Ensembles of bidirectional LSTM and GRU neural nets for predicting mother-infant synchrony in videos

Daniel Stamate[1,2,3], Pradyumna Davuloori[1*], Doina Logofatu[1,4], Evelyne Mercure[5], Caspar Addyman[6], Mark Tomlinson[6]

1. Data Science & Soft Computing Lab, London
2. Department of Computing, Goldsmiths, University of London, UK
3. School of Health Sciences, The University of Manchester, UK
4. Faculty of Computer Science and Engineering, Frankfurt University of Applied Sciences, Germany
5. Department of Psychology, Goldsmiths, University of London, UK
6. Department of Global Health, Stellenbosch University, South Africa

(* joint first-author)

Abstract. The importance of positive, healthy and reciprocal interactions between mother and infant cannot be understated as it leaves a lasting impact on the rest of the infant's life. One way to identify a positive interaction between two people is the amount of nonverbal synchrony - or spontaneous coordination of bodily movements, present in the interaction. This work proposes a neural network and ensemble learning based approach to automatically labelling a mother-infant dyad interaction as high versus low by predicting the level of synchrony of the interaction. Bidirectional Long Short-Term Memory (BiLSTM) and Bidirectional Gated Recurrent Unit (BiGRU) models were trained and evaluated on a dataset consisting of 25 key body position coordinates of mother and infant extracted with an AI specialised tool called OpenPose, from 58 different videos. Ensembles of 30 such bidirectional recurrent neural network base models were built and then post-processed via ROC analysis, to improve prediction stability and performance, both of which assessed in a Monte Carlo validation procedure of 30 iterations. The prediction performances on the unseen test samples for the ensembles of BiLSTM and ensembles of BiGRU models include a mean AUC of 0.781 and 0.796, a mean precision of 0.857 and 0.899, and a mean specificity of 0.817 and 0.872, respectively. In particular our models predict higher probability scores for the high synchrony class versus the low synchrony class in 80% of cases. Moreover the achieved high precision level indicates that 90% of mother-infant dyads predicted to be in the high synchrony class are predicted correctly, and the high specificity level indicates a detection rate of the mother-infant dyads with low synchrony in 87% of cases, suggesting these models' high capability for automatically flagging cases that may be clinically relevant for further investigation and potential intervention.

Keywords: Mother-infant synchrony detection, Recurrent neural networks, Bidirectional LSTM, Bidirectional GRU, Ensemble learning, Model post-processing optimisation, Monte Carlo validation, Video classification

1 Introduction

The relationship with their mother is arguably the most important one in a person's life. This relationship and the strength of it in infancy can shape the course of the person's social, emotional and mental wellbeing for the rest of their life. The early stages of life are paramount for babies' brain and emotional development, and the quality of interaction between mother and infant is critical in that period. If infants are denied attention and a positive interaction, they can struggle in later life with forming relationships, education and functioning in society [1, 10]. The strong bond and the early positive interactions between mother and infant can shape the social development of the latter [2]. On the other hand, infants who were neglected from the early stages of development face further social development difficulties [2, 21]. Moreover, authors of [22] found that parent-child closeness and affection are good predictors of adolescent mental health and self-worth.

One method to assess the quality of mother-infant interactions is nonverbal synchrony which is the spontaneous coordination of bodily movements. Synchrony can be a vital indicator of a positive, reciprocal mother-infant interaction and it also indicates a healthy relationship with familiarity between mother and infants, leading to positive developmental outcomes for the infant [23, 9]. 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 infant. Developing new methods for finding synchrony patterns can help to automate the process of assessing the mother-infant interaction quality.

One of the problems of interest in this context is the expert's assessment of the synchrony between mother and infant in videos capturing this interaction. 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 infant. This would constitute a useful tool supporting specialists in an early intervention in problematic mother-infant 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, 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-infant dyads, based on a dataset comprising 58 records with body part coordinates extracted from 58 videos capturing the interaction of these dyads. The approach proposed in this paper is based on Bidirectional Long Short-Term Memory and Bidirectional Gated Recurrent Unit neural network models [8, 13, 14], denoted BiLSTM and BiGRU, which are used as baseline models to build ensembles models, denoted BiLSTMens and BiGRUens, that enhance the base models' prediction performance and stability.

The rest of the paper is organised as follows: Section 2 provides a further discussion on synchrony related work. Section 3 introduces our proposed prediction modelling methodology, including data pre-processing, prediction models training and evaluation, and Monte Carlo validation. Section 4 presents and discusses our results, and Section 5 concludes the paper and indicates future research directions.

