EURASIP Journal on Audio, Speech, and Music Processing

REVIEW Open Access

Automatic detection of attachment style in married couples through conversation analysis

Tuğçe Melike Koçak[1*], Büşra Çilem Dibek[2], Esma Nafiye Polat[1], Nilüfer Kafesçioğlu[3] and Cenk Demiroğlu[1]

Abstract

Analysis of couple interactions using speech processing techniques is an increasingly active multi-disciplinary field that poses challenges such as automatic relationship quality assessment and behavioral coding. Here, we focused on the prediction of individuals' attachment style using interactions of recently married (1–15 months) couples. For low-level acoustic feature extraction, in addition to the frame-based acoustic features such as mel-frequency cepstral coefficients (MFCCs) and pitch, we used the turn-based i-vector features that are the commonly used in speaker verification systems. Sentiments, positive and negative, of the dialog turns were also automatically generated from transcribed text and used as features. Feature and score fusion algorithms were used for low-level acoustic features and text features. Even though score and feature fusion algorithms performed similar, predictions with score fusion were more consistent when couples have known each other for a longer period of time.

Keywords Attachment style, Acoustic features, I-vectors, Couple interaction

1 Introduction

Attachment theory first originating in the 1940s was developed to explain the role of early experiences in children's development [1]. The theory has later proven to be a rich framework to understand the adult close relationships as well [2]. According to the theory, infants' experiences with their primary caregivers especially during stressful times contribute to the development of expectations about one's worthiness and how others respond to one's valid physical and emotional needs. In turn, the history of attachment experiences with primary caregivers develop into patterns about one's needs, emotions, emotion regulation mechanisms, and interpersonal behaviors [1]. Later in life, romantic partners take the place of main attachment figures for adults. As in the infant caregiver dyad, in romantic dyads too, the stressful moments help to classify various attachment styles of individuals (e.g., secure, insecure). Studies have shown attachment styles to be predictive of significant psychological and relational outcomes for adults such as emotional wellbeing [3] and relationship satisfaction [4]. More recently, attachment style and intimate relationship history, as well as strong feelings such as shame, fear, and anxiety, are proposed to impact one's vocal characteristics [5]. We propose that the examination of voice features of romantic relationship partners can prove to be helpful in classifying their attachment styles.

In this study, acoustic features of spoken language [6] as they are relatively less controllable compared to other nonverbal behaviors [7] will be explored as indicators of speakers' attachment styles. Since [2] first proposed the extension of attachment theory to adult romantic relationships, several scales and interview techniques have been used to determine adult attachment styles. Even though scales are relatively easy to administer and score

*Correspondence:
Tuğçe Melike Koçak
tugce.kocak@ozu.edu.tr
[1] Department of Engineering, Ozyegin University, Istanbul, Turkey
[2] Department of Psychology, Bilgi University, Istanbul, Turkey
[3] Department of Psychology, Ozyegin University, Istanbul, Turkey

compared to the labor-intensive data collection and rating of interview transcripts, social desirability bias is in question when self-report measures are used. So, another goal of this study is to test a measurement technique which uses interview data but is less labor intensive to rate.

Behavioral and social signal processing (BSS) is the study of human behavior, particularly in distressed situations, through the analysis of signals such as audio or video. Surveys of social signal processing literature are given in [8] and [9]. In [10], a survey of some of the recent work in the BSS field is done within the context of couple interactions during therapy.

Psychologists that use observational methods in their research often rely on standardized rating systems for assessment of behavioral codes that are relevant to characterizing the domain and problem under consideration (e.g., in couple therapy,blame patterns are typically coded). In the context of couple therapy analysis, interactions are manually annotated by a psychologist which is an expensive and time-consuming process. In [11], manually coded behavior patterns, such as level of blame towards the other spouse, are predicted automatically using vocal interaction analysis.

Earlier study [12] have shown that, in couple therapy, approach style to problem solving is a strong indicator of the relationship quality. In [13], acoustic features extracted during couple interaction are used to predict the outcome of couple therapy, i.e., whether there is any improvement in the couple's relationship. Vocal entrainment, which is the process where the interlocutors naturally and mutually adapt their voices during the interaction, is measured in [14] and then used for affect analysis in married couple interactions.

Lexical features can be used in addition to acoustic features for interaction analysis. For example, spoken language data is used to identify interactional style in [15]. In [16], speech recognition-based lexical features are used in addition to the acoustic features which improved the prediction accuracy of blame level. Similarly, in [17], transcriptions of couple interactions during therapy were found to contain significant information regarding the behavioral codes even when the transcription processes are noisy.

Even though the whole interaction session contains relevant data, there are salient events that are enough to predict the session-level behavioral codes [18]. Multiple instance learning is used in [19] to extract features from the salient events to predict the effective states in couple interactions. In [20], spoken language-based features are used with in a two-state dynamical behavioral model framework to analyze each turn locally before reaching a global decision.

In this study, we focused on predicting the attachment styles of recently married couples using acoustic features. Our work is novel in two major aspects. In the first aspect, this is the first study to analyze vocal interactions, from a speech processing perspective, of distressed married couples (the range of being married ranged between 1 and 15 months). In the second aspect, this is the first study that uses acoustic features to predict adult attachment style according to the questionnaire.

2 Related work

2.1 Adult attachment

Attachment refers to the coherent patterns of behavior that specify the quality of the emotional bond within close relationships [21]. Attachment theory proposes that infants' early interactions with their caregivers generate certain internal working models of self and others in the form of expectations and beliefs that guide one's perceptions, emotions, and behaviors in significant relationships [1]. Individual differences in these internal working models have been linked to different attachment styles [22]. Differences in attachment styles have been conceptualized in line with two dimensions of attachment related anxiety and avoidance [23].

Attachment anxiety is characterized by fear of abandonment and rejection and excessive need for approval in line with the person's negative representation of self as unworthy. In contrast, attachment avoidance is characterized by discomfort with closeness and excessive self-reliance in line with the person's negative representation of others as untrustworthy [24]. Those individuals who hold positive representations of both self and others were referred to as securely attached. Securely attached individuals had a positive sense of self that is worthy of love, care, and attention while also felt safe with depending on others in times of need.

Utilizing the two dimensions of attachment anxiety and avoidance, in interaction with negative or positive views of self and others, Bartholomew [21] proposed four adult attachment styles: secure, preoccupied, dismissing, and fearful. According to this model, preoccupied individuals had a negative view of self but a positive view of others whereas individuals with dismissing style had a positive view of self but a negative view of others. Based on these dimensions, Bartholomew [21] also proposed a fearful attachment style that is characterized by negative views of both self and others. While both groups of preoccupied and fearful tend to have strong dependency needs, the fearful group avoids closeness similar to the dismissing group. Thus, the fearful group tends to have both autonomy and intimacy difficulties [21]. A summary of these classifications can also be found in Table 1.

Table 1 Attachment styles and their descriptions

Attachment Types	Descriptions
Secure	Thinks himself/herself as lovable and others as accepting
Preoccupied	Craves acceptance for himself/herself to gain self worth
Dismissing	Feels negative towards others, has a high self-worth
Fearful	Thinks himself/herself as unlovable and others as untrustworthy

2.2 Measuring attachment

Many studies have been conducted since the 1980s to develop measures of adult attachment and its manifestations in the couple's relationship. One of these measurement methods is questionnaires and the other is interview techniques. Both of these methods have certain limitations in understanding and interpreting attachment with an impact on both research and clinical assessment of attachment. Questionnaires are prone to self-report biases [25]. Furthermore, almost no concurrent validity has been found between questionnaires and interview-based assessments of adult attachment, adding to the construct validity controversy.

In contrast to the questionnaires, interview methods are relatively more difficult, costly, and lengthy to administer and score. They add additional layers to the process, that is, the manual transcription and coding of the interview. There is a training load required for both administering and scoring the interviews. This limitation might decrease the replicability of attachment studies, adding to psychology's current "replicability crisis." A different method as the one suggested in this study, not focusing on the content of the responses but focusing on the vocal characteristics of the responses may make it easier to interpret emotional arousal responses as attachment-related implications. Because studies have shown that emotional arousal reactions are manifestations of attachment security and so that can be uncovered by the composition of conversation [26, 27]. In order to understand and interpret emotional arousal reactions, it is necessary to focus on the way the message is conveyed, that is the sound characteristics of speech, and it is not sufficient to focus solely on the content of the message.

Since the internal working models as another attachment related implication are especially activated by significant others, the impact of the activation of the attachment system may be observed in partners' voice features such as jitter, shimmer, and their derivatives, which are some of the vocal features examined in this study [5]. When we focus only on the content of the narrative, it is difficult to make sense of these internalized working models' representations. Though the defensive inhibition strategy can manage what is said, it cannot manage how it is said [28], thus allowing prosody to provide a better understanding of emotion regulation strategies that individuals use when discussing their past [29]. This is because, as studies indicate, voice parameters are less manageable compared to other types of nonverbal behavior in a conversation [30].

2.3 Studies of vocal characteristics and attachment

As the first examples, mother-infant studies have exemplified the importance of the vocal characteristics in predicting attachment security. Mother-infant studies have shown that vocal synchrony predicts attachment security later in life [31–37]. Kolacz's [38] study on mother-infant communication and attachment has shown that the frequency and acoustic characteristics of babies' crying behaviors can be interpreted as the precursors of their developing neural systems in social interactions while the current period. The acoustic properties of these vocalizations were found to predict the infants' social adjustment in later childhood.

Likewise, studies have also focused on the vocal features of romantic partners in order to shed light on their attachment security. Synchronous vocal activities between romantic partners have been studied as indicators of co-regulation, which is considered to be an important element associated with attachment [37]. In relation to this, Sroufe [35] found that the security of attachment was associated with the emotional tone of adult romantic relationships. For example, in an observational study measuring hostility in romantic relationships, researchers found the signs that partners carry from their attachment stories. The manifestation of couples' synchronized systems is parallel with infant-mother studies showing reciprocal associations of behavioral or physiological systems [37, 39, 40]. And the prosodic, non-verbal, synchronicity in conversation provides important clues regarding the affective characteristics and adaptations of the interacting partners. [41].

3 Corpus description

3.1 Participants and procedures

The data were collected from 103 newlywed heterosexual couples from Turkey. The couples were all in their first marriages and without children. Marriage length of the participants ranged between 1 and 15 months with an average length of 6.06 months (SD = 3.43). Participants' ages ranged between 20 and 48 years with an average of 27.83 (SD = 3.80). On average, participants attended school for 16.36 years with a range between 8 and 26 years.

