

A foundation-model approach to pediatric headache classification from resting-state fMRI

Guilherme S. Imai Aldeia

GUILHERME.IMAIALDEIA@CHILDRENS.HARVARD.EDU

Computational Health Informatics Program, Boston Children’s Hospital, Boston, MA, USA Department of Pediatrics, Harvard Medical School, Boston, MA, USA

Clara Moon

CLARA.MOON@CHILDRENS.HARVARD.EDU

Pediatric Pain Pathway Lab; Department of Anesthesia, Critical Care and Pain Medicine; Department of Anesthesia, Boston Children’s Hospital, Boston, MA, USA

Julie Shulman

JULIE.SHULMAN@CHILDRENS.HARVARD.EDU

Department of Anesthesia, Critical Care and Pain Medicine; Department of Anesthesia; Pediatric Pain Rehabilitation Center, Boston Children’s Hospital, Boston, MA, USA

Navil Sethna

NAVIL.SETHNA@CHILDRENS.HARVARD.EDU

Department of Anesthesia, Critical Care and Pain Medicine; Department of Anesthesia; Pediatric Pain Rehabilitation Center, Boston Children’s Hospital, Boston, MA, USA

Allison Smith

ALLISON.SMITH@CHILDRENS.HARVARD.EDU

Department of Anesthesia, Critical Care and Pain Medicine; Department of Anesthesia; Pediatric Headache Program, Boston Children’s Hospital, Boston, MA, USA

Alyssa Lebel*

ALYSSA.LEBEL@CHILDRENS.HARVARD.EDU

Department of Anesthesia, Critical Care and Pain Medicine; Department of Anesthesia, Boston Children’s Hospital, Boston, MA, USA

William G. La Cava*

WILLIAM.LACAVA@CHILDRENS.HARVARD.EDU

Computational Health Informatics Program, Boston Children’s Hospital, Boston, MA, USA; Department of Pediatrics, Harvard Medical School, Boston, MA, USA

Scott Holmes*

SCOTT.HOLMES@CHILDRENS.HARVARD.EDU

Pediatric Pain Pathway Lab; Department of Anesthesia, Critical Care and Pain Medicine; Department of Anesthesia, Boston Children’s Hospital, Boston, MA, USA

Abstract

Headache is the most common neurological disorder in children and substantially affects quality of life. We investigated whether resting-state functional MRI (rs-fMRI) can support pediatric headache classification using machine learning. We encoded rs-fMRI data using NeuroSTORM, a recent foundation model, and fine-tuned it to distinguish healthy controls from children with headache and subsequently classify headache subtypes. We then compared NeuroSTORM with a standard neuroscience approach that uses functional connectivity (FC) matrices derived from brain activity as predictors. Using 189 rs-fMRI scans from 110 individuals collected across two visits (prevalence of any headache: 74%), NeuroSTORM achieved an area under the receiver operating characteristic curve (AUROC) of 0.82 (95% CI, 0.82–0.82) and an area under the precision-recall curve (AUPRC) of 0.93 (95% CI, 0.93–0.94) in discriminating headache from non-headache. In contrast, models

* Equal contribution

trained on FC matrices showed limited performance (AUROC, 0.67 [95% CI, 0.67–0.67]; AUPRC, 0.85 [95% CI, 0.85–0.85]). In a multiclass setting, when tasked with classifying individuals as healthy controls, individuals with chronic migraine, or individuals with non-chronic headaches (e.g., post-viral headache, new daily persistent headache, post-traumatic headache), NeuroSTORM achieved a macro-AUROC of 0.69 (95% CI, 0.68–0.69). The results suggest this approach can distinguish chronic migraine, the most common headache syndrome, but has difficulty differentiating other headache subtypes from chronic migraine. Overall, under limited-data conditions, NeuroSTORM appears to capture latent rs-fMRI representations that transfer to headache-related tasks. The findings provide proof-of-concept for fMRI-based prediction of pediatric headache using a foundation model without relying on functional connectivity features and highlight the potential of this approach for subtype identification. Further development of such tools may help clinicians improve diagnosis based on brain activity and tailor treatment strategies to individual patients.

1. Introduction

Headache is the most prevalent neurologic disorder in children, affecting at least 58.4% of the pediatric and adolescent population (Prezioso et al., 2022). Chronic daily headache impacts 1–3% of adolescents globally (Seshia, 2012; Cuvellier et al., 2008; Arruda et al., 2010), with U.S. prevalence reaching 6% (Pawlowski et al., 2019). Among subtypes, chronic migraine is the most common and recurrent headache syndrome (Youssef and Mack, 2020), with a much higher prevalence in girls than in boys during adolescence (Blume, 2012). These conditions can severely impact daily life, leading to school absence, cognitive difficulties, nausea, and photophobia (Langdon and DiSabella, 2017b).

Machine learning (ML) has become increasingly common in neuroimaging applications (Pereira et al., 2009; Khosla et al., 2019) as a data-driven approach for identifying complex patterns and supporting automated decision-making. Brain connectivity plays a key role in such applications, commonly measured using functional Magnetic Resonance Imaging (fMRI), a non-invasive technique that captures hemodynamic responses associated with neural activity (Logothetis, 2008), that is, the blood-oxygen-level dependent (BOLD) signal. Both fMRI and its derived functional connectivity (FC) matrices—pairwise correlation coefficients between brain regions—can be used to investigate how brain function relates to disorders, with FC often preferred because of its lower dimensionality and more interpretable features.

When acquired during resting state (rs-fMRI), these patterns reflect the brain’s intrinsic organization (Varoquaux and Craddock, 2013) and provide a valuable tool for investigating neurological disorders. Prior work in cases such as post-traumatic headache has shown the potential for structural MRI and FC matrices to be used in the context of differentiating patient cohorts (Dumkrieger et al., 2023; Joshi et al., 2024; Chong et al., 2017; Mao et al., 2025). Relatively few studies have explored FC-based analyses in pediatric cohorts (Holmes et al., 2022; Ofoghi et al., 2021). Such data-oriented approaches are particularly relevant in pediatric populations because young children often struggle to describe their symptoms (Langdon and DiSabella, 2017a), and diagnosis relies on semi-structured interviews and following the ICHD-3 (Headache Classification Committee of the International Headache Society (IHS), 2018) guidelines. Disability and quality of life can be assessed using questionnaires such as the HIT-6 (Rendas-Baum et al., 2014). When neuroimaging is

used for diagnosis, it primarily focuses on investigating potential structural characteristics, such as abnormalities, inflammation, and ischemia (Blume, 2012).

Despite advances in ML, fMRI-based applications still lack standardized pipelines and strong transferability. A key reason is that many brain-imaging studies are not replicable (Poldrack et al., 2017), involve too few participants to support robust conclusions (Button et al., 2013; Marek et al., 2022), and brain structures exhibit substantial inter-individual differences (Brett et al., 2002). Recent work has explored architectures such as the Swin Transformer (Liu et al., 2021) and the Mamba state space model (Gu and Dao, 2024) for fMRI data, and the emergence of foundation models trained on more than 28 million fMRI frames (Wang et al., 2026) has shown that direct processing of fMRI images can generate effective representations. However, these studies have been limited to adult populations and applications to pediatric health remain limited. Furthermore, the application of foundation models to fMRI remains a very recent development in the field.

Here, we investigate whether ML models can classify pediatric headache using either FC features or direct fMRI data encoded with a foundation model. We collected data from 110 individuals across two waves, totaling 189 scans from participants aged 8–22 years at the time of visit. Participants were either healthy controls ($n = 45$ individuals) or had a headache condition, with chronic migraine being the most frequent subtype ($n = 68$ individuals). We preprocessed the fMRI data, extracted FC matrices, and trained shallow ML methods to classify pediatric headache. We then trained a final model to differentiate headache subtypes based on the best-performing binary-classification setting.

