Predicting high vs low mother-baby synchrony with GRU-based ensemble models

Daniel Stamate[1,2], Riya Haran[1], Karolina Rutkowska[1], Pradyumna Davuloori[1], Evelyne Mercure[3], Caspar Addyman[4], Mark Tomlinson[4]

1. Data Science & Soft Computing Lab, Department of Computing, Goldsmiths, University of London, UK

2. School of Health Sciences, University of Manchester, UK

3. Department of Psychology, Goldsmiths, University of London, UK

4. Department of Global Health, Stellenbosch University, South Africa

d.stamate@gold.ac.uk

Abstract. The early stages of life are paramount for the baby's brain and emotional development, and the quality of interaction between mother and baby - measured as a dyadic synchrony score, is critical in that period. This study proposes the first machine learning prediction modelling approach, based on Gated Recurrent Unit - GRU ensemble models, to automatically differentiate high from low dyadic synchrony between mother and baby, using a dataset of videos capturing this interaction. The GRU ensemble models which were post-processed by maximising the Youden statistic in a ROC analysis procedure, show a good prediction capability on test samples, including a mean AUC of 0.79, a mean accuracy of 0.72, a mean precision of 0.87, a mean sensitivity of 0.64, a mean f1 performance of 0.72, and a mean specificity of 0.83. In particular the latter performance represents an 83% detection rate of the mother-baby dyads with low synchrony, suggesting these models' high capability for automatically flagging such cases that may be clinically relevant for further investigation and potential intervention. A Monte Carlo validation procedure was conducted to accurately estimate the above mean performance levels, and to assess the proposed models' stability. The statistical significance of the prediction ability of the models was also evaluated, i.e. mean AUC > 0.5 (p-value < 9.82×10^{-19}), and future research directions were discussed.

Keywords: Automating mother-baby synchrony detection, Gated Recurrent Units - GRU, Ensemble learning, ROC analysis, Monte Carlo validation

1 Introduction

The early stages of life are paramount for babies' brain and emotional development, and the quality of interaction between mother and baby is critical in that period. If a baby is denied the attention and a positive interaction, they can struggle in later life with forming relationships, education and functioning in society [1]. An increasing body of research shows that babies who were neglected from the early stages of

development face further social development difficulties [2]. In particular, research suggests that synchrony between the infant's behaviour and their caregivers play many functions in the infants' development, from co-regulations of exchanges in interactions to language acquisition [3]. A functional interaction between mother and baby is one in which the mother focuses her attention on the child and responds to their behaviour in a short time. Such an interaction can be described as synchronous. According to [4] synchrony between two people is defined as a state where they move together in the same or almost the same time with one another. Research suggests that synchrony in group interactions can have a later positive influence on forming social actions [5]. Synchrony is used to find patterns in movements of positive and negative interactions between mother and baby. Developing new methods for finding synchrony patterns can help to automate the process of assessing the mother-baby interaction quality.

Due to its vital role in the early stages of baby's development, expert assessment of the synchrony between mother and baby in videos capturing this interaction, is an important research question. Moreover, there is value in automating this assessment process using machine learning, as such automation could flag those videos which are more likely to capture a negative, lower synchrony between mother and baby, allowing early specialist intervention in problematic mother-baby interactions.

Predicting synchrony between participants in videos using machine learning models, was previously tackled in literature including works such as [6], in which the authors successfully trained a model based on Long Short-Term Memory (LSTM) recurrent neural networks [13, 14], on facial expressions data that had been extracted from pre-recorded videos representing a group of three interacting people. The proposed approach was used to predict synchrony score on a scale of 1 to 5, and the recurrent neural model's predictions were validated by comparison with predictions based on a random permutations baseline. In another machine learning study proposing the prediction of synchrony between a human arm and a robot arm, the final position of the human arm was predicted also with recurrent neural networks based on LSTM models [7].