Participants were recruited through flyers posted at different universities, municipalities, and institutions in a metropolitan city as well as posts on social media and various email groups. Individuals who were interested in the study were contacted by phone or email to give them further information about the study. If individuals agreed to participate in the study, they were sent an email with an ID number and a link to the online survey.

3.2 Measures

3.2.1 Attachment anxiety and avoidance

The two dimensions of attachment were measured by the Turkish version of Fraley, Waller, and Brennan's [42] Experiences in Close Relationships-Revised Scale [43]. Half of the items on the 36-item scale measures attachment avoidance referring to discomfort with closeness and depending on others, and the other half measures attachment anxiety referring to fear of rejection and abandonment. The attachment anxiety and avoidance subscales demonstrated adequate reliability with this sample. The attachment anxiety Cronbach's alpha value for women was .88 and .86 for men. The attachment avoidance Cronbach's alpha value for women was .89 and .87 for men.

3.2.2 Audio-visual data

The audio-visual data were collected in the Ozyegin University Relationship Research Lab which is equipped with private rooms with video cameras. The audio-visual data were collected via following the procedures of the Couples Interaction Coding System developed by Heavey, Gill, and Christensen [44]. CIRS is an observational coding scheme used to rate each spouse as they interact with their partner during a problem solving discussion. After arriving at the lab, according to the instructions of the CIRS, each spouse was asked to determine a topic that they thought was important in their relationship and that they had disagreements about. After each spouse independently determined their topic and approved of each other's topics, the couple were asked to have two discussions where they talked about one spouse's topic for 10 min and the others for another 10 min. Then, the couples were left alone in the room and the audio-video recording began. After the first 10 min, a research assistant knocked on the door and asked the couple to switch to the next topic. When the final 10 min were over, the assistant knocked again on the door and informed the couple that the recording was over.

Total collected data is 2060 min of recordings. After the Bi-Gaussian Voice Activity Detector (VAD) [45] process, the final amount is 1948 min. However, we use 1507 min of recording of secure and fearful people's conversations.

4 System description

4.1 Low-level acoustic features

Both prosodic and spectral features were used. Prosodic features reflect vocal characteristics relating to various behavioral aspects [11, 14, 18], and spectral features are informative in various tasks related to emotion recognition [46, 47] and behavioral signal processing [11, 14].

Pitch, intensity, and energy and its delta and delta-delta parameters were extracted as prosodic features. In addition, for voice features, jitter, shimmer, and their delta and delta-delta parameters were extracted. 15 Mel-frequency Cepstral Coefficients (MFCCs), 8 Mel-Filter Banks (MFBs), 8 Line Spectral Frequencies (LSFs), and their delta and delta-delta parameters were extracted as spectral features. All these low-level acoustic features are extracted every 10 ms with a 25 ms Hamming window using OpenSMILE [48]. For extracting spectral features, the parameters used shown in Table 2 were used. For voice features, including shimmer and jitter, The INTERSPEECH 2010 Paralinguistics Challenge (IS10) [49] config file is used.

Next, for each problem discussion session, six statistical functions (Minimum, Maximum, Range (Maximum - Minimum), Mean, Median, and Standard Deviation) of all low-level acoustic features were computed. Those statistical functions were computed separately for each of the speakers (husband and wife) across the problem discussion session, generating a set of session-level features for each low-level acoustic feature. For predicting the attachment style of husband and wife, the same approach was used for both of the two interactions that the couple was engaged in.

Table 2 Configuration parameters of spectral features

Feature type	Configuration parameters
MFCC	Hamming window with 25 ms audio frames with 10 ms sample rate is used.
MFB	The log of MFBs in the range between 20 and 6500 Hz.
LSP	The LSP features are computed with linear prediction with order of 8.

Note that sessions in the dataset are relatively long, around 10 min each, and turn-level computation of functionals could be used for a more fine-grained analysis. However, attachment style is a personality trait that we expected to observe in all turns. Moreover, turn-level emotions and style changes are captured with the i-vector and sentiment features discussed below. Furthermore, session-level features, even though not fine-grained, were found to contain significant information [13] for capturing the acoustic cues from each of the speakers individually. Thus, session-level functionals of low-level acoustic features are used here.

A summary of the low-level acoustic features and statistical functions that are used here are shown in Table 3.

4.2 I-vector features

I-vectors are commonly used in speaker verification systems. They are also useful for identifying changes in a person's voice depending on the context. The reason we use them here is to detect changes in the voice characteristics of couples during a conversation.

Compared to low-level acoustic features that are aggregated using functionals over the entire session, i-vectors are extracted at the turn-level. Variances of turn-level i-vector features over the entire session are computed for each gender and concatenated to generate the session-level features.

A brief description of the i-vector extraction algorithm is as follows. First, a Gaussian mixture model (GMM) is used to represent the acoustic feature space in most speaker verification systems. A universal background model (UBM) is first trained, and then speaker-specific models are obtained by adapting the UBM using a maximum a posteriori adaptation (MAP) approach.

The super vector of mean vectors in UBM is typically high dimensional. Factor analysis can be used to reduce the number of parameters to adapt to where the lower-dimensional speaker and channel vectors are used for verification. More recently, a total variability space (TVS) approach [50] is proposed that combines the speaker and interaction variabilities that are represented in a single total variability space. In the TVS approach,

$$m_s = m_0 + Tw_s \tag{1}$$

where m_s is speaker and channel dependent supervector, m_0 is speaker and channel independent mean supervector, T is a low rank rectangular matrix, and w_s is called an identity vector (i-vector). The T matrix, which represents the total variability space, is typically trained using a database where multiple sessions are available for each speaker. w_s is extracted for each turn of male and female speakers in a given session using a MAP approach [50]. The set of i-vectors from the male speaker can be represented with $W_m = \{w_m^{(1)}, w_m^{(2)}, ..., w_m^{(N_m)}\}$ where N_m is the number of turns when the male speaker speaks. Similarly, the set of i-vectors for the female speakers is $W_f = \{w_f^{(1)}, w_f^{(2)}, ..., w_f^{(N_m)}\}$. Covariance matrices of i-vectors for the male and female speakers, computed using W_f and W_m, are $\sum_f$ and $\sum_m$. The feature vector extracted from the male speaker is $\left[\sigma_{f,1}^{(2)}, \sigma_{f,2}^{(2)}, ..., \sigma_{f,d}^{(2)}\right]$ where $\sigma_{f,i}^2$ is the i^{th} diagonal element of $\sum_f$. Similarly, the feature extracted from the male speaker is $\left[\sigma_{m,1}^{(2)}, \sigma_{m,2}^{(2)}, ..., \sigma_{m,d}^{(2)}\right]$. Turn-level i-vector extraction for male and female speakers is shown in Fig. 1.

5 Method

Baseline features that were used for predicting the attachment style are shown in Table 3. 25-ms analysis window with a 10-ms frame rate was used for feature extraction. Silence segments were removed using VAD. Partitioning each interaction to turns (speaker diarization) was done manually. Overlapping segments were removed. Similarly, nonverbal segments, such as laughter, were removed.

5.1 I-vector extraction

We used MFCC, energy, and their first- and second-order derivatives for training the UBM which consists

Table 3 Low-level acoustic features and statistical functions

Feature type	Feature names
Spectral	Energy and its derivatives, intensity, pitch 15 MFCCs and their derivatives, 8 MFBs and their derivatives, 8 LSFs and their derivatives Jitter, shimmer, and their derivatives
Functionals	Minimum, Maximum Range (Maximum - Minimum) Mean, Median, Standard Deviation

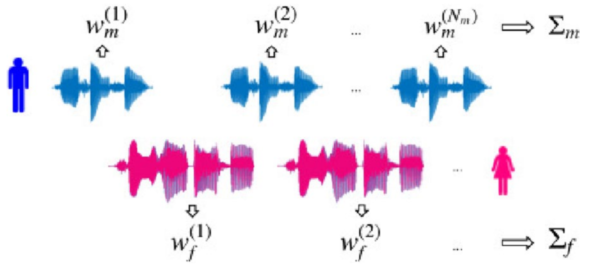

Fig. 1 Turn-level i-vector extraction and computation of covariance matrices of i-vectors for the male and female speakers

of a 256-component GMM model using all the training data available.

Two gender-dependent T matrices were trained. I-vectors were extracted for each turn in the sessions. All of the training data was used for training the T matrix in Eq. 1, but some of the turns are left out of training. For extracting more accurate i-vectors, we did not include the turns with less than 2 s of speech in training. Moreover, to avoid having speaker-biased T matrices, we excluded some of the turns in some of the sessions such that all interaction sessions have the same number of turns during training.

For tuning the rank of the T matrix, T matrices with the rank of 50, 100, and 400 were generated. The rank of the T matrix was set to 100 since that performed the best.

5.2 Sentiment analysis

Automatic sentiment analysis can be used to detect polarity (positive, negative, and neutral) and emotions (angry, happy, sad, etc.) within text. Because couple interaction-sessions are targeted around unresolved conflict points, conversations are typically intense and emotional. Thus, analysis of how the emotions change throughout the interaction can add significant value to detection of relationship parameters. In Table 4, there are example sentences, from our dataset, tagged as positive and negative emotions.

Here, the Google sentiment API was used to generate sentiment analysis scores. Transcripts of the Turkish couple interactions were manually generated and translated to English. Google sentiment analysis engine [51] was used to extract the score and the magnitude of the emotion from translated text.

Sentiment of each turn was computed independently. The score of each turn was represented with a number between $-$ 1.0 and 1.0 where -1 indicates negative and $+$ 1 indicates positive sentiment. Magnitude is the strength of an emotion present in text with a range between 0.0 and +inf. Thus, two sentiment features are extracted for each turn. Functionals of those two features are then computed to generate the session-level sentiment features. Functionals shown in Table 3 are used.

5.3 Feature selection

Compared to the training data, dimensionality of the extracted features is high. Thus, we performed feature selection to choose a subset of the original features that provide the maximum discriminatory information.

