OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

Evaluating Structural MRI and Alcohol-Use Covariates in Machine Learning Models of Heavy Cannabis Use: A Longitudinal Pilot Study

Brandon Kong
brandon.kong10@gmail.com

ABSTRACT

Machine learning is increasingly used to identify neuroimaging patterns associated with substance use, but small public datasets make it difficult to determine whether a model is learning brain-based signals or simpler non-imaging variables. This study evaluated whether T1-weighted structural magnetic resonance imaging (MRI) features improved the classification of heavy cannabis users versus healthy controls beyond demographic and alcohol-use covariates. T1-weighted MRI is an anatomical MRI scan type commonly used to visualize brain structure. In this paper, covariates refer to non-MRI variables that may help explain group differences, including age, gender, and baseline Alcohol Use Disorders Identification Test (AUDIT) score. A higher AUDIT score indicates greater alcohol-use risk or severity. The analysis used OpenNeuro ds000174, a public longitudinal dataset containing baseline and three-year follow-up MRI scans from 20 heavy cannabis users and 22 healthy controls, along with Cannabis Use Disorder Identification Test (CUDIT) scores, where higher CUDIT scores indicate greater cannabis-use problems. Baseline MRI and longitudinal MRI-change features were generated through intensity normalization, resizing, voxel flattening, and principal component analysis (PCA). Logistic regression was evaluated under stratified five-fold cross-validation. The covariate-only model achieved the strongest performance, with balanced accuracy of 0.620 and receiver operating characteristic area under the curve (ROC-AUC) of 0.608. Baseline MRI, longitudinal MRI change, combined MRI features, and MRI-plus-covariate features did not improve performance. These findings suggest that, in this limited public dataset, simple voxel-based structural MRI representations did not provide stable classification value beyond demographic and alcohol-use covariates.

Keywords: cannabis use; structural MRI; alcohol-use covariates; machine learning; longitudinal neuroimaging; principal component analysis

INTRODUCTION

Cannabis use is a complex public health and neurobiological issue because it intersects with behavior, development, mental health, and social context. Neuroimaging studies have attempted to determine whether heavy cannabis exposure is associated with measurable differences in brain structure or function. In this study, T1-weighted MRI refers to an anatomical MRI sequence that provides structural information about brain tissue and is commonly used in studies of brain anatomy. The interpretation of such findings remains difficult, however, because cannabis use may co-occur with alcohol use, demographic variables, psychiatric symptoms, tobacco use, or other behavioral factors. A model that distinguishes cannabis users from healthy controls may therefore not be learning a cannabis-specific neuroanatomical pattern. It may instead be capturing variables that differ between the groups for reasons only indirectly related to brain structure.

These additional variables are often called covariates. In this study, covariates are non-MRI variables that may influence, explain, or help predict group differences. Age and gender are demographic covariates, while the AUDIT score is an alcohol-use covariate. Covariates are important because they provide a comparison point for interpreting MRI-based classification. If cannabis users also have higher alcohol-use scores than healthy controls, then a classifier may separate the two groups partly through alcohol-use information rather than through MRI-derived brain features alone. Including covariate-only and MRI-plus-covariate models therefore helps test whether structural MRI features provide additional predictive value beyond simpler demographic and behavioral information.

Machine learning intensifies this concern because classification performance is sometimes treated as evidence that a model has discovered a meaningful biological signal. This interpretation can be misleading in neuroimaging. Structural MRI data are high-dimensional, often containing hundreds of thousands of voxel measurements per subject. A voxel is a small three-dimensional unit of an MRI image, similar to a pixel but representing a tiny volume of tissue. Public substance-use MRI datasets, by contrast, may contain only a few dozen participants. The resulting imbalance between feature dimensionality and sample size poses a substantial risk of overfitting, in which a model learns sample-specific noise rather than patterns that generalize to new participants. Dimensionality reduction and cross-validation can reduce this risk, but they do not by themselves show that imaging features provide incremental value, meaning additional predictive information, beyond simpler non-imaging variables.

