This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Inflated Performance in fMRI-Based ASD Diagnosis: A Reproducibility Study of Data Leakage in Feature Selection Pipelines

Mohamad Bashar Disoki[1] and Riad Sonbol[1]

Higher Institute for Applied Sciences and Technology (HIAST), Damascus, Syria
{mohamadbashar.disoki,riad.sonbol}@hiast.edu.sy

Abstract. Machine learning models applied to resting-state functional MRI (rs-fMRI) data have reported increasingly high classification accuracy for Autism Spectrum Disorder (ASD) diagnosis, with some recent works claiming accuracies above 97%. In this paper we investigate whether such figures reflect genuine generalisation or are artefacts of data leakage introduced during feature selection. We conduct a systematic reproducibility study of two peer-reviewed ASD classification pipelines built on the public ABIDE I dataset: *MADE-for-ASD* [1] and *XAI-for-ASD* [2]. We identify and correct concrete leakage faults in both codebases — global F-score-based feature selection prior to cross-validation in the former, and pre-fold application of supervised Recursive Feature Elimination (SVM-RFE) in the latter — and quantify the resulting performance inflation. After correction, accuracy drops from 74.04% to 66.35% for MADE-for-ASD and from 97.68% to 65.02% (±2.51%) for XAI-for-ASD, revealing gaps of up to 33 percentage points. Independent corroboration from Dong et al. [?], who corrected a separate leakage fault in EV-GCN [9] on ABIDE I and observed a drop from approximately 81% to approximately 68%, provides convergent evidence that leakage-induced inflation is a recurring rather than isolated occurrence in this literature, though its broader prevalence across the field remains to be established. Beyond diagnosis of the problem, we provide corrected, publicly available forks of both codebases and propose practical safeguards. We further argue that large language model (LLM)-based code review tools, used here as a first-pass diagnostic aid complementing expert validation, represent a cost-effective quality-assurance layer for research software.

Keywords: Data leakage · Autism Spectrum Disorder · fMRI · Reproducibility · Feature selection · Cross-validation · AI code review

1 Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition affecting approximately 1 in 100 individuals worldwide according to the World Health

Organization [10], characterised by challenges in social communication and restricted, repetitive behaviours. Objective biomarkers that complement subjective behavioural assessments are urgently needed; resting-state functional MRI (rs-fMRI), which captures BOLD-derived functional connectivity between brain regions, has emerged as a promising candidate.

Over the past decade, machine learning and deep learning models applied to the publicly available ABIDE I dataset [3] have reported steadily rising classification accuracies, from around 66% with classical methods [5] to figures exceeding 97% in recent deep-learning pipelines [2]. Such numbers are routinely cited as evidence of clinical readiness. However, a recurring concern in the broader machine-learning reproducibility literature is that high accuracy figures in medical imaging studies can be driven by *data leakage* — the inadvertent incorporation of information from held-out test samples into the model training or feature selection process — rather than genuine generalisation [4].

Data leakage in neuroimaging pipelines most commonly arises during dimensionality reduction steps. High-dimensional fMRI connectivity matrices (often >6,000 features per sample) require aggressive feature selection before deep networks can be trained effectively. If this selection step uses class labels and is applied to the *full* dataset *before* cross-validation splits are formed, the model effectively sees the test labels during training, producing optimistic and misleading performance estimates.

In this paper we investigate two recently published, peer-reviewed ASD classification pipelines that together illustrate the breadth of this problem: (1) *MADE-for-ASD* [1], a multi-atlas deep ensemble that employs F-score-based feature ranking, and (2) *XAI-for-ASD* [2], an explainable deep-learning model that employs Support Vector Machine Recursive Feature Elimination (SVM-RFE). We reproduce the original results, identify the leakage faults, apply corrections, and re-evaluate performance. The corrected results reveal dramatic drops in accuracy, demonstrating that the reported scores are largely artefactual.

We also address numerical-stability issues discovered during the reproduction exercise and propose a set of concrete coding safeguards. Importantly, we demonstrate that AI-powered code review — specifically, Copilot augmented with the Claude Haiku 4.5 model from Anthropic — can surface these problems efficiently, and we recommend its systematic adoption in the neuroimaging community.

2 Background and Related Work

2.1 The ABIDE I Dataset

ABIDE I [3] aggregates rs-fMRI data and phenotypic information from 1,112 individuals (505 with ASD, 530 typical controls) collected at 17 international sites. After standard preprocessing and quality control, most studies work with approximately 1,035 samples. The multi-site, heterogeneous nature of ABIDE I makes it a challenging benchmark, and performance above ∼75% on the full dataset has historically been considered strong.