We investigated Mutual Information Maximization (MIM), Joint Mutual Information (JMI), and minimum Redundancy Maximum Relevance (mRMR) as the feature selection algorithms. For feature selection, MIM uses mutual information to calculate relevance to each feature to class label [52]. mRMR selects the most relevant features for classification while minimizing redundancy within the selected set [53]. JMI was proposed for selecting features by jointly computing their redundancy and complementary effects [52]. For feature selection, the FEAST Toolbox was used [54]. Feature selection was done separately for males and females.

5.4 Attachment styles

We divided our data set into four attachment styles, following [42], which are secure, fearful, dismissive, and preoccupied. Those attachment styles are described in Table 1 [24].

Our sample consisted of 206 individuals of which 60 had secure, 39 had dismissive, 58 had fearful, and 49 had preoccupied attachment styles. According to Fraley [55], attachment styles can be categorized based on group median scores of attachment anxiety and attachment avoidance. Anxiety and avoidance scores that are both below the group median indicates secure attachment style. Similarly, fearful attachment is indicated by anxiety and avoidance scores that are both lower than the group median. Anxiety score that is above the threshold together with avoidance score is below the threshold indicates preoccupied attachment style. Finally, anxiety score that is below the threshold together with avoidance score is above the threshold indicates dismissive

Table 4 Some conversation examples' scores and magnitude values computed using Google Cloud Natural Language Analyzing Sentiment API

Sentence	Magnitude	Score
I mean, I love your family, you know we do not have any troubles when they come and stay	0.7	0.7
The problem with your answer is that it does not take this sarcasm and it only takes serious side of it and responds with extreme seriousness	0.7	-0.7
I am not interested. It is so boring	1.5	-0.7

Table 5 Range, mean, median, and standard deviation of attachment anxiety and attachment avoidance scores

Attachment style	Range	Median	Mean	STDev
Attachment anxiety	1–6.23	2.44	2.61	0.87
Attachment avoidance	1–6	1.61	1.85	0.77

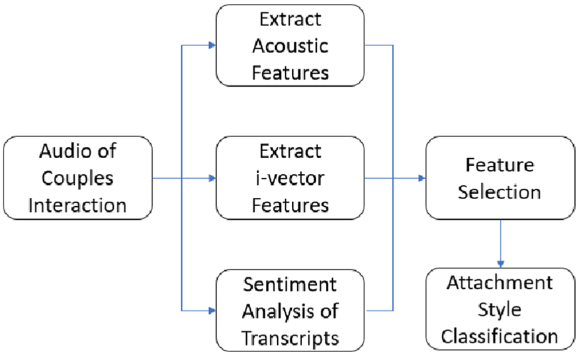
Extract Acoustic Features
Audio of Couples Interaction
Extract i-vector Features
Feature Selection
Sentiment Analysis of Transcripts
Attachment Style Classification

Fig. 2 Overview of the proposed automatic attachment style classification algorithm

attachment style. The statistic of attachment anxiety and attachment avoidance score is shown in Table 5.

Because we have limited data, we focused on discriminating between fearful and secure styles since those are the two most extreme cases in the spectrum and, hence, distinguishing among these two styles requires less data than dismissive and preoccupied styles. Therefore, two conversations from 118 people, with fearful and secure attachment styles, were used for analysis.

A summary of the method used in this paper to predict the attachment style of spouse is represented in Fig. 2. Classifiers that were used here are described in Fig. 2.

5.5 Classification algorithms

A range of classifiers was used to find the highest accuracy for the classification of attachment styles. All classifiers and were implemented and tested using Scikit-Learn library [56]. The performance was evaluated with F-1, recall, and precision scores. Description of the classifiers are given as follows.

5.5.1 Decision trees

Decision trees are often used when the training data is limited [57]. Here, five different decision tree classifiers were used: vanilla decision tree, random forest, gradient boosting tree, AdaBoost decision tree, and extra tree classifiers.

5.5.2 Random forests

Random forests are a combination of decision trees generated by using a sample of data to obtain low-bias trees [58]. Instead of a single decision tree, an ensemble of decision trees is trained where each tree is trained with a random subset of the training data and a random subset of all features. Majority voting is used to combine the decisions of the trees for final classification decision.

5.5.3 Gradient boosting

The working principle of the gradient boosted decision trees is to iteratively create multiple weak trees each of which attempts to reduce the errors of the previous tree [59].Thus, instead of training multiple parallel trees independently as in the random forest case, the trees are generated one at a time in a dependent manner with the gradient boosting approach. Each tree is created based on previous tree's residual values. The variable that is tuned in this algorithm is the target values. Gradient-descent optimization is used to update the target values for the next tree using the gradient of the loss function (cross-entropy in our case) with respect to the target values [60].

5.5.4 XGBoost

XGBoost which also called extreme gradient boost is a different implementation of gradient boosting method with more accurate and faster predictions [61]. It typically outperforms all other decision tree-based techniques in machine learning tasks. Besides several hardware optimizations for parallelizing and speeding-up the computations, it allows regularization (both L1 and L2) to avoid overfitting. Moreover, its learning algorithm is depth-first, i.e., it first creates the full tree and prunes backwards to the specified max-depth parameter. The gradient boosting algorithm, however, is greedy and it only stops splitting the nodes based on a threshold on the split gain.

5.5.5 Extra tree

Extra tree classifiers have many independent binary decision trees [62]. As opposed to random forests, all training samples are used for splitting the nodes of each tree. A random subset of the features are selected for each tree. Then, among those subset of features, at each node, a random feature is selected for splitting the node. The algorithm has performance similar to random forest but has a lower variance because of randomization steps in the algorithm.

5.5.6 Adaboost

Adaboost is similar to gradient boosting in the sense that it creates multiple weak decision trees sequentially to get a strong ensemble learning algorithm [63]. Each weak learner updates weights of the training samples based on the mistakes of the previously trained decision tree. Thus, the algorithm tries to put more effort into eliminating the errors that were done by the previous trees in the ensemble.

5.5.7 SVM

Another classification method support vector machine (SVM) was used to classify the problem. SVM classifier creates a hyperplane to divide data into two classes. It is a maximum margin classifier, which means it tries to minimize the distance of the closest data points to the hyperplane for both classes [64]. SVM is one of the best off-the-shelf classifiers when the amount of data is limited.

5.5.8 Convolutional networks

ConvNets architectures are useful for extracting high-level features from input data. They use a number of filters on input data, and generate output, namely feature map, using convolution. At the end of the convolution operation, the resultant feature maps are collected and a final output is generated. Then, the output is passed through a nonlinear activation function. After the convolution layer, max-pooling is used to reduce the number of parameters, preventing overfitting and accelerating the training process. After the convolutional and pooling layer, a fully connected layer is used to generate the final outcome. Regularization techniques such as batch normalization and dropout can also be used to avoid overfitting.

The neural network architecture used here is shown in Fig. 3. The simplified network had the input layer, a convolution layer with RELU activation function, a dropout layer, and a fully connected layer with the sigmoid function. Pytorch library was used for training, and Adam optimizer was used. There were 32 nodes in each layer.

Note that the network requires a sequence of features as input as opposed to the algorithms discussed above that require a single session-level feature vector. Thus, for the CNN network, instead of aggregating acoustic features over the entire session using functionals, we split each session to fixed 10-s-long chunks. Low-level acoustic features extracted from those chunks are then used as input to the network that generates a score for the chunk. Then, scores generated for all chunks are averaged to compute the final score. For feature fusion, the network is conditioned with the sentiment and i-vector features when they are available.

5.5.9 Sparsely connected and disjointly trained DNN

The amount of data is not sufficient for training complex neural networks. To circumvent the problem, the approach in [65] was used. The proposed approach follows a two-stage training process. The first stage requires training a classifier network on each feature subset independently. Knowledge-based feature subsets were used. Each subset had a significantly lower feature dimension compared to original feature set, which significantly reduced overfitting. In the second training stage, parameters of each independent network were frozen, and a randomly initialized fusion layer was trained. The fusion layer connects hidden layers of the networks trained in the first stage. The system in [65] allows reduction of the total number of parameters via sparseness and disjoint training while allowing full connectivity.

Adam optimizer is used together with binary cross-entropy loss. Five training epochs were used together with a fixed learning rate of 1e−3.

5.5.10 Hyperparameter tuning for tree-based algorithms

In this study, we optimized hyperparameters of tree-based methods using the leave-one-out cross validation approach. For each tree-based classifier, we optimized

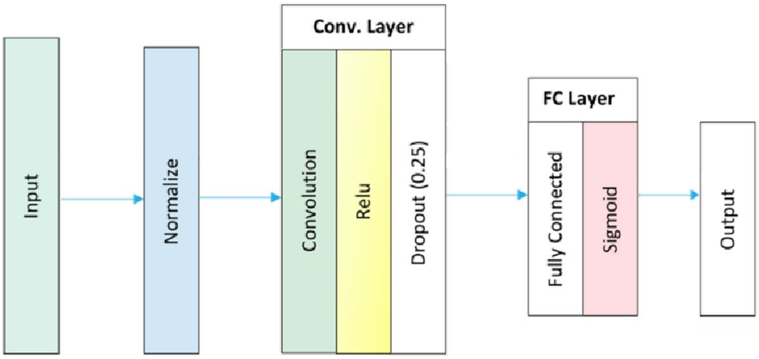

Fig. 3 Neural network architecture

hyperparameters for best classification results and reported average performance of all folds.

Two different leave-one-out strategies were used for the conversation- and individual-based approaches. For the conversation-based approach, one conversation is left out for each fold. For the individual-based approach, one individual was left out for each fold.

Note that we did not use a separate development set during cross-validation. Thus, the cross-validation approach used here is not nested, which means our results could potentially be slightly optimistic.

Hyperparameters that were tuned were learning rate, impurity criterion, estimator number, and the depth of the built trees. The learning rate controls the model's changes in regard to previous mistakes. Smaller learning size helps the model to learn better but raises the computing cost [66]. Both Gini and entropy are used for calculating the impurity of a node, which is used to split the node. Estimator number is the number of weak classification trees built by Adaboost, gradient boost, extra tree, and XGBoost classifiers. Controlled tree depth can raise the performance of the model but it could cause overfitting [67].

5.6 Score fusion and feature fusion

In some of our experiments, we used feature level fusion and score level fusion. The aim of feature fusion is to get discriminative information by combining multiple input patterns to make more accurate predictions. Score level fusion approaches are divided into two subcategories. The first one is non-trainable which is used to output scores of classifiers and merge the scores, and the second one is trainable which uses scores as input features and creates a new pattern classification problem [68]. In [69], score level fusion is used with SVM to boost system performance.