This paper addresses that methodological issue using OpenNeuro ds000174, a public longitudinal dataset of heavy cannabis users and healthy controls. Rather than presenting a high-accuracy diagnostic model, the study asks a narrower question: do simple T1-weighted structural MRI features improve classification of heavy cannabis users beyond demographic and alcohol-use covariates? This question is important because alcohol-use severity is not merely a nuisance variable in substance-use research. It may reflect

co-use, behavioral risk, or broader substance-use severity. If alcohol-use covariates perform as well as or better than MRI features, MRI-only interpretations should be treated cautiously.

The prediction task was binary classification: given a participant's available features, the model was trained to predict whether the participant belonged to the heavy cannabis-use group or the healthy-control group. The analysis compared five feature configurations: covariates only, baseline MRI only, longitudinal MRI change only, baseline plus longitudinal MRI, and MRI combined with covariates. Longitudinal MRI change refers to within-participant change from the baseline scan to the three-year follow-up scan, calculated by comparing each participant's follow-up MRI features with that same participant's baseline MRI features. Each configuration was evaluated under the same leakage-safe cross-validation framework so that the comparison focused on the predictive value of the input information rather than on differences in classifier architecture.

The main contributions of this study are as follows:

- It evaluates whether simple structural MRI features add predictive value beyond demographic and alcohol-use covariates in a public longitudinal cannabis-use dataset.
- It distinguishes 42 independent participants from 84 total MRI scans, reducing the risk of overstating the effective sample size.
- It compares baseline MRI features, longitudinal MRI-change features, combined MRI features, covariates, and MRI-plus-covariate feature sets under the same validation design.
- It shows that, in this dataset, age, gender, and baseline AUDIT score provided more stable classification performance than the simple voxel-based MRI representations.

METHODS

Dataset

This study used the OpenNeuro dataset ds000174[1,2], a public T1-weighted structural MRI dataset containing heavy cannabis users and healthy controls scanned at baseline and three-year follow-up. The dataset contains 42 participants and 84 anatomical MRI scans. The number 84 refers to scans, not independent participants, because each participant contributed one baseline scan and one three-year follow-up scan. The cohort included 20 heavy cannabis users and 22 healthy controls. Group status, coded as heavy cannabis user versus healthy control, served as the binary prediction target. The participants' file also included demographic and behavioral variables, including gender, age at baseline, CUDIT total scores, and AUDIT total scores. AUDIT is a screening instrument for hazardous and harmful alcohol consumption, with higher scores generally indicating greater alcohol-use risk or severity.[3] CUDIT is used to assess cannabis-use problems, with higher scores indicating greater cannabis-use-related concerns.[4]

Because the dataset is publicly available and de-identified, this project used secondary analysis only. No new human-subject data were collected. The analysis was not intended to diagnose cannabis use or infer

individual clinical status. Instead, it evaluated whether simple structural MRI features added predictive information beyond demographic and alcohol-use covariates.

Cohort Characteristics

Table 1 summarizes baseline cohort characteristics. Cannabis users were slightly younger than controls on average, although the difference was not statistically significant. Baseline AUDIT scores were higher in cannabis users than controls, suggesting greater alcohol-use severity in the cannabis-use group, although the group difference was marginal. CUDIT scores strongly differed between groups, as expected, because high CUDIT scores indicate more cannabis-use problems and the group label itself distinguishes heavy cannabis users from controls. CUDIT was therefore not used as a primary predictor of cannabis-group status, since doing so would partly encode the target label.

Variable	Healthy Controls (N=22)	Cannabis Users (N=20)	Group Comparison
Age at baseline	21.56 ± 2.45	20.53 ± 2.11	p=0.151
AUDIT baseline	4.41 ± 3.38	6.25 ± 3.35	p=0.084
CUDIT baseline	0.05 ± 0.21	12.70 ± 6.59	$p=5.77 \times 10^{-8}$
Female / male	8 / 14	5 / 15	descriptive

Table 1: Baseline cohort characteristics. Values are reported as mean ± standard deviation unless otherwise noted.

This table shows that the groups differed most clearly in cannabis-use severity, while alcohol-use severity showed a smaller but still relevant group difference.