In the present study we propose an innovative machine learning approach to predicting the categorical level of dyadic synchrony – high versus low, for 58 mother-baby dyads, based on a dataset comprising 58 records with body part coordinates extracted from 58 videos capturing the interaction of these dyads. Our approach is based on Gated Recurrent Unit (GRU) recurrent neural networks [8, 13] as baseline models, with a focus on ensembles of such models – with the purpose to enhance the models' prediction and stability on a relatively small number of record dataset. GRUs are similar to but involve a lesser complexity in training than LSTM models [13, 14] since they are able to store and filter the information using only two gates - *reset* and *update*, as opposed to three gates – *input*, *output* and *forget*, for LSTMs, respectively. GRU models are often capable of performance levels comparable to SLTM models, and due to their reduced relative complexity are preferred in this preliminary study on a dataset comprising a relatively small number of records. However, the volume of data extracted from videos is relatively large, overall, leading to a substantial computational cost.

The rest of the paper is organised as follows: Section 2 introduces our proposed prediction modelling approach's methodology, including data description and pre-processing, and model development, evaluation, and Monte Carlo validation. Section 3

presents and discusses our results, and Section 4 concludes the paper and outlines future research directions.

2 Methodology

2.1 Data description and pre-processing

This work was based on a sample of 60 videos from the SPEAKNSIGN dataset [20], each lasting more than 10 minutes with 25 frames per second, capturing a session of free-play between 4-7-month-old infants and their mothers. The videos were scored by experts with a dyadic synchrony score ranging from 2 (low) to 14 (high).

OpenPose library [19] was used to extract a 5D array based on coordinates of body part keypoints from each video. In particular, for the purpose of this analysis, data representation was adapted and simplified by extracting, for each frame, pairs of x and y coordinates for 25 body keypoints for each mother and her baby. Fig 1 illustrates the body part keypoints extracted by OpenPose from a single frame of the interaction video. 3D arrays were finally obtained for the analysis, comprising the record number corresponding to each video, the frame number, and the sum aggregation of the x and y coordinates. Two records were discarded as they did not meet the data quality requirements, leading to a dataset of 58 records in all. Records were categorized in two classes by using the dyadic synchrony scores: class 1 – high synchrony, and class 0 – low synchrony, containing the highest 60% scores and the lowest 40% scores in the dataset, respectively.

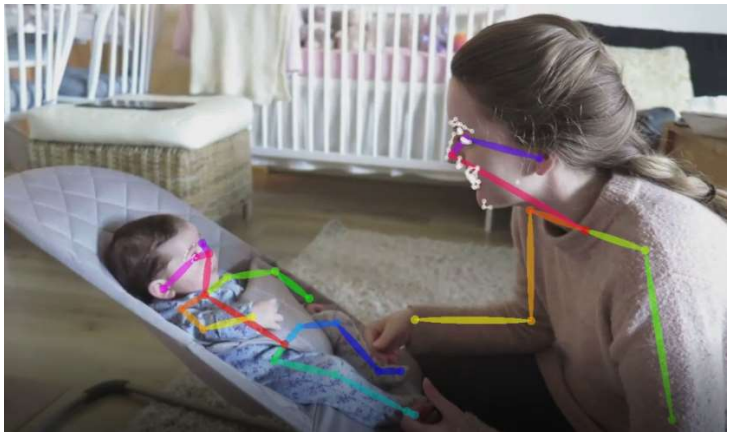


Fig. 1. Body part keypoints extracted by OpenPose from a single frame of the interaction video.

The dataset was cleaned with respect to missing values which were imputed with linear interpolation, and outliers were detected using criteria based on the range of 0.025 or 0.975 quantiles, and discarded. Data was normalised.

Fig. 2 illustrates, in a preliminary exploratory data analysis conducted in [20], a partial correlation between mother and baby as reflected by the whole body movement index aggregating differences in body coordinates in the frame sequence [20]. We note various levels of correlation of the body movement index between mother and baby, and these go as high as 0.84 in the four examples of mother-baby dyads illustrated here (see second plot).