2.2 Data Leakage in Machine Learning

Data leakage occurs when information unavailable at inference time is used during model development, producing inflated evaluation metrics that do not generalise [4]. In feature-selection pipelines, leakage arises when a supervised selector — one that uses the class label y — is fitted on the combined training and test data before cross-validation. The selector therefore encodes test-set label information, giving the downstream classifier an unfair advantage.

Kapoor and Narayanan [4] documented this pattern across hundreds of published machine-learning papers and found that reported accuracies were systematically inflated. In neuroimaging, the problem is acute because (i) the high feature-to-sample ratio incentivises aggressive dimensionality reduction, and (ii) both F-score and RFE are supervised techniques that directly exploit the class label.

2.3 Feature Selection Methods Studied

F-score. The Fisher discriminant ratio ranks feature i by

$$F(i) = \frac{(\bar{r}_i^{(+)} - \bar{r}_i)^2 + (\bar{r}_i^{(-)} - \bar{r}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(r_{k,i}^{(+)} - \bar{r}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(r_{k,i}^{(-)} - \bar{r}_i^{(-)})^2}, \tag{1}$$

where $\bar{r}_i$, $\bar{r}_i^{(+)}$, $\bar{r}_i^{(-)}$ are the global, positive-class, and negative-class means, respectively. Because both numerator and denominator depend on class membership across all samples, computing F-scores on the full dataset before splitting constitutes target leakage.

SVM-RFE. Recursive Feature Elimination fits a support vector machine on labelled data and iteratively removes features with low weight magnitudes. As a supervised method it is even more susceptible to leakage than F-score: fitting the SVM on the whole dataset exposes the selector to test-set labels and effectively pre-optimises the feature space for the exact test examples.

3 Case Study 1: MADE-for-ASD

3.1 Original Pipeline

MADE-for-ASD [1] is a weighted deep ensemble that combines three brain atlases (CC200, AAL, EZ) with a Stacked Sparse Denoising Autoencoder (SSDAE) and multi-layer perceptron (MLP), augmented with demographic features. Feature selection using F-score is applied to the full 1,035-sample dataset to retain the top 15% of connectivity features per atlas. Only afterwards is the data partitioned into 10-fold cross-validation splits.

Table 1. MADE-for-ASD: performance before and after data-leakage correction on the whole ABIDE I CPAC dataset (10-fold cross-validation). Original results are reported as point estimates to match the single-run reporting style of the original paper.

Condition	Accuracy	Precision	F1	Sensitivity	Specificity	ROC-AUC
Original (with leakage)	0.7404	0.7547	0.7477	0.7407	0.7400	0.8107
Corrected (no leakage)	0.6635	0.6939	0.6602	0.6296	0.7000	0.7715
Drop	**−0.0769**	**−0.0608**	**−0.0875**	**−0.1111**	**−0.0400**	**−0.0392**

3.2 Leakage Fault Identified

The F-score computation in `prepare_data.py` ingests the complete dataset — including future test folds — before any split. As a consequence, the top-ranked features are chosen with knowledge of test-set class labels, biasing the input space toward the test distribution and inflating measured accuracy.

3.3 Correction

The corrected pipeline moves F-score feature selection *inside* the cross-validation loop. For each fold, features are ranked using only the training partition; the resulting feature indices are then applied to select columns in the validation and test partitions. This ensures that test-fold samples are never seen by the feature selector.

3.4 Results

Table 1 reports performance before and after correction, obtained by re-running the original code (`python prepare_data.py -folds=10 -whole cc200 aal ez` followed by `python nn.py -whole cc200 aal ez`) and the corrected fork.

In the corrected experiments we deliberately retain the original hyperparameters. Re-tuning after removing leakage would introduce a second source of variation alongside the leakage correction itself, making it impossible to attribute the performance change specifically to leakage. Holding all other factors constant is required to isolate the effect.

The corrected accuracy of 66.35% is consistent with scores reported by comparable non-leaking methods on the same dataset (e.g., 70% by Heinsfeld et al. [6] and 70.8% by Almuqhim and Saeed [7]). It also falls within the 65–72% range of test accuracies reported by Dong et al. [8] in an independent standardised benchmark on ABIDE I using per-fold tuning across multiple architectures, lending further confidence to its validity as a realistic generalisation estimate. The 7.69 percentage-point drop confirms that a non-trivial fraction of the original score was artefactual.