The overview of the score fusion algorithm is as follows, each couple has two 10 min of conversations. Female and male features were extracted separately for each conversation. Then, the low-level acoustic features and i-vectors were used as inputs to the classifiers. Two scores are predicted for each person using the two conversations and those scores are averaged to predict whether a person has secure or fearful attachment style.

The overview of the feature fusion algorithm is that each person's low-level features and i-vector features extracted for each conversation. Then, those extracted features were fused to make a prediction of the attachment style.

For testing the feature selection algorithms, five different feature dimensions were used: 5, 10, 15, 30, and 50. For each test setup, performance of only the best performing feature size is reported for each feature selection algorithm.

Even though different classifiers seemed to have different performance results, comparison of classifiers was crucial to understand whether those differences were significant or not. Here, we used McNemar's Test [70] to analyze the statistical difference between classifiers.

6 Results

To predict attachment styles, we conducted two different sets of experiments. In the first set, a conversation-based approach was used where only the conversation with the topic picked by the target spouse was used. The second conversation was not used. Features extracted from both spouses for that conversation were fused to predict the spouse's attachment style. The rationale for this approach was that the spouses affect each other and the entanglement of emotions during conversation could be captured if both spouses' features are fused even though we were predicting the attachment style of only one of them.

In the second set of experiments, an individual-based approach was used where both conversations of a spouse are used for prediction. Features for each spouse were extracted separately for each conversation. Score fusion and feature fusion methods were then used for prediction to fuse information from each conversation.

The leave-one-out method was used for assess the performance. Accuracy, precision, recall, and F1 scores were used as the performance metrics. Calculation of precision, recall, and F1 scores are shown in Eqs. 2, 3, and 4where TP is true positives, FN is false negatives, and FP is false positives.

$$recall = \frac{TP}{TP + FN} \tag{2}$$

$$precision = \frac{TP}{TP + FP} \tag{3}$$

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \tag{4}$$

6.1 Performance of the conversation-based approach

Table 6 shows the accuracy, precision, recall, and F1 scores of the classifiers using the low-level and i-vector features with three different feature selection algorithms. SVM and vanilla decision tree algorithms performed worse than the other algorithms. Extra tree and XGboost algorithms performed the best. Difference between different feature selection algorithms were not significant in Table 6.

Table 6 Accuracy, precision, recall, and F1 scores of the classifiers with low-level acoustic features, i-vector features, and sentiment features are shown using MIM, mRMR, and JMI feature selection algorithms. For each feature selection algorithm, results for only the best performing feature sizes are shown. Feature fusion is used in the second and third columns. Experiments were done using the conversation-based approach. Systems with best F1 scores are shown in bold

	Low-level acoustic features					**Low-level acoustic and I-vector features**					**Low-level acoustic and sentiment features**				
	Feature type	**Accuracy**	**Precision**	**Recall**	**F1 score**	**Feature type**	**Accuracy**	**Precision**	**Recall**	**F1 score**	**Feature type**	**Accuracy**	**Precision**	**Recall**	**F1 score**
SVM	**MIM (50)**	0.7	0.69	0.69	0.69	**MIM (30)**	0.72	0.72	0.68	0.7	**MIM (15)**	0.68	0.68	0.67	0.68
	mRMR (50)	0.79	0.8	0.76	0.77	**mRMR (15)**	0.75	0.76	0.72	0.74	**mRMR (30)**	0.73	0.73	0.74	0.73
	JMI (50)	0.73	0.76	0.65	0.7	**JMI (30)**	0.7	0.69	0.7	0.7	**JMI (30)**	0.7	0.7	0.69	0.7
Decision tree	**MIM (10)**	0.72	0.73	0.69	0.7	**MIM (50)**	0.74	0.69	0.84	0.75	**MIM (5)**	0.72	0.76	0.65	0.7
	mRMR (30)	0.73	0.72	0.75	0.74	**mRMR (30)**	0.73	0.72	0.76	0.74	**mRMR (50)**	0.73	0.71	0.77	0.74
	JMI (15)	0.72	0.68	0.81	0.74	**JMI (5)**	0.73	0.72	0.74	0.73	**JMI (15)**	0.74	0.77	0.69	0.73
Random forest	**MIM (30)**	0.81	0.82	0.79	0.8	**MIM (15)**	0.8	0.8	0.79	0.8	**MIM (15)**	0.8	0.83	0.75	0.79
	mRMR (50)	0.81	0.85	0.75	0.8	**mRMR (30)**	0.8	0.83	0.76	0.79	**mRMR (10)**	0.8	0.83	0.75	0.79
	JMI (30)	0.81	0.85	0.75	0.8	**JMI (30)**	0.82	0.85	0.77	0.81	**JMI (15)**	0.8	0.83	0.76	0.79
AdaBoost	**MIM (30)**	0.82	0.81	0.83	0.82	**MIM (30)**	0.79	0.8	0.77	0.79	**MIM (10)**	0.78	0.79	0.74	0.77
	mRMR (10)	0.79	0.77	0.81	0.79	**mRMR (10)**	0.79	0.77	0.83	0.8	**mRMR (10)**	0.83	0.8	0.86	0.83
	JMI (30)	0.83	0.82	0.84	0.83	**JMI (10)**	0.8	0.8	0.81	0.8	**JMI (30)**	0.78	0.78	0.76	0.77
Gradient boosting	**MIM (15)**	0.8	0.84	0.74	0.79	**MIM (30)**	0.81	0.82	0.79	0.81	**MIM (30)**	0.81	0.83	0.77	0.8
	mRMR (15)	0.8	0.81	0.76	0.78	**mRMR (30)**	0.77	0.8	0.7	0.75	**mRMR (30)**	0.84	0.88	0.77	0.82
	JMI (15)	**0.83**	**0.84**	**0.84**	**0.84**	**JMI (15)**	0.79	0.8	0.76	0.78	**JMI (50)**	0.8	0.8	0.81	0.8
Extra tree	**MIM (15)**	0.8	0.8	0.79	0.8	**MIM (30)**	0.81	0.85	0.76	0.8	**MIM (30)**	0.82	0.84	0.79	0.81
	mRMR (50)	0.83	0.87	0.79	0.83	**mRMR (50)**	0.81	0.81	0.81	0.81	**mRMR (50)**	**0.84**	**0.85**	**0.81**	**0.83**
	JMI (30)	0.83	0.87	0.8	0.83	**JMI (30)**	0.84	0.87	0.79	0.83	**JMI (15)**	0.81	0.82	0.79	0.8
XGBoost	**MIM (30)**	0.77	0.77	0.76	0.76	**MIM (10)**	0.75	0.75	0.74	0.75	**MIM (10)**	0.79	0.77	0.83	0.8
	mRMR (30)	0.8	0.77	0.75	0.76	**mRMR (15)**	0.75	0.74	0.76	0.75	**mRMR (5)**	0.78	0.76	0.81	0.78
	JMI (15)	0.8	0.77	0.84	0.8	**JMI (10)**	**0.84**	**0.83**	**0.84**	**0.84**	**JMI (30)**	0.77	0.76	0.77	0.77
Artificial neural network	**MIM(50)**	0.71	0.72	0.67	0.7	**MIM (30)**	0.78	0.76	0.79	0.78	**MIM(30)**	0.76	0.76	0.76	0.76
	mRMR(50)	0.74	0.72	0.76	0.74	**mRMR (50)**	0.77	0.72	0.86	0.78	**mRMR(50)**	0.75	0.72	0.81	0.76
	JMI(50)	0.75	0.76	0.72	0.74	**JMI (50)**	0.71	0.68	0.75	0.72	**JMI(50)**	0.74	0.78	0.67	0.72
SD-DNN	-	0.8	0.92	0.63	0.75	-	-	-	-	-	-	-	-	-	-

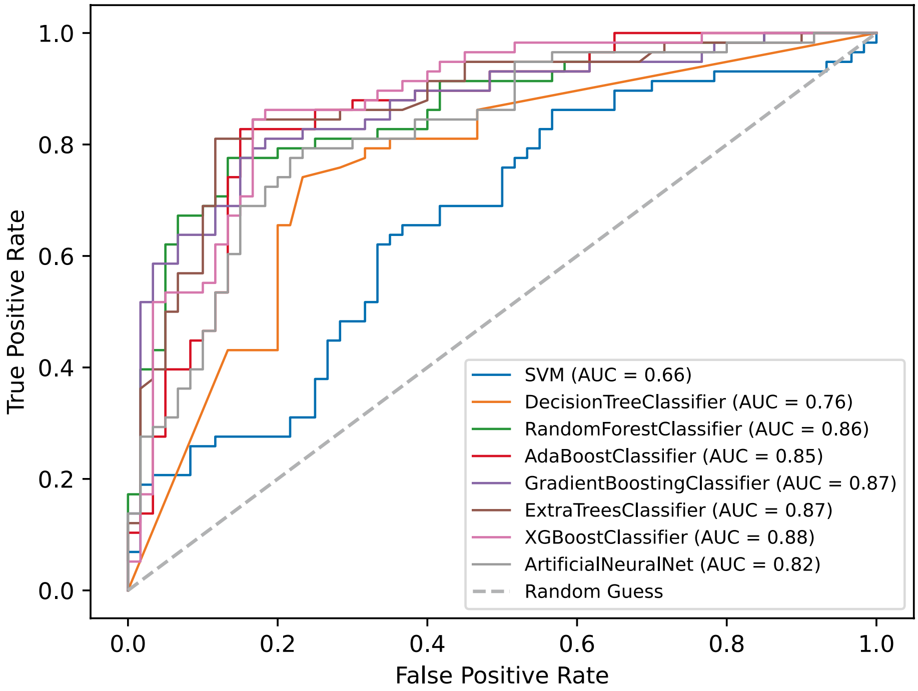

Fig. 4 ROC curves of the classifiers in Table 6 with low-level features and i-vector. For each classifier, output of the best performing feature selection algorithm is reported

The extra tree classifier performed the best for all three feature sets in Table 6. Extra tree model had an accuracy of 84% with fusion of low-level and i-vector fusion with JMI feature selection. Additionally, mRMR feature selection with extra tree model performed with higher precision and recall scores of 87%.