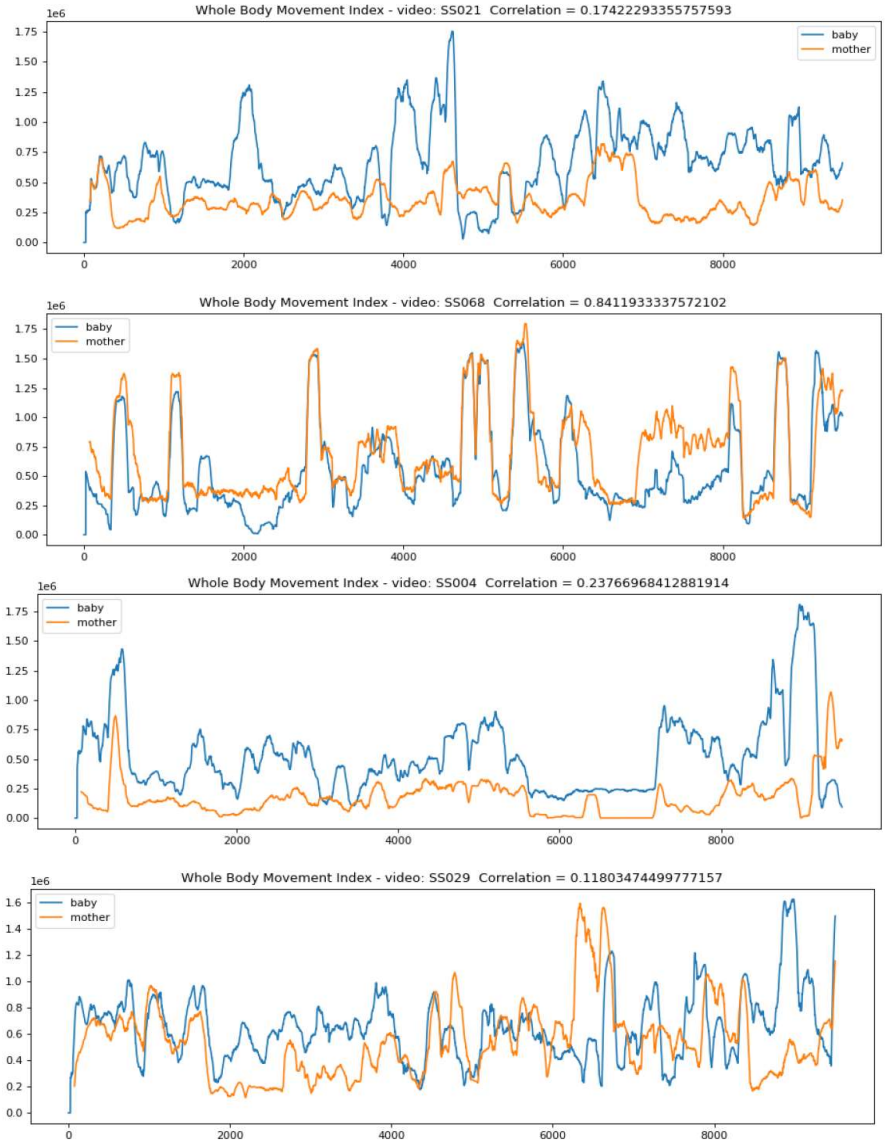


Fig. 2. Body movement index capturing a correlation in mother and baby's body movement [20].

2.2 Model development, evaluation and Monte Carlo validation

The baseline GRU neural network architectures [8,13] used in this work comprised 2 and 3 GRU layers with 200, 256, 300 nodes, and 1 hidden dense layer with 50, 64, 70 nodes, implemented in Keras and TensorFlow. As activation functions we employed *relu*, *prelu*, *elu*, *selu*, *softplus* for the hidden dense layer and *sigmoid* for the output layer, while for the GRU layers we used *tanh* as activation function and *sigmoid* as recurrent activation. As loss functions we employed *Binary crossentropy* and *binary focal crossentropy* (for the moderate data imbalance 60:40 present in the data). The constant learning rate of 0.001, and the exponential learning rate scheduling were used, together with *adam*, and *nadam* optimisers. To prevent overfitting, an early stopping with 4, 5, 7, 10 *patience*, and *L2 regularization* for the dense layer, were explored.