MRI Preprocessing

All MRI scans were processed using a reproducible Python-based pipeline. The goal was not to perform detailed anatomical segmentation, but to construct a simple voxel-level representation that could be tested against covariate-based models. Each scan was loaded as a three-dimensional image volume, converted to a floating-point numerical array, and cleaned of non-finite values, meaning missing, infinite, or undefined numerical entries. Nonzero intensities were clipped between the 1st and 99th percentiles to reduce the influence of extreme values. The nonzero voxels were then standardized within each scan:

$I'_v = (I_v - \mu\square z) / \sigma\square z$

where I_v is the clipped intensity at voxel v, and $\mu\square z$ and $\sigma\square z$ are the mean and standard deviation of nonzero voxels in the scan.

Each MRI volume was resized to a common 64 × 64 × 64 grid and flattened into a vector:
$x_i \in \mathbb{R}^p, \quad p = 64^3 = 262{,}144$

The resulting baseline MRI matrix was:
X_BL ∈ ℝ^(42 × 262144)

This means that each participant was represented by one baseline row, while each row contained 262,144 voxel values.

Longitudinal MRI-Change Features

Because the dataset includes both baseline and follow-up scans, longitudinal change features were generated. In this study, a longitudinal change feature represents how a participant's MRI voxel values changed from baseline to the three-year follow-up scan. For each subject i, the MRI-change vector was defined as:

$\Delta x_i = x_{i,}FU - x_{i,}BL$

where x_i, FU is the follow-up scan vector and x_i, BL is the baseline scan vector after identical preprocessing.

The longitudinal change matrix was:

X_Δ ∈ ℝ^(42 × 262144)

This representation was used to test whether three-year structural MRI changes distinguished cannabis users from healthy controls more effectively than baseline anatomy alone.

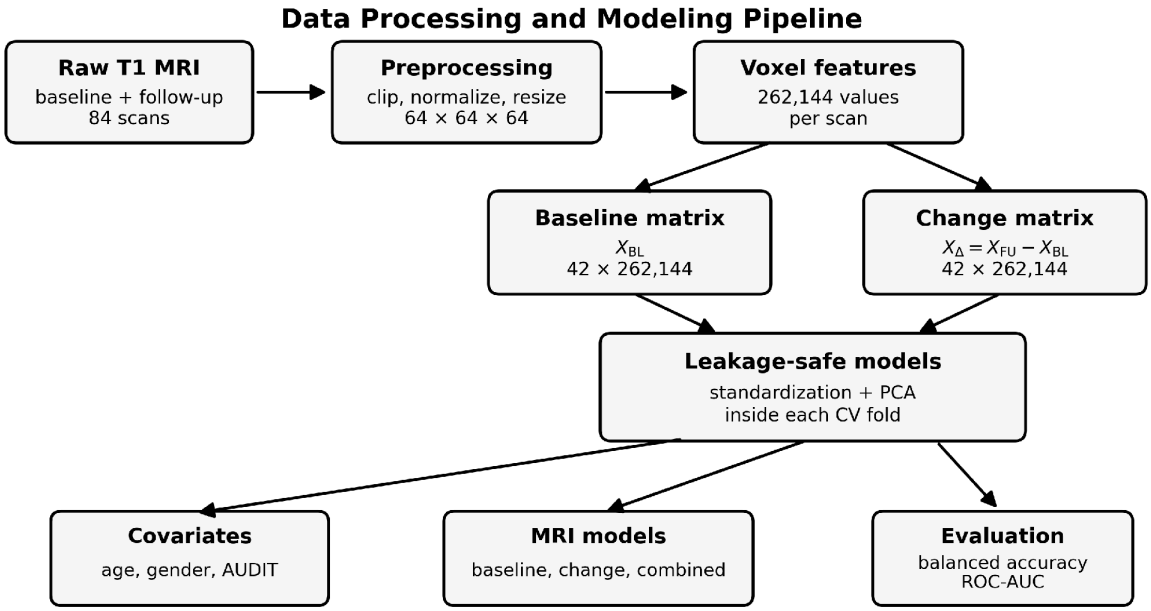


Figure 1: Overview of the analysis pipeline. Each T1-weighted MRI scan was normalized, resized, and converted into a voxel vector. Baseline and longitudinal-change matrices were evaluated alongside

demographic and alcohol-use covariates. Standardization and PCA were fit inside each cross-validation fold to reduce information leakage.

This pipeline shows how the raw MRI scans were converted into the baseline and change-based feature sets used for classification, while also showing where covariates entered the comparison.

