynamics can be decoded and interpreted at the level of individual events during naturalistic viewing, this study bridges research on face perception, naturalistic fMRI, and time-resolved neural analysis, and provides a foundation for future work aimed at understanding dynamic brain function in real-world settings.

An additional strength of this work is that this work relies on openly available data and provides open-source code, supporting transparency, reproducibility, and reuse by the broader research community. The complete preprocessing, modeling, and interpretability pipeline has been made openly available at https://github.com/ekstrandlab/NeuroFaceCNN, allowing researchers to

reproduce results, examine methodological choices, and adapt the framework to new datasets or cognitive domains.

3.5 LIMITATIONS

Several limitations of the present framework should be considered when interpreting the results. First, the voxel-by-time representation necessarily trades anatomical structure for computational simplicity. Flattening the spatial dimensions makes event-level learning feasible and allows the model to scale to naturalistic datasets with many participants and long recordings. However, this representation reduces explicit modeling of local spatial neighborhoods and may obscure fine-grained spatial patterns within face-selective cortex. As a result, while the framework is well suited for capturing distributed and temporally structured neural signals, it may be less sensitive to subtle spatial distinctions that would be better captured by fully spatial or hybrid spatiotemporal models.

Second, the use of 1.5 T MRI data represents a methodological constraint. Lower field strength typically yields reduced signal-to-noise ratio and spatial specificity compared to higher-field acquisitions (e.g., 3 T or 7 T). Stronger field strength could therefore improve sensitivity to face-selective responses and potentially enhance decoding performance. Future work should evaluate whether similar models benefit from higher-field naturalistic fMRI datasets.

Third, a methodological limitation of the present framework is that in our proposed representation, the three-dimensional spatial structure of the brain was flattened into a one-dimensional voxel axis prior to CNN input. Although voxel identity and anatomical alignment were preserved using a conservative group-level mask, explicit 3D spatial adjacency between neighboring voxels was not maintained for model learnings. Consequently, the convolutional

filters operated over ordered voxel indices rather than true anatomical neighborhoods, which may have reduced sensitivity to fine-grained local spatial organization and potentially constrained peak classification accuracy. However, this design choice reflected a deliberate trade-off. Implementing full 3D or 4D convolutions over whole-brain fMRI data would have substantially increased parameter count and memory demands, raising the risk of overfitting given the limited number of event-level samples. By flattening the spatial dimension while preserving the temporal evolution of the BOLD signal, the model prioritized distributed spatiotemporal patterns and computational efficiency. Importantly, the reconstructed attribution maps still aligned with canonical face-selective regions such as the FFA, OFA, and pSTS, suggesting that meaningful spatial information was implicitly captured despite the absence of explicit 3D convolutions.

A further limitation is that the CNN was evaluated on a single dataset (NNDb v2.0; Aliko et al., 2020), despite the use of multiple films and a large number of participants. Thus, generalization was tested across subjects and movies within the same acquisition and preprocessing pipeline, but not across truly independent datasets with different scanners, acquisition parameters, preprocessing choices, or annotation pipelines. This limits conclusions about robustness to broader domain shifts, such as changes in TR, spatial resolution, noise characteristics, or preprocessing strategies. This issue is common in neuroimaging machine learning, where models often generalize well within a dataset but degrade when applied out of sample. Extending this framework to additional naturalistic fMRI datasets will be critical for assessing replicability and broader applicability.

Another limitation concerns the event labels used for training. Face onset and offset events were derived from automated face detection tools (AWS Rekognition and Google Cloud Video Intelligence), which indicate only the presence of at least one face on the screen. These labels do

not capture what participants were actually attending to at a given moment. During naturalistic viewing, participants may fixate away from faces, blink, attend to other objects, or process scenes at different cognitive levels (e.g., background context versus facial identity). In complex social scenes, faces may also be present but not perceptually central. Consequently, the model cannot disentangle different components of face processing, such as gaze direction, number of faces, identity, gender, emotional expression, or social context.

Finally, the dataset exhibits an imbalance between face onset and face offset events. Naturalistic movies tend to include frequent or prolonged face exposure, resulting in fewer long intervals without faces. In this study, the limited number of usable face offset events constrained the dataset and prevented randomization of offset events within movies. Although this could, in principle, bias the model toward better prediction of offset events, such an effect was not observed. Nonetheless, this imbalance restricts the diversity of training samples and may underrepresent variability in face-related conditions (e.g., lighting, viewing angle, or scene dynamics). More broadly, face offset events are inherently heterogeneous, as the absence of faces can coincide with a wide range of scene content (e.g., landscapes, objects, text, or motion), increasing within-class variability and making this condition more difficult to model.