Generalizable Insights about Machine Learning in the Context of Healthcare

This work presents the first proof-of-concept of effective pediatric headache classification from rs-fMRI data. The results are made possible by the novel application of a pre-trained foundation model (NeuroSTORM) to directly encode fMRI data, which outperformed traditional FC-based features in a limited-data setting. More broadly, it suggests that the recent emergence of large-scale, pre-trained neuroimaging models provides a practical starting point for clinical ML applications in which site-specific datasets are too small to support training high-capacity models from scratch. At the same time, our results highlight the need for careful evaluation of subgroup performance, interpretability, and prospective validation before such methods can be used as clinical decision-support tools.

2. Related Work

Prior work has leveraged rs-fMRI and FC for migraine classification in adults (58 migraine individuals and 50 healthy controls), applying ML to distinguish individuals using 33 brain regions manually pre-selected based on findings from the pain and migraine literature (Chong et al., 2017). The authors reported that migraine is easier to classify in individuals older than 14 years, with classification accuracy decreasing from 96% to 82% for individuals aged 14 years or younger.

In a related line of work, Holmes et al. (2022) investigated post-traumatic headache (PTH) in a pediatric and young cohort (64 PTH individuals and 34 healthy controls), a condition with still unclear underlying mechanisms. Using ML techniques combined with data reduction, feature selection, and clustering, the authors reduced 70 regions of interest

to 14 salient features. Similarly, [Mao et al. \(2025\)](#) studied PTH in adults (73 individuals), aiming to predict clinical improvement three months after mild traumatic brain injury (mTBI). Complementary, [Ofoghi et al. \(2021\)](#) studied PTH after mTBI in children, using FC-based features and general linear models. [Qiu et al. \(2023\)](#) examined new daily persistent headache (NDPH) (35 individuals with NDPH and 40 healthy controls) from a functional connectivity perspective using magnetoencephalography in an adult cohort. These studies predominantly focus on adult populations and typically rely on derived FC representations rather than direct rs-fMRI images.

More recently, foundation models have begun to be applied across healthcare domains. These models are trained on large-scale, heterogeneous datasets and learn generalizable representations of the data of interest — in this case, fMRI. In neuroscience, this paradigm has led to models such as fMRI-LM, which embeds fMRI-derived tokens into a language space to enable high-level, natural language interpretation of low-level neuroimaging signals ([Wei et al., 2025](#)). Another example is SwiFT ([Kim et al., 2023](#)), a model based on the Swin Transformer ([Liu et al., 2021](#)) that learns brain dynamics directly from fMRI volumes. SwiFT was trained on large cohorts, including some of the largest publicly available datasets: the Human Connectome Project (HCP) ([Van Essen et al., 2012](#)), Adolescent Brain Cognitive Development (ABCD) ([Casey et al., 2018](#)), and UK Biobank (UKB) ([Sudlow et al., 2015](#)); and validated on downstream tasks such as brain age prediction. More recently, NeuroSTORM ([Wang et al., 2026](#)) was proposed and trained over more than 28 million fMRI frames, replacing the standard attention mechanism with a state space model (Mamba) ([Gu and Dao, 2024](#)), and demonstrating strong performance across phenotype and diagnosis prediction tasks.

In this work, we bridge the gap between prior ML-based approaches for headache prediction and the emerging use of foundation models for fMRI headache prediction in pediatric and youth populations. Leveraging a sample size of 189 scans, we utilize the encoder of the pre-trained NeuroSTORM model to predict pediatric headache outcomes, and contrast its performance with traditional ML approaches commonly used in the neuroscience literature.

3. Materials and methods

We tackle two classification tasks: (1) binary classification to distinguish healthy controls from pediatric headache patients, and (2) multi-class classification to distinguish healthy controls, chronic migraine, and non-chronic migraine using a one-vs-all scheme. We focus on differentiating chronic migraine because it is the most common and well-defined headache subtype. Other headache subtypes are more susceptible to labeling noise because their diagnostic criteria are less well defined and their symptoms often overlap.

We use two representations to compare approaches: one uses traditional FC matrices derived from cortical brain regions, while the other uses embeddings from a foundation model (NeuroSTORM) applied directly to fMRI data. This comparison allows us to evaluate whether the foundation model outperforms the standard approach and to quantify how well traditional FC-based methods perform on this headache-classification problem. Figure 1 illustrates the overall experimental pipeline.

The raw fMRI data were first preprocessed using the CONN toolbox ([Whitfield-Gabrieli and Nieto-Castanon, 2012](#)) functional preprocessing pipeline with default parameters and

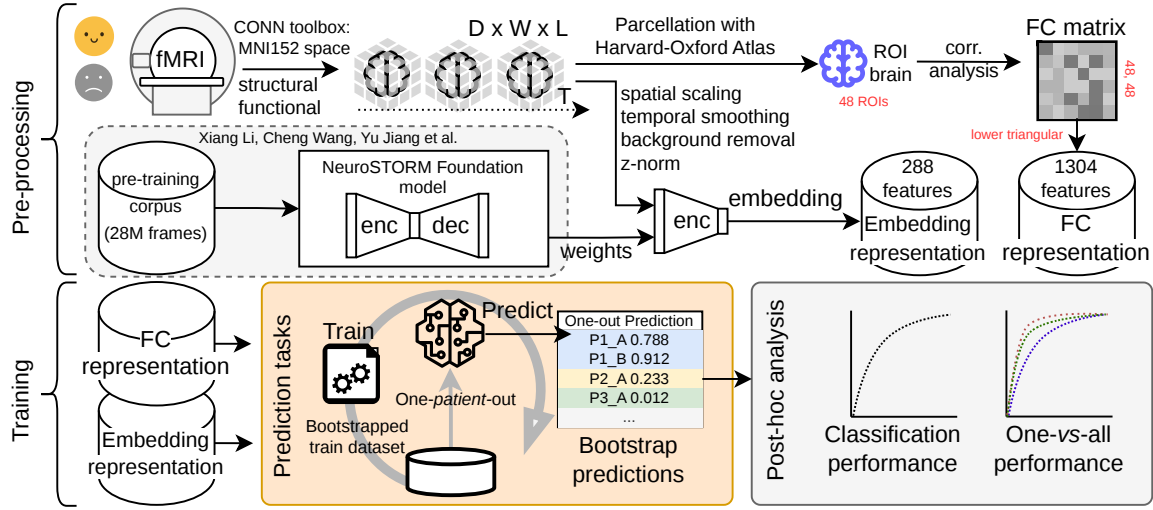


Figure 1: Experimental pipeline. The data are preprocessed to generate two distinct representations, one using features extracted from functional connectivity matrices and the other using embeddings extracted directly from the fMRI data using a foundation model. Both feature representations are used as inputs for training machine learning methods using a bootstrapped leave-one-out approach.

then mapped to MNI152 space, followed by the default denoising pipeline. Mapping to MNI152 space is common in neuroscience and was also used by the NeuroSTORM authors when preparing the training data for the foundation model. The CONN default pipeline includes brain extraction, functional realignment and normalization, slice-timing correction, outlier identification, smoothing with a kernel of 8 mm FWHM, and bandpass filtering between [0.008 Hz, 0.09 Hz] to remove BOLD fluctuations that may be generated by physiological processes, head motion, and other noise sources. Brain extraction helps remove confounding signal outside the brain, and smoothing increases the signal-to-noise ratio in the time series. The resulting data were used to generate FC and embedding representations.

For the FC representation, the Harvard-Oxford atlas is used for parcellation, containing 48 cortical and 21 subcortical regions of interest (ROIs). Given the sample size, we kept only the cortical ROIs to obtain a more favorable feature-to-sample ratio. We chose not to perform feature selection on the FC features. Although feature selection may increase statistical power, it can produce cohort-dependent feature sets and reduce generalizability. Pairwise correlations were computed between all ROI time series, yielding a symmetric 48x48 functional connectivity matrix. The upper triangular portion and main diagonal were discarded, resulting in 1,128 unique connectivity features. Including subcortical regions would increase the feature space to 2,346 features.