Receiver operating characteristic (ROC) curve of the classifiers in Table 6 are shown in Figs. 4 and 5. The results are inline with the observations above. XGboost and extra tree algorithm have the largest area under curve (AUC). The other multiple randomized tree algorithms also performed close to those two algorithms.

To check statistical difference between classifiers, we used the McNemar's significant test. We have found that the extra tree algorithm, when low-level and i-vector features are used, is significantly better than SVM and the decision tree model with p-value below 0.05.

Performance of classifiers using the low-level and sentiment features with three different feature selection algorithms are shown in the last column of Table 6. Sentiment features with low-level acoustic features performed slightly worse than the low-level and i-vector features as shown in Table 6. Similarly, adding sentiment features degraded results compared to using low-level features only. Even though the sentiment features slightly degraded the performance, difference is insignificant in most cases (Figs. 4 and 5).

Through further manual analysis of our sentiment features, we found out that the sentiment features were sometimes wrong. Some of the errors were due to automatic prediction engine, some of them had to do with sarcastic speech, and some of the errors were due to mistakes while translating Turkish text to English. Thus, even though we have anecdotal evidences that sentiment features can be useful for the task, automatic detection errors diminished their impact in our tests.

Another aspect of attachment style detection is the gender-dependent performance of the classifiers. For example, Leaper and Robnett showed that women are more likely than men to use tentative speech forms [71]. Hence, besides pooling male and female features, we also experimented with them separately. Results are shown in Table 7. Performance of the classifiers for the two genders were not consistently and significantly different. Using a combination of features from both genders consistently outperformed each gender even though the differences were not statistically significant.

6.2 Performance of the individual-based approach

Table 8 shows multiple classification methods' accuracy, precision, recall, and F1 scores using feature fusion method with low-level acoustic and i-vector features.

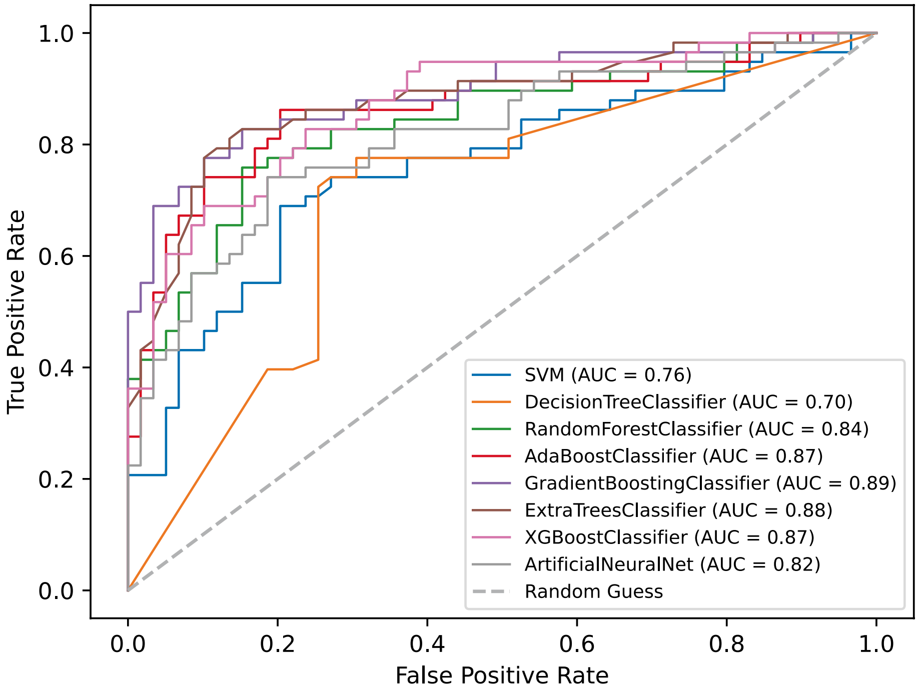

Fig. 5 ROC curves of the classifiers in Table 6 with low-level acoustic features and sentiment. For each classifier, output of the best performing feature selection algorithm is reported

SVM, vanilla decision tree, and neural network algorithms performed worse than the other classifiers. Adaboost, extra tree, and gradient boosting decision tree algorithms performed the best with F1 scores of 84% with i-vector and low-level features together. Additionally, Adaboost model showed a high accuracy of 85% with only low-level acoustic features. XGBoost and random forest algorithms performed close to them with F1 scores of 82% and 80%, respectively. Performance differences between those five algorithms were not significant.

In Fig. 6, ROC curves of feature fusion is shown for all classifiers. AUC in Fig. 6 is inline with results in Table 8. AUC for gradient boosting and Adaboost are highest and AUC for decision tree and SVM are lowest.

Table 8 shows the accuracy of the score fusion algorithm with the low-level and i-vector features. Adaboost get the highest accuracy by 81%. However, there was no significant difference between the algorithms that are based on multiple randomized trees.

In Fig. 7, ROC curves of classifiers are shown for score fusion. Although Adaboost gets the highest accuracy in Table 8, gradient boosting and XGBoost classifiers have the largest area under curve.

Even though feature fusion performed slightly better in terms of accuracy, no significant difference was found between score fusion and feature fusion when best performing classifiers were compared.

7 Discussion

Our goal in this work was to show the effectiveness of low-level acoustic features, i-vector features, and sentiment scores for classification of attachment style of adults in couple interactions. In conversation-based prediction, we see that fusion of sentiment scores and low-level acoustic features did not boost the accuracy. At least part of the reason why sentiment scores did not increase the accuracy scores is the occasional mismatch between the original and translated text. The conversations were conducted in Turkish, the native language of the spouses. During translation of text to English, meaning of some of the sentences were lost, which caused mismatch. Those errors, together with the errors that normally occur in the automatic sentiment detection engine, reduced the accuracy of the scores.

Even though sentiment scores did not improve the accuracy when fused with other features, how much information they contain is also an important question. To assess that, experiments were conducted using only the sentiment features. The results are presented in Table 9. Indeed, performance is far below what was obtained with the acoustic features, which explains why sentiment features did not help improve the performance of the acoustic features.

Additionally, i-vectors did not improve the performance. In both conversation-based approach and

Table 7 Accuracy, precision, recall and F1 scores with i-vectors and low-level features are shown for the investigated classifiers and feature selection algorithms. Female and male features are used together and separately

	Algorithm	Gender	Accuracy	Precision	Recall	F1 score
SVM	**mRMR(30)**	Male	0.68	0.66	0.72	0.69
	JMI (50)	Female	0.66	0.67	0.63	0.65
	mRMR (15)	Both	0.75	0.76	0.72	0.74
Decision tree	**mRMR(10)**	Male	0.69	0.68	0.68	0.68
	JMI(15)	Female	0.68	0.69	0.65	0.67
	MIM (5)	Both	0.68	0.68	0.67	0.68
Random forest	**JMI(10)**	**Male**	**0.77**	**0.79**	**0.72**	**0.75**
	JMI(10)	**Female**	**0.77**	**0.79**	**0.72**	**0.75**
	JMI (10)	Both	0.77	0.8	0.72	0.76
AdaBoost	**JMI(10)**	Male	0.72	0.71	0.70	0.71
	JMI(30)	Female	0.74	0.73	0.75	0.74
	MIM (30)	Both	0.79	0.79	0.77	0.78
Gradient boosting	**JMI(10)**	Male	0.74	0.75	0.7	0.73
	JMI(15)	Female	0.75	0.76	0.72	0.74
	JMI (10)	Both	0.76	0.76	0.76	0.76
Extra tree	**JMI(5)**	**Male**	**0.77**	**0.83**	**0.69**	**0.75**
	JMI(15)	**Female**	**0.77**	**0.76**	**0.77**	**0.77**
	JMI (50)	Both	0.79	0.79	0.77	0.78
XGBoost	**JMI(10)**	Male	0.76	0.75	0.77	0.76
	mRMR(15)	Female	0.76	0.77	0.74	0.75
	JMI (10)	**Both**	**0.84**	**0.76**	**0.79**	**0.78**
Artificial neural network	**mRMR(50)**	Male	0.68	0.67	0.69	0.68
	JMI(50)	Female	0.71	0.71	0.69	0.7
	MIM (30)	Both	0.78	0.76	0.79	0.78

individual-based approach, we see that fusion of i-vector and acoustic features' F1 scores are similar to low-level acoustic features. To assess the information in i-vectors, we conducted experiments with i-vector features. The results are shown in Table 10. Even though i-vector features were not as successful as the low-level acoustic features, the difference between F1 scores of best systems is not large. Therefore, we concluded that the reason why i-vector features, which are extracted using MFCC features, did not increase the performance is because the information contained within them was already available in the low-level features.

7.1 Difference between classifiers

We explored state-of-the-art classification algorithms for detecting secure and fearful attachment styles. We found that attachment styles of couples could be predicted to some extent with decision trees that use randomized multiple trees. SVM, neural network, and vanilla decision tree algorithms performed significantly worse using McNemar's significance tests.

Even though neural networks achieve state of the art results in many tasks, they did not perform as good in our experiments because of the limited amounts of data that we have. We used data normalization, data shuffling, and dropout regularization to mitigate the issue. We also preferred a relatively lower complexity network in the tests to reduce the risk of overfitting. Those methods improved the performance of neural network classifier. However, it still could not perform significantly better than SVM and vanilla decision tree algorithms.

SD-DNN method was explored for overcome the issue of limited data. This approach reduces the total number of parameters which avoids risk of overfitting. F1 scores improved with SD-DNN method.

7.2 Self-reported baseline measures

Even though our model achieves good performance, the classifies failed in assessment of some people's attachment styles in a quite confident way. In other words, prediction was wrong and the classifiers were confident about their mistakes. To further analyze the root-causes of those, we manually analyzed those particular recordings.