2 Related work

Prior research has defined synchrony as the coordination of movements, patterns, rhythm and timing between two people in a well-established or budding relationship. It is known to aid in building rapport and understanding. Synchrony can also manifest itself as people imitating each other's speech patterns, through a phenomenon termed speech convergence (or linguistic convergence) [24]. [3] by Delaherche et al. offers valuable additional details about synchrony that merits consideration. The authors propose the following formal definition of synchrony: "Synchrony is the dynamic and reciprocal adaptation of the temporal structure of behaviours between interactive partners". A distinction is also made between mirroring and synchrony, stating that mirroring is the coordination of actions or behaviours, while synchrony is the coordination of behaviours at the same time. Mirroring and synchrony are interrelated, but not equal, with synchrony being more dynamic and fluid in time. With time being a pivotal factor for synchrony, the authors argue that there is but a limited window of time for a person to produce behaviour matching that of their partner, thereby achieving synchrony. Delaherche et al. [3] also delve into the topic of synchrony in mother-infant interactions. First, the authors state that a strong sense of synchrony with their mother is essential in early infancy as the infant uses these moments of connection with their mother to build confidence in their ability to interact with others. Synchrony with the mother also builds a sense of secure attachment in the infant and helps them learn languages.

Nguyen et al. [25] discuss neural synchrony between mother and infant. The authors define neural synchrony as "the temporal coordination of concurrent rhythmic brain activities between individuals". Related research stated by the authors showed neural synchronisation of the left inferior frontal cortex during conversations, measured through a technology named hyperscanning. The main focus of [25] was to understand neural synchrony between mother and infant during conversations and to study the impact of factors like turn-taking and conversation topics on said synchrony. Wavelet Coherence was used to calculate synchrony. The work found mother and child to have synchronised brain activity during conversations, with turn-taking proving a strong

influence on increasing synchrony. This could be because, when the mother engages in proper turn-taking, she is listening intently and is allowing the child enough room to communicate and express themselves, thereby fostering good communication and synchrony. This study is enlightening on neural synchrony during mother-child conversations.

Egmose et al. in [26], investigated the impact of bodily movements including upper body, arms and head on synchrony between mothers and infants. The study included infants aged 4 months and 13 months old. An eight-camera optoelectronic motion registration system was used to capture the bodily movements of mothers and infants. The quality of interaction between the mother and infant was coded using the Coding Interactive Behaviour scale [27]. For data analysis purposes, Matlab and SPSS were used to process the data. The study found a few valuable insights. Firstly, it was found that when the mother coordinated her head movements with the infant's, the interaction had a higher rating for synchrony. Secondly, the correlation between bodily movements and synchrony ratings was higher for 4-month-old infants compared to 13-month-old infants.

3 Prediction modelling methodology

3.1 Data pre-processing

This work was based on a sample of 60 videos from the SPEAKNSIGN dataset [28, 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 to 14. Out of the 60 videos available, 2 videos were discarded as they were lacking in clarity. The distribution of the synchrony scores of the remaining 58 videos is illustrated in Fig.1.

The OpenPose library [19] was used to extract a 5D array including information such as the video number, camera number, frame number, number of people present in a video, pairs of x and y coordinates and their confidence intervals of 25 body keypoints. Fig. 2 illustrates a single frame of a mother-infant dyad interaction video, with body part keypoints extracted with OpenPose.

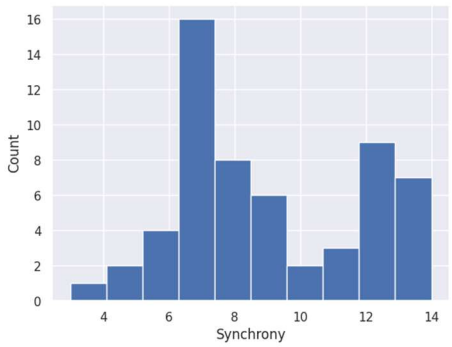


Fig. 1. Distribution of the mother-infant dyad synchrony scores.

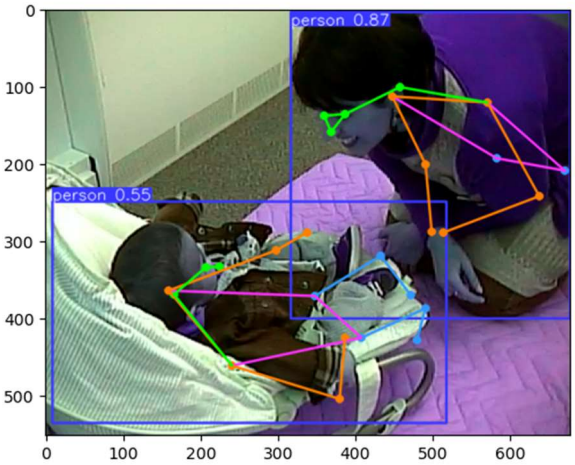


Fig. 2. Body part keypoints extracted with OpenPose from a single frame of an interaction video.

