5. Transformer and multi-tasking to detect ASD on rs-fMRI

5.1. Introduction

Functional MRI (fMRI) marked a revolution in neuroimaging in the 1990s, enabling unambiguous visualisation of human brain dynamics with MR for the first time. This technology opened new avenues for psychology and neurology research but also posed new analysis challenges given fMRI's lower spatial resolution yet higher dimensionality. Key questions arose around modelling relationships between brain regions across the newly captured time dimension.

While specific to autism spectrum disorder (ASD), analysing fMRI data also reflects methodological approaches applicable across domains. So far, machine learning models built on fMRI have shown greater ASD classification performance versus structural MRI approaches [21].

Traditional fMRI analysis entails extensive preprocessing, compressing information into derived neuroimaging features like regional homogeneity [22] or intrinsic connectivity [23-24]. While neuroscientifically meaningful, this may discard predictive signals during the initial processing pipeline.

A key machine learning challenge is determining appropriate data sizes for given models to obtain truly generalizable neuroimaging biomarkers (He et al., 2020; Traut et al., 2021). Standard fMRI preprocessing may introduce biases that exacerbate this issue (Traut et al., 2021; Dadi et al, 2019). For example, Churchill et al. (2012) found optimising pipelines individually revealed activation patterns absent under fixed preprocessing, demonstrating the significant impact of pipeline choices.

Unintended replicability issues and non-reproducible fMRI findings present major concerns (He et al., 2019; Churchill et al. 2012; Dadi et al, 2019; Traut et al., 2021), compounded by small sample sizes as large-scale data remains scarce.

Deep learning (DL) approaches show promise for ASD classification, at times outperforming traditional machine learning techniques [21, 25-40]. Unlike ML, DL philosophy entails utilising less preprocessed data (LeCun et al., 2015).

However, reviewing the literature reveals many DL studies still train on derived connectivity matrices. Common workflows parcellate 4D scans into regions of interest (ROIs) per atlases like AAL [41], extract mean time series within ROIs, and compute correlation matrices (Biswal et al., 1995; Dadi et al., 2019).

Viewed as graphs or concatenated 3D images, connectivity matrices have been classified with graphical networks or 3D CNNs for ASD [25, 27, 29, 30, 35, 36]. For instance, [25] used 3D CNNs on connectivity "fingerprint" images. [29] constructed ASD and neurotypical connectivity graphs as spectral convolution network templates.

Pearson correlation matrices may discard important temporal patterns like phase shifts in comparing time series, where asynchronous responses could be meaningful.

Several studies have incorporated temporal dynamics for ASD prediction using DL approaches. For example, [42] generated rs-fMRI time series embeddings with LSTMs for classification. [43-45] applied high-dimensional 3D/4D convolutional networks. Capturing spatiotemporal representations is also critical in natural language processing (NLP), where Transformer architectures currently reign supreme (cite chapter 2)[46]. Inspired by their success, Transformers have been applied in medical imaging [47-49] including fMRI analysis.

Notably, [49] compressed 4D task-fMRI into 3D embeddings using a 3D CNN, and fed these to a Transformer encoder to determine important frames per task. [50] proposed predicting task fMRI brain states from time series sequences using a Transformer, compressing spatial data similarly to unsupervised methods like ICA (cites). A second model takes the latent representation for state prediction. While effective for task fMRI, the learned embeddings may lack interpretability and spatial relationships important for resting state modelling. [51] incorporated spatial and temporal dynamics via a cross-window

Transformer with a learned CLS token summarising latent features for classification. [52] used a 3D CNN autoencoder to compress volumes into input representations for a downstream Transformer. However, this may discard informative spatial interactions across timeframes, better suited to task fMRI. [54] also applied self-supervised Transformers to infer functional networks in space and time. [55] compared various Transformer architectures on fMRI, finding pre-training on broad neuroimaging data improved generalisation for mental state decoding over training from scratch. Causal modelling outperformed other approaches. [53] fed connection profiles of mean time series from known ROIs into a Transformer encoder. An orthonormal clustering projection enhanced discriminability for downstream prediction.

In this project, I aimed to model interactions between brain regions that may underlie autistic functioning at rest. This requires analysing spatial relationships across time series, which [49] does not enable by compressing space via CNN. I explored Transformers for rs-fMRI, extracting time series from Craddock parcellation (Craddock et al., 2012) to represent meaningful brain regions, akin to words in a sentence. Like Transformers find linguistic relationships, I hypothesised they could decode relationships between region activities related to Autism. I investigated whether auxiliary prediction tasks like gender and age could improve autism classification, similar to multi-task learning. The rationale was that optimising for additional relevant variables may help the model learn more useful representations. My experiments comparing single and multi-task Transformer architectures showed no clear performance differences yet. In the discussion, I consider refinements like loss weighting, hyperparameter optimization, and augmented data that may better demonstrate the potential of each modelling approach.

5.2. Methods

5.2.1. Data preparation

fMRI data was preprocessed with the C-PAC pipeline (version 0.4.0 for HBN and version 0.3.9 for ABIDE 1), with global signal correction and band-pass filtering (0.01-0.1Hz). A functional parcellation - Craddock 200 (Craddock et al., 2012) - was performed, and mean time series were extracted for each region. The preprocessed data is open source for ABIDE 1 and, for HBN, available for researchers authorized to use the database. In total, 1102 time-series files were available in the ABIDE 1 dataset, and 1096 were available in the HBN dataset.

For ABIDE 1, manual quality control annotations were provided. I retained only the scans where at least one rater assessed the scan to be of good quality. 1022 scans remained after this step. No quality control annotations were provided with the preprocessed HBN data. However, the functional MRI data underwent automated quality control using tools like MRIQC to identify issues with coverage, motion, signal loss, etc. Low quality fMRI scans were excluded based on thresholds set for these QC metrics.

For the set of time-series for each participant, I checked that all the time-series had at least one non-zero value, that the time-series lengths were sufficiently long (the minimum length of 100 frames was chosen arbitrarily to retain as many time series as possible), and that there were 200 time-series (following the Craddock 200 parcellation). I excluded participant data that did not meet these conditions. Hence, I generated time-series harmonised in length (100 frames) across the whole dataset. Next, each time-series was normalised separately by removing the standard deviation and dividing by the mean of the time-series.

For ABIDE 1, I chose the full sample 1 from the University of Michigan data collection site to be the independent test set. In the HBN data, I took 10% of the dataset as the independent test set, where each site was represented in proportion to its representation in the full dataset.