Dimensionality Reduction and Model Families

The MRI matrices contained far more voxel features than participants. To make model fitting feasible and reduce overfitting, principal component analysis (PCA) was used to compress the high-dimensional voxel matrices into a smaller set of summary features. PCA identifies directions of variation in the training data and represents each participant using scores along those directions. These scores are called principal components and serve as compressed MRI features for the classifier.[5]

The analysis tested five feature configurations, referred to as model families for clarity. All model families used logistic regression as the classifier; the difference between them was the information supplied to the classifier. The covariate-only family used age, gender, and baseline AUDIT score. The baseline MRI family used baseline MRI-derived PCA features. The MRI-change family used longitudinal change-derived PCA features. The baseline-plus-change MRI family combined both MRI feature sets. The MRI-plus-covariate family combined MRI-derived features with the demographic and alcohol-use covariates.

Model family	Covariates	Baseline MRI	MRI Change
Covariates only	Yes	No	No
Baseline MRI	No	Yes	No
MRI change	No	No	Yes
Baseline + change MRI	No	Yes	Yes
MRI + covariates	Yes	Yes	Yes

Figure 2: Feature-design summary across the five primary model families. Each row represents a different feature configuration evaluated with the same general classifier.

This feature design helped test the central question of the study by keeping the classifier consistent while changing whether the model received covariates, MRI features, or both.

Model Evaluation

Models were evaluated using stratified five-fold cross-validation, a validation method that splits the dataset into five class-balanced folds and repeatedly trains on four folds while testing on the held-out fold. In each fold, all preprocessing steps, including standardization and PCA, were fit only on the training fold. This leakage-safe design reduces the risk that information from the test fold influences the training process.

Performance was summarized using accuracy, balanced accuracy, and ROC-AUC. Accuracy measures the overall fraction of correct predictions. Balanced accuracy averages performance across the two classes, which is useful when class sizes are not perfectly equal:

Balanced Accuracy = 1/2 × [TP/(TP+FN) + TN/(TN+FP)]
where TP, FN, TN, and FP refer to true positives, false negatives, true negatives, and false positives, respectively. ROC-AUC measures how well predicted probabilities rank cannabis users above healthy controls across possible decision thresholds. A ROC-AUC of 0.5 indicates chance-level discrimination, meaning the model ranks the two groups no better than random guessing, while a ROC-AUC of 1.0 indicates perfect discrimination.

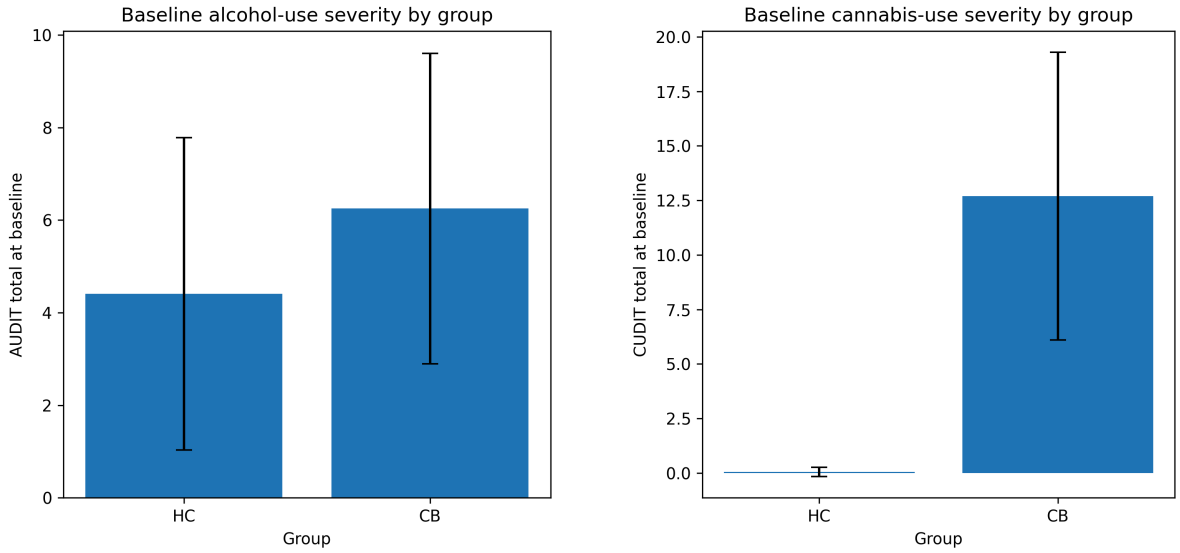
The pipeline was implemented in Python using NumPy, pandas, nibabel, scipy, matplotlib, and scikit-learn.[6] Mean and standard deviation across folds were reported.