3.6 FUTURE DIRECTIONS

Future research can build on the present framework by addressing its limitations and extending its neuroscientific scope. Although face perception served as a validation target, the voxel-by-time representation is domain-agnostic. Any event with temporal structure embedded in continuous experience-emotion, language, social interaction, narrative shifts-could in principle be decoded using the same framework. A key next step is evaluating generalization across independent naturalistic fMRI datasets collected at different sites and with varying scanners,

acquisition parameters, and preprocessing pipelines. Such cross-dataset validation would help determine whether the observed temporal decoding patterns reflect robust neural signatures of face perception or are partly driven by dataset-specific factors. Large, multi-site public datasets would also enable more detailed analyses of individual variability in temporal dynamics.

Translational applications represent an additional promising direction. Many psychiatric and neurological disorders-including depression and schizophrenia-are characterized by altered neural dynamics rather than focal structural abnormalities visible on conventional MRI. Temporally preserved decoding frameworks may therefore provide sensitivity to dysfunctional processing patterns that static analyses fail to capture. Extending this approach to clinical populations could clarify whether abnormal trajectory patterns emerge during naturalistic cognition.

Methodological extensions could further improve sensitivity to spatial organization while preserving computational efficiency. Hybrid architectures that combine voxel-by-time representations with spatially informed constraints, such as region-based grouping or limited convolutional structure, may better capture fine-grained spatial patterns within face-selective cortex. Direct comparisons between temporally focused and fully spatiotemporal models would clarify trade-offs among interpretability, efficiency, and spatial precision.

Data quality represents another important direction. Applying this approach to higher-field naturalistic fMRI (e.g., 3 T or above) may enhance signal-to-noise ratio and spatial specificity, enabling clearer characterization of how face-selective responses evolve over time across ventral visual regions. Systematic comparisons across field strengths could inform best practices for deep learning–based decoding in naturalistic paradigms.

Improving perceptual ground truth is also critical. Incorporating eye-tracking during naturalistic viewing would help resolve whether participants were attending to faces during labeled events, strengthening links between visual input, attention, and neural responses. More detailed behavioral or scene-level annotations could further disentangle different components of face processing, such as gaze direction, number of faces, identity, emotional expression, and social context, moving beyond coarse face presence toward finer-grained cognitive interpretations.

Beyond face perception, the voxel-by-time representation offers opportunities to study other dynamic cognitive domains. Applying the framework to events such as emotional expressions, social interactions, narrative structure, or action perception would test whether temporally preserved representations generalize across domains. Comparing temporal profiles across cognitive systems may reveal shared principles of time-varying neural organization or domain-specific dynamics.

Multimodal extensions provide another promising avenue. Combining fMRI with electrophysiological measures such as EEG or MEG could link slow, spatially distributed BOLD dynamics with faster neural processes, offering a more complete account of how cognitive events unfold over time (Mulert, 2013). Additional physiological measures, such as pupil dilation or heart rate, could help dissociate perceptual, attentional, and arousal-related contributions to observed neural patterns.

Finally, the sliding-window decoding strategy demonstrated here aligns naturally with the requirements of real-time state inference systems. The representational principle of preserving short-term trajectories could be applied to faster modalities such as EEG or MEG, where low-latency cognitive state detection is feasible. Such extensions may contribute to future brain-computer interface (BCI) systems designed to monitor or adapt to continuously evolving mental

states. Together, these directions position the present approach as a scalable foundation for studying moment-to-moment neural dynamics during complex, real-world experiences, bridging controlled experimental designs with the richness of naturalistic cognition.

3.7 FINAL CONCLUSION

This thesis introduced a computationally efficient deep learning framework for decoding event-level cognitive states from naturalistic fMRI while preserving short-term temporal structure. By reorganizing 4D fMRI segments into voxel-by-time matrices, the approach retained meaningful BOLD dynamics without relying on computationally intensive 3D or 4D convolutions. Using this representation, a lightweight convolutional neural network reliably distinguished face-onset from face-offset events across participants and diverse movie stimuli, demonstrating that temporal information alone carries discriminative signals for dynamic perceptual processing.

