A Time Series Classification Pipeline for Detecting Interaction Ruptures in HRI Based on User Reactions: **Appendix**

Lennart Wachowiak, Peter Tisnikar, Andrew Coles,
Gerard Canal, Oya Celiktutan

This is the appendix to our paper *A Time Series Classification Pipeline for Detecting Interaction Ruptures in HRI Based on User Reactions*. In this document, we provide further information about the data set (Section 1), as well as further results on feature importance, model parameters, and learning curve for the user awkwardness and robot error prediction tasks (Section 2).

1 Data

Here, we provide some further detail on what data was used and how we pre-processed it. In the following section, you can find more information about the individual folds, a list of all input features, participant statistics, and a visualization of the sliding window approach.

Data Folds

Table 1 provides additional information about the label distribution of the 4 folds used for cross-validation (18 participants) and the final hold-out test set (5 participants). While the proportion of errors is very similar across all folds, the test set contains noticeably more moments of user awkwardness than any of the cross-validation folds (+ 8 percentage points).

Fold	Sessions	Proportion User Awkwardness	Proportion Robot Error	Proportion Interaction Rupture
1	19	16%	16%	23%
2	20	17%	18%	27%
3	16	12%	16%	23%
4	16	17%	15%	24%
1, 2, 3, 4	71	16%	16%	24%
Hold-out test	18	23%	17%	34%

Table 1: Label distribution for different data folds.

List of Input Features

Table 2 provides a description of all features that are available for all participants in the dataset. Features are grouped by type: facial action units from OpenFace, speech features from openSMILE, body pose features from OpenPose, speaker diarization features, and the frame. There are 94 features in total, of which 20 are binary, and 74 are continuous.

[1] `https://github.com/TadasBaltrusaitis/OpenFace/wiki/Action-Units`
[2] `https://github.com/audeering/opensmile`
[3] `https://en.wikipedia.org/wiki/Mel-frequency_cepstrum`
[4] `https://github.com/CMU-Perceptual-Computing-Lab/openpose`