RESULTS

Baseline Alcohol and Cannabis Severity

The baseline cohort showed clear separation in cannabis-use severity, as expected from the group definitions. Cannabis users had much higher baseline CUDIT scores than healthy controls. AUDIT scores were also higher in cannabis users than controls, although the difference was marginal rather than conventionally significant. This pattern is important because it suggests that alcohol-use severity may be relevant when interpreting cannabis-related group differences.

Figure 3 visualizes baseline AUDIT and CUDIT scores by group. The CUDIT difference confirms the expected cannabis-use severity contrast because higher CUDIT scores indicate greater cannabis-use problems. The AUDIT difference is more relevant to the modeling question because AUDIT measures alcohol-use severity, a non-cannabis behavioral covariate that may partially distinguish groups.



(a) Baseline AUDIT score (b) Baseline CUDIT score

Figure 3: Baseline alcohol-use and cannabis-use severity by group. HC refers to healthy controls, and CB refers to heavy cannabis users. Higher AUDIT scores indicate greater alcohol-use risk or severity, while higher CUDIT scores indicate greater cannabis-use problems. AUDIT scores were higher in cannabis users than in controls, although the difference was marginal in this sample. CUDIT scores strongly separated the groups, as expected from the group definition.

Together, these results show why covariates were important for this study. CUDIT confirmed the expected cannabis-use separation between groups, while AUDIT showed that alcohol-use severity also differed to some extent, meaning that classification performance could reflect behavioral covariates rather than MRI features alone.

Cross-Validated Classification Performance

The primary leakage-safe model-family comparison is summarized in Table 2. The covariate-only model achieved the strongest overall performance, with a mean balanced accuracy of 0.620 and ROC-AUC of 0.608. Baseline MRI alone achieved balanced accuracy of 0.590, but its ROC-AUC was close to chance at 0.522 and showed substantial fold-to-fold variation. Longitudinal MRI change performed similarly, with balanced accuracy of 0.565 and ROC-AUC of 0.528.

Combining MRI features did not improve classification. The baseline-plus-change MRI model had balanced accuracy of 0.510 and ROC-AUC of 0.520. The model combining MRI features with demographic and alcohol-use covariates also failed to improve performance, producing the same mean balanced accuracy and ROC-AUC as the combined MRI model. These results suggest that, under the current feature representation and validation scheme, MRI features did not add stable predictive value beyond the non-imaging covariates.

Model family	Accuracy	Balanced Accuracy	ROC-AUC
Covariates only	0.625 ± 0.154	0.620 ± 0.156	0.608 ± 0.162
Baseline MRI	0.586 ± 0.210	0.590 ± 0.214	0.522 ± 0.275
MRI change	0.550 ± 0.071	0.565 ± 0.056	0.528 ± 0.061
Baseline + change MRI	0.497 ± 0.125	0.510 ± 0.132	0.520 ± 0.156
MRI + covariates	0.497 ± 0.125	0.510 ± 0.132	0.520 ± 0.156

Table 2: Leakage-safe cross-validated model performance. Values are the mean ± standard deviation across five folds.

Classification improvement was judged by comparing each model's balanced accuracy and ROC-AUC against the covariate-only model and against chance-level performance. Balanced accuracy and ROC-AUC values near 0.5 indicate limited discrimination between heavy cannabis users and healthy controls. Because none of the MRI-based feature configurations exceeded the covariate-only model on these metrics, the results did not show improved classification from adding MRI features.

Figure 4 shows the same comparison using balanced accuracy. Although baseline MRI and MRI-change models were slightly above chance in balanced accuracy, neither exceeded the covariate-only model. Figure 5 shows that ROC-AUC remained close to chance for all MRI-based models.