Importantly, interpretability analyses provided biological validation for the model's predictions. Integrated Gradients consistently emphasized canonical face-selective regions, including the FFA, OFA, and pSTS. Attribution strength followed a temporal profile consistent with expected hemodynamic dynamics, indicating that classification decisions were driven by neurobiologically plausible signals rather than trivial confounds.

Beyond event-level classification, the framework also generalized beyond isolated event windows. When applied to continuous fMRI time courses from unseen participants and movies, the model produced moment-to-moment predictions that tracked face presence over extended viewing periods. This result demonstrates that representations learned from short, event-centered segments capture dynamic neural signatures that persist during continuous naturalistic stimulation.

More broadly, this work highlights the importance of representation choices that respect the temporal nature of brain signals. Rather than increasing architectural complexity, the study shows that preserving short-term temporal structure enables efficient, interpretable decoding while remaining aligned with known neural organization. By integrating temporal fidelity, computational efficiency, and biological interpretability, this thesis provides a proof-of-concept framework for studying how cognitive states emerge and evolve in real-world settings and offers a foundation for future investigations of dynamic brain function using naturalistic neuroimaging data.

REFERENCES

Aliko, S., Huang, J., Gheorghiu, F., Meliss, S., & Skipper, J. I. (2020). A naturalistic neuroimaging database for understanding the brain using ecological stimuli. *Scientific Data*, *7*(1), 347. https://doi.org/10.1038/s41597-020-00680-2

Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2018). *Towards better understanding of gradient-based attribution methods for Deep Neural Networks* (No. arXiv:1711.06104). arXiv. https://doi.org/10.48550/arXiv.1711.06104

Atteveldt, N. M. van, Kesteren, M. van, Braams, B., & Krabbendam, L. (2018). Neuroimaging of learning and development: Improving ecological validity. *Frontline Learning Research*, *6*(3), 186–203. https://doi.org/10.14786/flr.v6i3.366

Avberšek, L. K., & Repovš, G. (2022). Deep learning in neuroimaging data analysis: Applications, challenges, and solutions. *Frontiers in Neuroimaging*, *1*. https://doi.org/10.3389/fnimg.2022.981642

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering Event Structure in Continuous Narrative Perception and Memory. *Neuron*, *95*(3), 709-721.e5. https://doi.org/10.1016/j.neuron.2017.06.041

Bedel, H. A., & Çukur, T. (2023). *DreaMR: Diffusion-driven Counterfactual Explanation for Functional MRI* (No. arXiv:2307.09547). arXiv. https://doi.org/10.48550/arXiv.2307.09547

Belyaeva, I., Bhinge, S., Long, Q., & Adali, T. (2021). Taking the 4D Nature of fMRI Data into Account Promises Significant Gains in Data Completion. *IEEE Access*, *9*, 145334–145362. https://doi.org/10.1109/ACCESS.2021.3121417

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex (New York, N.Y.: 1991)*, *19*(12), 2767–2796. https://doi.org/10.1093/cercor/bhp055

Biswal, B., Yetkin, F. Z., Haughton, V. M., & Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magnetic Resonance in Medicine*, *34*(4), 537–541. https://doi.org/10.1002/mrm.1910340409

Boynton, G. M., Engel, S. A., Glover, G. H., & Heeger, D. J. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *16*(13), 4207–4221. https://doi.org/10.1523/JNEUROSCI.16-13-04207.1996

Brennan, J. R., Stabler, E. P., Van Wagenen, S. E., Luh, W.-M., & Hale, J. T. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language*, *157–158*, 81–94. https://doi.org/10.1016/j.bandl.2016.04.008

Buckner, R. L., Bandettini, P. A., O'Craven, K. M., Savoy, R. L., Petersen, S. E., Raichle, M. E., & Rosen, B. R. (1996a). Detection of cortical activation during averaged single trials of a cognitive task using functional magnetic resonance imaging. *Proceedings of the National*

Academy of Sciences of the United States of America, *93*(25), 14878–14883. https://doi.org/10.1073/pnas.93.25.14878

Buckner, R. L., Bandettini, P. A., O'Craven, K. M., Savoy, R. L., Petersen, S. E., Raichle, M. E., & Rosen, B. R. (1996b). Detection of cortical activation during averaged single trials of a cognitive task using functional magnetic resonance imaging. *Proceedings of the National Academy of Sciences*, *93*(25), 14878–14883. https://doi.org/10.1073/pnas.93.25.14878