Out of the 3 available cameras from which data was available for each video, only data from camera one was retained. Moreover, only data belonging to mother and infant were preserved in two separate arrays. To keep the shape of the data consistent for each video in these two arrays for mother and infant, a total of 9000 frames of data, starting from frame 500 and ending with frame 9,500 were retained and the rest of the frames were discarded. Moreover, the x and y coordinates of body keypoints were preserved while the confidence intervals for each coordinate, which were also estimated by OpenPose, were removed as they were not necessary for this study. 3D arrays were finally obtained for each, mother and infant, comprising the record number corresponding to each video, the frame number, and the sum aggregation of the x and y coordinates for body keypoints. Data cleaning included also the treatment of missing values which were imputed via linear interpolation [11], and the detection and removal of outliers using criteria based on the range of 0.025 or 0.975 quantiles [15, 16]. Data was normalised using the L2 norm.

Records corresponding to videos 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.

3.2 Classification and model post-processing via Receiver Operating Characteristic (ROC) analysis and optimisation

A bidirectional LSTM (BiLSTM) layer and bidirectional GRU (BiGRU) layer is a recurrent neural network layer that learns bidirectional long-term dependencies or patterns present in the input sequence data. They are based on Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) layers [12, 13, 14] whose computations are described by the equations below, which calculate the outputs y(t) from inputs x(t) as follows:

LSTM layer equations:

$$i(t) = \sigma(W_{xi}{}^{T}x(t) + W_{hi}{}^{T}h(t-1) + b_i)$$
$$f(t) = \sigma(W_{xf}{}^{T}x(t) + W_{hf}{}^{T}h(t-1) + b_f)$$
$$o(t) = \sigma(W_{xo}{}^{T}x(t) + W_{ho}{}^{T}h(t-1) + b_o)$$
$$g(t) = tanh(W_{xg}{}^{T}x(t) + W_{hg}{}^{T}h(t-1) + b_g)$$
$$c(t) = f(t) \otimes c(t-1) + i(t) \otimes g(t)$$
$$y(t) = h(t) = o(t) \otimes tanh(c(t))$$

GRU layer equations:

$$z(t) = \sigma(W_{xz}{}^{T}x(t) + W_{hz}{}^{T}h(t-1) + b_z)$$
$$r(t) = \sigma(W_{xr}{}^{T}x(t) + W_{hr}{}^{T}h(t-1) + b_r)$$
$$g(t) = tanh(W_{xg}{}^{T}x(t) + W_{hg}{}^{T}(r(t) \otimes h(t-1)) + b_g)$$
$$y(t) = h(t) = z(t) \otimes h(t\text{-}1) + (1\text{-}z(t)) \otimes g(t)$$

where σ is the sigmoid function, $W_{..}$ are the weight matrices, and $b_{.}$ are the bias terms.

For illustration purposes, an LSTM cell which is a more complex version of a GRU cell, is depicted in Fig. 3 below [17]:

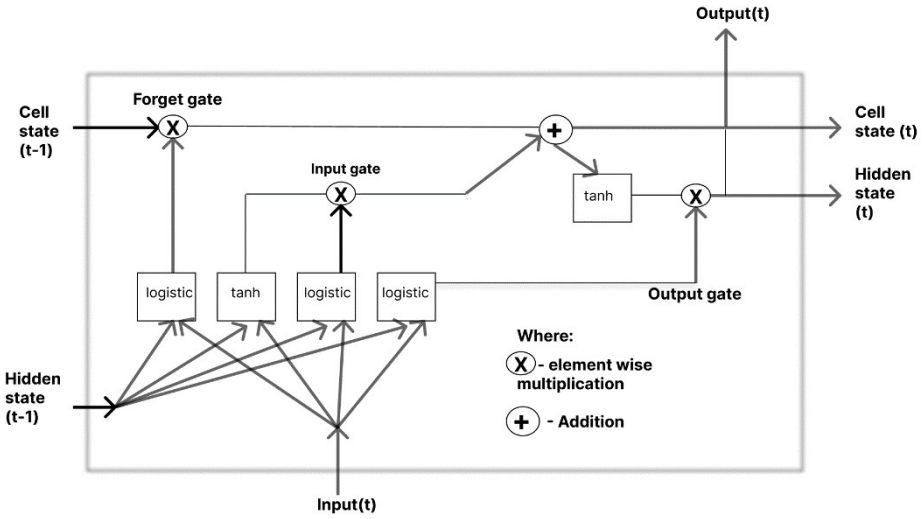


Fig. 3. An LSTM cell architecture where Input(t), Output(t), Cell state(t), and Hidden state(t) are the x(t), y(t), c(t), and h(t) quantities, respectively, appearing in the LSTM layer equations above.