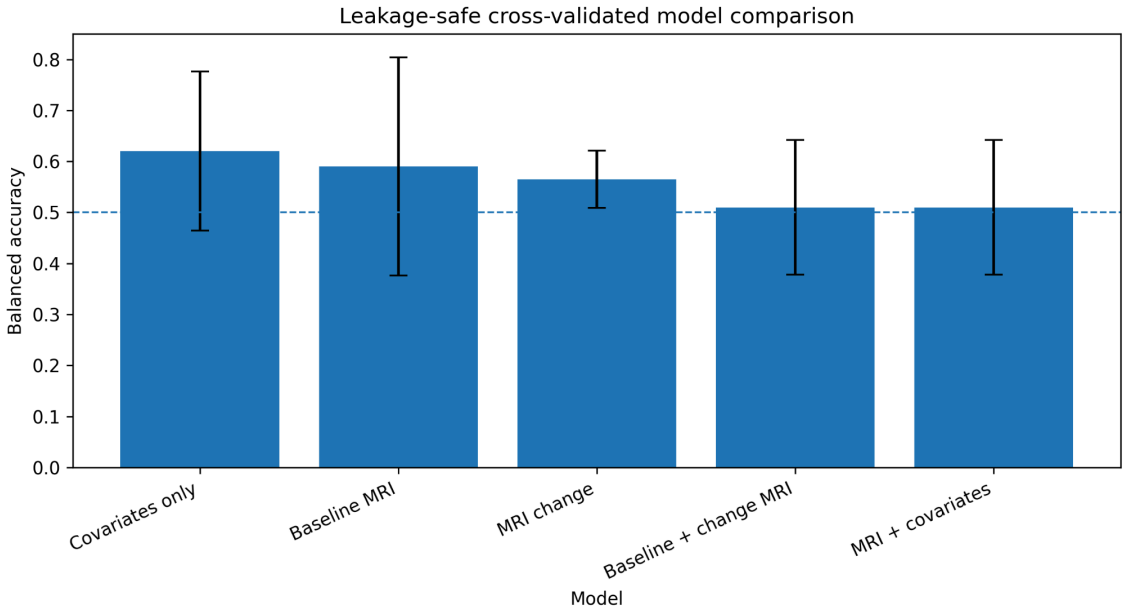


Figure 4: Leakage-safe cross-validated balanced accuracy across model families. The dashed line marks chance-level performance. Error bars indicate standard deviation across folds.

This figure supports the main result because the covariate-only model had the highest mean balanced accuracy, while the MRI-based models were close to chance or showed substantial fold-to-fold variability.

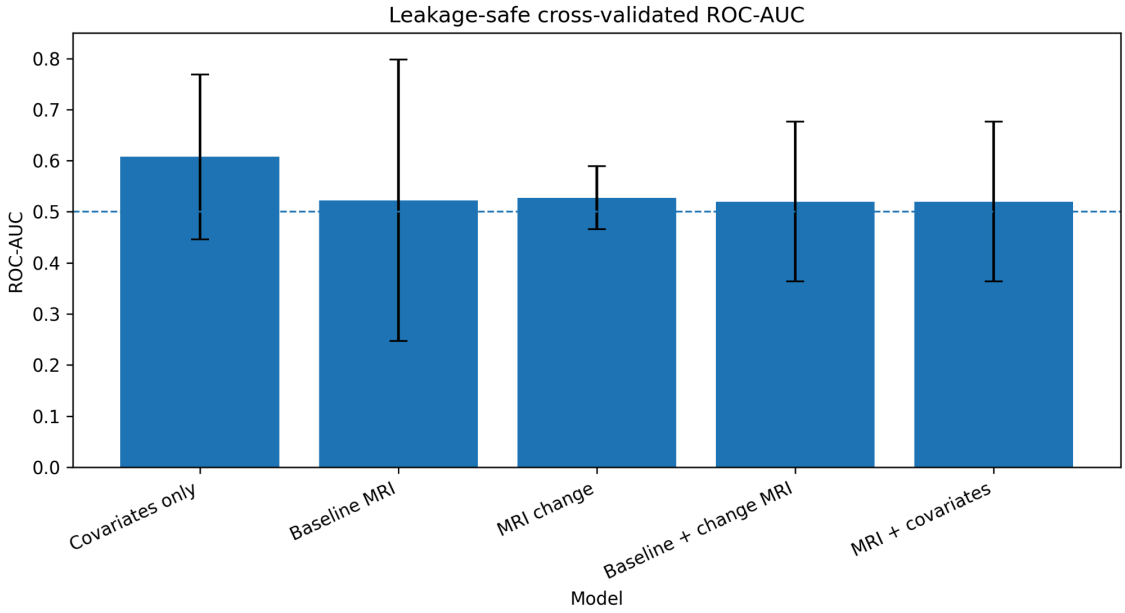


Figure 5: Leakage-safe cross-validated ROC-AUC across model families. MRI-based models remained close to chance-level discrimination, suggesting limited incremental predictive value from the simple voxel-based MRI representations. Error bars indicate standard deviation across folds.