Buxton, R. B., Wong, E. C., & Frank, L. R. (1998). Dynamics of blood flow and oxygenation changes during brain activation: The balloon model. *Magnetic Resonance in Medicine*, *39*(6), 855–864. https://doi.org/10.1002/mrm.1910390602

Calder, A. J., & Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews. Neuroscience*, *6*(8), 641–651. https://doi.org/10.1038/nrn1724

Cox, R. W. (1996). AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages. *Computers and Biomedical Research*, *29*(3), 162–173. https://doi.org/10.1006/cbmr.1996.0014

Cox, R. W., & Hyde, J. S. (1997). Software tools for analysis and visualization of fMRI data. *NMR in Biomedicine*, *10*(4–5), 171–178. https://doi.org/10.1002/(SICI)1099-1492(199706/08)10:4/5%253C171::AID-NBM453%253E3.0.CO;2-L

Dale, A. M., & Halgren, E. (2001). Spatiotemporal mapping of brain activity by integration of multiple imaging modalities. *Current Opinion in Neurobiology*, *11*(2), 202–208. https://doi.org/10.1016/S0959-4388(00)00197-5

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017a). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, *152*, 184–194. https://doi.org/10.1016/j.neuroimage.2016.10.001

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017b). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, *152*, 184–194. https://doi.org/10.1016/j.neuroimage.2016.10.001

Eitel, F., Lebenswissenschaftlichen, J. von B. D. der, Gutachter, C. U., Ritter, K., Stober, S., & Markett, S. (2022). *Explainable deep learning classifiers for disease de-tection based on structural brain MRI data*.

Furl, N., Garrido, L., Dolan, R. J., Driver, J., & Duchaine, B. (2011). Fusiform gyrus face selectivity relates to individual differences in facial recognition ability. *Journal of Cognitive Neuroscience*, *23*(7), 1723–1740. https://doi.org/10.1162/jocn.2010.21545

Glover, G. H. (2011). Overview of Functional Magnetic Resonance Imaging. *Neurosurgery Clinics of North America*, *22*(2), 133–139. https://doi.org/10.1016/j.nec.2010.11.001

Gobbini, M. I., & Haxby, J. V. (2007). Neural systems for recognition of familiar faces. *Neuropsychologia*, *45*(1), 32–41. https://doi.org/10.1016/j.neuropsychologia.2006.04.015

Güçlü, U., & van Gerven, M. A. J. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *The Journal of*

Neuroscience: The Official Journal of the Society for Neuroscience, *35*(27), 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004a). Intersubject Synchronization of Cortical Activity During Natural Vision. *Science*, *303*(5664), 1634–1640. https://doi.org/10.1126/science.1089506

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004b). Intersubject synchronization of cortical activity during natural vision. *Science (New York, N.Y.)*, *303*(5664), 1634–1640. https://doi.org/10.1126/science.1089506

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000a). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*(6), 223–233. https://doi.org/10.1016/s1364-6613(00)01482-0

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000b). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*(6), 223–233. https://doi.org/10.1016/s1364-6613(00)01482-0

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000c). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*(6), 223–233. https://doi.org/10.1016/S1364-6613(00)01482-0

Hay, L., Duffy, A. H. B., Gilbert, S. J., & Grealy, M. A. (2022). Functional magnetic resonance imaging (fMRI) in design studies: Methodological considerations, challenges, and recommendations. *Design Studies*, *78*, 101078. https://doi.org/10.1016/j.destud.2021.101078

Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews. Neuroscience*, *7*(7), 523–534. https://doi.org/10.1038/nrn1931

Heinze, H. J., Mangun, G. R., Burchert, W., Hinrichs, H., Scholz, M., Münte, T. F., Gös, A., Scherg, M., Johannes, S., & Hundeshagen, H. (1994). Combined spatial and temporal imaging of brain activity during visual selective attention in humans. *Nature*, *372*(6506), 543–546. https://doi.org/10.1038/372543a0

Hillman, E. M. C. (2014). Coupling Mechanism and Significance of the BOLD Signal: A Status Report. *Annual Review of Neuroscience*, *37*(1), 161–181. https://doi.org/10.1146/annurev-neuro-071013-014111

Horikawa, T., & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, *8*(1), 15037. https://doi.org/10.1038/ncomms15037