The remainder was used to train the model in a 100-fold cross-validation (CV) fashion. The CV was stratified on the ASD/non-ASD labels. I used the

StratifiedKFold class from scikit-learn's model_selection module in Python to generate the folds, with the random state set to 42.

	Training - Validation sets		**Testing set**	
	ABIDE 1	*HBN*	*ABIDE 1*	*HBN*
Model 1	773 (353 ASD)	/	94 (42 ASD)	/
Model 2	773 (353 ASD, 659 males)	/	94 (42 ASD, 70 males)	/
Model 3	773 (353 ASD, 277 aged between 10-15)	/	94 (42 ASD, 50 aged between 10-15)	/
Model 4	847 (413 ASD)	975 (67 ASD)	105 (51 ASD)	108 (6 ASD)
Model 5	/	975 (67 ASD)	/	108 (6 ASD)

Table X. Description of data used in training-validation sets (100-folds CV) and in testing set for each model

5.2.2. Models

I first trained a binary classification model for ASD diagnosis using a Transformer encoder followed by a fully connected layer block (add fig). For this initial experiment, I utilised data from participants with no diagnosis and those diagnosed only with ASD.

Next, I designed a simple multitask model, with a shared Transformer encoder as the common component and separate fully connected blocks to predict different targets (add fig).

In this study, the simple classification models were implemented with a cross-entropy loss function:

$$H(P^{*} \mid P) = \sum_{i} P^{*}(y|x_{i}) log(P(y|x_{i}; \theta)$$

Where:

- H is the cross-entropy between the true class distribution P^{*} and the predicted class distribution P
- y is the class
- x_{i} is an input instance
- θ are the parameters of the model

For the multitask models, a weighted sum of cross-entropy computations was used as the total loss function:

$$H_{1}(P_{1}^{*} \mid P_{1}) = \sum_{i} P_{1}^{*}(y|x_{i}) log(P_{1}(y|x_{i}; \theta_{1})$$

$$H_{2}(P_{2}^{*} \mid P_{2}) = \sum_{i} P_{2}^{*}(y|x_{i}) log(P_{2}(y|x_{i}; \theta_{2})$$

$$H_{sum} = \alpha H_{1}(P_{1}^{*}|P_{1}) + (1 - \alpha) H_{2}(P_{2}^{*} \mid P_{2})$$

Where:

- H_{1} is the cross-entropy between the true class distribution P_{1}^{*} and the predicted class distribution P_{1}
- H_{2} is the cross-entropy between the true class distribution P_{2}^{*} and the predicted class distribution P_{2}
- H_{sum} is the loss criterion of the model
- θ_{1} are the parameters of the encoder + the FC block 1 (see **Figure X**)
- θ_{2} are the parameters of the encoder + the FC block 2 (see **Figure X**)
- y is the class
- x_{i} is an input instance
- α = 0,5 (arbitrary choice)

I primarily monitored accuracy and AUROC as model performance metrics, and the mean of these two scores to evaluate overall model balance.

The Adam optimizer (Kingma et al., 2017) was used for training with a learning rate of 10^{-3} and weight decay of 10^{-7}.

I arbitrarily set the time-series representation dimension to 16.

The input embedder consists of one linear layer to project the input data into a 16-dimensional embedding space, plus a positional encoding layer similar to Vaswani et al. (2017).

The encoder block (Figure X) comprises 3 encoder layers, each containing 4 multi-head attention modules.

Post-encoding, the 200x16 representation of each input is flattened and passed through a fully connected block with 3 layers. Finally, a softmax function is applied to produce output probabilities.

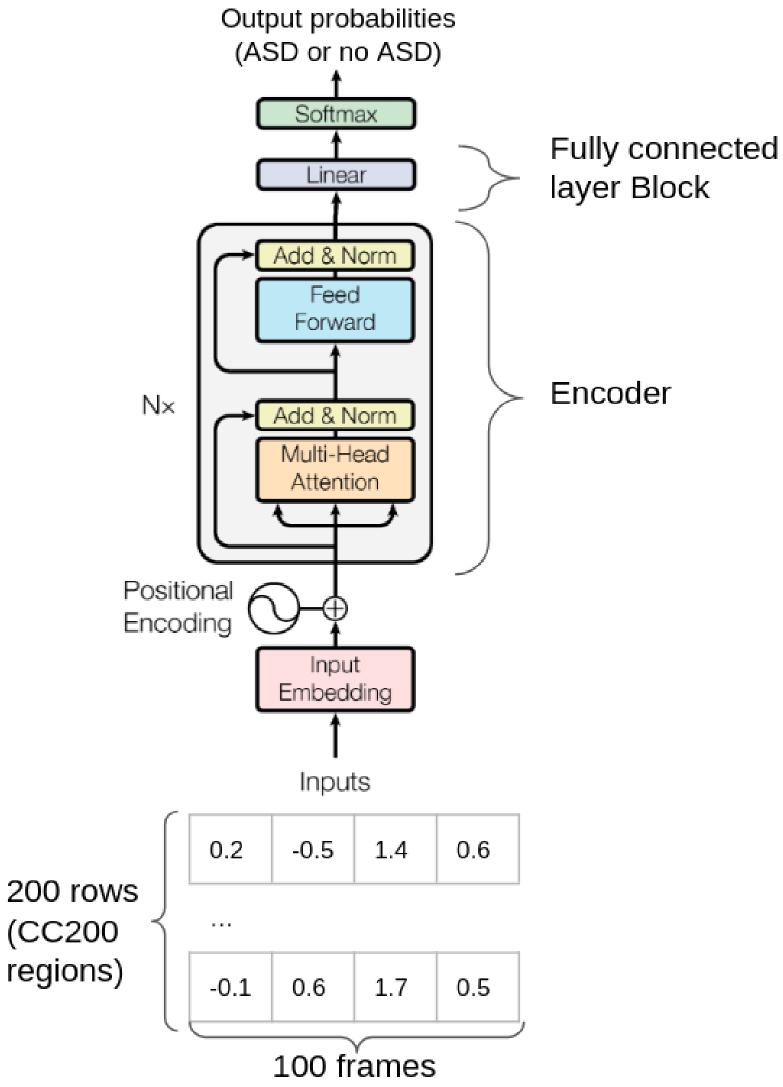


Figure X. Architecture of the models 1, 4 and 5. Inputs are the 200 extracted mean time series (CC200 atlas - Craddock et al., 2012) cropped to 100 non-null frames. The Encoder part is similar to a classical Transformer encoder (Vaswani et al., 2017) and returns an intermediate representation of the inputs. The Fully connected layer block processes the representation flattened and returns the probability to be diagnosed with ASD.