The ROC-AUC results further support the same conclusion: MRI-based models did not rank cannabis users and healthy controls more effectively than the covariate-only model.

Exploratory PCA Feature Models

As a secondary exploratory analysis, precomputed baseline and longitudinal MRI change principal components were combined with covariates in tabular logistic regression models. These analyses were not treated as the primary results because the PCA features were computed before cross-validation, which can introduce information leakage. Nevertheless, they were useful as an additional signal analysis. The exploratory models did not identify a stronger MRI-based classifier. Covariates alone again performed best, while models using baseline principal components, change principal components, or all MRI principal components plus covariates remained near chance. The exploratory analysis supported the same overall conclusion: MRI-based models did not outperform the covariate-only model.

DISCUSSION

This study found that simple voxel-based structural MRI features did not improve the classification of heavy cannabis users versus healthy controls in a limited public longitudinal dataset. The strongest model family used only age, gender, and baseline AUDIT score. In contrast, baseline MRI features, longitudinal

MRI-change features, and combined MRI features performed near chance or showed unstable cross-validation behavior. These findings suggest that, for this dataset and modeling pipeline, alcohol-use and demographic covariates carried more stable predictive information than simple structural MRI representations.

The result should not be interpreted as evidence that MRI is irrelevant to cannabis-use research. Instead, it shows that one particular representation of MRI, namely downsampled voxel intensities reduced by PCA, did not provide robust classification value in this small sample. Structural MRI effects associated with cannabis use may be subtle, spatially localized, developmentally dependent, or better captured by region-level morphometric features such as cortical thickness, gray-matter volume, or surface area. The current pipeline did not use segmentation-derived neuroanatomical measures. It also did not use functional MRI, diffusion MRI, or task-based measures, which may capture different aspects of neural function or connectivity.

The more important implication concerns how covariates should be interpreted. In this dataset, cannabis users had somewhat higher AUDIT scores than controls. Although this difference was not significant at the conventional $p < 0.05$ threshold, it was large enough to matter in predictive modeling. The covariate-only model outperformed MRI-based models, suggesting that alcohol-use severity should not be ignored in cannabis neuroimaging analyses. This does not mean that AUDIT caused the group difference or that alcohol use explains all cannabis-related variation. Rather, it means that a simple alcohol-use measure carried more stable predictive information than the voxel-based MRI features in this dataset. If a model distinguishes cannabis users from controls, the signal may reflect cannabis-specific neural differences, alcohol-use patterns, demographic variation, or some combination of these factors.

This study also illustrates the value of cautionary results in biomedical machine learning. Many imaging projects emphasize model accuracy, but high accuracy in small datasets can be misleading when validation is not carefully designed. Here, the most scientifically meaningful result was not a high-performing classifier. It was the absence of stable MRI-added value under a leakage-safe validation scheme. Such findings can help prevent overinterpretation and encourage more careful modeling practices in public neuroimaging datasets.

LIMITATIONS

Several limitations should be emphasized. First, the sample size was limited, with only 42 subjects. Although the dataset includes 84 scans, those scans come from two time points per participant and should not be interpreted as 84 independent subjects. Cross-validation estimates from such a sample can have high variance, and model performance may be sensitive to fold composition. Second, MRI preprocessing was intentionally simple. The analysis used downsampled voxel intensities rather than brain-region measures derived from tools such as FreeSurfer or FastSurfer. Third, the feature representation may have removed useful anatomical information or preserved variation unrelated to cannabis exposure. Fourth, the

study evaluated the classification of group status rather than the causal effects of cannabis use. The design cannot establish whether cannabis use caused any brain differences. Fifth, alcohol use was represented through AUDIT scores, but other potentially relevant covariates, including psychiatric symptoms, tobacco use, socioeconomic variables, or medication history, were not modeled here.