Due to the relatively small number of records available in the dataset, i.e. 58, which may increase the variance of the model performance and hence negatively affect the model stability, we built ensembles of 10 and 20 GRU models whose predicted probabilities were averaged. After splitting the dataset into a test set and a non-test set, the GRU models in each ensemble were obtained by repeatedly further splitting the non-test set into validation and train set, for 10 and 20 times, and training the models in each case. The ensemble of 10 and 20 models was then evaluated on the test set. Data splitting was stratified, and the following proportions were used for the test, validation, and training set, respectively: (0.3, 0.3, 0.4), (0.25, 0.3, 0.45), (0.25, 0.25, 0.5).

For this binary classification problem with a moderately imbalanced dataset, the primary performance in evaluating the models was the ROC Area Under Curve, denoted *AUC*. We utilised the *Youden statistic* maximisation method in a ROC analysis procedure [18] for estimating the optimal probability threshold on the non-test (i.e. training and validation) set of records, in order to apply this threshold on the test set to predict the high and low synchrony classes. With this optimal threshold we computed *accuracy, precision*, *sensitivity*, *specificity*, and *f1* performances.

Moreover, for each model, we computed the *Cohen's kappa statistic* and *MCC* (*Matthews correlation coefficient*) whose positive values, when sufficiently far from 0, suggest the existence of predictive pattern in the data that is captured by models. Model predictiveness was established also by running a one-side T-test, inferring statistically that the model's AUC is significantly larger than 0.5 which corresponds to a random prediction model.

Such evaluations are useful also when working with a relatively small number of records, which usually increases the range of variation of models' performance at the point of overlapping with the performance range of a random prediction. With the same rationale in mind, we conducted a Monte Carlo validation (MCV) based on 30 experiments, each of which consisting of: (a) a test / non-test data set split; and (b) building the ensemble model as explained above in this subsection, and then evaluating it on the test set using the performances mentioned above.

Building a GRU ensemble model especially on a large data volume extracted from videos is a computationally expensive procedure (even if the number of records is relatively small as in our approach). Moreover, an MCV multiplies this computational cost by the number of experiments (i.e. 30). However, this is beneficial in our case to reliably assess the model prediction performances and stability, given the relatively small number of records at our disposal in this study (i.e. 58).

2.3 Software and hardware

The data analysis was conducted using Python, with libraries Numpy, Pandas, TensorFlow, Keras, Sklearn and Seaborn. Videos were initially processed with OpenPose library to detect the body, hand, facial, and foot keypoints coordinates.

The hardware consisted of 3 Linux servers with Xeon 6-cores processors and 96GB RAM each, for data exploration and pre-processing, and for code prototyping, and 2 Linux servers with Intel 9 10-cores and AMD Ryzen 16-cores with 128GB RAM each, and Titan RTX 24GB and 3090 RTX 24GB GPUs, for GRU and ensemble model training and MCV intensive computation procedures for building and assessing the models' performances and stability.

3 Results and discussion

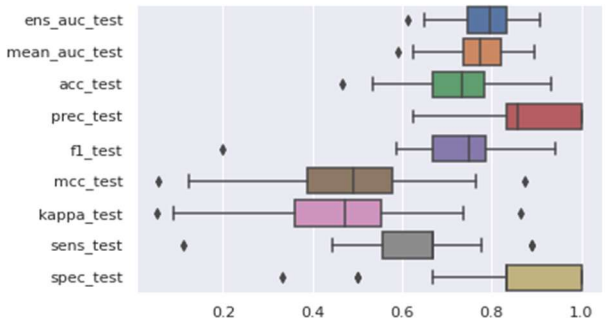
The results in the Monte Carlo validation (MCV) illustrated in Fig. 3, reveal the following aspects:

a) The mean AUC of 0.79 of the GRU ensemble models (ens_auc_test) shows a good prediction level for the relatively small number of records in the dataset.