Bidirectional LSTM (BiLSTM) and Bidirectional GRU (BiGRU) networks follow the general structure of a bidirectional recurrent neural network illustrated in Fig. 4 [13, 14].

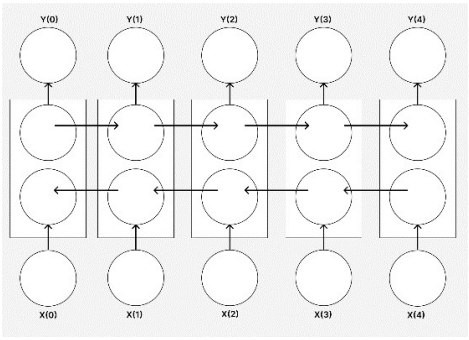


Fig. 4. A bidirectional recurrent neural network architecture, in which information from an input sequence data x is fed and processed in both directions as indicated by the horizontal arrows from left to right and from right to left, in order to compute the network's outputs y.

The two types of recurrent neural network architectures employed in this work comprised, in this order, an input layer, 3 BiLSTM layers (3 BiGRU layers, respectively) with a number of units between 300 and 350 (the same for all 3 layers), and 1 hidden dense layer with a number of units between 50 and 70 and with *elu*, *gelu, mish,* or *swish* activations. The networks had an output layer with 1 unit with *sigmoid* activation for binary classification. As loss functions we employed *Binary focal crossentropy* with default parameter values. The *adam* optimiser with default parameter values was used, together with the exponential learning rate scheduling starting with the base learning rate of 0.003 which was decreased by a factor of 0.85 per epoch. To prevent overfitting, early stopping with 4 steps of *patience* was used on the validation set, and a *Gaussian noise layer* with standard deviation of 0, 0.01 and 0.02 was inserted after the input layer. To tune the hyperparameters for the respective value ranges mentioned above, we performed a random search with 10 iterations. Optimal hyperparameter values in the two architectures introduced above, are not constant and depended on the multiple non-test datasets used in the Monte Carlo procedure introduced in the next subsection. But some typical values were: (a) for the BiLSTM based architecture: 310 units for the BiLSTM layers, 62 units and *elu* activation for the dense layer, and a standard deviation of 0.1 for the Gaussian noise layer; and (b) for the BiGRU based architecture: 300 units for the BiGRU layers, 60 units and *elu* activation for the dense layer, and a standard deviation of 0 for the Gaussian noise layer.

Due to the relatively reduced 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 30 BiLSTM models, and ensembles of 30 BiGRU models. We denote these ensemble models by BiLSTMens and BiGRUens, respectively. More precisely, BiLSTMens (BiGRUens) was built as follows: after splitting the dataset into test set and a non-test set, the base BiLSTM (BiGRU) models in each ensemble were obtained by first computing the optimal hyperparameter values using the non-test set as explained in the paragraph above, and then by further randomly splitting the non-test set into validation and train sets, 30 times, and by training 30 BiLSTM (BiGRU) models with the determined optimal hyperparameter values. The predicted probabilities were averaged among the 30 BiLSTM (BiGRU) models. Regarding data splitting, the following proportions were used: 25% test set, 75% non-test set. The non-test set was further split, 30 times, in 67% train and 33% validation.

For the BiLSTMens (BiGRUens) model post-processing, we utilised the *Youden index* maximisation method in a ROC analysis procedure [18, 29] for estimating the optimal probability threshold using the model's ROC curve calculated on the non-test (i.e. training plus validation) data set of records. For each generic probability threshold $t \in [0,1]$ of a model that discriminates 2 classes such as high synchrony versus low synchrony (t is 0.5 by default), there is a point $P(t)$ on the model's ROC curve computed on the non-test set, whose x and y coordinates are $1\text{-}Specificity(t)$, and $Sensitivity(t)$, respectively [29]. In this case, for each probability threshold $t \in [0,1]$, $Sensitivity(t)$ and $Specificity(t)$ indicate the model's rates of detection of the high synchrony and low synchrony classes, respectively, on the non-test dataset. Similarly to [29], the probability threshold based *Youden index* $Y(t)$ is defined as:

$$Y(t) = Sensitivity(t) + Specificity(t) - 1$$

We maximised the distance $D(t)$ from the point $P(t)$ to the main diagonal of the ROC curve ($D(t)$ is computed below and intuitively indicates how far our model is from the random guess model). Hence, we optimised the probability threshold t as follows:

$$D(t) = \sin(\pi/4) \times Y(t); \qquad p = \operatorname{argmax}_{t \in [0,1]} D(t) = \operatorname{argmax}_{t \in [0,1]} Y(t)$$