Hutchison, R. M., Womelsdorf, T., Allen, E. A., Bandettini, P. A., Calhoun, V. D., Corbetta, M., Della Penna, S., Duyn, J. H., Glover, G. H., Gonzalez-Castillo, J., Handwerker, D. A., Keilholz, S., Kiviniemi, V., Leopold, D. A., de Pasquale, F., Sporns, O., Walter, M., & Chang, C. (2013). Dynamic functional connectivity: Promise, issues, and interpretations. *NeuroImage*, *80*, 360–378. https://doi.org/10.1016/j.neuroimage.2013.05.079

Janoos, F., Machiraju, R., Singh, S., & Morocz, I. Á. (2011). Spatio-temporal models of mental processes from fMRI. *NeuroImage*, *57*(2), 362–377. https://doi.org/10.1016/j.neuroimage.2011.03.047

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *The Journal of Neuroscience*, *17*(11), 4302–4311. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997

Kanwisher, N., & Yovel, G. (2006). The fusiform face area: A cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *361*(1476), 2109–2128. https://doi.org/10.1098/rstb.2006.1934

Kingma, D. P., & Ba, J. (2017). *Adam: A Method for Stochastic Optimization* (No. arXiv:1412.6980). arXiv. https://doi.org/10.48550/arXiv.1412.6980

Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, *21*(9), 1148–1160. https://doi.org/10.1038/s41593-018-0210-5

Kwong, K. K., Belliveau, J. W., Chesler, D. A., Goldberg, I. E., Weisskoff, R. M., Poncelet, B. P., Kennedy, D. N., Hoppel, B. E., Cohen, M. S., & Turner, R. (1992). Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proceedings of the National Academy of Sciences*, *89*(12), 5675–5679. https://doi.org/10.1073/pnas.89.12.5675

Laarhoven, T. van. (2017). *L2 Regularization versus Batch and Weight Normalization* (No. arXiv:1706.05350). arXiv. https://doi.org/10.48550/arXiv.1706.05350

Lee, H., & Chen, J. (2022). Predicting memory from the network structure of naturalistic events. *Nature Communications*, *13*(1), 4235. https://doi.org/10.1038/s41467-022-31965-2

Li, H., Satterthwaite, T. D., & Fan, Y. (2018). BRAIN AGE PREDICTION BASED ON RESTING-STATE FUNCTIONAL CONNECTIVITY PATTERNS USING CONVOLUTIONAL NEURAL NETWORKS. *Proceedings. IEEE International Symposium on Biomedical Imaging*, *2018*, 101–104. https://doi.org/10.1109/ISBI.2018.8363532

Logothetis, N. K. (2008, June). What we can do and what we cannot do with fMRI. In *Nature* (Vol. 453, Issue 7197, pp. 869–878). Nature Publishing Group. https://doi.org/10.1038/nature06976

Mahmoudi, M., Tachibana, A., Goldstone, A. B., Woo, Y. J., Chakraborty, P., Lee, K. R., Foote, C. S., Piecewicz, S., Barrozo, J. C., Wakeel, A., Rice, B. W., Bell III, C. B., & Yang, P. C. (2016). Novel MRI Contrast Agent from Magnetotactic Bacteria Enables In Vivo Tracking of iPSC-derived Cardiomyocytes. *Scientific Reports*, *6*(1), 26960. https://doi.org/10.1038/srep26960

Nastase, S. A., Goldstein, A., & Hasson, U. (2020). Keep it real: Rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage*, *222*. https://doi.org/10.1016/j.neuroimage.2020.117254

Ogawa, S., Lee, T. M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of*

Sciences of the United States of America, *87*(24), 9868–9872. https://doi.org/10.1073/pnas.87.24.9868

Parhi, R., & Nowak, R. D. (2020). The Role of Neural Network Activation Functions. *IEEE Signal Processing Letters*, *27*, 1779–1783. https://doi.org/10.1109/LSP.2020.3027517

Parry, A., & Matthews, P. M. (2002). Functional magnetic resonance imaging: A window into the brain. *Interdisciplinary Science Reviews*, *27*(1), 50–60. https://doi.org/10.1179/030801802225002908

Pitcher, D., & Ungerleider, L. G. (2021). Evidence for a Third Visual Pathway Specialized for Social Perception. *Trends in Cognitive Sciences*, *25*(2), 100–110. https://doi.org/10.1016/j.tics.2020.11.006

Polimeni, J. R., & Lewis, L. D. (2021). Imaging faster neural dynamics with fast fMRI: A need for updated models of the hemodynamic response. *Progress in Neurobiology*, *207*, 102174. https://doi.org/10.1016/j.pneurobio.2021.102174