For the embedding representation, the rs-fMRI volumes undergo the preprocessing proposed by the NeuroSTORM authors, with spatial scaling to fit the default input size of $96 \times 96 \times 96$ and temporal smoothing to match the input requirement of the pre-trained foundation model (20 time frames). Background voxels are removed, and z-normalization

is applied before passing the data through the encoder, yielding a 288-dimensional feature embedding.

Figure 2 illustrates the data modalities used in this study. We do not use raw fMRI directly because of its low signal-to-noise ratio and because it contains structural information unrelated to the brain (e.g., eyes and other tissues). Regardless of whether one uses fMRI or FC representation, standard preprocessing is routine in the neuroscience literature before model development.

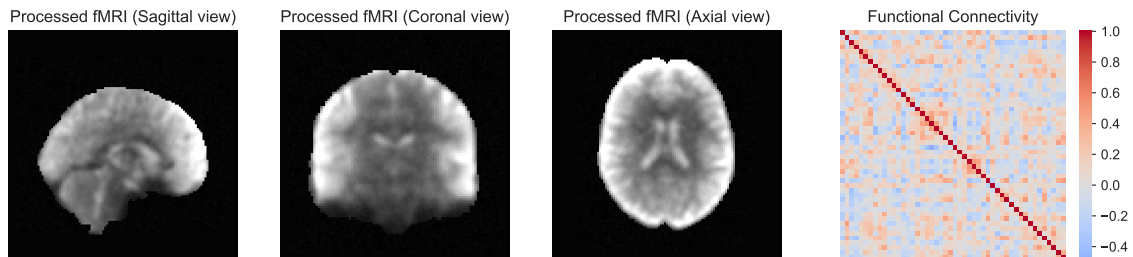


Figure 2: Illustrative example of the data used in the experiments. The first three images show three views of a 4-dimensional image after applying the preprocessing pipeline, which includes denoising, MNI152 space alignment, and smoothing. The last image shows a functional connectivity matrix depicting pairwise correlations among 48 ROIs over time.

Using these two representations, we train a range of shallow machine learning classifiers to distinguish between healthy controls and headache participants. Given the limited sample size, we use leave-one-out (LOO) cross-validation, in which each individual is held out once while the model is trained on the remaining participants. When an individual is held out for testing, all of that participant’s visits are excluded from the training data to avoid leakage. The intent of this approach is to use all available data while ensuring that no visit from a test participant appears in the training set. This LOO procedure is repeated 100 times with different random seeds to capture uncertainty in the model’s predicted probabilities for each participant.

After identifying the best-performing model for the binary classification task, a final experiment is conducted to predict headache subtypes. This stage follows the same LOO cross-validation and bootstrapping procedure used for the first experiment, in a one-vs-all schema.

3.1. Eligibility Criteria and Recruitment

A total of 110 participants were recruited from the Greater Boston area. Data were collected in two waves, yielding 189 scans: 104 acquired at the first visit and 85 acquired at the second visit. 79 participants completed both visits. Of the 189 scans, 144 were acquired from patients with a formal headache diagnosis who were evaluated through Boston Children’s Hospital’s Chronic Headache Program, Sports Medicine Clinic, Adolescent/Young Adult Medicine, Pediatric or Young Adult Pain Rehabilitation Centers, or Post-Acute Sequelae of SARS-CoV-2 clinic. Following diagnosis, eligible patients were invited to participate during or after their clinical visits.

Eligibility was determined through pre-screening questionnaires that assessed headache history and screened for exclusion criteria, including concurrent pain conditions and mental health disorders. Participants were eligible if they were 8–22 years old; of any gender, including non-binary and transgender; able to speak and read English sufficiently; and free of other acute or chronic medical or pain conditions that could confound the data. Additional exclusion criteria included metallic implants, pregnancy, and claustrophobia. Healthy controls ($n = 45$) were pain-free individuals from the hospital or the local community with no history of headache, recruited through advertisements, fliers, the hospital’s internal website, and word of mouth.

All persons with a headache were evaluated clinically with oversight by study physician (AL). All persons were evaluated within the Boston Children’s Hospital system with clinical mention of headache as their primary (e.g., chronic migraine, new daily persistent headache), or secondary (e.g., post-traumatic headache, post-viral headache) symptoms.

The Boston Children’s Hospital Institutional Review Board approved the data-acquisition protocol used in this study and waived the requirement for informed consent.

3.2. Imaging

Brain imaging was acquired on a 3T Siemens Prisma scanner equipped with a 32-channel phased-array head coil. High-resolution T1-weighted magnetization-prepared rapid acquisition gradient echo (MPRAGE) images were acquired. Resting-state images were acquired using the following parameters: isotropic voxel size = $2.4 \times 2.4 \times 2.4 \text{ mm}^3$, TR = 3420 ms, TE = 33.0 ms, and slice thickness = 2.4 mm.

3.3. Clinical and Demographic Information

Participants completed demographic questionnaires reporting age, assigned sex at birth, and years of education, along with a brief clinical assessment battery. This battery included the Head Impact Test-6 (HIT-6), which evaluates headache symptoms and severity, and the Central Sensitization Inventory (CSI), which assesses symptoms related to central sensitization and pain amplification in the central nervous system. The HIT-6 score can range from 36 to 78, with higher scores indicating greater impact on daily life due to headache. The CSI can range from 0 to 100 and uses a Likert scale with responses from 0 (never) to 4 (always) to score symptoms commonly associated with central sensitization syndrome. Demographic information is summarized in Table 1.

Distributions of age, CSI scores, and HIT-6 scores for healthy controls (HC, $n=45$, 23.81%) and the headache subtypes are presented in Figure 3, split by assigned sex at birth. In this figure, statistical tests were performed between the HC group and each headache subtype, as well as pairwise comparisons between each headache subtype. Bonferroni correction was applied. Non-significant results were omitted.

The final cohort included data from patients with chronic migraine (CM; 35.98%, 68 of 189 scans) and other identified headache subtypes (40.21%, 76 of 189 scans), including post-viral headache (PVH), new daily persistent headache (NDPH), and post-traumatic headache (PTH).

Table 1: Demographic statistics of the study population. Values are reported as mean (SD) unless otherwise specified. HIT-6 and CSI scores are reported by cohort only.

			Cohort	
			Healthy	Headache
Overall				
n, counts (%)			189	45 (23.8) 144 (76.2)
Assigned sex at birth	F, counts (%)	133 (70.4)	21 (46.7)	112 (77.8)
	M, counts (%)	56 (29.6)	24 (53.3)	32 (22.2)
Age		15.7 (2.8)	15.5 (4.2)	15.7 (2.3)
HIT-6		–	44.0 (6.1)	61.3 (7.0)
CSI		–	18.6 (15.6)	43.9 (16.7)

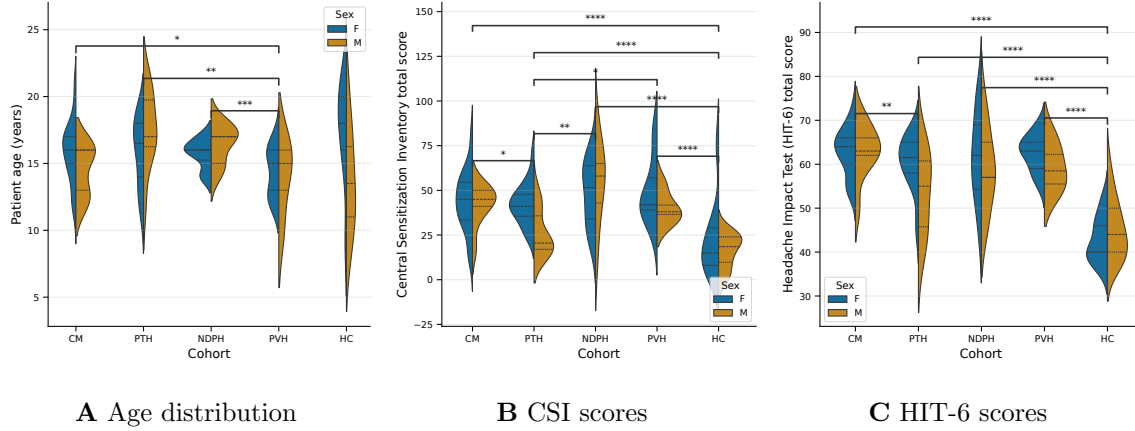


Figure 3: Clinical and demographic measurements by cohort. (A) Age distribution across healthy controls and headache patients. (B) Central Sensitization Inventory (CSI) scores. (C) Head Impact Test-6 (HIT-6) scores. * $10^{-2} < p \leq 5 \times 10^{-2}$, ** $10^{-3} < p \leq 10^{-2}$, *** $10^{-4} < p \leq 10^{-3}$, **** $p \leq 10^{-4}$. non-significant differences are omitted.