Feature	Description	Data Type
OpenFace	AUs encoded as intensity or binary activation value[1]	
AU01	Inner brow raiser	float/binary
AU02	Outer brow raiser	float/binary
AU04	Brow lowerer	float/binary
AU05	Upper lid raiser	float/binary
AU06	Cheek raiser	float/binary
AU07	Lid tightener	float/binary
AU09	Nose wrinkler	float/binary
AU10	Upper lip raiser	float/binary
AU12	Lip corner puller	float/binary
AU14	Dimpler	float/binary
AU15	Lip corner depressor	float/binary
AU17	Chin raiser	float/binary
AU20	Lip stretcher	float/binary
AU23	Lip tightener	float/binary
AU25	Lips part	float/binary
AU26	Jaw drop	float/binary
AU28	Lip suck	float
AU45	Blink	float/binary
openSMILE	Spectral and energy-based features of sound. Smoothed with a moving average of 3 frames. [2]	
Loudness	Perceived audio signal loudness	float
alphaRatio	Energy ratio and a feature of phonation type	float
hammarbergIndex	Spectral balance index	float
slope0-500	Low-frequency spectrum tilt	float
slope500-1000	Mid-frequency spectrum tilt	float
spectralFlux	Measure of spectral change over time	float
mfcc1	First MFCC[3], a measure of energy in spectrum	float
mfcc2	Balance between low and high frequencies	float
mfcc3/mfcc4	Spectral shape feature	float/float
F0semitoneFrom27.5Hz	Pitch tracking expressed as semitones above 27.5 Hz	float
jitterLocal	Pitch stability indicator	float
shimmerLocaldB	Loudness stability detector	float
HNRdBACF	Ratio of energy between harmonics and noise	float
logRelF0-H1-H2	Logarithmic difference in amplitude between F0 and second harmonic	float
logRelF0-H1-A3	Logarithmic difference in amplitude between F0 and third formant	float
F1frequency	Frequency of first formant	float
F1bandwidth	Spread of frequencies around F1	float
F1amplitudeLogRelF0	Logarithmic ratio between F1 and F0	float
F2frequency	Frequency of second formant	float
F2bandwidth	Spread of frequencies around F2	float
F2amplitudeLogRelF0	Logarithmic ratio between F2 and F0	float
F3frequency	Frequency of third formant	float
F3bandwidth	Spread of frequencies around F3	float
F3amplitudeLogRelF0	Logarithmic ratio between F3 and F0	float
OpenPose	Distances and velocities between different 2D-body-keypoints [4]	
dist_4_7/vel_dist_4_7	Right wrist to left wrist	float/float
dist_4_2/vel_dist_4_2	Right wrist to right shoulder	float/float
dist_4_5/vel_dist_4_5	Right wrist to left shoulder	float/float
dist_4_1/vel_dist_4_1	Right wrist to neck	float/float
dist_4_17/vel_dist_4_17	Right wrist to right ear	float/float
dist_4_15/vel_dist_4_15	Right wrist to right eye	float/float
dist_4_18/vel_dist_4_18	Right wrist to left ear	float/float
dist_4_16/vel_dist_4_16	Right wrist to left eye	float/float
dist_7_2/vel_dist_7_2	Left wrist to right shoulder	float/float
dist_7_5/vel_dist_7_5	Left wrist to left shoulder	float/float
dist_7_1/vel_dist_7_1	Left wrist to neck	float/float
dist_7_17/vel_dist_7_17	Left wrist to right ear	float/float
dist_7_15/vel_dist_7_15	Left wrist to right eye	float/float
dist_7_18/vel_dist_7_18	Left wrist to left ear	float/float
dist_7_16/vel_dist_7_16	Left wrist to left eye	float/float
Speaker Diarization	Binary values indicating who speaks	
Robot	Robot speaks	binary
Participant	Participant speaks	binary
Pause	No one speaks	binary
Others		
Frame	Timer going up with a rate of 30fps	integer

Table 2: Descriptions of all input features.

Participant Statistics

Participant statistics for the dataset used [30, 31]:

- 6 women, 1 non-binary person, 19 men
- age: 18-25 (7 participants), 26-35 (11), 36-45 (4), 46-55 (4)
- employees of Cambridge Consultants

Of these 26 participants, 3 were excluded due to recording issues; however, as a competition team, information on participant exclusion was not made available to us.

Processing and Classification Pipeline

Figure 1 visualizes the process of chunking the original, full-length time series data into individual intervals that are fed to the classification model. In the second step, The sliding window splits the data into equal-length intervals and moves forward by a distance defined through a stride parameter s. At the same time, labels are created for the individual intervals based on the last s frames of each interval besides the first. Lastly, the image shows how the first interval's data is padded so that the same labeling process can be applied uniformly.

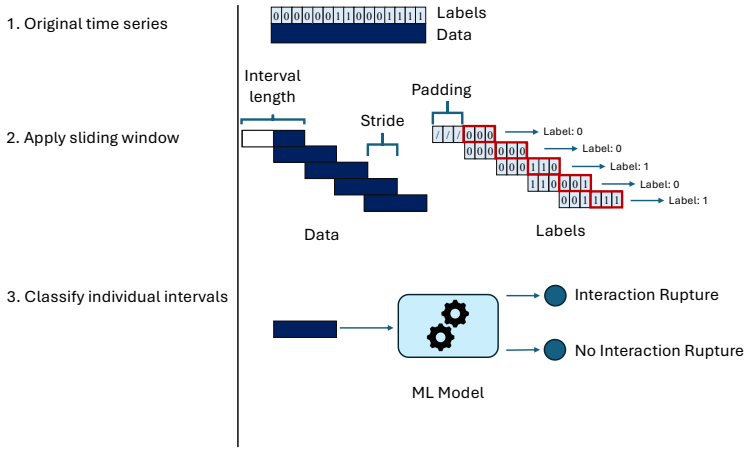


Figure 1: Sliding window visualization. From (1) the original data and labels to (2) the individual intervals with one label per interval and start padding to (3) classifying the individual intervals with the trained model.