b) The ROC analysis estimating optimal probability thresholds for classification by maximising the Youden statistic [18], led to good levels of mean accuracy (acc_test) 0.72, mean precision (prec_test) 0.87, mean f1 performance (f1_test) 0.72, as well as positive, far from 0, mean Mathews correlation coefficient (mcc_test) 0.48 and mean Kappa coefficient (kappa_test) 0.44.

c) Given the mean precision (prec_test), mean sensitivity (sens_test), and mean specificity (spec_test) levels achieved by the models, we can infer that 87% of mother-infant dyads predicted as being in the high synchrony class, were predicted correctly by the ensemble models, and that these ensemble models detected 64% of the high synchrony cases; More importantly, these ensemble models detected also 83% of mother-infant dyads with low synchrony. This suggests our models' capability for automatically flagging such cases that may be clinically relevant for further investigation and potential intervention.

The performance values in Fig. 3 are means computed in the Monte Carlo validation on 30 test sets randomly sampled from the dataset (more precisely, via random training, validation, test stratified splits). Due to the relatively small number of records and the data splitting required for building and evaluating the models, which make the training and test sets even smaller, the model stability has some expected limitations as suggested by the various performance boxplots illustrated in Fig. 3 and by the AUC performance histogram depicted in Fig. 4, both of which showing a significant variation of such performances across the Monte Carlo validation procedure.



ens_auc_test	0.79
mean_auc_test	0.77
acc_test	0.72
prec_test	0.87
f1_test	0.72
mcc_test	0.48
kappa_test	0.44
sens_test	0.64
spec_test	0.83

Fig. 3. Left: Boxplots of ensemble model performances on 30 test sets in Monte Carlo validation. Right: mean performances in Monte Carlo validation.

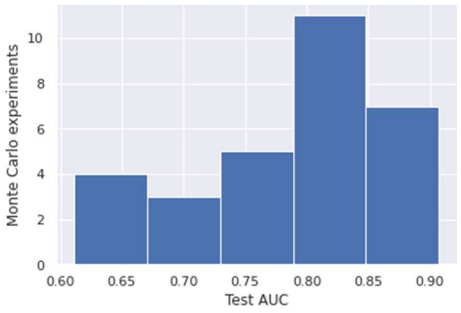


Fig. 4. Histogram of AUC performances on 30 test sets in Monte Carlo validation.

We also conducted a T-Test based on the AUC results obtained in the 30 experiments in the Monte Carlo validation, which led to establishing, with a significant p-value < 9.82×10^{-19}, the alternative hypothesis that mean AUC > 0.5. This proves also statistically that the models proposed in this approach predict better than chance.

4 Conclusion and future research directions

To our knowledge, this work represents the first machine learning based approach in literature, predicting the categorical level of dyadic synchrony – high versus low, in

mother-baby interactions captured in a dataset of videos. We processed the videos with OpenPose library for extracting coordinates from the mother and baby body movements, expected to inform mother-baby dyadic synchrony. Using the dataset of extracted coordinates, this work proposed a novel and substantially high-performing prediction modelling approach, by developing GRU models and ensembles of such models, which were studied in terms of exploring various model architectures, and of assessing prediction performances and model stability with a Monte Carlo Validation procedure.

The GRU ensemble models showed a good prediction capability on test samples, including a mean AUC of 0.79, a mean accuracy of 0.72, a mean precision of 0.87, a mean sensitivity of 0.64, a mean f1 performance of 0.72, and a mean specificity of 0.83. In particular the latter performance represents an 83% detection rate of the mother-baby dyads with low synchrony, suggesting these models' very good capability for automatically flagging such cases that may be clinically relevant for further investigation and potential intervention.