We first trained a demographic logistic regression model using only age and assigned sex at birth to assess whether performance in the binary task could be explained by cohort composition rather than neuroimaging signal. This comparison is important because the cohorts differ in sex distribution, while age differences are modest, making sex a potential confounder. Age was discretized into six bins: 18 months–3 years, 3–5 years, 5–10 years, 10–15 years, 15–18 years, and above 18 years. We then fit an unregularized logistic regression model using discretized age and sex as demographic predictors.

3.4. Machine learning algorithms

We evaluated a range of ML methods, considering both interpretable models and black-box approaches. Interpretable models included decision trees (DT) and logistic regression

(LR), while black-box models comprised ensemble methods, namely random forests (RF) and gradient boosting (GB). All models were implemented in scikit-learn v1.8.0.

Preliminary experiments were conducted to constrain model complexity for the FC-based approach, whose features are clinically meaningful, while avoiding extensive tuning of the foundation-model pipeline. Given the limited sample size and the risk of overfitting, we intentionally avoided extensive hyperparameter optimization. Instead, we restricted tuning primarily to parameters that directly affect model complexity, and primarily relied on standard configurations.

Performance degraded for DT when attempting to limit the maximum depth to 2, 3, or 5, and thus we retained the default hyperparameters. Similarly, reducing the number of estimators in RF and GB below the default value of 100 also resulted in worse performance. We tested 100, 1000, and 5000 maximum iterations for logistic regression methods. Increasing maximum iterations from 100 to 1000 improved performance, but not beyond 1000. A detailed description of the model configurations used to obtain the reported results is provided in Appendix A.

Finally, because embedding encoders are designed to produce linearly separable representations, we focus on logistic regression with L1 regularization for the embedding-based representation rather than more complex models such as neural networks.

4. Results

Group differences were assessed using Welch’s two-sample t -test. Consistent with the descriptive statistics in Table 1, participants with headache exhibited substantially higher HIT-6 scores (headache: 61.3 ± 7.0 ; healthy: 44.0 ± 6.1), a difference that was highly significant ($t = 14.84$, $p = 2.00 \times 10^{-33}$). CSI scores were also significantly elevated in the headache group (headache: 43.9 ± 16.7 ; healthy: 18.6 ± 15.6 ; $t = 9.02$, $p = 2.19 \times 10^{-16}$). Mean age differed modestly between cohorts (headache: 15.7 ± 2.3 ; healthy: 15.5 ± 4.2), but this difference was not statistically significant ($t = 0.43$, $p = 0.68$).

Classification performance is summarized in Table 2. For each metric, we report the mean performance along with bootstrapped 95% confidence intervals.

Table 2: Classification performance for the models evaluated with functional connectivity (DT, LR L1/L2, GB, and RF) and embedding-based features (NeuroSTORM). AUROC: Area Under Receiver Operating Characteristic Curve; AUPRC: Area Under Precision-Recall Curve. Values are reported as (mean, 95% CI). The highlighted value in each column depicts the best performing model for that specific measure.

	AUROC	AUPRC
Demographic Model	0.61 [0.61, 0.61]	0.82 [0.80, 0.83]
DT	0.53 [0.48, 0.56]	0.77 [0.75, 0.79]
LR L1	0.67 [0.67, 0.67]	0.85 [0.85, 0.85]
LR L2	0.58 [0.58, 0.58]	0.79 [0.79, 0.79]
GB	0.40 [0.38, 0.43]	0.71 [0.69, 0.73]
RF	0.54 [0.50, 0.57]	0.78 [0.75, 0.80]
NeuroSTORM	0.82 [0.82, 0.82]	0.93 [0.93, 0.94]

Among the FC-based classifiers, decision trees yielded low performance (AUROC = 0.53, AUPRC = 0.77), close to a random classifier. Ensemble methods (i.e., GB, RF) did not outperform linear models, with the best-performing ensemble achieving AUROC = 0.54 and AUPRC = 0.78. Logistic regression with L1 and L2 regularization exhibited very tight confidence intervals, with variations detected only at the fourth decimal place.

The best performance was observed for NeuroSTORM, which uses logistic regression with L1 regularization on the embedding representation, achieving an AUROC of 0.82 and an AUPRC of 0.93. These results were robust to variation in the random seed, with no observable differences within the 95% confidence interval. The strong performance of a simple linear classifier on the embeddings suggests that the foundation model captures resting-state dynamics in a representation that transfers well to this downstream task. More broadly, the narrow confidence intervals suggest that, in this experimental setting, LR converges to a highly similar solution across seeds despite the stochastic solver.

Given the poor performance from the GB model, we inspected the training logs in detail and noticed that AUROC training performance was always equal to one. Inspecting feature importance obtained with SHAP (Lundberg and Lee, 2017) explanations across runs, we additionally noted that there was large variation in the most important features. In this light, we believe the model is overfitting to spurious correlations, resulting in an LOO performance worse than guessing.

The demographic logistic regression model for predicting any headache achieved an AUROC of 0.61 and an AUPRC of 0.82. Coefficient inspection showed that non-zero weights were concentrated in age-bin indicators. To evaluate the robustness of different models to potential demographic confounding, we performed a stratified analysis by sex and computed AUROC and reported the results in Table 3. Notably, NeuroSTORM maintained performance across demographic groups, whereas the FC-based approaches exhibited a greater degree of performance degradation. These results suggest that performance is not due to confounding by sex.

Table 3: AUROC performance for the models when stratified by assigned sex at birth. Values are reported as (mean, 95% CI). The highlighted value in each column depicts the best performing model for that specific measure.

	Male		Female	
DT	0.55	[0.49, 0.62]	0.50	[0.45, 0.55]
LR L1	0.71	[0.71, 0.71]	0.60	[0.60, 0.60]
LR L2	0.59	[0.59, 0.59]	0.51	[0.51, 0.51]
GB	0.38	[0.35, 0.44]	0.36	[0.33, 0.38]
RF	0.58	[0.51, 0.65]	0.45	[0.39, 0.51]
NeuroSTORM	0.80	[0.79, 0.81]	0.81	[0.80, 0.81]

Precision–recall and ROC curves are presented in Figure 4. All models trained on the FC representation exhibit performance close to that of a random classifier (the diagonal dashed gray line in the ROC plot). We note that, because these curves are estimated using leave-one-out (LOO) validation on a limited dataset, individual points may correspond to predictions generated by different model instances rather than a single, fixed model.