2 Additional Results

Here, we provide some further details of the experiment results. We provide tables indicating the best hyperparameters for each model type as well as the parameters used for the models submitted to the ICMI ERR@HRI competition. We provide further performance analysis for different model types on the interaction rupture task, and provide a more detailed feature importance analysis for the best-performing MiniRocket model, as well as feature analysis for all three tasks.

MiniRocket Configurations

Table 3 provides values for data- and model-related hyperparameters used to create the MiniRocket ensembles that won the ERR@HRI competition. The right-most column shows the parameters for the MiniRocket ensemble that ended up performing best during cross-validation after the competition ended.

The difference between the best MiniRocket model and all competition models is that it has a longer interval window, which was enabled by context padding at the beginning of the interaction. Furthermore, the best MiniRocket removes all OpenPose features as opposed to just the velocity-based features. The best MiniRocket model reaches an accuracy of around 84% and a macro F1 score of 0.76 in cross-validation, improving on the submitted MiniRocket model by around 2% both in accuracy as well as macro F1 score.

	Interaction Rupture (submitted)	**Robot Error** (submitted)	**User Awkwardness** (submitted)	**Interaction Rupture** (best MiniRocket)
Data Param.				
Interval Length	15s	16s	15s	25s
Stride Train	4s	4s	4s	6s
Stride Eval	2.25s	2.25s	2.25s	3s
FPS	25	25	25	25
Removed Features	vel_dist, c_openface	openpose, c_openface	vel_dist, c_openface	openpose, c_openface
Label Creation	stride_eval	stride_eval	stride_eval	stride_eval
NaN Handling	avg	avg	avg	avg
Oversampling Rate	0.15	0.2	0.1	0.1
Undersampling Rate	0.05	0.0	0.05	0.1
Rescaling	normalization	none	none	none
Zero Padding	False	False	False	True
Model Param.				
Number of Estimators	25	20	20	10
Max Dilations per Kernel	64	32	64	32
Class Weight	None	None	None	None
Performance				
Accuracy (Cross-Val.)	0.82	0.89	0.84	0.84
Macro F1 (Cross-Val.)	0.74	0.77	0.55	0.76
Accuracy (Test)	0.80	0.87	0.76	N/A
Macro F1 (Test)	0.75	0.73	0.55	N/A

Table 3: Parameters used for the MiniRocket ensembles that we submitted to the three sub-challenges of the ERR@HRI competition. In the last column, the parameters that led to the overall best performance based on further experiments after the competition ended.

Other Model Configurations

Table 4 contains the parameters resulting in the best cross-validation performances for each other model type, i.e., ConvTran, TST, and Random Forest. While both deep models perform best with longer interval lengths and lower framerates, the Random Forest model performs best with shorter intervals and a higher framerate. This can be attributed to the fact that Random Forest computes a summary statistic (i.e., the mean) of the data within the interval, compressing them into a single point on which it then performs the prediction. All models exclude OpenPose data and OpenFace's binary AUs. All models utilize some undersampling and oversampling to reduce the data imbalance.

The performance of the ConvTran model is comparable with that of MiniRocket, with Random Forest trailing by about 2% in accuracy, and 0.05 in macro F1. TST, on the other hand, only managed to match Random Forest's macro F1 score, while being even less accurate at around 80%.