4 Case Study 2: XAI-for-ASD

4.1 Original Pipeline

XAI-for-ASD [2] combines a Stacked Sparse Autoencoder (SSAE) with a softmax classifier for ASD classification on ABIDE I (884 samples after framewise-displacement filtering). SVM-RFE reduces 6,670 functional connectivity features to 1,000 before training. Seven interpretability methods are benchmarked using the Remove-and-Retrain (ROAR) framework, and results are reported under 5-fold cross-validation. The reported peak accuracy is 98.2%.

4.2 Leakage Fault Identified

SVM-RFE is applied globally to all 884 samples before cross-validation folds are created. Because SVM-RFE is a supervised selector, it uses the complete label vector $\mathbf{y}$ to rank features. Test-fold labels therefore influence which 1,000 features are retained for every fold, introducing systematic leakage. Additionally, the `StandardScaler` normalization was fitted on the whole dataset, further leaking test-fold statistics into training.

4.3 Additional Numerical Issue

During code inspection we identified that Fisher Z transformation of Pearson correlation coefficients was applied without clipping, allowing values of ± 1.0 to produce infinite outputs and destabilise training. We corrected this by clipping coefficients to ± 0.9999 prior to transformation.

4.4 Corrections Applied

The complete set of corrections applied to the forked repository is as follows:

- **Per-fold feature extraction:** RFE is fitted exclusively on each fold's training partition via a new `fit_rfe_on_fold` function and subsequently applied to validation and test partitions.
- **Per-fold scaling:** `StandardScaler` is fitted on training features only and applied to validation and test sets.
- **Validation split integrity:** Validation splits are drawn from the training fold; test-fold samples are never used for early stopping or hyperparameter selection.
- **Fisher Z clipping:** Correlation coefficients are clipped to $[-0.9999, 0.9999]$ before transformation.
- **Early stopping improvement:** `EarlyStopping` now records and restores the best model state when triggered.
- **RFE step-size parameter:** `rfe_step` is exposed as a configurable argument to allow reproducible experiments.

Table 2. XAI-for-ASD: performance before and after data-leakage and numerical-stability corrections on ABIDE I (5-fold cross-validation). Corrected results are reported as mean ± std across folds to match the original paper's multi-run reporting style.

Condition	Accuracy	Sensitivity	Specificity	Precision	F1
Original (with leakage)	97.68%,± 1.22%	0.98 ± 0.011	0.97 ± 0.023	0.97 ± 0.024	0.98 ± 0.011
Corrected (no leakage)	$65.02\% \pm 2.51\%$	0.658 ± 0.039	0.642 ± 0.050	0.660 ± 0.030	0.658 ± 0.026
Drop	**−32.66 pp**				

Table 3. Performance inflation summary. Leakage-free benchmarks are included for comparison.

Study	Method	Reported Acc.	Corrected Acc.
Abraham et al. [5]	SVC (no leakage)	66.8%	—
Heinsfeld et al. [6]	AE+DNN (no leakage)	70.0%	—
Almuqhim & Saeed [7]	SAE+MLP (no leakage)	70.8%	—
MADE-for-ASD [1]	SSDAE+MLP *(leakage: F-score)*	74.04%	66.35%
XAI-for-ASD [2]	SSAE+softmax *(leakage: SVM-RFE)*	97.68%	65.02%

4.5 Results

Table 2 compares the originally reported figures against the corrected evaluation (5-fold cross-validation, mean ± std).

The corrected accuracy of 65.02% is in line with baseline results achievable on ABIDE I without leakage, confirming that the originally reported 97.68% was almost entirely an artefact of the pre-fold SVM-RFE and scaler. Confidence intervals from 5-fold evaluation further indicate that performance is indistinguishable from the leakage-free baseline of ∼66.8% reported by Abraham et al. [5] using conventional SVM classifiers.

5 Summary of Performance Inflation

Table 3 places both case studies in the context of the broader ABIDE I literature by contrasting original claimed scores, corrected scores, and representative leakage-free benchmarks.

The two case studies demonstrate that leakage mechanisms of differing severity — a comparatively mild global F-score ranking versus a highly supervised global SVM-RFE — produce proportionally different amounts of inflation (7.7 vs. 32.7 percentage points). Critically, neither corrected result surpasses the leakage-free deep-learning baseline of ∼71%, suggesting that the architectural advances claimed in both papers do not survive rigorous evaluation.