We employed the optimal probability threshold p and took $t=p$ to obtain the post-processed model, which was used to compute the test *accuracy, precision*, *sensitivity*, *specificity*, and *f1* performances [29]. Moreover, we computed *Cohen's kappa statistics* and *Matthews correlation coefficient MCC,* defined below, whose positive values, when sufficiently far from 0, indicate that the model predicts better than chance. In particular, *kappa* focuses more on the positive class while *MCC* treats classes equally.

$$kappa = 2 \times (TP \times TN - FN \times FP) / ((TP+FP) \times (FP+TN) + (TP+FN) \times (FN+TN))$$
$$MCC = (TP \times TN - FP \times FN) / ((TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN))^{0.5}$$

The model's capability to predict better than chance was investigated also statistically, by running a one-side T-test, in order to prove that the model's general performance for binary classification, called Area Under ROC Curve, denoted simply by AUC, and defined with the conditional probability below, is significantly larger than 0.5 which is the performance of a random guess model. More precisely:

$$AUC = Pr(S(r_1) > S(r_2) \mid r_1 \in H, r_2 \in L)$$

where r_1 and r_2 are arbitrary records from the high synchrony class H, and low synchrony class L, respectively, and S is the score (i.e. probability to belong to the positive/ high synchrony class) outputted by the model for each record.

3.3 **Monte Carlo validation for assessing models' performance and stability**

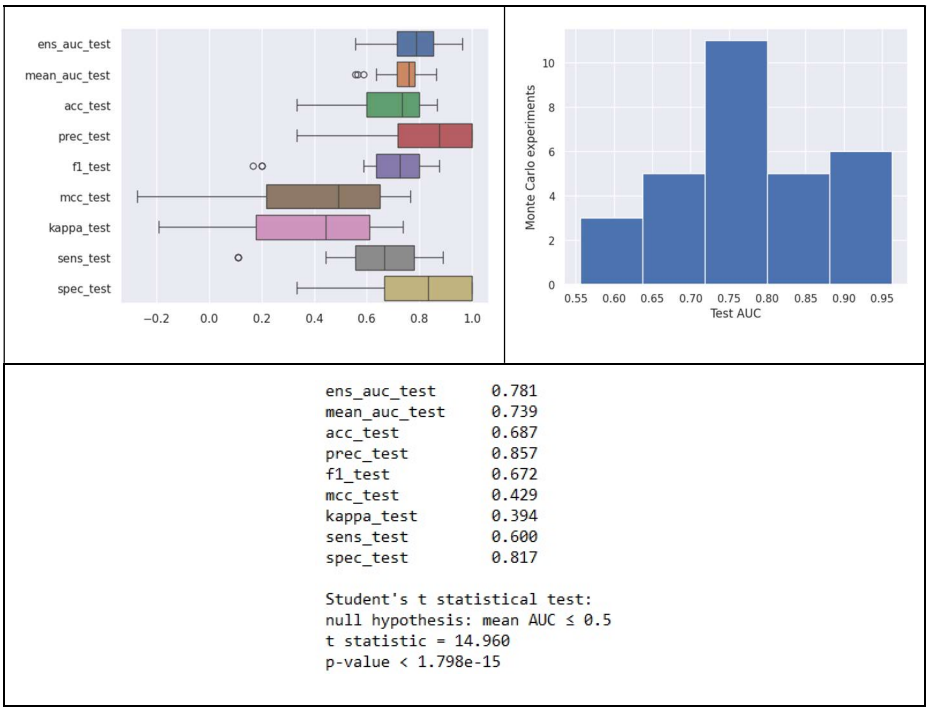
To assess our models' predictive capability, we conducted a Monte Carlo validation based on 30 iterations, which allows to reliably evaluate the models' performance and stability.

3.4 Software and hardware

Videos were initially processed with OpenPose library to detect the body keypoints coordinates. Data preprocessing and prediction modelling [17] were conducted in Python with Numpy, Scipy, Pandas, TensorFlow, Keras, Sklearn and Seaborn libraries, using 5 Linux servers with up to 128GB RAM per machine, and Titan RTX 24GB, RTX 3090 24GB, and RTX 4090 24GB GPUs, for training the BiLSTM and BiGRU base models and building the BiLSTMens and BiGRUens ensemble models, as well as for assessing the models' performances and stability in computationally intensive Monte Carlo validation procedures of 30 iterations.