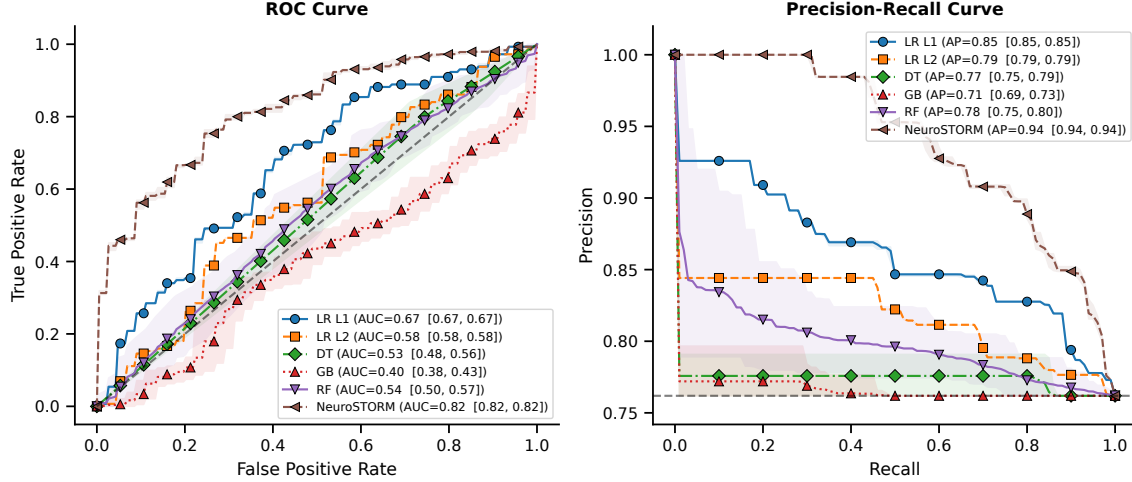


Figure 4: Precision-Recall and Receiver Operating Characteristic (ROC) Curves for the ML methods evaluated in the experiments, with the Area Under the Curve (AUC) and Average Precision (AP) reported for each curve. Confidence intervals are estimated using the leave-one-out evaluation procedure repeated 100 times with different random seeds.

Finally, we trained the best-performing model from the binary classification task to predict headache subtypes using the same experimental pipeline. Receiver Operating Characteristic and Precision-recall curves are presented in Figure 5. The model achieved a macro-AUROC of 0.69 (95% CI, 0.68 – 0.69) and a weighted AUROC = 0.66 (95% CI, 0.66 – 0.67). Overall, performance was lower for subtype discrimination than for binary classification, indicating that this is a more challenging task.

The bottom panels show the detailed precision–recall curves. We observe that Migraine classification drive most of the overall performance, whereas the non-migraine types are less separable. The classifier performed best for healthy controls (AUROC = 0.81; AUPRC = 0.55, prevalence 0.24) and chronic migraine (AUROC = 0.67; AUPRC = 0.57, prevalence 0.36), but with a decreased performance for other headache (AUROC = 0.58; AUPRC = 0.50, prevalence 0.40).

5. Discussion

Machine learning algorithms can help identify imaging patterns that differentiate pediatric patients with headache from those without a clinical headache history. To date, however, most ML studies in neuroimaging have focused on cohort discrimination in adult populations and have primarily relied on seed-based FC matrices as input features. In the current study, we evaluate several ML algorithms for differentiating pediatric patients with headache from healthy controls and compare FC features with foundation model embeddings

Headache and its subtypes remain challenging to diagnose, as they rely heavily on subjective self-report and criteria that are often difficult to operationalize, particularly in pediatric populations. Using HIT-6 and CSI scores, we confirmed that participants in the headache cohort exhibited elevated headache burden and symptoms consistent with central

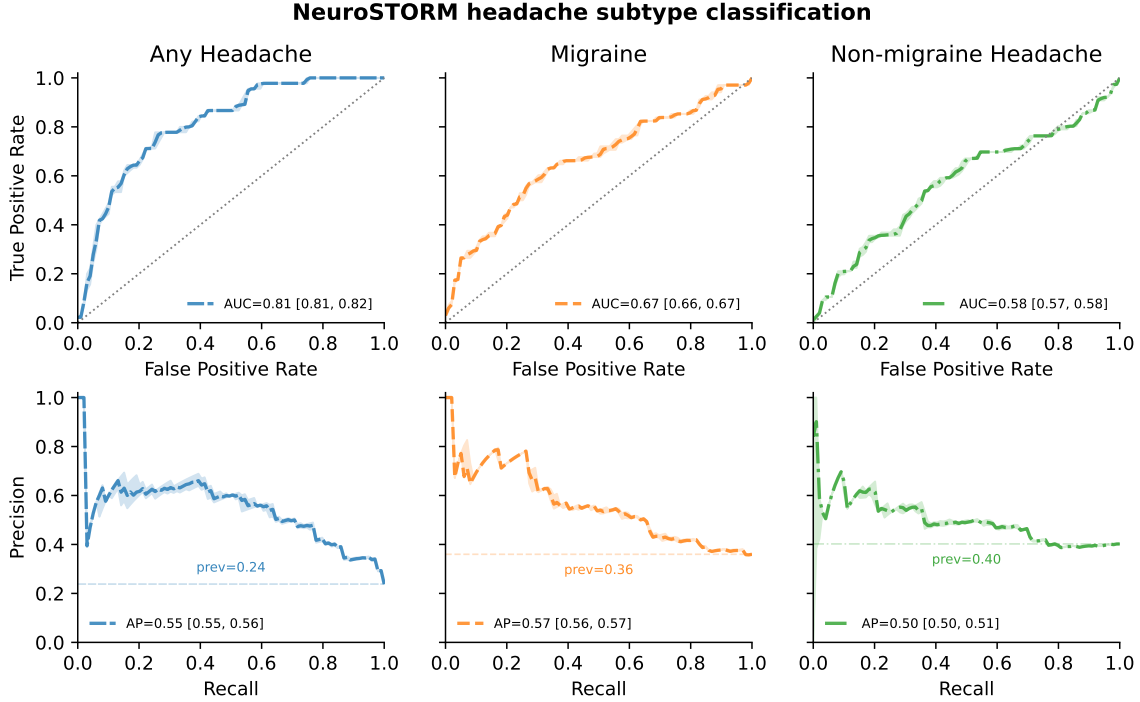


Figure 5: Receiver Operating Characteristic (ROC) and Precision-Recall (PR) Curves for the NeuroSTORM method for predicting headache subtypes. The top row shows the ROC curves, with the Area Under the Curve (AUC) reported for each curve, while the bottom row shows the PR curves, with Average Precision (AP) reported for each curve

sensitization, suggesting that our participants experienced headache symptoms and had a significant centralized component. Together with previous findings from [Holmes et al. \(2022, 2023\)](#), these studies highlight the value of integrating neuroimaging with clinical assessments to obtain objective markers of headache-related pathology.

In this work, we show that a foundation model used to encode pediatric rs-fMRI data can identify headache status with stronger discriminative performance than traditional methods based on FC matrices. We further compare this approach with standard FC-based methods without incorporating prior domain knowledge to restrict the feature space used for training. Although FC-based methods underperformed in our experiments, they offer clinically interpretable features, in contrast to the NeuroSTORM embedding representation, which lacks direct interpretability.

Ultimately, the value of ML and AI models in clinical settings depends on their ability to produce outputs that are clinically interpretable and actionable. While the embeddings are not directly interpretable, future work can adapt tools such as GradCAM ([Gildenblat and contributors, 2021](#); [Selvaraju et al., 2019](#)) for generating heatmaps to identify regions the foundation model focuses on for generating the encoded representation of the input fMRI.

The limited subtype performance may partly reflect the training data used for the foundation model, which predominantly consist of large cohorts of healthy individuals and, among disorder cohorts, populations such as pediatric ADHD (e.g., ADHD200), early psy-

chosis, and other psychiatric diagnoses. As a result, the learned representations may transfer imperfectly to headache-specific outcomes. In addition, survey-based measures such as CSI and HIT-6 are inherently noisy because they depend on individual symptom perception and reporting.

When examining headache subtypes, [Chong et al. \(2017\)](#) found that chronic migraine individuals with longer disease durations were more accurately classified. They also observed a marked effect of age: classification accuracy was substantially higher for individuals older than 14 years compared to younger patients, with performance dropping by 14% in the younger age group. Notably, their cohort had a mean age of 36.3 years, whereas our pediatric cohort averages 15.7 years.