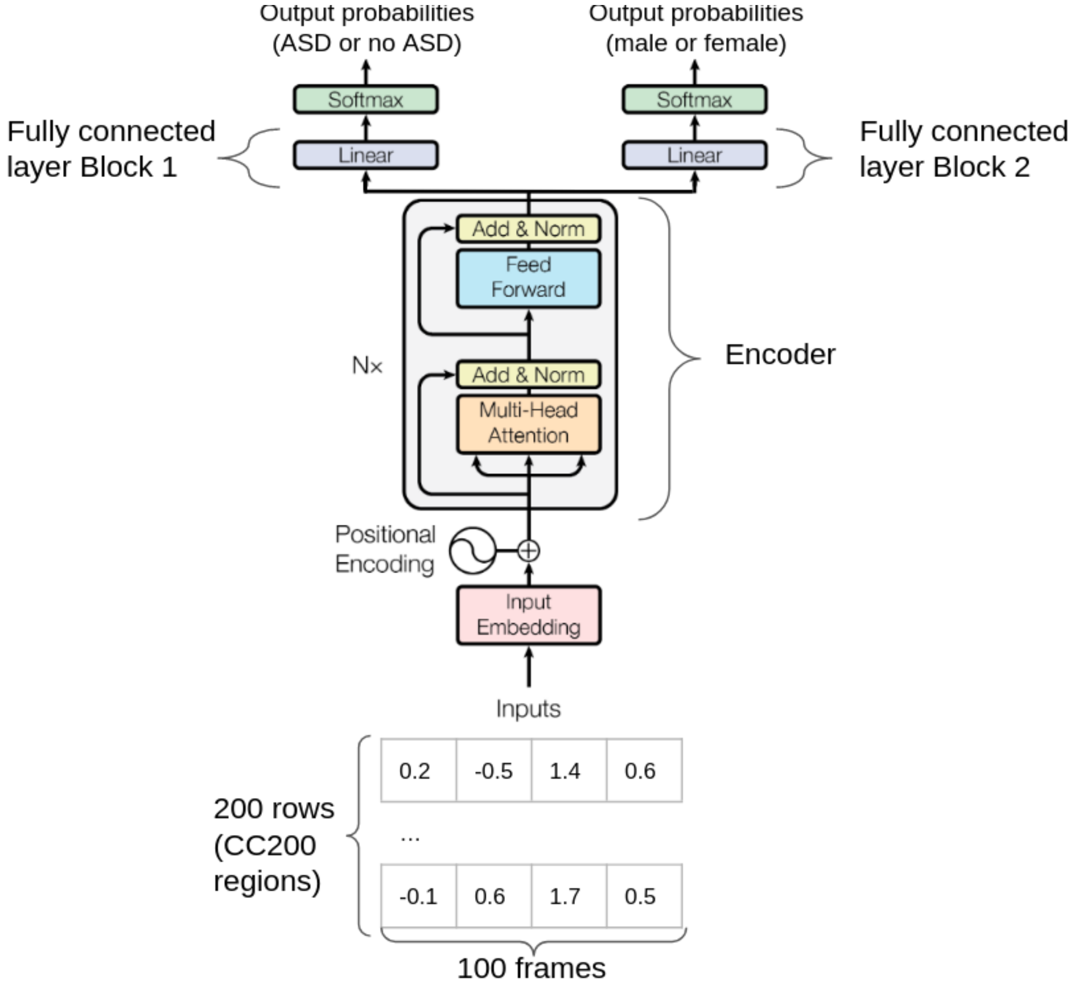


Figure X. Architecture of the multitask models 2 and 3. Inputs are the 200 extracted mean time series (CC200 atlas - Craddock et al., 2012) cropped to 100 non-null frames. The Encoder part is similar to a classical Transformer encoder (Vaswani et al., 2017) and returns an intermediate representation of the inputs. The Fully connected layer block 1 processes the representation flattened and returns the probability to be diagnosed with ASD. The Fully connected layer block 2 processes the same representation flattened and returns the probability to be a male (for model 2, or aged between 10-15 for model 3). The two FC blocks are optimised separately while the Encoder is optimised taking into account the two tasks.

5.2.3. Interpretation methods

To evaluate model fit, I plotted train versus validation accuracy and AUROC score curves over epochs to check for underfitting or overfitting.

I defined a metric as the mean of accuracy and AUROC scores on the validation set per epoch per fold. The best metric over all epochs for each fold corresponds to the optimal model. Comparing best metrics across folds assessed multitask improvements. Paired t-tests or Mann-Whitney U tests were used after verifying assumptions.

I evaluated the best models (optimal per fold) on independent test sets to assess generalisation.

To visualise representations, I computed the mean post-encoder representation across subjects (Figure X). Pearson correlations between all 200 regions were calculated, with correlations above 0.6 visualised in a chord diagram.

Additionally, I implemented LIME (cite) to locally interpret model predictions. For a given input, LIME approximates the decision boundary and feature importance for that observation. These explanations may provide insights into model behaviour.

Results

This section presents outcomes for five differently designed models shown in Figure X.

Model 1 comprised a Transformer encoder followed by a fully connected block to predict ASD status (ASD or non-ASD). Model 1 was trained only on a subset of ABIDE 1 data, excluding participants with comorbid diagnoses.

Models 4 and 5 had identical architecture to Model 1, but were trained on ABIDE 1 and HBN, and HBN only, respectively. These datasets included participants with diagnoses other than just ASD.

Models 2 and 3 were intended as multitask models, with a shared Transformer encoder and separate fully connected blocks to predict ASD status plus another binary target (gender for Model 2, age 10-15 years or not for Model 3). The training data for Models 2 and 3 matched Model 1, excluding comorbidities.

Models were trained for 50 epochs with validation performed each epoch to obtain accuracy and AUROC scores on the validation fold. As a reminder, 100-fold stratified cross-validation was used, so each model was trained 100 times.

Figure X shows the evolution of accuracy and AUROC over epochs, aggregating repeated fold values to plot mean and 95% confidence intervals for training and validation sets. The models exhibit overfitting - fast convergence on training with stagnating, near-random validation performance. Models 4 and 5 have higher validation scores, likely due to imbalanced classes in the HBN dataset (~6% ASD).