4 Results and discussion

In this section we present and discuss the results of the analyses conducted following the lines of methodology introduced in Section 3.



ens_auc_test	0.781
mean_auc_test	0.739
acc_test	0.687
prec_test	0.857
f1_test	0.672
mcc_test	0.429
kappa_test	0.394
sens_test	0.600
spec_test	0.817

Student's t statistical test:
null hypothesis: mean AUC ≤ 0.5
t statistic = 14.960
p-value < 1.798e-15

Fig. 5. Top: Boxplots (left) and histogram (right) of performances on the test data of Bidirectional LSTM ensemble models (BiLSTMens) in Monte Carlo validation of 30 iterations. Bottom: Mean performances, and one-tailed Student's t-test proving the alternative hypothesis: mean AUC > 0.5, demonstrating statistically the BiLSTMens models' prediction capability.

The distribution of classification performances of BiLSTMens models, provided as boxplots in Fig. 5 (top left), comprises performances computed in the 30 iterations of Monte Carlo validation. More precisely they are evaluated using the 30 BiLSTMens models, each of which produced in 1 iteration of the Monte Carlo validation procedure, and the corresponding 30 test sets issued from the stratified splitting of the dataset. The performances evaluated are the AUC (ens_auc_test), accuracy (acc_test), precision (prec_test), f1 (f1_test), Matthews correlation coefficient MCC (mcc_test), kappa statistic (kappa_test), sensitivity (sens_test) and specificity (spec_test). Moreover, for comparison, we included also the distribution of the mean AUC of the 30 BiLSTM base models (mean_auc_test) which are the components of one BiLSTMens ensemble model. The means of the above mentioned performances are provided in Fig. 5 (bottom). On the other hand, the distribution of the AUC of the BiLSTMens models is represented additionally as a histogram in Fig. 5 (top right). The mean AUC of 0.781 represents a good capability of the BiLSTMens models to distinguish between mother-infant dyads with low vs. high synchrony. The distribution of AUC illustrated in Fig. 5 (top right) indicates a substantial variation of this performance across the Monte Carlo validation's 30 iterations, which may be explained by the relatively reduced number of records (videos) in the dataset that expectedly increases variance. When we conducted a one-tailed Student's t-test for the null hypothesis: mean AUC ≤ 0.5 (Fig. 5, bottom), we obtained the significant p-value $< 1.798e\text{-}15$ proving statistically that the BiLSTMens models predict better than random guess models, which is also illustrated by the AUC performance distributions in Fig. 5 (top right), and by the kappa_test and mcc_test performances in Fig. 5 (bottom).

exp	count	mean	std	min	25%	50%	75%	max
0	30.0	0.741975	0.055007	0.611111	0.722222	0.750000	0.777778	0.814815
1	30.0	0.756790	0.049531	0.574074	0.726852	0.759259	0.796296	0.814815
2	30.0	0.688272	0.034759	0.611111	0.666667	0.703704	0.722222	0.722222
3	30.0	0.640741	0.069902	0.518519	0.592593	0.648148	0.699074	0.740741
4	30.0	0.782099	0.100691	0.370370	0.759259	0.796296	0.833333	0.888889
5	30.0	0.812346	0.110442	0.425926	0.777778	0.824074	0.884259	0.981481
6	30.0	0.762346	0.080071	0.500000	0.740741	0.777778	0.810185	0.870370
7	30.0	0.635802	0.049912	0.462963	0.615741	0.648148	0.666667	0.685185
8	30.0	0.854321	0.040024	0.740741	0.837963	0.861111	0.870370	0.907407
9	30.0	0.782716	0.076111	0.629630	0.726852	0.777778	0.851852	0.907407
10	30.0	0.766358	0.050987	0.592593	0.759259	0.796296	0.796296	0.814815
11	30.0	0.863889	0.100114	0.370370	0.856481	0.870370	0.907407	0.944444
12	30.0	0.676543	0.067164	0.537037	0.634259	0.675926	0.722222	0.796296