Samek, W., Wiegand, T., & Müller, K.-R. (2017, August 28). *Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models*. arXiv.Org. https://arxiv.org/abs/1708.08296v1

Serences, J. T. (2004). A comparison of methods for characterizing the event-related BOLD timeseries in rapid fMRI. *NeuroImage*, *21*(4), 1690–1700. https://doi.org/10.1016/j.neuroimage.2003.12.021

Shakil, S., Lee, C.-H., & Keilholz, S. D. (2016). Evaluation of sliding window correlation performance for characterizing dynamic functional connectivity and brain states. *NeuroImage*, *133*, 111–128. https://doi.org/10.1016/j.neuroimage.2016.02.074

Shibasaki, H. (2008). Human brain mapping: Hemodynamic response and electrophysiology. *Clinical Neurophysiology*, *119*(4), 731–743. https://doi.org/10.1016/j.clinph.2007.10.026

Shrikumar, A., Greenside, P., & Kundaje, A. (2019). *Learning Important Features Through Propagating Activation Differences* (No. arXiv:1704.02685). arXiv. https://doi.org/10.48550/arXiv.1704.02685

Simony, E., & Chang, C. (2020). Analysis of stimulus-induced brain dynamics during naturalistic paradigms. *NeuroImage*, *216*, 116461. https://doi.org/10.1016/j.neuroimage.2019.116461

Soares, J. M., Magalhães, R., Moreira, P. S., Sousa, A., Ganz, E., Sampaio, A., Alves, V., Marques, P., & Sousa, N. (2016). A Hitchhiker's Guide to Functional Magnetic Resonance Imaging. *Frontiers in Neuroscience*, *10*. https://doi.org/10.3389/fnins.2016.00515

Sonkusare, S., Breakspear, M., & Guo, C. (2019). Naturalistic Stimuli in Neuroscience: Critically Acclaimed. *Trends in Cognitive Sciences*, *23*(8), 699–714. https://doi.org/10.1016/j.tics.2019.05.004

Specht, K. (2020, January). Current Challenges in Translational and Clinical fMRI and Future Directions. In *Frontiers in Psychiatry* (Vol. 10). Frontiers Media S.A. https://doi.org/10.3389/fpsyt.2019.00924

St-Yves, G., Allen, E. J., Wu, Y., Kay, K., & Naselaris, T. (2023). Brain-optimized deep neural network models of human visual areas learn non-hierarchical representations. *Nature Communications*, *14*, 3329. https://doi.org/10.1038/s41467-023-38674-4

Sundararajan, M., Taly, A., & Yan, Q. (2017a). *Axiomatic Attribution for Deep Networks*. http://arxiv.org/abs/1703.01365

Sundararajan, M., Taly, A., & Yan, Q. (2017b). *Axiomatic Attribution for Deep Networks* (No. arXiv:1703.01365). arXiv. https://doi.org/10.48550/arXiv.1703.01365

Thomas, A. W., Müller, K.-R., & Samek, W. (2019). Deep Transfer Learning for Whole-Brain FMRI Analyses. In L. Zhou, D. Sarikaya, S. M. Kia, S. Speidel, A. Malpani, D. Hashimoto, M. Habes, T. Löfstedt, K. Ritter, & H. Wang (Eds.), *OR 2.0 Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging* (Vol. 11796, pp. 59–67). Springer International Publishing. https://doi.org/10.1007/978-3-030-32695-1_7

VanRullen, R., & Reddy, L. (2019). Reconstructing faces from fMRI patterns using deep generative neural networks. *Communications Biology*, *2*, 193. https://doi.org/10.1038/s42003-019-0438-y

Warren, S. L., & Moustafa, A. A. (2023). Functional magnetic resonance imaging, deep learning, and Alzheimer's disease: A systematic review. *Journal of Neuroimaging*, *33*(1), 5–18. https://doi.org/10.1111/jon.13063

Zhao, Y., Li, X., Huang, H., Zhang, W., Zhao, S., Makkie, M., Zhang, M., Li, Q., & Liu, T. (2020). Four-Dimensional Modeling of fMRI Data via Spatio-Temporal Convolutional Neural Networks (ST-CNNs). *IEEE Transactions on Cognitive and Developmental Systems*, *12*(3), 451–460. https://doi.org/10.1109/TCDS.2019.2916916