The demographic model achieved an AUROC of 0.61 and an AUPRC of 0.82, indicating that age provide non-trivial discriminative signal even in the absence of neuroimaging features. The imaging-based representation provides substantial additional predictive value in this cohort, as evidenced by the NeuroSTORM-based model, which achieved an AUROC of 0.82 and an AUPRC of 0.93.

The challenging performance for subtype classification, such as post-traumatic headaches (PTH), may reflect inherent diagnostic and phenotypic complexities. Prior work ([Dumkrieger et al., 2023](#)) has argued that migraine and PTH exhibit similar neurobiological phenotypes, which could explain the lower discriminability between these subtypes. Notably, that study required extensive FC-based feature pre-selection (reducing 5,624 features to 1,521), suggesting that feature engineering may be a critical factor for distinguishing these phenotypically similar conditions. In our work, the superior performance for chronic migraine and lower performance for non-migraine headache may also be influenced by the nature of labeling: chronic migraine, being more well-defined clinically, may be more consistently labeled, while other diagnoses may carry greater uncertainty. This measurement variability in the target labels could contribute to differential classification performance across headache subtypes, given the NeuroSTORM performance drop from binary classification to subtype classification.

This study demonstrates the feasibility of leveraging modern AI methods to handle high-dimensional neuroimaging data such as fMRI. Although fMRI is resource intensive, it provides additional information about the temporal dynamics of the brain that cannot be captured by measures used in standard of care. As an initial step, the results support continued efforts to expand data collection, increase sample size, and adapt pre-trained models more directly to headache-related tasks.

Another contribution of this work is the extension beyond binary classification to headache subtype identification, a more challenging and clinically relevant problem that better mirrors real-world diagnostic settings, although performance remains modest at this task. Prior neuroimaging studies have predominantly focused on binary discrimination tasks: distinguishing patients with a specific headache type from healthy controls.

6. Limitations

There are limitations to the present work that should be discussed. First, our cohort size was relatively small for ML, particularly for subtype analysis, or to perform an comprehensive hyperparameter tuning for the baseline models. To partly offset this, we used a

lower-dimensional ROI representation for whole-brain analyses rather than a larger atlas with hundreds of regions. We chose to evaluate model performance using a whole-brain approach rather than seed-based ROI analyses that could further reduce dimensionality. Second, the headache cohort included participants with different headache diagnoses. This heterogeneity is clinically realistic, but it may also have dampened model performance by reducing label homogeneity. Third, although the foundation approach achieves strong performance, it offers limited mechanistic interpretability compared to the lower-performing approach based on functional connectivity matrices. Lastly, the current work used only rs-fMRI. This modality is intended to capture intrinsic brain-network organization rather than task-evoked responses; however, uncontrolled variation in participants' internal mental state during scanning may have introduced noise that we could not measure directly. Promising next steps include leveraging larger cohorts, conducting prospective validation to support clinical translation, and fine-tuning the foundation model while developing more informative attribution analyses.

Declaration of competing interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability Study data was extracted from the BCH's electronic health record and cannot be shared publicly to protect the privacy of the participants. However, it can be shared upon request and subject to relevant approvals. All our source code is available at <https://github.com/cavalab/pediatric-headache-foundation-models>.

Acknowledgments

We would like to thank study physicians at Boston Children's Hospital Department of Sports Medicine, as well as Dr. Alicia Johnston and Dr. Catherine Lachenauer for their assistance with recruitment.

Funding Support

This work was supported in part by Boston Children's Hospital (BCH) and the National Institutes of Health (NIH) Grant Number R01NS125265. The neuroimaging research was supported by the Office of the Director, National Institutes of Health under Award Number S10OD025111. G.S.I.A. is supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) fellowship Finance Code 001.

References

- Marco A Arruda, Vincenzo Guidetti, Federica Galli, Regina CAP Albuquerque, and ME Bigal. Frequent headaches in the preadolescent pediatric population: a population-based study. *Neurology*, 74(11):903–908, 2010.
- Heidi K. Blume. Pediatric headache: A review. *Pediatrics In Review*, 33(12):562–576, December 2012. ISSN 1526-3347. doi: 10.1542/pir.33.12.562. URL <http://dx.doi.org/10.1542/pir.33.12.562>.

- Matthew Brett, Ingrid S. Johnsrude, and Adrian M. Owen. The problem of functional localization in the human brain. *Nature Reviews Neuroscience*, 3(3):243–249, March 2002. ISSN 1471-0048. doi: 10.1038/nrn756. URL <https://www.nature.com/articles/nrn756>.
- Katherine S. Button, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376, April 2013. ISSN 1471-0048. doi: 10.1038/nrn3475. URL <http://dx.doi.org/10.1038/nrn3475>.
- B.J. Casey, Tariq Cannonier, May I. Conley, Alexandra O. Cohen, Deanna M. Barch, Mary M. Heitzeg, Mary E. Soules, Theresa Teslovich, Danielle V. Dellarco, Hugh Garavan, Catherine A. Orr, Tor D. Wager, Marie T. Banich, Nicole K. Speer, Matthew T. Sutherland, Michael C. Riedel, Anthony S. Dick, James M. Bjork, Kathleen M. Thomas, Bader Chaarani, Margie H. Mejia, Donald J. Hagler, M. Daniela Cornejo, Chelsea S. Sicut, Michael P. Harms, Nico U.F. Dosenbach, Monica Rosenberg, Eric Earl, Hauke Bartsch, Richard Watts, Jonathan R. Polimeni, Joshua M. Kuperman, Damien A. Fair, and Anders M. Dale. The adolescent brain cognitive development (abcd) study: Imaging acquisition across 21 sites. *Developmental Cognitive Neuroscience*, 32:43–54, August 2018. ISSN 1878-9293. doi: 10.1016/j.dcn.2018.03.001. URL <http://dx.doi.org/10.1016/j.dcn.2018.03.001>.
- Catherine D Chong, Nathan Gaw, Yinlin Fu, Jing Li, Teresa Wu, and Todd J Schwedt. Migraine classification using magnetic resonance imaging resting-state functional connectivity data. *Cephalalgia*, 37(9):828–844, August 2017. ISSN 0333-1024, 1468-2982. doi: 10.1177/0333102416652091. URL <http://journals.sagepub.com/doi/10.1177/0333102416652091>.
- J-C Cuvelier, J-M Cuisset, and L Vallée. Les céphalées chroniques quotidiennes de l’enfant et de l’adolescent. *Archives de pédiatrie*, 15(12):1805–1814, 2008.
- Gina Dumkrieger, Catherine D Chong, Katherine Ross, Visar Berisha, and Todd J Schwedt. The value of brain MRI functional connectivity data in a machine learning classifier for distinguishing migraine from persistent post-traumatic headache. *Frontiers in Pain Research*, 3:1012831, January 2023. ISSN 2673-561X. doi: 10.3389/fpain.2022.1012831. URL <https://www.frontiersin.org/articles/10.3389/fpain.2022.1012831/full>.
- Jacob Gildenblat and contributors. Pytorch library for cam methods. <https://github.com/jacobgil/pytorch-grad-cam>, 2021.
- Albert Gu and Tri Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces, May 2024. URL <http://arxiv.org/abs/2312.00752>. arXiv:2312.00752 [cs].
- Headache Classification Committee of the International Headache Society (IHS). The international classification of headache disorders, 3rd edition. *Cephalalgia*, 38(1):1–211, 2018. doi: 10.1177/0333102417738202. URL <https://ichd-3.org/>.