Parameter	ConvTran	TST	RandomForest
Data Parameters			
Interval Length	25s	25s	7s
Stride Train	3s	3s	2s
Stride Eval	3s	6s	3s
FPS	25	25	50
Removed Features	openpose, c_openface	openpose, c_openface	openpose, c_openface
Label Creation	stride_eval	stride_eval	stride_eval
NaN Handling	zeros	zeros	zeros
Oversampling Rate	0.1	0.1	0.2
Undersampling Rate	0.1	0.1	0.1
Rescaling	normalization	normalization	standardization
Start Padding	true	true	false
Summary Technique	N/A	N/A	mean
Model Parameters			
d_model	8	128	N/A
n_heads	8	10	N/A
dim_ff	32	N/A	N/A
encoder_dropout	0.2	0.2	N/A
fc_dropout	0.1	0.4	N/A
n_layers (sub-encoder)	N/A	2	N/A
Batch Size	16	16	N/A
Learning Rate	0.041	0.00003	N/A
Loss	focal loss	cross-entropy loss	N/A
n_estimators	N/A	N/A	220
max_depth (per tree)	N/A	N/A	40
criterion (for split)	N/A	N/A	entropy
max_features (considered for split)	N/A	N/A	log2
Performance			
Accuracy	0.84	0.80	0.82
Macro F1	0.77	0.72	0.72

Table 4: Comparison of parameters and performance metrics for best-performing ConvTran, TST, and RandomForest models in interaction rupture detection.

Feature Importance

In this section, we present some additional feature importance analysis. We first present additional results on different feature groups' impact on model performance in terms of macro F1 score. Then, we focus our analysis on the best-performing MiniRocket model and perform a more fine-grained feature analysis, where we analyze the performance of all possible feature combinations between modalities. Furthermore, we perform additional feature analysis on the user awkwardness and robot error prediction tasks. Finally, we verify that the speaker diarization data (i.e., who speaks when), given their strong predictive power, do not hold a simple linear relationship with the task labels.

Macro F1 Score

Complementary to Figure 6 in the main paper which shows the change in accuracy over the baseline majority classifier for the interaction rupture task, Figure 2 shows the change in macro F1 score when training models with or without certain feature groups. Whereas certain feature combinations decreased accuracy compared to the majority classifier baseline, the macro F1 score is positively affected by all features except for the frame, which does not improve the macro F1 score of any model except for Random Forest. The best-performing feature combination in terms of macro F1 improvement is the one that excludes OpenPose and OpenFace's binary AUs, as the improvement is, on average, the biggest. This is followed closely by features excluding OpenPose. Somewhat surprisingly, providing only speaker diarization features to the model also results in big improvements in performance for the

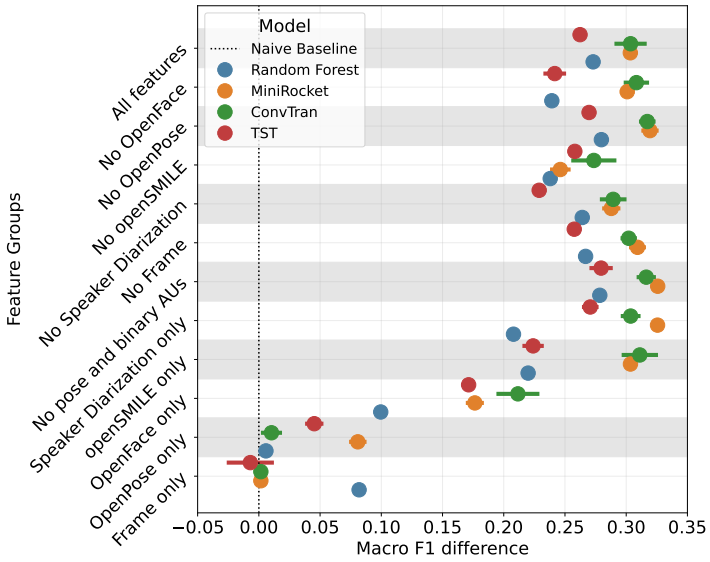


Figure 2: Changes in F1 score compared to naive baseline across different feature combinations.

convolutions-based models (MiniRocket and ConvTran).

