Data leakage in feature selection invalidates reported classification accuracy for autism diagnosis from fMRI

Semyon Kurkin* and Alexander Hramov

Research Institute of Applied Artificial Intelligence and Digital Solutions, Plekhanov Russian University of Economics, Moscow, Russia

eClinicalMedicine
2026;■: 104049

Published Online XXX
https://doi.org/10.1016/j.eclinm.2026.104049

We read with interest the article by Vidya et al.[1] reporting 98.2% accuracy for autism spectrum disorder (ASD) classification using deep learning on functional MRI data from the ABIDE I dataset. While the authors' focus on interpretability is commendable, we identified a critical methodological flaw that invalidates the reported performance metrics.

The authors employed Support Vector Machine with Recursive Feature Elimination (SVM-RFE) to reduce dimensionality from 6670 to 1000 features before training their Stacked Sparse Autoencoder classifier. Critically, this feature selection was performed on the entire dataset prior to cross-validation, constituting a well-documented form of data leakage known as "selection bias" or "feature selection bias."[2,3]

We verified this by examining the authors' publicly available code (https://github.com/v1dya/XAI-for-ASD). The implementation confirms our concern:

Evidence from the code (main.py):

1. **Lines 96–102** define the SVM-RFE function that fits on complete data:

```
def get_top_features_from_SVM_RFE
(X, Y, indices, N, step):
    svm = SVC(kernel = "linear")
    rfe = RFE (estimator = svm, n_featur-
    es_to_select = N, step = step)
    rfe.fit (X, Y) # Fits on ALL samples
```

2. **Lines 1376–1384** show that feature selection occurs before any data splitting:

```
# top_features, top_rois = get_top_
features_from_SVM_RFE (
# feature_vecs, labels, feature_vec_
indices, 1000, RFE_step)
top_features = np.loadtxt (f'data/
{pipeline}/sorted_top_features...')
```

3. **Lines 675–681** confirm cross-validation receives pre-selected features:

```
def train_and_eval_model (top_features,
labels_from_abide, ...):
    skf = StratifiedKFold (n_splits = 5, ...)
    for train_idx, test_idx in skf.split
    (top_features, labels_from_abide):
```

This methodological error means that information from test samples influenced which features were selected for training. The SVM-RFE algorithm identified the 1000 features that best discriminate between ASD and control groups across the entire dataset, including samples that would later serve as test data. Consequently, the model was trained exclusively on features already known to separate the test samples, artificially inflating classification accuracy.

This issue is particularly problematic given the high-dimensional nature of fMRI connectivity data (6670 features) relative to sample size (884 participants). Under such conditions, there exist numerous feature subsets that can achieve spuriously high accuracy on any given dataset through overfitting, but these features may not generalise to independent samples.

The correct procedure requires performing feature selection independently within each cross-validation fold, using only training data. Previous methodological studies have demonstrated that such leakage can inflate accuracy estimates by 10–30 percentage points,[2,3] which aligns with the unusually high performance reported here compared to typical ABIDE classification results (70–85%).[4]

Given the severity of this methodological error, the reported classification accuracy of 98.2% cannot be considered valid. The claim of "state-of-the-art performance" (p. 7) is unfounded, as is the assertion that the model "captured genuine neurobiological ASD markers rather than overfitting to dataset artifacts" (p. 1). The identified biomarkers, including the visual processing regions highlighted as key findings, may be artifacts of the data leakage rather than true neural signatures of ASD.

We conclude that the primary results of this study are invalid and should not be used to inform clinical practice or future research directions. Reanalysis should be conducted with proper nested cross-validation (SVM-RFE performed within each fold using only training data). We urge the authors to conduct this reanalysis and publish corrected findings.

DOIs of original articles: https://doi.org/10.1016/j.eclinm.2026.104052, https://doi.org/10.1016/j.eclinm.2025.103452

*Corresponding author.
E-mail address: kurkinsa@gmail.com (S. Kurkin).

Correspondence

Contributors

SK and AH are contributors to this work, and have read and approved the final version to be published. KS was responsible for conceptualisation, formal analysis, investigation, methodology, software, and writing—original draft. AH was responsible for investigation, project administration, resources, supervision, validation, and writing—review & editing.

Declaration of interests

SK and AH declare no competing interests.

Acknowledgements

No funding was received for this work.

References

1 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*. 2025;88:103452.
2 Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinf*. 2006;7:91.
3 Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res*. 2010;11:2079–2107.
4 Heinsfeld AS, Franco AR, Craddock RC, Buchweitz A, Meneguzzi F. Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *Neuroimage Clin*. 2018;17:16–23.