Future research directions to further develop the current study concern: (a) Extending the analysis to a superset of the current dataset, comprising additional videos not available in the present analysis, and incorporating further derived variables exploiting correlations similar to those illustrated in Fig 1; (b) Expanding the machine learning prediction modelling methodology including the application of autoencoders [13, 14] for alternative feature extraction and representation, and of transfer learning [17] based on other similar datasets, as further enhancements of the approach proposed in this study; (c) Developing explanatory models for getting insights of the prediction, and for performance comparison with the black-box models presented in this study; (d) Expanding and evaluating the generalisability of this methodology by employing alternative video based data capturing the interaction between parents and children in other various joint activities.

Acknowledgments: This work was supported by Goldsmiths, University of London, and Global Parenting Initiative (Funded by The LEGO Foundation).

References

1. R. Winston, R. Chicot. The importance of early bonding on the long-term mental health and resilience of children, London Journal of Primary Care, 8:1, 12-14, 2016
2. R. Feldman. The relational basis of adolescent adjustment: Trajectories of mother–child interactive behaviors from infancy to adolescence shape adolescents' adaptation. Attachment & Human Development, 12(1-2), 2010
3. E. Delaherche, M. Chetouani, et al.. Interpersonal synchrony: A survey of evaluation methods across disciplines. IEEE Trans. on Affective Computing, 3(3), 2012
4. Merriam-Webster. (n.d.). Synchrony. In Merriam-Webster.com dictionary.
5. S. Wiltermuth, C. Heath. Synchrony and cooperation. Psychological science, 20(1), 2009
6. N. Watkins, I. Nwogu. Computational Social Dynamics: Analyzing the Face-level Interactions in a Group. arXiv preprint arXiv:1807.06124., 2018
7. R. Chellali, Z.Li, Predicting Arm Movements A Multi-Variate LSTM Based Approach for Human-Robot Hand Clapping Games, Proceedings of 27th IEEE International Symposium on Robot and Human Interactive Communication, 2018

8. K. Cho; B. van Merrienboer, et al.. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014
9. C. Leclère, M. Avril, S. Viaux-Savelon, N. Bodeau, C. Achard, S. Missonnier, et al. Interaction and behaviour imaging: a novel method to measure mother–infant interaction using video 3D reconstruction. Translational Psychiatry, 6(5), 2016
10. A. Guedeney, S. Matthey, K. Puura. Social withdrawal behavior in infancy: a history of the concept and a review of published studies using the Alarm Distress baby scale. Infant Mental Health Journal, 34(6), 516-531, 2013
11. M. N. Noor, A. S. Yahaya, N. A. Ramli, A. M. M. Al Bakri. Filling missing data using interpolation methods: study on the effect of fitting distribution, Key Engineering Materials Volumes 594-595, 2013
12. R. Dey, F. M. Salem. Gate-variants of Gated Recurrent Unit (GRU) neural networks. Proceedings of IEEE 60th International Midwest Symposium on Circuits and Systems, 2017
13. C. Aggarwal. Neural networks and deep learning: A textbook. Springer, 2018.
14. I. Goodfellow, Y. Bengio, A. Courville. Deep Learning, MIT Press, 2016
15. C. Bishop. Pattern Recognition and Machine Learning, Springer, 2006
16. T. Hastie , R. Tibshirani , J. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 2009
17. A. Geron, Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, O'Reilly, 2019
18. I. Unal. Defining an Optimal Cut-Point Value in ROC Analysis: An Alternative Approach. J. Computational and Mathematical Methods in Medicine, Vol 2017, 2017
19. Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, Y. Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields, J. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 43, 2021
20. K. Rutkowska. Automated measurement of nonverbal synchrony in infant-mother interaction using machine learning, MSc dissertation, supervisors D. Stamate, C. Addyman, Data Science & Soft Computing Lab and Computing Department, Goldsmiths College, University of London, 2020