- Scott Holmes, Joud Mar'i, Laura E. Simons, David Zurakowski, Alyssa Ann LeBel, Michael O'Brien, and David Borsook. Integrated Features for Optimizing Machine Learning Classifiers of Pediatric and Young Adults With a Post-Traumatic Headache From Healthy Controls. *Frontiers in Pain Research*, 3:859881, May 2022. ISSN 2673-561X. doi: 10.3389/fpain.2022.859881. URL <https://www.frontiersin.org/articles/10.3389/fpain.2022.859881/full>.
- Scott Holmes, Nicholas Reyes, Jaxon J. Huang, Anat Galor, Pradip M. Pattany, Elizabeth R. Felix, and Eric A. Moulton. Disentangling the neurological basis of chronic ocular pain using clinical, self-report, and brain imaging data: use of K-means clustering to explore patient phenotypes. *Frontiers in Neurology*, 14:1265082, November 2023. ISSN 1664-2295. doi: 10.3389/fneur.2023.1265082. URL <https://www.frontiersin.org/articles/10.3389/fneur.2023.1265082/full>.
- Amogh Joshi, Md Mahfuzur Rahman Siddiquee, Jay Shah, Todd Schwedt, Catherine Chong, Baoxin Li, and Teresa Wu. Prediction of Headache Improvement Using Multimodal Machine Learning in Patients with Acute Post-traumatic Headache (P5-12.002). *Neurology*, 102(7_supplement_1):3391, April 2024. doi: 10.1212/WNL.0000000000205109. URL <https://www.neurology.org/doi/10.1212/WNL.0000000000205109>. Publisher: Wolters Kluwer.
- Meenakshi Khosla, Keith Jamison, Gia H Ngo, Amy Kuceyeski, and Mert R Sabuncu. Machine learning in resting-state fmri analysis. *Magnetic resonance imaging*, 64:101–121, 2019.
- Peter Yongho Kim, Junbeom Kwon, Sunghwan Joo, Sangyoon Bae, Donggyu Lee, Yoonho Jung, Shinjae Yoo, Jiook Cha, and Taesup Moon. SwiFT: Swin 4D fMRI Transformer, October 2023. URL <http://arxiv.org/abs/2307.05916>. arXiv:2307.05916 [cs].
- Raquel Langdon and Marc T. DiSabella. Pediatric headache: An overview. *Current Problems in Pediatric and Adolescent Health Care*, 47(3):44–65, 2017a. ISSN 1538-5442. doi: <https://doi.org/10.1016/j.cppeds.2017.01.002>. URL <https://www.sciencedirect.com/science/article/pii/S1538544217300305>.
- Raquel Langdon and Marc T DiSabella. Pediatric headache: an overview. *Current Problems in Pediatric and Adolescent Health Care*, 47(3):44–65, 2017b.
- Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, August 2021. URL <http://arxiv.org/abs/2103.14030>. arXiv:2103.14030 [cs].
- Nikos K Logothetis. What we can do and what we cannot do with fmri. *Nature*, 453(7197): 869–878, 2008.
- Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

- Lingchao Mao, Jing Li, Todd J. Schwedt, Teresa Wu, Katherine Ross, Gina Dumkrieger, Dani C. Smith, and Catherine D. Chong. Identifying and predicting headache trajectories among those with acute post-traumatic headache. *Headache: The Journal of Head and Face Pain*, 65(7):1124–1133, July 2025. ISSN 0017-8748, 1526-4610. doi: 10.1111/head.14955. URL <https://headachejournal.onlinelibrary.wiley.com/doi/10.1111/head.14955>.
- Scott Marek, Brenden Tervo-Clemmens, Finnegan J. Calabro, David F. Montez, Benjamin P. Kay, Alexander S. Hatoum, Meghan Rose Donohue, William Foran, Ryland L. Miller, Timothy J. Hendrickson, Stephen M. Malone, Sridhar Kandala, Eric Feczko, Oscar Miranda-Dominguez, Alice M. Graham, Eric A. Earl, Anders J. Perrone, Michaela Cordova, Olivia Doyle, Lucille A. Moore, Gregory M. Conan, Johnny Uriarte, Kathy Snider, Benjamin J. Lynch, James C. Wilgenbusch, Thomas Pengo, Angela Tam, Jianzhong Chen, Dillan J. Newbold, Annie Zheng, Nicole A. Seider, Andrew N. Van, Athanasia Metoki, Roselyne J. Chauvin, Timothy O. Laumann, Deanna J. Greene, Steven E. Petersen, Hugh Garavan, Wesley K. Thompson, Thomas E. Nichols, B. T. Thomas Yeo, Deanna M. Barch, Beatriz Luna, Damien A. Fair, and Nico U. F. Dosenbach. Reproducible brain-wide association studies require thousands of individuals. *Nature*, 603(7902):654–660, March 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-04492-9. URL <http://dx.doi.org/10.1038/s41586-022-04492-9>.
- Zahra Ofoghi, Christiane S Rohr, Deborah Dewey, Signe Bray, Keith Owen Yeates, Melanie Noel, and Karen M Barlow. Functional connectivity of the anterior cingulate cortex with pain-related regions in children with post-traumatic headache. *Cephalalgia Reports*, 4, January 2021. ISSN 2515-8163. doi: 10.1177/25158163211009477. URL <http://dx.doi.org/10.1177/25158163211009477>.
- Callie Pawlowski, Cierra Buckman, Dmitry Tumin, Aimee W. Smith, and Jennifer Crotty. National trends in pediatric headache and associated functional limitations. *Clinical Pediatrics*, 58(14):1502–1508, 2019. doi: 10.1177/0009922819875560. URL <https://doi.org/10.1177/0009922819875560>. PMID: 31522542.
- Francisco Pereira, Tom Mitchell, and Matthew Botvinick. Machine learning classifiers and fmri: a tutorial overview. *Neuroimage*, 45(1):S199–S209, 2009.
- Russell A. Poldrack, Chris I. Baker, Joke Durnez, Krzysztof J. Gorgolewski, Paul M. Matthews, Marcus R. Munafò, Thomas E. Nichols, Jean-Baptiste Poline, Edward Vul, and Tal Yarkoni. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, 18(2):115–126, January 2017. ISSN 1471-0048. doi: 10.1038/nrn.2016.167. URL <http://dx.doi.org/10.1038/nrn.2016.167>.
- Giovanni Prezioso, Agnese Suppiej, Valentina Alberghini, Patrizia Bergonzini, Maria Elena Capra, Ilaria Corsini, Alessandro De Fanti, Elisa Fiumana, Martina Fornaro, Lucia Marangio, et al. Pediatric headache in primary care and emergency departments: consensus with rand/ucla method. *Life*, 12(2):142, 2022.
- Dong Qiu, Zhaoli Ge, Yanliang Mei, Wei Wang, Zhonghua Xiong, Xiaoshuang Li, Ziyu Yuan, Peng Zhang, Mantian Zhang, Xin Liu, Yaqing Zhang, Xueying Yu, Hefei Tang,