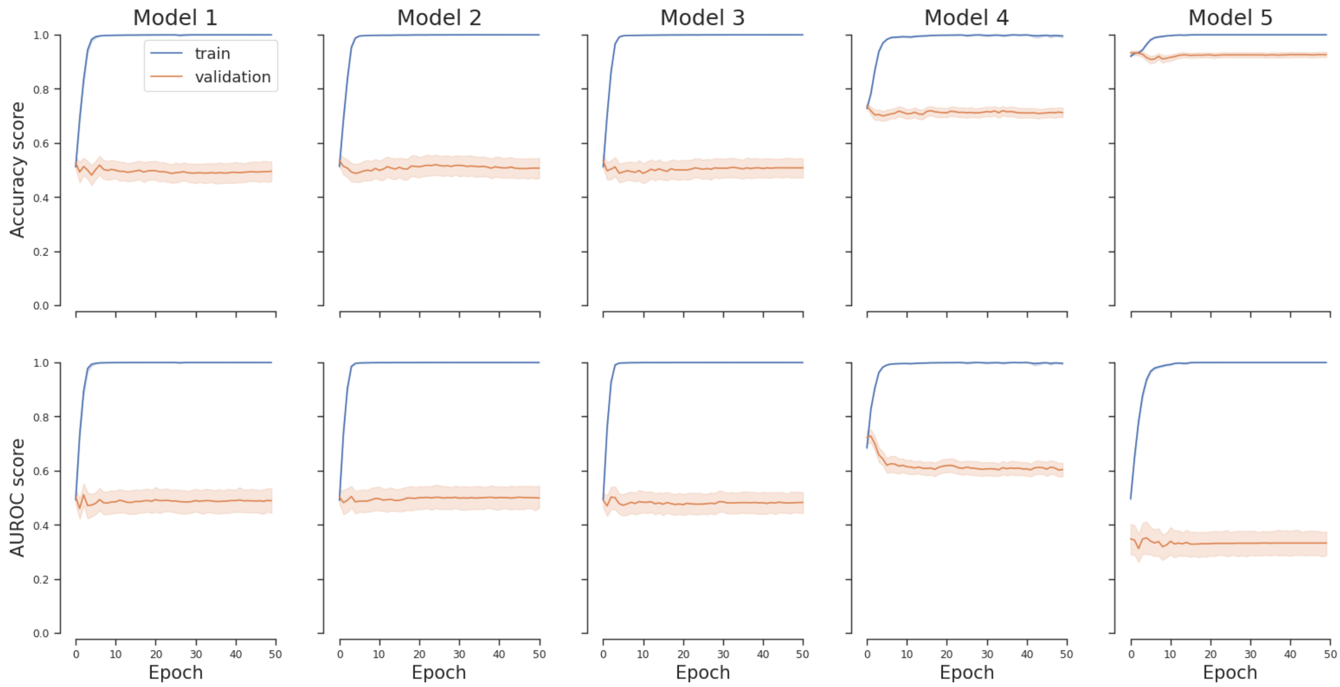
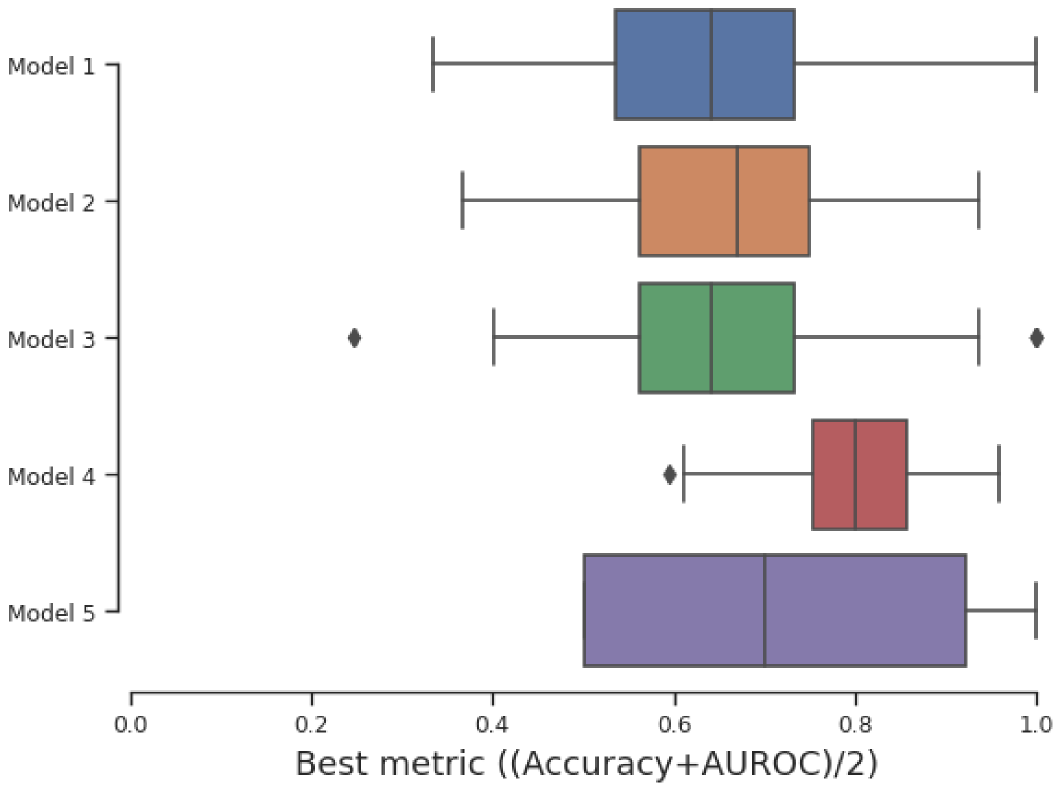


Fig. X. shows a comparison of the best metrics (computed on validation sets as (accuracy+AUROC)/2) on all the folds between the five models. The boxplots represent the distribution of the best metrics of all the folds for each model. From Fig. X., it appears that there is no clear difference between Model 1, 2 and 3 best metrics, that have respectively a mean of m_1=0.644, m_2=0.660, m_3=0.644, and a standard-deviation of s_1=0.151, s_2=0.148, s_3=0.144. However, Model 4 and 5 appear different from the others, and have respectively a mean of m_4=0.794, m_5=, and a standard-deviation of s_4=0.078, s_5=.



To confirm observations, I performed statistical tests comparing models under two conditions (simple vs multitask, ASD-only vs ASD+comorbidities, multitask gender vs age). Paired t-tests were suitable for comparing metric differences across folds.

Verifying normality assumptions, all differences passed Shapiro-Wilk except Models 4 and 5 vs 1 (Appendix). Thus, I used Mann-Whitney U tests for those pairs.