Table 8 Accuracy, precision, recall, and F1 scores of classifiers with low-level acoustic features, and i-vector features features are shown using MIM, mRMR, and JMI feature selection algorithms. Feature fusion is used in the first and second columns. Score fusion is used in third column. Experiments were done using the individual-based approach. Systems with best F1 scores are shown in bold

	Low-level acoustic feature fusion					Low-level acoustic and I-vector feature fusion					Low-level acoustic and I-vector score fusion				
	Feature type	Accuracy	Precision	Recall	F1 score	Feature type	Accuracy	Precision	Recall	F1 score	Feature type	Accuracy	Precision	Recall	F1 score
SVM	**MIM (30)**	0.73	0.72	0.72	0.72	**MIM (30)**	0.73	0.72	0.72	0.72	**MIM (10)**	0.61	0.62	0.52	0.57
	mRMR (15)	0.68	0.69	0.65	0.67	**mRMR (15)**	0.69	0.69	0.66	0.67	**mRMR (30)**	0.57	0.56	0.53	0.55
	JMI (15)	0.67	0.65	0.7	0.68	**JMI (15)**	0.67	0.65	0.71	0.68	**JMI (5)**	0.65	0.65	0.62	0.64
Decision tree	**MIM (5)**	0.75	0.74	0.74	0.76	**MIM (50)**	0.72	0.7	0.74	0.72	**MIM (30)**	0.75	0.75	0.74	0.75
	mRMR (30)	0.74	0.77	0.69	0.73	**mRMR (30)**	0.79	0.79	0.77	0.78	**mRMR (10)**	0.64	0.63	0.67	0.65
	JMI (30)	0.71	0.7	0.72	0.71	**JMI (10)**	0.73	0.73	0.71	0.72	**JMI (10)**	0.74	0.72	0.77	0.75
Random forest	**MIM (30)**	0.8	0.81	0.76	0.78	**MIM (30)**	0.8	0.81	0.76	0.78	**MIM (30)**	0.76	0.75	0.77	0.76
	mRMR (15)	0.8	0.84	0.72	0.78	**mRMR (15)**	0.8	0.81	0.76	0.78	**mRMR (30)**	0.79	0.76	0.83	0.79
	JMI (15)	0.78	0.8	0.74	0.77	**JMI (10)**	0.77	0.76	0.77	0.77	**JMI (30)**	0.77	0.78	0.74	0.76
AdaBoost	**MIM (30)**	**0.85**	**0.87**	**0.81**	**0.84**	**MIM (30)**	**0.84**	**0.84**	**0.83**	**0.83**	**MIM (50)**	0.8	0.81	0.79	0.8
	mRMR (30)	0.82	0.84	0.79	0.81	**mRMR (30)**	0.82	0.82	0.81	0.82	**mRMR (30)**	0.79	0.77	0.81	0.79
	JMI (50)	0.8	0.83	0.76	0.79	**JMI (30)**	0.8	0.8	0.8	0.8	**JMI (15)**	**0.81**	**0.8**	**0.83**	**0.81**
Gradient boosting	**MIM (30)**	0.81	0.83	0.77	0.8	**MIM (30)**	**0.84**	**0.84**	**0.83**	**0.83**	**MIM (30)**	0.78	0.77	0.79	0.78
	mRMR (50)	0.81	0.82	0.79	0.81	**mRMR (50)**	0.83	0.85	0.79	0.82	**mRMR (30)**	0.8	0.81	0.79	0.8
	JMI (30)	0.82	0.82	0.81	0.82	**JMI (30)**	0.82	0.81	0.83	0.82	**JMI (15)**	0.8	0.8	0.81	0.8
Extra tree	**MIM (30)**	0.84	0.85	0.81	0.83	**MIM (30)**	**0.84**	**0.84**	**0.83**	**0.83**	**MIM (30)**	0.8	0.8	0.77	0.79
	mRMR (15)	0.81	0.82	0.79	0.8	**mRMR (15)**	0.81	0.82	0.79	0.81	**mRMR (30)**	0.78	0.78	0.76	0.77
	JMI (15)	0.78	0.81	0.74	0.77	**JMI (15)**	0.79	0.82	0.74	0.77	**JMI (30)**	0.77	0.79	0.72	0.76
XGBoost	**MIM (30)**	0.82	0.82	0.81	0.82	**MIM (30)**	0.82	0.82	0.81	0.82	**MIM (30)**	0.79	0.77	0.81	0.79
	mRMR (30)	0.79	0.78	0.79	0.79	**mRMR (30)**	0.79	0.79	0.79	0.78	**mRMR (30)**	0.79	0.78	0.79	0.79
	JMI (15)	0.77	0.75	0.79	0.77	**JMI (15)**	0.77	0.77	0.76	0.76	**JMI (30)**	0.75	0.75	0.73	0.74
Artificial neural network	**MIM(15)**	0.7	0.71	0.63	0.67	**MIM(15)**	0.69	0.71	0.64	0.67	**MIM(50)**	0.73	0.73	0.71	0.72
	mRMR(15)	0.7	0.7	0.65	0.68	**mRMR(15)**	0.69	0.7	0.65	0.68	**mRMR(50)**	0.68	0.69	0.62	0.65
	JMI(15)	0.63	0.63	0.62	0.62	**JMI(15)**	0.63	0.63	0.62	0.63	**JMI(50)**	0.77	0.78	0.74	0.76
SD-DNN	-	0.85	0.97	0.72	0.83	-	-	-	-	-	-	-	-	-	-

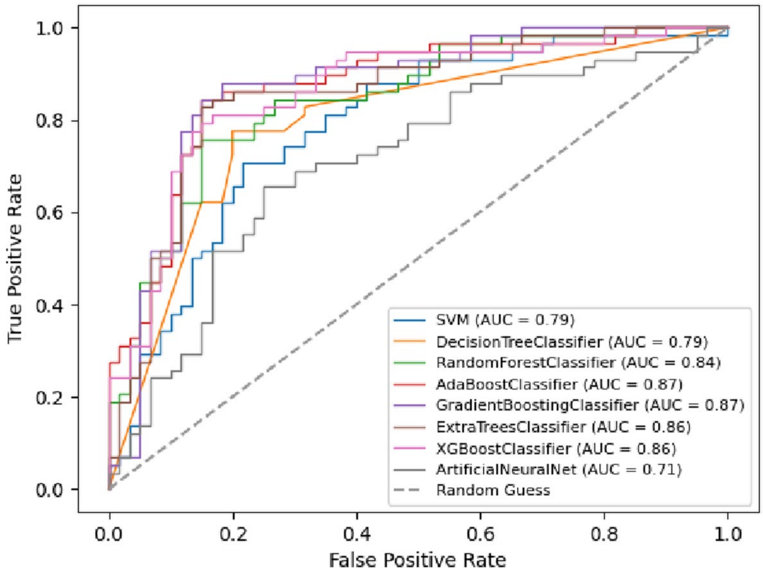

Fig. 6 ROC curves of the classifiers in Table 8 with low-level acoustic and i-vector feature fusion. For each classifier, output of the best performing feature selection algorithm is reported

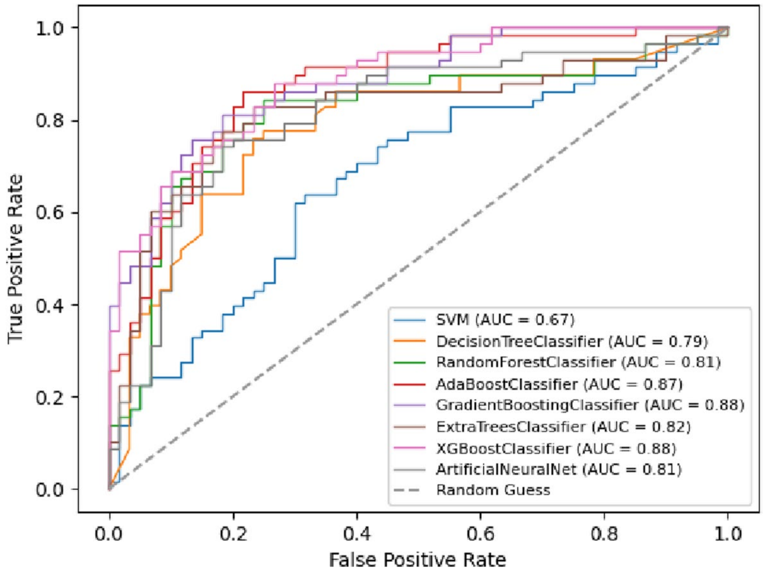

Fig. 7 ROC curves of the classifiers in Table 8 with low-level acoustic and i-vector score fusion. For each classifier, output of the best performing feature selection algorithm is reported

Note that our models use self-reported measures of attachment avoidance and attachment anxiety to make predictions with speech features. Thus, one explanation for some of those confidently wrong predictions may be the social desirability aspect of the self-reported measures. Social desirability refers to the participants' tendency to respond to questionnaire items with a concern for looking good rather than accurate [72]. Thus, the participants with mismatch between automatic prediction and self-assessment may have reported themselves in a socially desirable fashion while responding to the items of the self-report questionnaire. On the other hand, they might have acted differently during their problem discussion with their partner once their attachment system was activated in real-time as reflected in their voice features. That was found to be one of the contributors of the errors in our manual analysis.

Table 9 Accuracy, precision, recall, and F1 scores of the classifiers with sentiment features are shown. Experiments were done using the conversation-based approach

	Sentiment features			
	Accuracy	**Precision**	**Recall**	**F1 score**
SVM	0.55	0.55	0.49	0.52
Decision tree	0.65	0.64	0.67	0.65
Random forest	0.6	0.6	0.6	0.6
AdaBoost	0.58	0.58	0.59	0.58
Gradient boosting	0.6	0.6	0.56	0.58
Extra tree	0.61	0.6	0.63	0.62
XGBoost	0.57	0.56	0.55	0.56

Table 10 Accuracy, precision, recall, and F1 scores of the classifiers with i-vector features are shown using MIM, mRMR, and JMI feature selection algorithms. For each feature selection algorithm, results for only the best performing feature sizes are shown. Experiments were done using the conversation-based approach. Systems with best F1 score are shown in bold

	I-vector features				
	Feature type	**Accuracy**	**Precision**	**Recall**	**F1 score**
SVM	**MIM (30)**	0.64	0.66	0.57	0.61
	mRMR (10)	0.62	0.62	0.58	0.6
	JMI (15)	0.72	0.75	0.65	0.7
Decision tree	**MIM (15)**	0.76	0.79	0.7	0.74
	mRMR (5)	0.74	0.75	0.72	0.74
	JMI (50)	0.65	0.65	0.62	0.64
Random forest	**MIM (30)**	0.76	0.75	0.76	0.76
	mRMR (15)	0.74	0.74	0.72	0.73
	JMI (50)	0.77	0.77	0.76	0.76
AdaBoost	**MIM (15)**	0.77	0.77	0.76	0.76
	mRMR (30)	0.7	0.69	0.72	0.7
	JMI (15)	0.79	0.77	0.81	0.79
Gradient boosting	**MIM (15)**	0.79	0.77	0.81	0.79
	mRMR (15)	0.73	0.72	0.72	0.72
	JMI (15)	**0.81**	**0.82**	**0.79**	**0.8**
Extra tree	**MIM (30)**	0.79	0.77	0.81	0.79
	mRMR (15)	0.76	0.75	0.77	0.76
	JMI (30)	0.8	0.8	0.77	0.79
XGBoost	**MIM (10)**	0.73	0.72	0.74	0.73
	mRMR (10)	0.73	0.73	0.7	0.72
	JMI (15)	0.76	0.78	0.72	0.75

7.3 Prediction difference between conversations

In the score fusion approach, we tested whether a person's own conversation topic or his/her spouse's topic is more informative for the attachment style of that person. To that end, subjects were divided into four groups: secure females, secure males, fearful females, and fearful males. Figure 8 shows the means and variances of predictions for each group when its their own topics and their spouse's topic. We used t-test to measure the statistical difference for each group. Even though there are some differences between secure males and between fearful males, the p-value of those differences were larger than 0.05, and the differences were not significant. Also, we tested if there is a significant difference between predictions based on between one's self topic and their spouses' topic. Using the t-test, the difference was not significant.