6 Safeguards and Recommendations

6.1 Technical Safeguards

Based on the corrections described above, we distil the following checklist for practitioners building fMRI classification pipelines:

1. **Feature selection inside the CV loop.** Any supervised selector (F-score, SVM-RFE, ANOVA, mutual information) must be fitted exclusively on each fold's *training* partition.
2. **Scaling inside the CV loop.** Normalization statistics (`StandardScaler`, `MinMaxScaler`) must be fitted on training data only.
3. **Validation from training only.** Early stopping and hyperparameter selection must use a validation subset carved from the training fold, never from the test fold.
4. **Numerical stability.** Transformations such as Fisher Z should include domain clipping to prevent infinite values.
5. **Ablation of the feature selector.** Performance should be reported with and without feature selection to isolate its contribution.

6.2 AI-Assisted Code Review

A key observation from this work is that both leakage faults are detectable by static code inspection — they appear as dataset-level operations placed before cross-validation loops — and therefore amenable to automated review.

We recommend that researchers and developers utilise AI tools not only as code-writing aids but also as systematic **code reviewers**. In this study, code review was conducted using **Copilot as an AI-assisted code reviewer**, which proved effective at identifying: (i) supervised feature selectors applied outside cross-validation loops, (ii) data scalers fitted on combined train/test splits, (iii) validation sets drawn from test data, and (iv) numerical instability in transformation functions.

The review process follows the checklist in Section 6.1, translated into structured prompts submitted to the AI reviewer. For example, a prompt of the form *"Identify any supervised feature selection (e.g. F-score, SVM-RFE, mutual information) that is applied to the full dataset before cross-validation splits are created"* directly targets the leakage pattern identified in both case studies. Table 4 provides representative examples of the prompt templates used.

AI code reviewers can flag these patterns far more quickly than manual inspection, and at scale — making them a practical *first-pass diagnostic tool*, not a substitute for expert validation, for the neuroimaging community. We encourage journal editors and conference programme committees to consider requiring AI-assisted code review as part of the reproducibility checklist for submitted software. The corrected codebases produced in this study are available as publicly maintained forks:

- XAI-for-ASD (corrected): `https://github.com/mhdbashar/XAI-for-ASD`
- MADE-for-ASD (corrected): `https://github.com/mhdbashar/MADE-for-ASD`

Table 4. Example structured prompts used in the AI-assisted code review workflow. Each prompt targets a specific leakage or quality-assurance pattern from the checklist in Section 6.1.

Target pattern	Example prompt
Pre-fold supervised feature selection	"Identify any supervised feature selection (F-score, SVM-RFE, ANOVA, mutual information) applied to the full dataset before cross-validation splits are created."
Pre-fold data scaling	"Check whether any scaler (StandardScaler, MinMaxScaler) is fitted on combined train and test data rather than exclusively on training folds."
Validation leakage	"Verify that validation sets used for early stopping or hyperparameter selection are drawn only from training folds, not from held-out test data."
Numerical instability	"Identify transformations (e.g. Fisher Z, log) applied without input clipping that could produce infinite or NaN values."

7 Discussion

Prevalence and impact. The two cases studied here provide concrete evidence of a broader problem. The reproducibility literature documents data leakage across hundreds of published machine-learning papers in medical imaging [4]. In ASD classification specifically, the combination of high-dimensional connectivity data and small sample sizes creates strong incentives for aggressive dimensionality reduction, amplifying the probability of leakage. Independent corroboration comes from Dong et al. [8], who report that EV-GCN [9] drops from approximately 81% to ∼68% once an information leakage issue is resolved and fold-wise model selection is applied — consistent with the inflation magnitudes we observe. Together, these converging results suggest that portions of the ABIDE I leaderboard may overstate the true state of the art, though we note that not all high-performing results have been independently audited.

Broader implications for reported biomarkers. Beyond performance numbers, data leakage distorts interpretability analyses. In XAI-for-ASD, the Integrated Gradients analysis that identified visual-cortex regions (calcarine sulcus, BA 17; cuneus, BA 17–18) as key biomarkers was conducted on a model whose feature space was pre-optimised using test-set labels, meaning those biomarker attributions lose statistical validity after correction. Recomputing feature importance under leakage-free pipelines falls outside the scope of this reproducibility study. Notably, Dong et al. [8] report leakage-free feature importance analyses in broadly comparable regions, suggesting that principled reanalysis remains feasible. We therefore scope interpretability reanalysis as an important direction

for future work, and encourage the community to treat biomarker findings from pipelines with unverified data partitioning with appropriate caution.