All Feature Group Combinations

To better understand the importance of different features, we split the features into six sub-groups: OpenPose, OpenFace's binary AUs, OpenFace's continuous AUs, openSMILE, Speaker diarization, and frame. We computed all combinations of these six sub-groups and trained the best-performing MiniRocket model on each one with cross-validation. We report the cross-validation accuracy scores in Figure 3.

As seen in the figure, the best-performing feature combination is the one that contains openSMILE, speaker diarization, and frame features. This is an interesting result which we further discuss in the following sections. Furthermore, all best-performing models were trained on some or all speech-based features (i.e., on either openSMILE, or speaker diarization, or both). The feature combination that we found during our hyperparameter search, which excludes OpenPose data and OpenFace's binary AUs, is the fifth-best-performing feature combination.

The exhaustive search over all feature combinations also confirmed that OpenPose and OpenFace data were among the least informative features, indicating that they are noisy and do not hold a lot of predictive power. This is further manifested in the fact that if all features are included, the performance of the model is worse than excluding any of the two or both. Furthermore, frame information, although helpful in conjunction with other features, does not hold any predictive power on its own and also lowers performance when paired with OpenPose.

Feature Importance Across All Tasks

To understand whether these trends hold across all tasks, we ran a feature combinations search for the best-performing MiniRocket model on all three tasks: user awkwardness prediction, interaction rupture prediction, and robot error prediction. We ran each MiniRocket model with 5 random seeds and computed an average cross-validation accuracy and an average cross-validation macro F1 score.