To understand why some people were predicted as secure in one conversation and fearful in other, we conducted 2 different experiments. In the first experiment, we compared the wrongly predicted couples' anxiety and avoidance scores measured at 3 different times over 2.5 years. When we did Student's t-test, there was no significant difference. Thus, people were consistent with their attachment style over the years, which is also reported in the literature [73]. In the second experiment, we compared subjects who were predicted correctly in one conversations and predicted false in the second conversation. We tested the effect of how many months they have known each other before marriage and for how many months they were married. Using an unpaired t-test, we found that those two variables had a significant effect on the outcome. Thus, we concluded that people who know each other longer before marriage can be predicted more consistently in terms of attachment style.

8 Conclusions

In this study, we focused on attachment style prediction using interactions of recently married distressed couples. In addition to usual frame-level acoustic features such as MFCCs and pitch, and their functionals, we used i-vectors and sentiment-based features. Experimental results showed that i-vectors and sentiment features did not further improve the performance when fused with the low-level features. Similarly, difference between score and feature fusion algorithms was not significant.

There was no significant difference between genders. In fact, pooling all genders together performed better, which can, however, potentially be related to having more

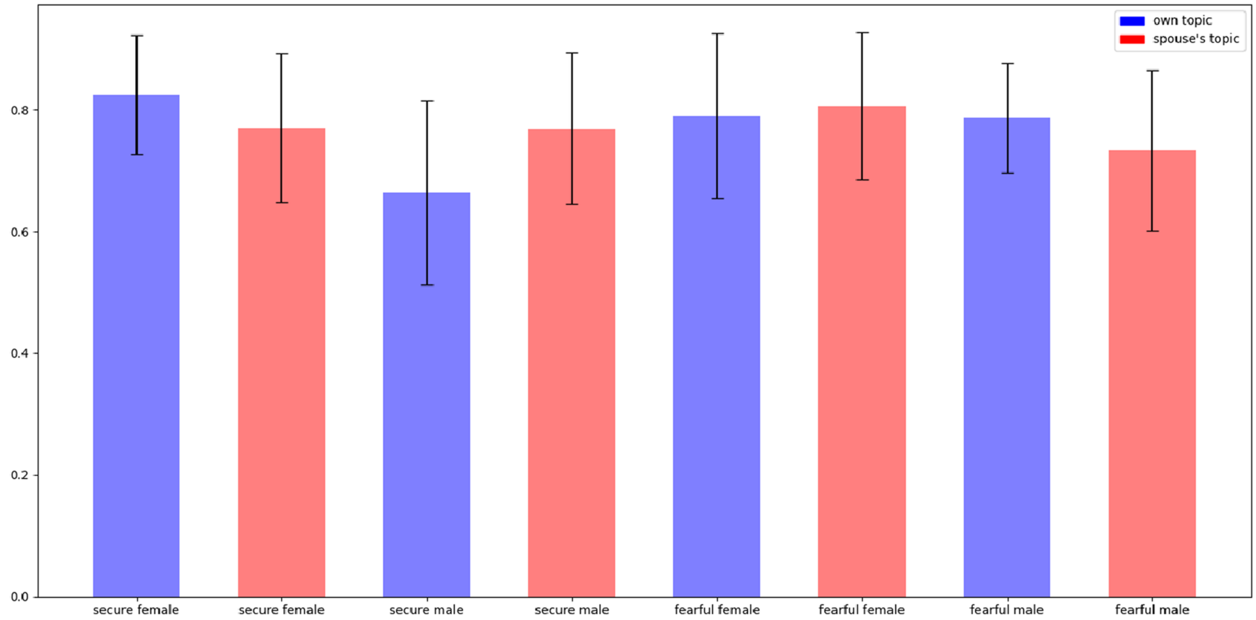

Fig. 8 Means of prediction of each attachment style groups their own conversation topic and their spouse's topic

data in the pool. Moreover, we found that people's attachment styles do not change over the 2.5 years which is consistent with the relationship literature in the psychology field. Also, people who know each other longer could be predicted more accurately in different conversations apart from who picks the topic.

Acknowledgements
Not applicable.

Authors' contributions
TMK and ENP performed the experiments. BÇD and NK contributed the psychological parts. CD and NK supervised the research. All authors read and approved the final manuscript.

Funding
The data of this study were collected in a project funded by TUBITAK 1001 Program (The Scientific and Research Council of Turkey; Project No: 113K538).

Availability of data and materials
The dataset used and analyzed during the current study are not publicly available but are available from the corresponding author on reasonable request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 28 December 2021 Accepted: 17 May 2023
Published online: 31 May 2023

References

1. J. Bowlby, Attachment and loss: retrospect and prospect. Am. J. Orthopsychiatry **52**(4), 664 (1982)
2. C. Hazan, P. Shaver, Romantic love conceptualized as an attachment process. J. Personal. Soc. Psychol. **52**(3), 511 (1987)
3. B.L. Hankin, J.D. Kassel, J.R. Abela, Adult attachment dimensions and specificity of emotional distress symptoms: prospective investigations of cognitive risk and interpersonal stress generation as mediating mechanisms. Personal. Soc. Psychol. Bull. **31**(1), 136–151 (2005)
4. O.S. Candel, M.N. Turliuc, Insecure attachment and relationship satisfaction: a meta-analysis of actor and partner associations. Personality Individ. Differ. **147**, 190–199 (2019)
5. E. Monti, D.C. Kidd, L.M. Carroll, E. Castano, What's in a singer's voice: the effect of attachment, emotions and trauma. Logoped. Phoniatr. Vocol. **42**(2), 62–72 (2017)
6. K.J. Tusing, J.P. Dillard, The sounds of dominance. Vocal precursors of perceived dominance during interpersonal influence. Hum. Commun. Res. **26**(1), 148–171 (2000)
7. M. Zuckerman, R. Klorman, D.T. Larrance, N.H. Spiegel, Facial, autonomic, and subjective components of emotion: the facial feedback hypothesis versus the externalizer-internalizer distinction. J. Personal. Soc. Psychol. **41**(5), 929 (1981)
8. A. Vinciarelli, M. Pantic, H. Bourlard, Social signal processing: survey of an emerging domain. Image Vis. Comput. **27**(12), 1743–1759 (2009)
9. S. Narayanan, P.G. Georgiou, Behavioral signal processing: deriving human behavioral informatics from speech and language. Proc. IEEE **101**(5), 1203–1233 (2013)
10. Georgiou, P.G., Black, M.P., Narayanan, S.S.: Behavioral signalprocessing for understanding (distressed) dyadic interactions: Somerecent developments. in *Proceedings of the 2011 Joint ACMWorkshop on Human Gesture and Behavior Understanding*. J-HGBU'11, pp. 7–12. Association for Computing Machinery, (New York,USA, 2011). https://doi.org/10.1145/2072572.2072576
11. M.P. Black, A. Katsamanis, B.R. Baucom, C.C. Lee, A.C. Lammert, A. Christensen, P.G. Georgiou, S.S. Narayanan, Toward automating a human behavioral coding system for married couples' interactions using speech acoustic features. Speech Commun. **55**(1), 1–21 (2013)
12. M. Sevier, K. Eldridge, J. Jones, B.D. Doss, A. Christensen, Observed communication and associations with satisfaction during traditional and integrative behavioral couple therapy. Behav. Ther. **39**(2), 137–150 (2008)
13. M. Nasir, W. Xia, B. Xiao, B. Baucom, S.S. Narayanan, P.G. Georgiou, in *Sixteenth Annual Conference of the International Speech Communication*

Association (ISCA). Still together?: The role of acoustic features in predicting marital outcome, (Dresden, Germany, 2015). https://doi.org/10.21437/interspeech.2015-539