A final limitation is that the study should not be interpreted as a clinical diagnostic model. The models were evaluated for methodological comparison in a public research dataset, not for individual-level diagnosis. The modest performance values, limited participant sample, and simple MRI representation all argue against clinical application.

FUTURE DIRECTIONS

Future research should test whether these findings hold in larger cannabis-use neuroimaging datasets with more participants and a broader set of covariates. Larger samples would allow more stable cross-validation estimates and would make it easier to evaluate whether MRI features add predictive value beyond demographic and substance-use variables. Future studies should also consider region-level anatomical measures, such as cortical thickness, gray-matter volume, or surface area, rather than relying only on downsampled voxel intensities. These features may better capture subtle structural differences that simple voxel-based PCA representations could miss. Finally, future work should examine additional covariates, including tobacco use, psychiatric symptoms, medication history, and socioeconomic variables, because these factors may also influence the interpretation of cannabis-related neuroimaging models.

CONCLUSION

In this limited public longitudinal structural MRI dataset, simple voxel-based MRI features did not robustly improve classification of heavy cannabis users versus healthy controls. A model family using age, gender, and baseline AUDIT score modestly outperformed baseline MRI, longitudinal MRI-change, combined MRI, and MRI-plus-covariate feature sets. These results suggest that alcohol-use severity and demographic variables may be more informative than crude structural MRI representations in limited cannabis-use neuroimaging datasets. More broadly, the study cautions against interpreting MRI-only machine learning models as evidence of cannabis-specific brain signatures unless covariates, sample size, and validation design are carefully addressed.

DATA AND CODE AVAILABILITY

The MRI data used in this study are publicly available through the OpenNeuro dataset ds000174. The analysis pipeline was implemented in Python using NumPy, pandas, nibabel, scipy, matplotlib, and scikit-learn. The project code generated baseline voxel matrices, longitudinal change matrices, PCA features, cross-validated model comparisons, and final figures. A reproducible code archive can be provided upon request or included as supplemental material if permitted by the journal.

ACKNOWLEDGMENTS

The author thanks the researchers who made the ds000174 dataset publicly available through OpenNeuro. This study was conducted as an independent secondary analysis of public data.

REFERENCES

Koenders, L., Cousijn, J., Vingerhoets, W. A. M., van den Brink, W., Wiers, R. W., Meijer, C. J., Machielsen, M. W. J., Veltman, D. J., Goudriaan, A. E., and de Haan, L. "T1-Weighted Structural MRI Study of Cannabis Users at Baseline and 3 Years Follow Up." *OpenNeuro Dataset ds000174*, 2015, https://openneuro.org/datasets/ds000174.

Markiewicz, C. J., Gorgolewski, K. J., Feingold, F., Blair, R., Halchenko, Y. O., Miller, E., Hardcastle, N., Wexler, J., Esteban, O., Goncalves, M., Jwa, A., and Poldrack, R. A. "The OpenNeuro Resource for Sharing of Neuroscience Data." *eLife*, vol. 10, 2021, e71774, https://doi.org/10.7554/eLife.71774.

Saunders, J. B., Aasland, O. G., Babor, T. F., de la Fuente, J. R., and Grant, M. "Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO Collaborative Project on Early Detection of Persons with Harmful Alcohol Consumption II." *Addiction*, vol. 88, no. 6, 1993, pp. 791–804, https://doi.org/10.1111/j.1360-0443.1993.tb02093.x.

Adamson, S. J., and Sellman, J. D. "A Prototype Screening Instrument for Cannabis Use Disorder: The Cannabis Use Disorders Identification Test (CUDIT) in an Alcohol-Dependent Clinical Sample." *Drug and Alcohol Review*, vol. 22, no. 3, 2003, pp. 309–315, https://doi.org/10.1080/0959523031000154454.

Pearson, K. "On Lines and Planes of Closest Fit to Systems of Points in Space." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, 1901, pp. 559–572, https://doi.org/10.1080/14786440109462720.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and

Duchesnay, E. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research*, vol. 12, 2011, pp. 2825–2830, https://jmlr.org/papers/v12/pedregosa11a.html.