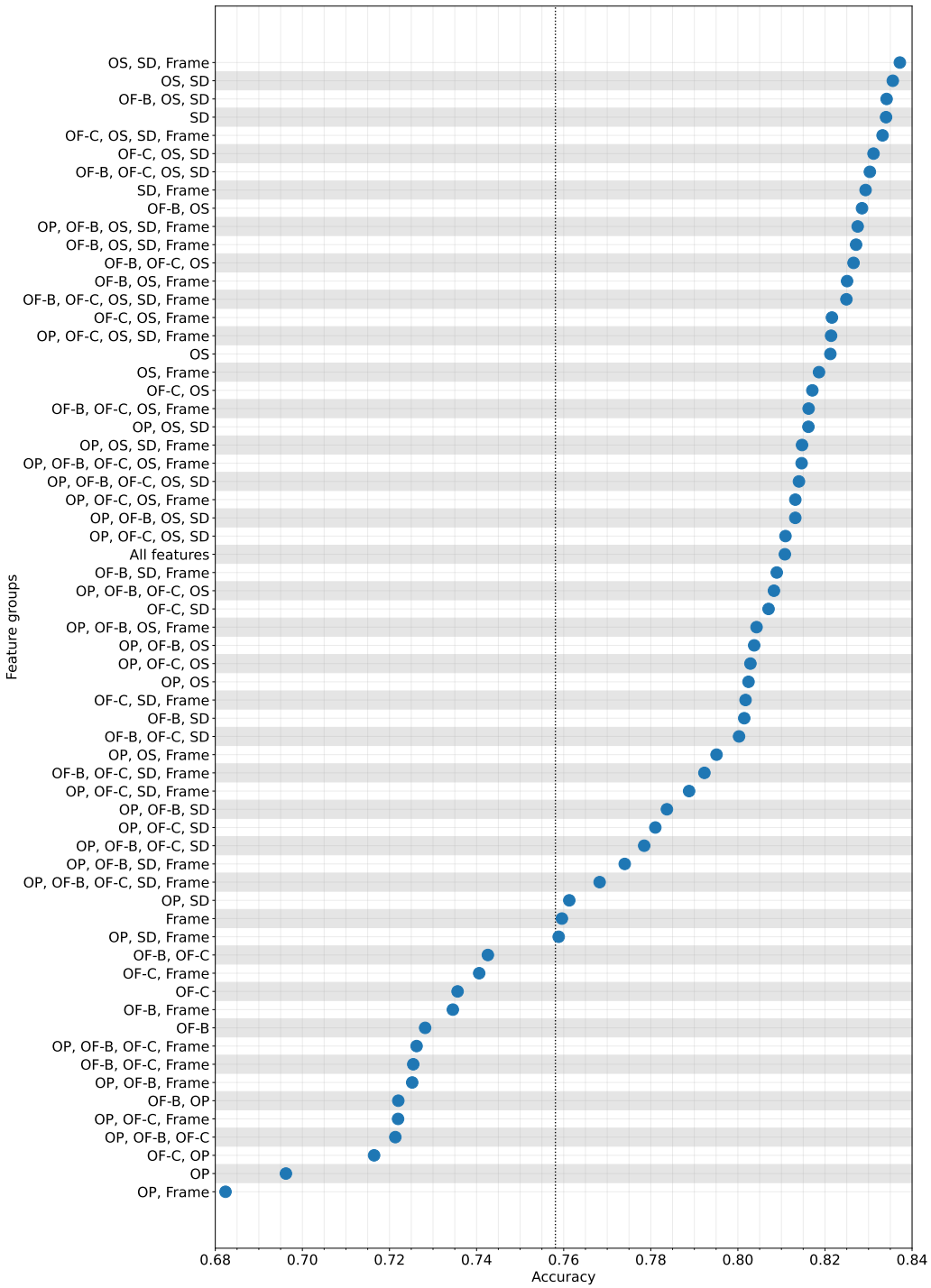
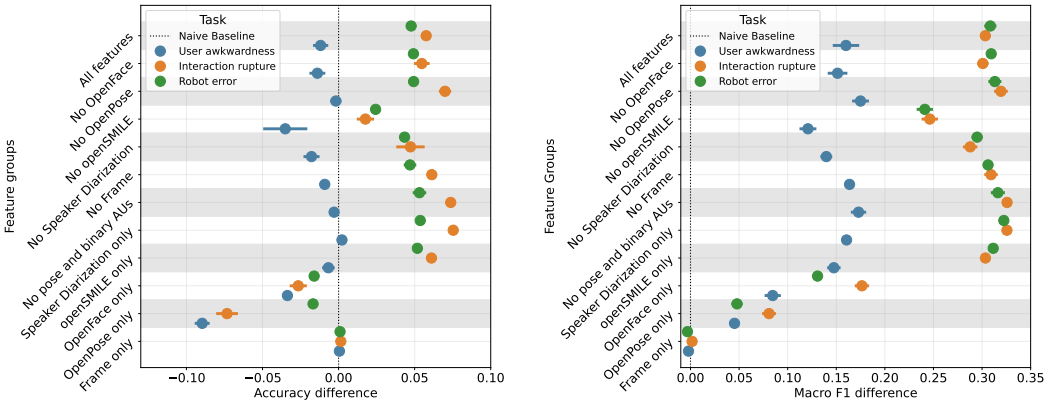


Figure 3: Accuracy for a MiniRocket model trained on all possible feature group combinations between six feature groups: OpenPose (**OP**), OpenFace's binary Activation Units (**OF-B**), OpenFace's continuous Activation Units (**OF-C**), openSMILE (**OS**), Speaker Diarization (**SD**), and Frame. The MiniRocket model was trained on the Interaction Rupture task. The dashed line indicates the performance of a majority classifier.



(a) Accuracy difference.

(b) Macro F1 difference.

Figure 4: Accuracy and Macro F1 differences compared to the majority classifier baseline across the three tasks

Figure 4a shows the accuracy difference to the majority classifier baseline (whose accuracy was approximately 76% for interaction rupture detection and approximately 84% for user awkwardness and robot error detection).

One immediate observation is the fact that a big improvement in being able to predict interaction rupture also means a big improvement in predicting robot errors. The relationship between accuracy in predicting user awkwardness and interaction rupture also seems to follow a similar pattern, but user awkwardness appears to be difficult to predict in general, with only one feature group even beating the naive baseline's accuracy.