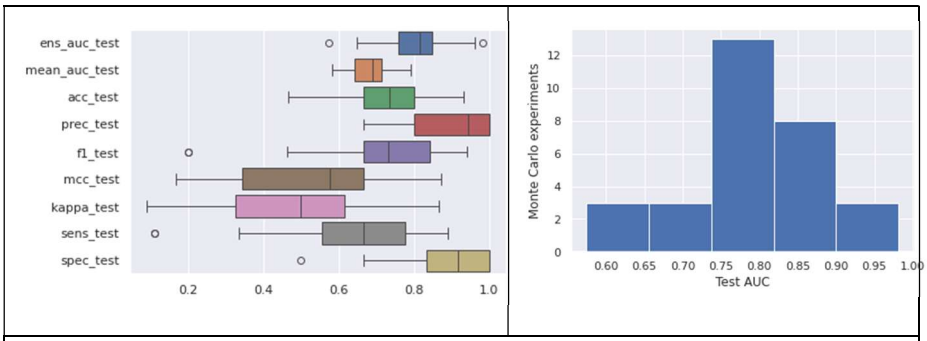
Fig. 6. Sample of the Monte Carlo validation iterations: aggregation in 7 basic statistics, including mean, std, min, max and 25%, 50%, 75% quartiles of test AUC of the 30 BiLSTM base models grouped in one row, which are the components of one BiLSTMens model.

Fig. 6 illustrates a sample of the 30 Monte Carlo iterations, each row representing the aggregation in 7 basic statistics including mean, standard deviation (std), min, max and the 25%, 50% and 75% quartiles, of test AUC of the 30 BiLSTM base models

which are the components of each ensemble BiLSTMens model produced in the Monte Carlo procedure. Note the substantial std of AUC across the 30 base models, especially in rows such as 11 with std around 0.1, (see Fig. 6), and the min AUC around 0.37 which corresponds to a base model that is clearly counter-performing. On all rows of the table in Fig. 6, all the quartiles, max and mean values of AUC indicate figures substantially above 0.5 (which is the expected AUC of a random prediction model), but there are multiple rows with min AUC values under 0.5, which clearly demonstrates that a prediction modelling approach *based on a single base model* does not constitute a viable solution in this case. These aspects justify our choice to propose of a prediction modelling approach relying on *ensembles* of base models despite the higher volume of computation for training the 30 base models forming an ensemble. As Fig. 5 illustrates with the histogram of AUC of ensemble models, all AUC values are above 0.5 hence all ensemble models predict better than chance. Moreover, the ensemble models have a synergistic effect in this case, as their mean AUC of 0.781 is larger than the mean AUC of all the base models, which is 0.739 (see, in Fig. 5 bottom, ens_auc_test and mean_auc_test, corresponding to the ensemble models and base models, respectively).

The results from the conducted analysis regarding the BiGRU base models and the BiGRUens ensemble models, are summarised and presented in Fig. 7 below, similarly to the presentation of BiLSTM and BiLSTMens models in Fig. 5. The explanations regarding the results in Fig. 7 are similar to those provided for the results in Fig. 5. The main difference between the BiLSTMens and BiGRUens models is that the latter achieve better performances in the Monte Carlo validation procedure on this dataset. Indeed, the BiGRUens models showed a better prediction capability on the test sets compared with BiLSTMens, including a mean AUC of 0.796 vs 0.781, a mean accuracy of 0.729 vs 0.687, a mean precision of 0.899 vs 0.857, a mean f1 of 0.715 vs 0.672, a mean Matthews correlation coefficient MCC of 0.518 vs 0.429 (indicating models' performances are substantially different from just random prediction, from both classes' perspective), a mean kappa statistic of 0.479 vs 0.394 (indicating again that models' performances are substantially different from just random prediction), a mean sensitivity of 0.633 vs 0.6, and a mean specificity of 0.872 vs 0.817. In particular, the high AUC result of 0.796 entails that our BiGRUens models predict in average higher probability scores for the high synchrony class versus the low synchrony class in about 80% of cases. Moreover, the achieved high precision level indicates that 90% of mother-infant dyads predicted to be in the high synchrony class are predicted correctly. On the other hand, the high specificity level of 0.872 of these models indicates a detection rate of the mother-infant dyads with low synchrony in about 87% of cases, suggesting these models' high capability for automatically flagging cases that may be clinically relevant for further investigation and potential intervention.

As a further remark on the BiGRU based models, the ensemble models have a synergistic effect in this case too, as their mean AUC of 0.796 is larger than the mean AUC of all the base models, of 0.685 (see, in Fig. 7 bottom, ens_auc_test and mean_auc_test, corresponding to the ensemble models and base models, respectively).



ens_auc_test	0.796
mean_auc_test	0.685
acc_test	0.729
prec_test	0.899
f1_test	0.715
mcc_test	0.518
kappa_test	0.479
sens_test	0.633
spec_test	0.872

Student's t statistical test:
null hypothesis: mean AUC ≤ 0.5
t statistic = 17.863
p-value < 1.714e-17

Fig. 7. Top: Boxplots (left) and histogram (right) of performances on the test data of Bidirectional GRU ensemble models (BiGRUens) in Monte Carlo validation of 30 iterations. Bottom: Mean performances, and one-tailed Student's t-test proving the alternative hypothesis: mean AUC > 0.5, demonstrating statistically the BiGRUens models' prediction capability.