Results (Table X) show no significant difference between Models 1, 2, and 3 ($p_{m1\text{-}m2}$=0.435, $p_{m1\text{-}m3}$=0.984, $p_{m2\text{-}m3}$=0.199). However, significant differences exist between Model 1 and Model 4 (p_{m1_m4}=0.0158 < 5%), and Model 1 and Model 5 ($p_{m1\_m5} < 10^{-13}$).

Paired T-test	**T**	**dof**	**Alternative**	**p_val**	**CI95%**	**cohen-d**	**BF10**	**power**
Model 1 - Model 2	0.783	99	2-sided	0.435	[-0.02, 0.05]	0.103	0.149	0.176
Model 1 - Model 3	-0.0206	99	2-sided	0.984	[-0.04, 0.04]	0.003	0.111	0.050
Model 2 - Model 3	1.293	99	2-sided	0.199	[-0.01, 0.04]	0.109	0.248	0.189
MW U-test	**U-val**		**Alternative**	**p_val**	**RBC**	**CLES**		
Model 1 - Model 4	8081.0		2-sided	5,17. 10^{-14}	-0.616	0.808		
Model 1 - Model 5	5981.5		2-sided	0.0158	-0.193	0.598		

Table X. Statistical tests on the best metrics of the validation fold results between the models.

Inference on test set:

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**
Accuracy	50%	52.1%	47.9%	67.1%	**85.2%**
AUROC	0.581	0.551	**0.599**	**0.599**	0.368
(Acc. + AUROC)/2	0,541	0,536	0,539	**0,635**	0,61
Specificity	0.442	0.808	0.327	0.776	**0.902**
Sensitivity	0.571	0.167	**0.667**	0.386	0.0

Table X presents test set results for each model, including accuracy, AUROC, specificity, and sensitivity. Model 5 achieved the highest accuracy (85.2%), while Models 3 and 4 showed the top AUROC scores (0.599). Model 5 had the highest specificity (0.902) and Model 3 the highest sensitivity (0.667). A drop in the mean metric (Accuracy+AUROC)/2 is observed, with Model 4 having the best value (0.635).

The data description in Table X shows highly imbalanced classes for Models 4 and 5 (~26% and ~6% ASD prevalence) compared to Models 1-3 (~45% ASD). Despite higher accuracy, Model 4's performance cannot be directly compared to Models 1-3 due to this imbalance. For example, Model 5 has 0 sensitivity but 85.2% accuracy, correctly predicting only non-ASD participants.

In summary, class imbalance introduces biases making accuracy metrics non-comparable between models. Future work should incorporate calibration strategies for balanced benchmarking. No model emerges as singularly optimal, but refinements to both approaches show promise in advancing ASD prediction.

Visualisation - Interpretation

The 200 Craddock atlas (cite) time series undergo transformations through the Transformer encoder, resulting in a 16-feature representation per region before the fully connected block. I computed Pearson correlations between regions under this implicit 16-dimensional encoding.

To simplify visualisation, I translated the 200 Craddock regions into the 7-network Yeo atlas. Chord diagrams in Figure X show correlation results in this Yeo space. "Glass brain" visualisations in Figures X-X depict these correlations mapped onto brain networks.

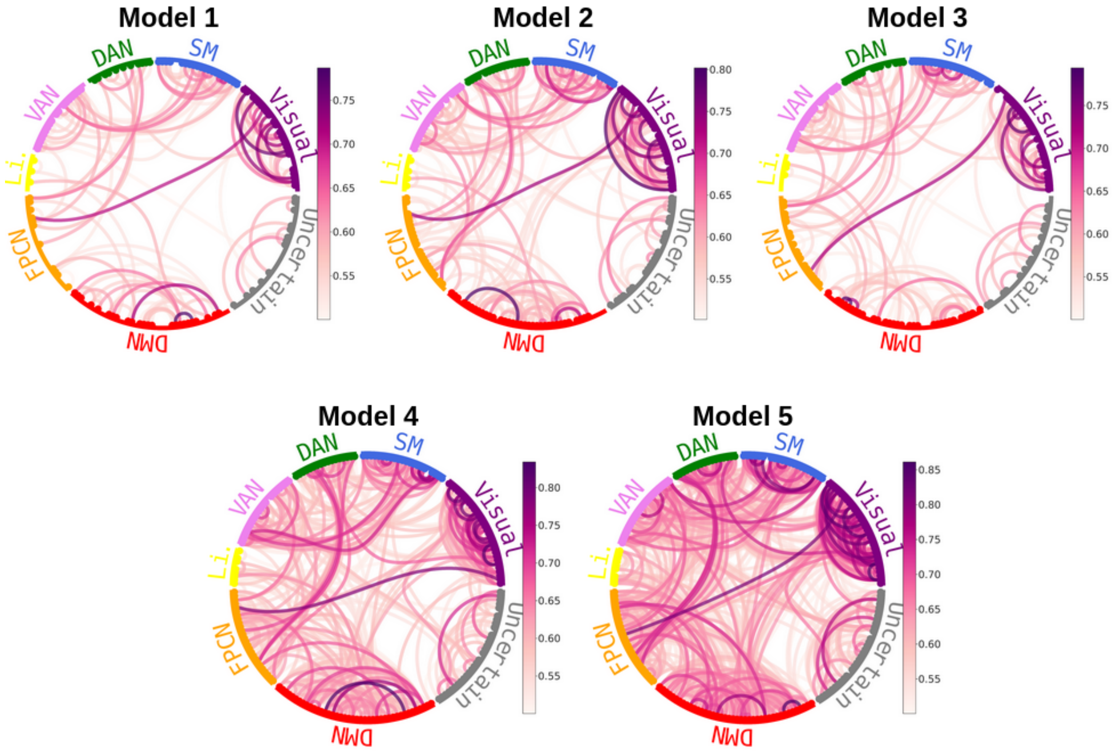
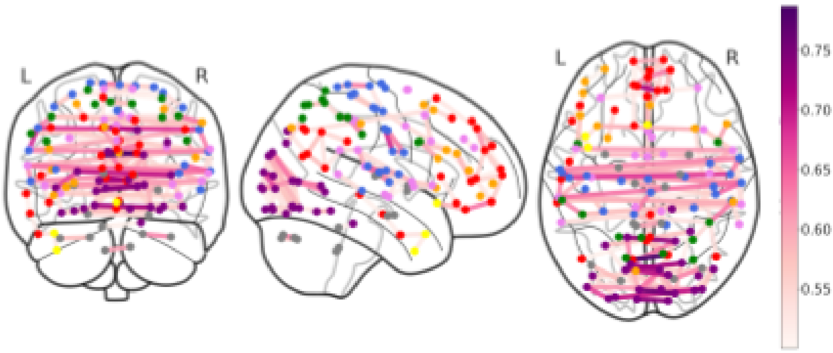
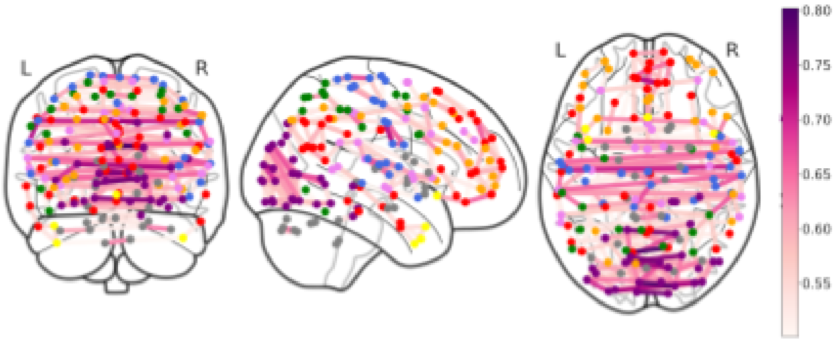