- and Yonggang Wang. Mapping brain functional networks topological characteristics in new daily persistent headache: a magnetoencephalography study. *The Journal of Headache and Pain*, 24(1):161, December 2023. ISSN 1129-2377. doi: 10.1186/s10194-023-01695-x. URL <https://thejournalofheadacheandpain.biomedcentral.com/articles/10.1186/s10194-023-01695-x>.
- Regina Rendas-Baum, Min Yang, Sepideh F Varon, Lisa M Bloudek, Ronald E DeGryse, and Mark Kosinski. Validation of the headache impact test (hit-6) in patients with chronic migraine. *Health and Quality of Life Outcomes*, 12(1), August 2014. ISSN 1477-7525. doi: 10.1186/s12955-014-0117-0. URL <http://dx.doi.org/10.1186/s12955-014-0117-0>.
- Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2): 336–359, October 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01228-7. URL <http://dx.doi.org/10.1007/s11263-019-01228-7>.
- Shashi S Seshia. Chronic daily headache in children and adolescents. *Current pain and headache reports*, 16:60–72, 2012.
- Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):e1001779, March 2015. ISSN 1549-1676. doi: 10.1371/journal.pmed.1001779. URL <http://dx.doi.org/10.1371/journal.pmed.1001779>.
- D.C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T.E.J. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S.W. Curtiss, S. Della Penna, D. Feinberg, M.F. Glasser, N. Harel, A.C. Heath, L. Larson-Prior, D. Marcus, G. Michalareas, S. Moeller, R. Oostenveld, S.E. Petersen, F. Prior, B.L. Schlaggar, S.M. Smith, A.Z. Snyder, J. Xu, and E. Yacoub. The human connectome project: A data acquisition perspective. *NeuroImage*, 62(4): 2222–2231, October 2012. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2012.02.018. URL <http://dx.doi.org/10.1016/j.neuroimage.2012.02.018>.
- Gaël Varoquaux and R Cameron Craddock. Learning and comparing functional connectomes across subjects. *NeuroImage*, 80:405–415, 2013.
- Cheng Wang, Yu Jiang, Zhihao Peng, Chenxin Li, Chang-Bae Bang, Lin Zhao, Wanyi Fu, Jinglei Lv, Jorge Sepulcre, Carl Yang, Lifang He, Tianming Liu, Xue-Jun Kong, Quanzheng Li, Daniel S. Barron, Anqi Qiu, Randy Hirschtick, Byung-Hoon Kim, Hongbin Han, Xiang Li, and Yixuan Yuan. Towards a general-purpose foundation model for functional mri analysis, April 2026. ISSN 2157-846X. URL <http://dx.doi.org/10.1038/s41551-026-01666-y>.
- Yuxiang Wei, Yanteng Zhang, Xi Xiao, Chengxuan Qian, Tianyang Wang, and Vince D. Calhoun. fMRI-LM: Towards a Universal Foundation Model for Language-Aligned

fMRI Understanding, December 2025. URL <http://arxiv.org/abs/2511.21760>. arXiv:2511.21760 [cs].

Susan Whitfield-Gabrieli and Alfonso Nieto-Castanon. Conn: A functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain Connectivity*, 2(3):125–141, June 2012. ISSN 2158-0022. doi: 10.1089/brain.2012.0073. URL <http://dx.doi.org/10.1089/brain.2012.0073>.

Paul E Youssef and Kenneth J Mack. Episodic and chronic migraine in children. *Developmental Medicine & Child Neurology*, 62(1):34–41, 2020. doi: <https://doi.org/10.1111/dmcn.14338>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/dmcn.14338>.

Appendix A. Additional Experimental Details

This appendix reports the different parameter values tested for each model in preliminary experiments, as well as detailed information about data preparation before generating the different dataset representations.

A.1. Data preprocessing

First, the fMRI data were processed using the CONN toolbox and mapped to MNI space. The data obtained after this preprocessing stage were used to generate the functional connectivity matrices.

The NeuroSTORM documentation states that fMRI data should be preprocessed and aligned to MNI152 using a preprocessing tool, which was performed in the previous step. We then adopted the preprocessing pipeline provided by NeuroSTORM by applying spatial interpolation to dimensions smaller than the required size and center-cropping dimensions larger than 96 voxels.

A z-score normalization was then applied after masking out-of-brain regions. The mask was also used to completely nullify out-of-brain regions, avoiding additional noise from areas irrelevant to the task.

Finally, the temporal dimension was reduced to 20 non-overlapping frames using average pooling.

A.2. Model Configurations

As stated in §3.4, given the low sample size relative to the number of features, as well as the use of bootstrapping and LOO cross-validation, extensive hyperparameter tuning could easily lead to overfitting. Therefore, we intentionally avoided extensive hyperparameter optimization and restricted tuning primarily to parameters that directly affect model complexity. Table 4 reports the parameters used for the ML models. Parameters not listed were left at their library defaults using scikit-learn v1.8.0. During the final experiments, training metrics and SHAP values were recorded for model inspection.

The ML model using the embedding representation consisted of Logistic Regression with L1 regularization, whose performance was higher in terms of AUROC when trained with data obtained from the functional connectivity representation.

A.3. Finetuning NeuroSTORM

We performed an exploratory fine-tuning experiment with NeuroSTORM by unfreezing different combinations of the `norm2` layers in the encoder architecture and combining them with a classification head. We illustrate the NeuroSTORM encoder architecture in Figure 6.

First, we performed the training of NeuroSTORM with a classification head using the fixed parameters reported in Table 5 and unfreezing every possible combination of the `norm2` layers at the end of each block (4 blocks in total; 16 possible configurations).

The setting that yielded the highest performance in terms of AUROC was unfreezing only the final layer of the initial block, where spatial reduction is minimal. We then boot-

Table 4: Parameters of the baseline machine learning models. Values in brackets represents different tested configurations.

Model	Parameter	Value
Logistic Regression L1 (LR L1)	l1_ratio	1.0 (equivalent to penalty='l1')
	solver	saga
	max_iter	{100, 1000, 5000}
Logistic Regression L2 (LR L2)	l1_ratio	0.0 (equivalent to penalty='l2')
	solver	liblinear
	max_iter	{100, 1000, 5000}
Decision Tree Classifier (DT)	max_depth	{None, 2, 3, 5}
Gradient Boosting Classifier (GB)	n_estimators	{5, 100}
	max_depth	{3, default}
	min_samples_leaf	{10, default}
Random Forest Classifier (RF)	n_estimators	{5, 100}
	max_depth	{None, 3}
	min_samples_leaf	{10, default}

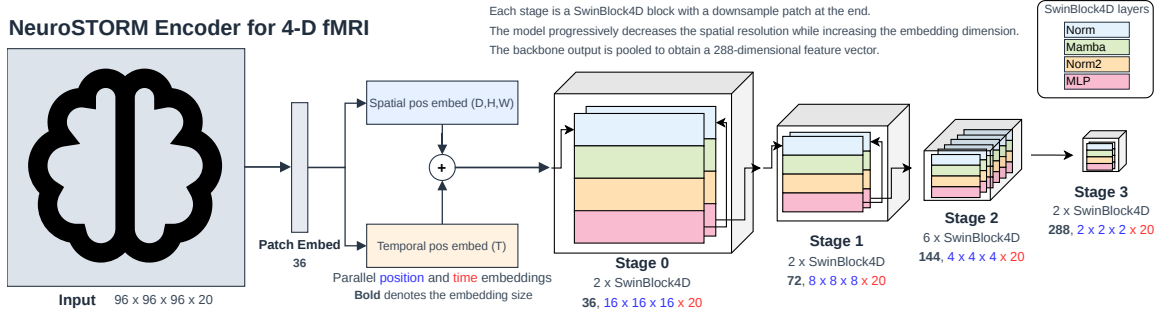


Figure 6: Architecture of the NeuroSTORM encoder backbone. A preprocessed 4D fMRI sequence ($96 \times 96 \times 96 \times 20$) is first partitioned into non-overlapping spatial patches and projected into an embedding space. The resulting features are processed through four hierarchical Mamba-based blocks, where the spatial resolution decreases ($16^3 \rightarrow 8^3 \rightarrow 4^3 \rightarrow 2^3$) while the embedding dimension increases ($36 \rightarrow 72 \rightarrow 144 \rightarrow 288$). Global pooling produces a 288-dimensional feature representation for downstream prediction.

strapped the predictions of the best performing configuration, achieving an AUROC of 0.74 ± 0.04 and an AUPRC of 0.91 ± 0.02 , outperforming the demographic model but underperforming a linear classifier trained on NeuroSTORM embeddings. We found that this fine-tuning approach actually decreased performance, which could be due to the small sample size inherent to this task. We expect that, given the complexity of pediatric headache, meaningful gains may require fine-tuning on larger pediatric-specific datasets.

Table 5: Parameters of the NeuroSTORM classifier head.

Parameter	Value
hidden_dim	68
dropout	0.1
learning rate	1×10^{-3}
weight decay	1×10^{-4}
batch size	16
epochs	100
optimizer	AdamW
feature scaling	StandardScaler
loss function	BCEWithLogitsLoss (binary) / CrossEntropyLoss (multiclass)