5 Conclusion and future research directions

In this paper we proposed a bidirectional LSTM and bidirectional GRU recurrent neural network approach to predicting mother-infant synchrony classes in 58 videos capturing the interaction between mother and their babies. To improve the level of prediction performance and stability, the base models have been integrated in ensemble models composed of 30 base models each, and we showed that this solution was an effective mitigation to the problem of substantial variation of the AUC performance of the base models due to the relatively reduced number of records (videos) in the dataset (i.e. 58). This research extends on previous work in [28] which proposed GRU ensemble based models for predicting high vs low mother-baby synchrony. In the new approach presented in this study, we explored bidirectional LSTM and bidirectional GRU base models and ensemble models, taking advantage of the bidirectional structure of these models which is suitable in this framework for analysing data extracted from the frames of the videos capturing the mother-infant interaction. Indeed, we detected better the patterns of synchrony in the mother-infant dyads by exploring, in both directions (i.e. in the normal sequence and reverse sequence of frames) the data extracted from the video. In the new approach presented here we improved the AUC, accuracy, precision, Matthews correlation coefficient, kappa statistics, and specificity,

the latter of which ensuring in particular a better rate of detection of the mother-infant dyads with low synchrony from 83% in [28] to 87% in this study, suggesting a higher models' capability for automatically flagging cases that may be clinically relevant for further investigation and potential intervention.

Future research directions include: expanding the machine learning prediction modelling methodology with classes of autoencoders for (a) alternative, more effective feature extraction and representation, and for (b) researching an effective approach to synthetic data generation and data augmentation, given the relatively reduced number of videos (58) used in this research. Moreover, the methodology proposed here is to be further extended to the study of predicting the interaction between parents and children, in other activities such as book reading which makes the object of related research work we develop.

Acknowledgments: This work was supported by the Global Parenting Initiative (Funded by The LEGO Foundation), by University of London - Goldsmiths College, and by the University of Manchester. Data collection was supported by an ESRC Future Research Leader Fellowship to EM (ES/K001329/1). The authors would like to thank Harriet Bowden-Howl & Rudi Dallos for their expert rating of dyadic synchrony.

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". Proc. of IEEE 60th International Midwest Symposium on Circuits and Systems, 2017.

13. C. Aggarwal. “Neural networks and deep learning”, 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, Data Science & Soft Computing Lab and Computing Department, Goldsmiths College, University of London, 2020.
21. L. Steinberg. “We Know Some Things: Parent-Adolescent Relationships in Retrospect and Prospect,” J. Res. Adolesc., vol. 11, no. 1, pp. 1–19, March 2001.
22. T. A. McAdams et al., “Associations between the parent-child relationship and adolescent self-worth: a genetically informed study of twin parents and their adolescent children”, J. Child Psychol. Psychiatry, vol. 58, no. 1, pp. 46–54, 2017.
23. C. Leclère et al.. “Why Synchrony Matters during Mother-Child Interactions: A Systematic Review”, PLoS ONE, vol. 9, no. 12, Dec. 2014.
24. L. Wade. “What makes us subconsciously mimic the accents of others in conversation”, The Conversation, May 20, 2022.
25. T. Nguyen, H. Schleihauf, E. Kayhan, D. Matthes, P. Vrtička, and S. Hoehl. “Neural synchrony in mother–child conversation: Exploring the role of conversation patterns,” Soc. Cogn. Affect. Neurosci., vol. 16, no. 1–2, Jan. 2021.
26. I. Egmose et al.. “Relations between Automatically Extracted Motion Features and the Quality of Mother-Infant Interactions at 4 and 13 Months”, Front. Psychol., vol. 8, Dec. 2017.
27. A. C. Stuart, I. Egmose, J. Smith-Nielsen, S. Reijman, K. I. Wendelboe, and M. S. Væver, “Coding Interactive Behaviour Instrument: Mother-Infant Interaction Quality, Construct Validity, Measurement Invariance, and Postnatal Depression and Anxiety,” J. Child Fam. Stud., vol. 32, no. 6, pp. 1839–1854, Jun. 2023.
28. D. Stamate, R. Haran, K. Rutkowska, S. Davuloori, E. Mercure, C. Addyman and M. Tomlinson, “Predicting High vs Low Mother-Baby Synchrony with GRU-Based Ensemble Models”, Proc. 32nd Int. Conf. on Artificial Neural Networks, ICANN, LNCS vol 14262, Springer, 2023.
29. M. Kuhn, K. Johnson, “Applied Predictive Modeling”, Springer, 2013.