Figure X. Chord representations of the correlations between the implicit representations of 200 regions by 16 features after the last layer of the Transformer encoder part of each model. The regions were translated into the Yeo Network (add ref) to simplify the plot: “Visual” (purple) is for the Visual Network; “SM” (blue) is for the Somatomotor Network; “DAN” (green) is for the Dorsal Attention Network; “VAN” (violet) is for the Ventral Attention Network; “Li.” (yellow) is for the Limbic Network; “FPCN” (orange) is for the Frontoparietal Control Network; “DMN” (red) is for the Default Mode Network; “Uncertain” (grey) is for regions in the CC200 atlas that did not match any region in the Yeo atlas. Inside the circles, I represented as pink lines with varying intensity levels the correlations between 0.5 and 1 (no negative correlation was found to have an absolute value greater than 0.5). Add ref

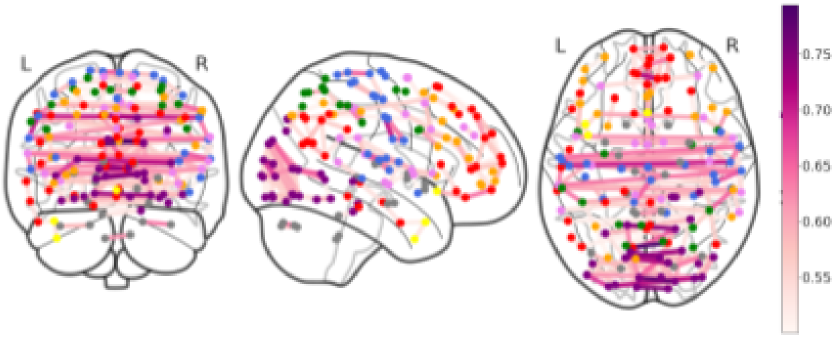
Model 1



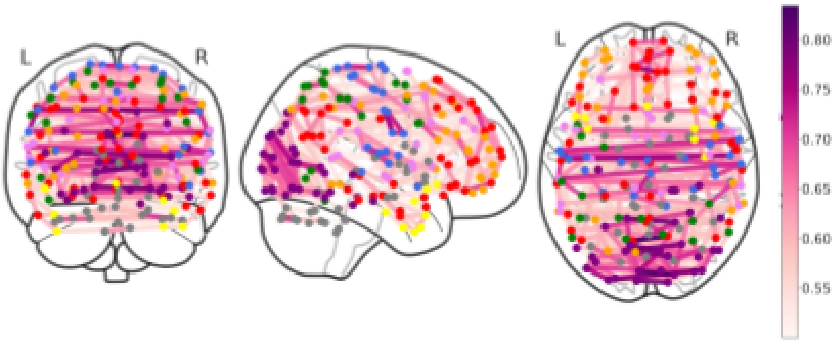
Model 2



Model 3



Model 4



Model 5

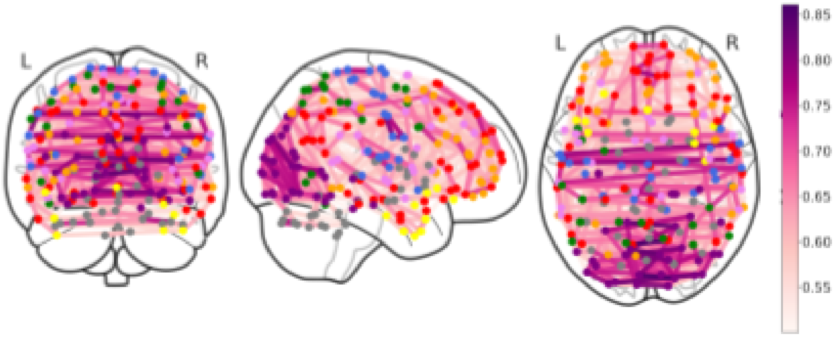
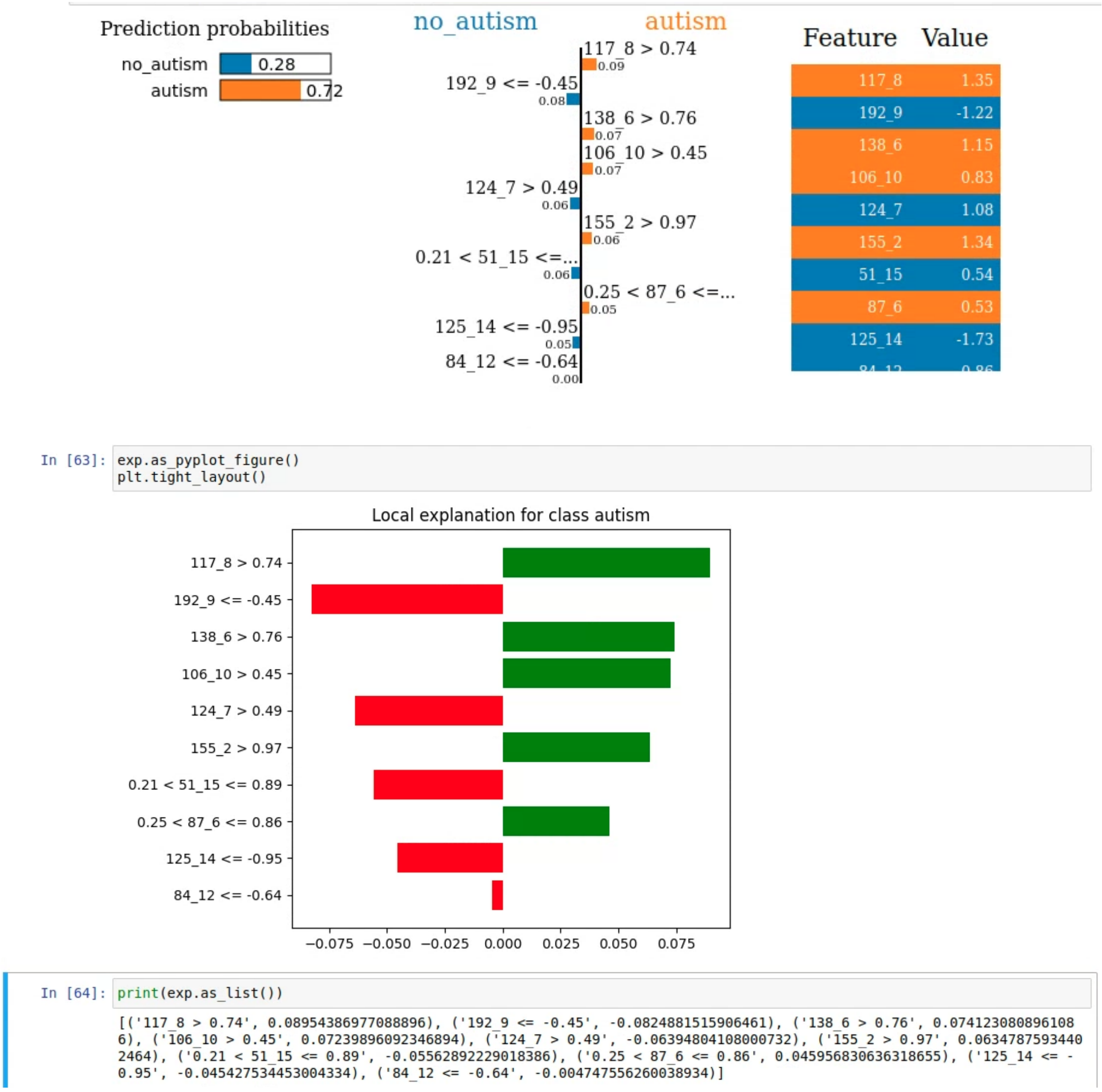


Figure X: Brain representations of the correlations between the implicit representations of 200 regions by 16 features after the last layer of the Transformer encoder part of each model. The regions were translated into the Yeo Network (add ref) to simplify the plot. Each CC200 region is represented by dot points and coloured in function of their correspondence with the Yeo atlas: Purple is for the Visual Network; Blue is for the Somatomotor Network; Green is for the Dorsal Attention Network; Violet is for the Ventral Attention Network; Yellow is for the Limbic Network; Orange is for the Frontoparietal Control Network; Red is for the Default Mode Network; Grey is for regions in the CC200 atlas that did not match any region in the Yeo atlas. I represented as pink lines with varying intensity levels the correlations between 0.5 and 1 (no negative correlation was found to have an absolute value greater than 0.5).

In addition to visualisations, I implemented the model-agnostic algorithm LIME (cite) to locally interpret model predictions. For a given input, LIME approximates the decision boundary and weights feature importance for that sample. These local explanations can provide insights into model behaviour.

I did not conduct a full LIME analysis across all participants, such as separating by diagnosis or segmenting by age, gender, and comorbidities. Proper LIME implementation requires optimising many parameters including data normalisation, model settings, and LIME hyperparameters (e.g. number of random projections). This extensive tuning enables robust feature importance mapping and represents an exciting area for future work to elucidate how models predict autism (and auxiliary targets like age and gender for multitask models).