A similar trend can also be observed in Figure 4b, where all feature combinations, except for only frame, improve the macro F1 score over the baseline. The combinations that most improve classifier performance across all three tasks all exclude OpenPose data, further confirming our findings that it doesn't hold much predictive power for the tasks at hand.

Linear Relationships with Speaker Data

	Interaction Rupture	**User Awkwardness**	**Robot Error**
predict during pause	66.2%	61.4%	67.9%
predict when robot speaks	52.7%	57.8%	54.4%
predict when user speaks	56.9%	65.2%	61.7%
logistic regression	75.8%	84.4%	84.0%
majority classifier baseline	75.8%	84.4%	84.0%

Table 5: Analysis results of linear relationships between speaker diarization data and labels, computed across all 71 sessions. Relationships were constructed on a frame-by-frame basis, as opposed to being based on the time series processing pipeline.

In the feature importance analysis, speaker diarization data turned out to be the most important modality for model performance — to such an extent that models only trained on speaker diarization data outperformed most models trained on any other data mixture.

To test whether the speaker patterns hold any simple linear relationships with the labels, we evaluate the performance of linear classifiers that always predict interaction rupture or no interaction rupture when the robot, participant, or no one speaks. However, no such relationship holds, as performance is always below that of a simple majority class baseline. Additionally, a logistic regression model fit to the three types of speaker diarization data

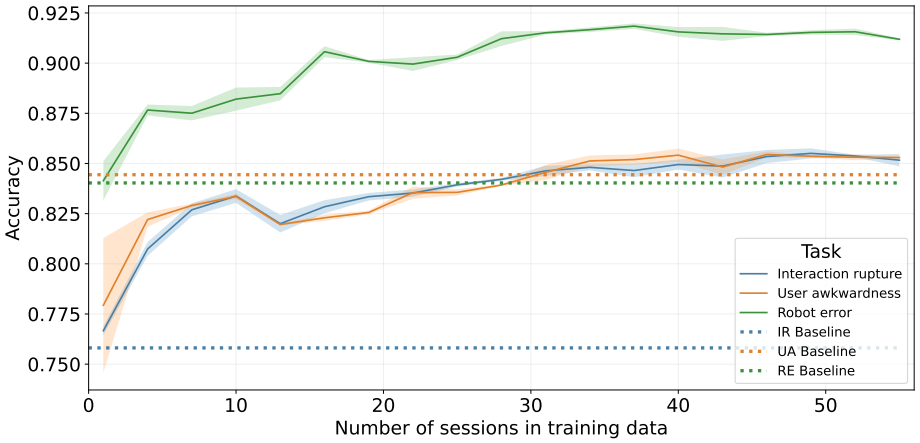


Figure 5: Learning curves for each of the sub-tasks, showing how the accuracy score depends on the number of sessions in the training data.

matches, but cannot outperform the majority class baseline. Exact numbers can be found in Table 5.

The classifiers were fit to all available training data and evaluated on the same. The models' inability to match the baseline holds true across all three tasks and, thus, shows that the formulation as a time series classification problem is crucial to pick up on the temporal aspects of the speech patterns.

Additional Learning Curves

As the main paper only showed the learning curve for interaction rupture detection, Figure 5 additionally shows the MiniRocket learning curves for robot error detection and user awkwardness detection. The curve for robot error detection is similar to that of interaction rupture detection, albeit starting out with a higher accuracy score when trained on a single session, which makes sense given the equally higher baseline accuracy of a majority class classifier, caused by an even greater class imbalance. The two curves for interaction rupture and robot error detection surpass their respective baseline almost immediately. While the interaction rupture detection score then rises a bit quicker than the robot error score, the overall gains are similar across all tasks.

On the other hand, the learning curve for user awkwardness detection only surpasses its majority class classifier baseline when trained on 30 or more sessions. This indicates that whilst more data helps the classifier performance, predicting user awkwardness remains a challenging task even as more data becomes available. This can potentially be attributed to the relatively small predictive power of facial and pose features that could be indicative of a person's affective state.