BOKI results

I S Plank

2025-10-17

# Results

## Distinguishing ASD-involved from BPD-involved interactions

Four of our eight base models were able to distinguish between ASD-involved and BPD-involved interactions based on permutation testing, with the exception of the *BODYsync*, *CROSSsync*, *HEADsync* and *INTRAsync* models (*BODYsync*: *pFDR* = 1; *CROSSsync*: *pFDR* = 0.136; *HEADsync*: *pFDR* = 0.936; *INTRAsync*: *pFDR* = 1; all other models *pFDR* < 0.05). Of the base models, the *CROSSturn* model performed best, with a balanced accuracy of 76.7%, weighing sensitivity and specificity equally, which were 67.6% and 85.7%, respectively. The other base models performing above chance were the *Speech* model with a balanced accuracy of 70.4% (64.7% sensitivity; 76.2% specificity), followed by the *MovEx* model with 68.7% (58.8% sensitivity; 78.6% specificity) and the *FACEsync* model with 68% (76.5% sensitivity; 59.5% specificity). The stacking model outperformed all base models, although the difference between the stacking model and the *CROSSturn* model was not significant (see [!T]). The stacking model achieved a balanced accuracy of 81.7% (70.6% sensitivity; 92.9% specificity; see Figure [!X]), meaning that only two interaction partners from BPD-involved interactions were misclassified as ASD-involved and ten were misclassified as BPD-involved despite being ASD-involved. The other 64 interaction partners were classified correctly by the stacking model (see Figure [!X]). For performance parameters and FDR-corrected *p*-values for all models, please consult the supplementary materials [!S].

## Distinguishing ASD-involved from non-clinical interactions

All base models except the *HEADsync* and the *INTRAsync* models performed significantly better than chance according to permutation tests (*HEADsync*: *pFDR* = 1; *INTRAsync*: *pFDR* = 1). For this comparison, the *FACEsync* model achieved the highest balanced accuracy, specifically 74.9% (79.4% sensitivity; 70.5% specificity), closely followed by the *CROSSsync* model with 71.3% balanced accuracy (67.6% sensitivity; 75% specificity). The stacking model had a significantly higher balanced accuracy of 80.6% (79.4% sensitivity; 81.8% specificity). Here, seven ASD-involved interactions were misclassified as non-clinical, while zero non-clinical interactions were mistaken as ASD-involved interactions. The rest of the interaction partners, 63 in total, were classified accurately. The stacking did not perform significantly better than the best base model (*FACEsync*: *pFDR* = 0.112) but outperformed all other base models (see [!T]).

## Exploring influential features

To explore the features in the interaction partners themselves, we plotted the three most influential features for the distinction between ASD-involved and BPD-involved as well as for the distinction between ASD-involved and non-clinical. We focused on the *CROSSturn*, *FACEsync*, *MovEx* and *Speech* models and determined the most influential features based on their feature weights. For the *Speech* model, we focused on individual features. Visual exploration of these features revealed COMP interaction partners in the ASD- or BPD-involved dyads adapting their behaviour to their clinical counterpart in some of the features (see Figure [!X]). For instance, group comparisons revealed that COMP interaction partners did not differ from their clinical interaction partners in the variance of their speech intensity in the hobbies conversation (ASD-involved: *pHSD* = 0.711; BPD-involved: *pHSD* = 0.999), but COMP interaction partners’ behaviour differed between COMP interaction partners in BPD-involved and COMP interaction partners in ASD-involved interactions (*pHSD* < 0.05). Furthermore, COMP interaction partners in non-clinical interactions behaved comparably to COMP interaction partners in both clinical interactions (ASD-involved: *pHSD* = 0.2; BPD-involved: *pHSD* = 0.203). A similar pattern can be observed for the standard deviation of synchronisation of AU26, which measures jaw dropping, in the hobbies conversation with COMP interaction partners in clinical interactions not differing from their interaction partners (ASD-involved: *pHSD* = 0.999; BPD-involved: *pHSD* = 0.936), but COMP interaction partners in BPD-involved interactions differing from COMP interaction partners in ASD-involved interactions (*pHSD* < 0.05). COMP interaction partners also showed comparable total head motion in the hobbies conversation to their interaction partners (ASD-involved: *pHSD* = 1; BPD-involved: *pHSD* = 0.151), but here they also did not differ between clinical interactions (*pHSD* = 0.57). However, COMP interaction partners in non-clinical interactions exhibited more head motion than COMP interaction partners in ASD-involved interactions (*pHSD* = 0.04), but not from COMP interaction partners in BPD-involved interactions (*pHSD* = 0.726).

## Exploring multi-group classification

Arguably, the above results already show that distinguishing between the two clinical labels involves a different pattern of models compared to distinguishing between the ASD-involved from the non-clinical label, which also reflects clinical decision making (a clinical decision between clinical labels is derived on different information than the decision between the presence and absence of a clinical label). Nevertheless, we also explored the use of a multi-group classification which uses the output of the three one-vs-one classifiers to predict one of the three possible labels: ASD-involved, BPD-involved or non-clinical. All base models performed significantly above chance except for the *BODYsync* model (*pFDR* = 1). The best base model was the *Speech* model which achieved a balanced accuracy of 63.9%, followed by the *MovEx* model which reached 60.8% balanced accuracy. The stacking model performed comparable to those two base models with 63.2% balanced accuracy, while outperforming all other base models (see [!T]). The confusion matrix of the stacking model revealed that many of BPD-involved and non-clinical interactions were misclassified as ASD-involved, 47.6% and NaN% respectively, while ASD-involved interactions were detected accurately in 61.8% of the cases, possibly reflecting our focus on ASD associated features.

| comparison | BODYsync | CROSSsync | CROSSturn | FACEsync | HEADsync | INTRAsync | MovEx | Speech |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ASD-inv vs. BPD-inv | 0.000\* | 0.000\* | 0.314 | 0.004\* | 0.000\* | 0.000\* | 0.006\* | 0.002\* |
| ASD-inv vs. COMP | 0.000\* | 0.013\* | 0.000\* | 0.112 | 0.000\* | 0.000\* | 0.025\* | 0.002\* |
| BPD-inv vs. COMP | 0.000\* | 0.000\* | 0.000\* | 0.007\* | 0.227 | 0.000\* | 0.269 | 0.015\* |
| Multi-Group | 0.000\* | 0.000\* | 0.000\* | 0.000\* | 0.000\* | 0.000\* | 0.067 | 0.462 |