As an initial example, Figure X shows the top 10 features driving autism classification for one participant with only an ASD diagnosis, explaining the importance via decision thresholds on continuous variables.



Feature	Value
117_8	1.35
192_9	-1.22
138_6	1.15
106_10	0.83
124_7	1.08
155_2	1.34
51_15	0.54
87_6	0.53
125_14	-1.73
84_12	0.86

```
In [63]: exp.as_pyplot_figure()
         plt.tight_layout()
```

```
In [64]: print(exp.as_list())

         [('117_8 > 0.74', 0.08954386977088896), ('192_9 <= -0.45', -0.0824881515906461), ('138_6 > 0.76', 0.074123080896108
         6), ('106_10 > 0.45', 0.07239896092346894), ('124_7 > 0.49', -0.06394804108000732), ('155_2 > 0.97', 0.0634787593440
         2464), ('0.21 < 51_15 <= 0.89', -0.05562892229018386), ('0.25 < 87_6 <= 0.86', 0.045956830636318655), ('125_14 <= -
         0.95', -0.045427534453004334), ('84_12 <= -0.64', -0.004747556260038934)]
```

Figure X. LIME algorithm executed on one autistic participant data: it explains the decisions of the Fully Connected Layer of the Model considered (e.g. Model 1). The feature names are provided with numbers. For instance, “117_8” is for region 117 of CC200 atlas and encoding feature number 8 (among the 16 ones) after the Transformer encoder part of the model. For this feature, the value is strictly greater than 0.74 that is the threshold found by the LIME algorithm, meaning that it is consistent with a prediction of Autism for LIME. The two types of graphs are displayed in a Python 3 Jupyter Notebook.

Discussion

This study explored innovative applications of the Transformer algorithm (Vaswani et al., 2017) to analyse brain activity patterns in resting-state functional MRI data for autism classification. We developed several modelling approaches: Model 1 was a basic binary classifier trained on data from individuals with autism and neurotypical controls. Models 2 and 3 took a multi-task learning approach, jointly predicting autism diagnosis along with gender or age group on the same dataset. This was motivated by known gender differences (cite) and age-related changes (cite) in autism phenotypes. Models 4 and 5 expanded the binary autism classifier to include individuals with comorbid diagnoses like ADHD, anxiety, and depression. Such comorbidities are common in autism (cite) and many psychiatric disorders show overlapping neural correlates and symptoms (cite). Overall, this work aimed to explore the potential for Transformer architectures to capture informative patterns in brain activity time series and improve autism classification. The multi-task and comorbidity-inclusive approaches were creative ways to incorporate additional relevant phenotypic information to potentially enhance model performance. This study provides promising initial results and directions for further developing neuroimaging-based classifiers using state-of-the-art DL methods.