Corrected performance in context. After correction, both pipelines yield accuracies in the 65–67% range, consistent with the leakage-free deep-learning baseline established on ABIDE I This does not imply that fMRI-based ASD classification is limited to ∼66%; rather, genuine improvements must be demonstrated under rigorous evaluation conditions. The rationale for retaining the original hyperparameters in our corrected experiments — rather than re-tuning after leakage removal — is to isolate the effect of leakage as the sole source of variation. Retuning would introduce a confounding factor, making it difficult to attribute performance changes specifically to the correction.

Limitations. The present study focuses on two specific pipelines and one dataset; the observed leakage patterns are illustrative rather than exhaustive. We do not claim that all high-performing ABIDE I results are artefactual, but encourage independent reproducibility checks across the literature.

8 Conclusion

We have demonstrated, through rigorous reproduction experiments, that two peer-reviewed ASD classification pipelines suffer from data leakage in their feature selection stages. After correcting F-score ranking (MADE-for-ASD) and SVM-RFE (XAI-for-ASD) to operate within cross-validation folds, accuracy drops by 7.7 and 32.7 percentage points respectively, bringing both systems to leakage-free baseline levels. The magnitude of these drops highlights the importance of principled experimental design in neuroimaging machine learning. These findings are further corroborated by independent evidence from Dong et al. [8], who report a comparable leakage-induced inflation in EV-GCN on the same dataset.

We further recommend the adoption of AI-assisted code review — demonstrated here using Copilot with Claude Haiku 4.5 — as a scalable and practical tool for catching leakage patterns and other methodological errors before publication. Corrected codebases have been made publicly available. Together, these contributions provide the community with concrete tools and protocols for producing trustworthy ASD classification results.

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Liu, X., Hasan, M.R., Gedeon, T., Hossain, M.Z.: MADE-for-ASD: A multi-atlas deep ensemble network for diagnosing autism spectrum disorder. Comput. Biol. Med. **182**, 109083 (2024). https://doi.org/10.1016/j.compbiomed.2024.109083
2. Vidya, S., Gupta, K., Aly, A., Wills, A., Ifeachor, E., Shankar, R.: Identification of critical brain regions for autism diagnosis from fMRI data using explainable AI: an observational analysis of the ABIDE dataset. eClinicalMedicine **88**, 103452 (2025). https://doi.org/10.1016/j.eclinm.2025.103452
3. Di Martino, A., Yan, C.G., Li, Q., et al.: The Autism Brain Imaging Data Exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol. Psychiatry **19**(6), 659–667 (2014). https://doi.org/10.1038/mp.2013.78
4. Kapoor, S., Narayanan, A.: Leakage and the reproducibility crisis in machine-learning-based science. Patterns **4**(9), 100804 (2023). https://doi.org/10.1016/j.patter.2023.100804
5. Abraham, A., Milham, M.P., Di Martino, A., Craddock, R.C., Samaras, D., Thirion, B., Varoquaux, G.: Deriving reproducible biomarkers from multi-site resting-state data: an autism-based example. NeuroImage **147**, 736–745 (2017). https://doi.org/10.1016/j.neuroimage.2016.10.045
6. Heinsfeld, A.S., Franco, A.R., Craddock, R.C., Buchweitz, A., Meneguzzi, F.: Identification of autism spectrum disorder using deep learning and the ABIDE dataset. NeuroImage Clin. **17**, 16–23 (2018). https://doi.org/10.1016/j.nicl.2017.08.017
7. Almuqhim, F., Saeed, F.: ASD-SAENet: A sparse autoencoder and deep neural network model for detecting autism spectrum disorder using fMRI data. Front. Comput. Neurosci. **15**, 654315 (2021). https://doi.org/10.3389/fncom.2021.654315
8. Dong, Y., Batalle, D., Deprez, M.: A framework for comparison and interpretation of machine learning classiers to predict autism on the ABIDE dataset. Hum. Brain Mapp. 46(5), e70190 (2025). https://doi.org/10.1002/hbm.70190
9. Huang, Y., Chung, A.C.S.: Edge-variational graph convolutional networks for uncertainty-aware disease prediction. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12267, pp. 576–585. Springer (2020).
10. World Health Organization: Autism spectrum disorders. Fact sheet (2023). https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders
11. Wang, C., Xiao, Z., Wang, B., Wu, J.: Identification of autism based on SVM-RFE and stacked sparse autoencoder. IEEE Access **7**, 118030–118036 (2019).
12. Craddock, C., Benhajali, Y., Chu, F., et al.: The Neuro Bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. Front. Neuroinform. **7**, 27 (2013).
13. Chen, Y.W., Lin, C.J.: Combining SVMs with various feature selection strategies. In: Feature Extraction: Foundations and Applications, pp. 315–324. Springer (2006).