The results demonstrate that no single model emerged as the unambiguous top performer for autism spectrum disorder (ASD) prediction. Imbalances between datasets introduced biases that made direct accuracy comparisons between certain models unfair. Future work could incorporate strategies to calibrate models trained on imbalanced data to enable fairer benchmarking.

Models 1, 2, and 3 were reasonably comparable overall. Interestingly, the multi-task approach did not boost global performance, though it did impact specificities and sensitivities considerably (sp1 = 0.442, sp2 = 0.808, sp3 = 0.327; se1 = 0.571, se2 = 0.167, se3 = 0.667). This aligns with expectations that auxiliary tasks would modulate model learning. However, multi-task models did not converge to more stable performance like Model 1.

Several factors may explain these observations. Multitask models summed task losses with arbitrary balancing (α=0.5). Optimising α more systematically could help. Additionally, complex inter-task loss relationships beyond a linear sum likely exist. Learning rates were fixed across tasks; optimising these independently may improve outcomes.

In a nutshell, while not improving overall accuracy, multi-task learning impacted model performance in nuanced ways. With refined loss weighting, learning rates, and other enhancements, multi-task and single-task approaches show promise for distilling insights about brain function from neuroimaging data to advance ASD prediction.

The use of 100-fold cross-validation for model training enabled robust performance estimation, although at the cost of greater computational demands compared to a standard train-test split. The large number of folds likely improved the reliability of the evaluated metrics. However, this extensive cross-validation constrained the extent of parameter optimization completed within project timelines. Many architectural and training hyperparameters warrant deeper investigation in future work, including encoder layer count, attention heads, positional encoding, learning rates, regularisation, and loss weighting.

In particular, the dimensionality of the time series representations may significantly impact model performance. The extracted 100-frame time series were compressed to 16 feature vectors, aiming to maximise information density. However, this compressed size could overlook important signals or relationships in the resting state data. Optimising representation dimensionality could better capture the complexity of whole-brain dynamics. Overall, this study developed a solid computational framework and baseline modelling results. With expanded hyperparameter tuning, the Transformer-based architectures show strong promise for decoding meaningful spatiotemporal patterns from rs-fMRI in relation to autism diagnosis (add refs).

This study highlights several interesting areas for future investigation. The 100-frame input sequences, though substantial, may not fully capture complex spatiotemporal dynamics across diverse brain networks at rest. Longer inputs could better model these interactions. Additionally, Transformers thrive on large datasets - the scale here, though sizable by fMRI standards, is small relative to typical Transformer applications. Given Autism's heterogeneity and inter-individual variability, larger data may be key.

There are also open questions around input data characteristics. Our ABIDE preprocessing retained most scans, but some artefacts likely remain. Heavy spatial registration could potentially distort signals. Resting-state alone may not provide sufficient signal. Multimodal integration (e.g. sMRI, PET, EEG, genetics) could add explanatory power. Controlling for factors like acquisition parameters and participant demographics may also be beneficial.

This study presented initial interpretability analyses, but Transformer model explanations remain challenging with multiple layers and attention heads. We visualised the final encoder representations to extract region-to-region relationships. However, studying intermediate representations could provide additional insights into how spatial patterns and dynamics evolve through the network. LIME highlighted influential regions for single-subject predictions, but generalising these local explanations across the whole dataset is an important next step. The fully-connected block offers limited interpretability; replacing it with more interpretable algorithms like regression or decision trees is an interesting idea.

Compared to recent Transformer studies, my models achieved lower performance, though [add ref simple TF+OCRead] noted stability challenges with ABIDE and addressed this via stratified splitting by site, age, and gender. While I balanced age and gender overall, fold-level stratification could improve robustness. As in [add ref Thomas], more extensive hyperparameter optimization is needed. Positional encoding choices also strongly impact models - [add ref simple TF+OCRead] found adjacency matrices superior to the original Vaswani et al. formulation for fMRI. Ultimately, multimodal fusion of structural and functional data may hold the most promise, as the top model in

[Traut et al., 2022] combined sMRI and rs-fMRI. Incorporating complementary anatomical patterns could enhance accuracy and interpretability

In summary, this work implemented a reasonable starting point for Transformer model interpretation in this novel application area. As a foundation for future research, we identified several promising directions such as: analysing intermediate representations, aggregating local explanations, and substituting alternative interpretable modules. Enhancing interpretation methods will lead to greater knowledge of how these models encode predictive fMRI signals related to autism, advancing applicability in healthcare settings. Expanding the datasets, input features, model capacity, and controllable variables represent exciting opportunities to build on these foundations in future studies. Leveraging the full breadth of neuroimaging, clinical, and demographic data could ultimately yield more robust and generalizable models. Overall this study introduced innovative DL architectures for decoding brain dynamics, laying groundwork for an explainable AI system to augment understanding of neuroimaging biomarkers.

Conclusion

This study pioneered applications of Transformer neural networks, a leading DL architecture, for decoding predictive patterns in resting-state fMRI data related to ASD. We developed a methodological framework encompassing data preprocessing, cross-validation strategies, multi-task learning, and interpretability analyses.

While accuracy gains over single-task models were modest, multi-task approaches significantly altered model behaviours in nuanced ways, demonstrating the value in joint training. This establishes a strong basis for refinements like loss weighting and learning rate optimization. With hyperparameter tuning and expanded datasets, both approaches show promise for distilling insights about spatiotemporal brain dynamics.

For model interpretation, we implemented reasonable initial techniques including representation visualisation and LIME relevance mapping. Analysing intermediate layers, aggregating local explanations, and integrating alternative interpretable modules offer exciting future directions. Enhanced interpretation can uncover how predictive fMRI patterns are encoded.

This work overall provides a springboard for developing robust and explainable DL systems that leverage diverse neuroimaging, clinical, and demographic data. Resulting models could offer invaluable biomarkers to advance autism prediction, personalised diagnosis, and treatment in healthcare applications. By pioneering innovative machine learning architectures for decoding brain activity, this study lays the groundwork for augmented understanding of neuroimaging signatures in ASD.

Appendix:

diff_Models	**W**	**p_val**	**normal**
diff_M2_M1	0,988	0,500	True
diff_M3_M1	0,987	0,409	True
diff_M3_M2	0,991	0,708	True
diff_M4_M1	0,974	0,047	False
diff_M5_M1	0,970	0,023	False

Table X. Shapiro-Wilk tests of normality of the differences between the models