14. C.C. Lee, A. Katsamanis, M.P. Black, B.R. Baucom, A. Christensen, P.G. Georgiou, S.S. Narayanan, Computing vocal entrainment: a signal-derived PCA-based quantification scheme with application to affect analysis in married couple interactions. Comput. Speech Lang. **28**(2), 518–539 (2014)
15. D. Jurafsky, R. Ranganath, D. McFarland, in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. Extracting social meaning: Identifying interactional style in spoken conversation, Colorado, (USA. 638–646, 2009). https://doi.org/10.3115/1620754.1620847
16. M.P. Black, P.G. Georgiou, A. Katsamanis, B.R. Baucom, S. Narayanan, in *Twelfth Annual Conference of the International Speech Communication Association (INTERSPEECH)*. "You made me do it": Classification of Blame in Married Couples' Interactions by Fusing Automatically Derived Speech and Language Information, (Florence, Italy, 2011). https://doi.org/10.21437/interspeech.2011-23
17. P.G. Georgiou, M.P. Black, A.C. Lammert, B.R. Baucom, S.S. Narayanan, in *International Conference on Affective Computing and Intelligent Interaction (ACII)*. "That's aggravating, very aggravating": is it possible to classify behaviors in couple interactions using automatically derived lexical features? (Springer, Berlin, Germany, 2011), pp. 87–96. https://doi.org/10.1007/978-3-642-24600-5_12
18. J. Gibson, A. Katsamanis, M.P. Black, S. Narayanan, in *Twelfth Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Automatic identification of salient acoustic instances in couples' behavioral interactions using diverse density support vector machines, (Florence, Italy, 2011). https://doi.org/10.21437/interspeech.2011-470
19. C.C. Lee, A. Katsamanis, M.P. Black, B.R. Baucom, P.G. Georgiou, S.S. Narayanan, in *International Conference on Affective Computing and Intelligent Interaction (ACII)*. Affective state recognition in married couples' interactions using PCA-based vocal entrainment measures with multiple instance learning (Springer, Berlin, Germany, 2011), pp. 31–41, https://doi.org/10.1007/978-3-642-24571-8_4
20. S.N. Chakravarthula, R. Gupta, B. Baucom, P. Georgiou, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A language-based generative model framework for behavioral analysis of couples' therapy (IEEE, South Brisbane, Australia 2015), pp. 2090–2094. https://doi.org/10.1109/icassp.2015.7178339
21. K. Bartholomew, Avoidance of intimacy: an attachment perspective. J. Soc. Pers. Relatsh. **7**(2), 147–178 (1990)
22. L. Campbell, J.A. Simpson, J. Boldry, D.A. Kashy, Perceptions of conflict and support in romantic relationships: the role of attachment anxiety. J Personal. Soc. Psychol. **88**(3), 510 (2005)
23. K.A. Brennan, C.L. Clark, P.R. Shaver, Self-report measurement of adult attachment: An integrative overview, In J. A. Simpson & W. S. Rholes (Eds.), Attachment theory and close relationships, (The Guilford Press, 1998), pp. 46–76
24. K. Bartholomew, L.M. Horowitz, Attachment styles among young adults: a test of a four-category model. J. Personal. Soc. Psychol. **61**(2), 226 (1991)
25. P.M. Podsakoff, S.B. MacKenzie, J.Y. Lee, N.P. Podsakoff, Common method biases in behavioral research: a critical review of the literature and recommended remedies. J. Appl. Psychol. **88**(5), 879 (2003)
26. P.N. Juslin, P. Laukka, Communication of emotions in vocal expression and music performance: different channels, same code? Psychol. Bull. **129**(5), 770 (2003)
27. K.R. Scherer, Vocal communication of emotion: a review of research paradigms. Speech Commun. **40**(1–2), 227–256 (2003)
28. M. Spinelli, M. Fasolo, G. Coppola, T. Aureli, It is a matter of how you say it: verbal content and prosody matching as an index of emotion regulation strategies during the adult attachment interview. Int. J. Psychol. **54**(1), 102–107 (2019)
29. J.A. Crowell, D. Treboux, Y. Gao, C. Fyffe, H. Pan, E. Waters, Secure base behavior in adulthood: measurement, links to adult attachment representations, and relations to couples communication skills and self-reports. Dev. Psychol. **38**, 679–693 (2002)
30. L. Galili, O. Amir, E. Gilboa-Schechtman, Acoustic properties of dominance and request utterances in social anxiety. J. Soc. Clin. Psychol. **32**(6), 651–673 (2013)
31. M. Jasnow, S. Feldstein, Adult-Like Temporal Characteristics of Mother-Infant Vocal Interactions. Child Development **57**(3), 754–761 (1986). https://doi.org/10.2307/1130352
32. R. Feldman, C.W. Greenbaum, N. Yirmiya, Mother-infant affect synchrony as an antecedent of the emergence of self-control. Dev. Psychol. **35**(1), 223 (1999)
33. J. Jaffe, B. Beebe, S. Feldstein, C.L. Crown, M.D. Jasnow, P. Rochat, D.N. Stern, Rhythms of Dialogue in Infancy: Coordinated Timing in Development. Monographs of the Society for Research in Child Development **66**(2), i–149 (2001)
34. R. Feldman, Infant–mother and infant–father synchrony: the coregulation of positive arousal. Infant Ment. Health J. Off. Publ. World Assoc. Infant Ment. Health **24**(1), 1–23 (2003)
35. L.A. Sroufe, Attachment and development: a prospective, longitudinal study from birth to adulthood. Attach. Hum. Dev. **7**(4), 349–367 (2005)
36. W.R. Mills-Koonce, J.L. Gariepy, C. Propper, K. Sutton, S. Calkins, G. Moore, M. Cox, Infant and parent factors associated with early maternal sensitivity: a caregiver-attachment systems approach. Infant Behav. Dev. **30**(1), 114–126 (2007)
37. M. Harma, Utility of Vocal Synchrony Measure as an Indicator of Coregulation in Adult Attachment, Jan. 2020, https://doi.org/10.31234/osf.io/dncke
38. J. Kolacz, Investigating infant crying persistence and cry acoustic features as early risk indicators for social adjustment: developmental associations with infant vagal tone and attachment stress. Ph.D. thesis, The University of North Carolina at Chapel Hill (2016)
39. R. Feldman, Parent-infant synchrony and the construction of shared timing; physiological precursors, developmental outcomes, and risk conditions. J. Child Psychol. Psychiatr. **48**(3–4), 329–354 (2007)
40. R. Feldman, R. Magori-Cohen, G. Galili, M. Singer, Y. Louzoun, Mother and infant coordinate heart rhythms through episodes of interaction synchrony. Infant Behav. Dev. **34**(4), 569–577 (2011)
41. W. Tschacher, G.M. Rees, F. Ramseyer, Nonverbal synchrony and affect in dyadic interactions. Front. Psychol. **5**, 1323 (2014)
42. R.C. Fraley, N.G. Waller, K.A. Brennan, An item response theory analysis of self-report measures of adult attachment. J. Personal. Soc. Psychol. **78**(2), 350 (2000)
43. E. Selçuk, G. Günaydin, N. Sümer, A. Uysal, Yetişkin Bağlanma Boyutlan İçin Yeni Bir Ölçüm: Yakin İlişkilerde Yaşantilar Envanteri-II'nin Türk Örnekleminde Psikometrik Açıdan Değerlendirilmesi [A New Scale Developed to Measure Adult Attachment Dimensions: Experiences in Close Relationships-Revised (ECR-R) - Psychometric Evaluation in a Turkish Sample]. Türk Psikoloji Yazilari **8**(16), 1–11 (2005)
44. C. Heavey, D. Gill, A. Christensen, *Couples interaction rating system 2 (CIRS2)* (University of California, Los Angeles, 2002)
45. G. Gravier, M. Betser, M. Ben, AudioSeg: Audio segmentation toolkit release 1.2, IRISA, Jan. 2010.
46. C.C. Lee, E. Mower, C. Busso, S. Lee, S. Narayanan, Emotion recognition using a hierarchical binary decision tree approach. Speech Commun. **53**(9–10), 1162–1171 (2011)
47. O.W. Kwon, K. Chan, J. Hao, T.W. Lee, in *Eighth European Conference on Speech Communication and Technology (Eurospeech)*. Emotion recognition by speech signals, (Geneva, Switzerland, 2003). https://doi.org/10.21437/eurospeech.2003-80
48. F. Eyben, F. Weninger, F. Gross, B. Schuller, in *Proceedings of the 21st ACM international conference on Multimedia (MM)*. Recent developments in opensmile, the munich open-source multimedia feature extractor, (Barcelona, Spain, 2013), pp. 835–838. https://doi.org/10.1145/2502081.2502224
49. B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, S.S. Narayanan, in *Eleventh Annual Conference of the International Speech Communication Association (Interspeech)*. The INTERSPEECH 2010 paralinguistic challenge, Chiba, Japan, (2010). https://doi.org/10.21437/interspeech.2010-739
50. N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. **19**(4), 788–798 (2010)

51. Natural language api basics. (2020). https://cloud.google.com/natural-language/docs/basics. Accessed 10 Jul. 2020
52. H. Peng, Y. Fan, Feature selection by optimizing a lower bound of conditional mutual information. Inf. Sci. **418**, 652–667 (2017)
53. H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern. Anal. Mach. Intell. **27**(8), 1226–1238 (2005)
54. G. Brown, A. Pocock, M.J. Zhao, M. Luján, Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. J. Mach. Learn. Res. **13**(1), 27–66 (2012)
55. R.C. Fraley. Information on the experiences in close relationships-revised (ecr-r) adult attachment questionnaire. http://labs.psychology.illinois.edu/ rcfraley/measures/ecrr.htm. Accessed 9 Apr. 2020
56. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg et al., Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
57. X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, S.Y. Philip et al., Top 10 algorithms in data mining. Knowl. Inf. Syst. **14**(1), 1–37 (2008)
58. R. Díaz-Uriarte, S.A. De Andres, Gene selection and classification of microarray data using random forest. BMC Bioinformatics **7**(1), 3 (2006)
59. J. Franklin, The elements of statistical learning: data mining, inference and prediction. Math. Intell. **27**(2), 83–85 (2005)
60. J.H. Friedman, Stochastic gradient boosting. Comput. Stat. Data Anal. **38**(4), 367–378 (2002)
61. T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, Xgboost: extreme gradient boosting. R Package version 0.4-2. 1–4 (2015)
62. P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees. Mach. Learn. **63**(1), 3–42 (2006)
63. Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997)
64. A. Mukhopadhyay, U. Maulik, Unsupervised pixel classification in satellite imagery using multiobjective fuzzy clustering combined with svm classifier. IEEE Trans. Geosci. Remote. Sens. **47**(4), 1132–1138 (2009)
65. H. Li, B. Baucom, P. Georgiou, Sparsely connected and disjointly trained deep neural networks for low resource behavioral annotation: acoustic classification in couples' therapy. (2016). arXiv preprint arXiv:1606.04518
66. R.A. Berk, Statistical Learning as a Regression Problem, Springer Series in Statistics, pp. 1–48, 2008, https://doi.org/10.1007/978-0-387-77501-2_1
67. M.R. Segal, Machine learning benchmarks and random forest regression. Technical Report, Center for Bioinformatics & Molecular Biostatistics, University of California, San Francisco (2004)
68. A.M. Basbrain, J.Q. Gan, A. Sugimoto, A. Clark, in *2018 10th Computer Science and Electronic Engineering (CEEC)*. A neural network approach to score fusion for emotion recognition (IEEE, Colchester, UK 2018), pp. 180–185. https://doi.org/10.1109/ceec.2018.8674191
69. D.A. Ramli, S.A. Samad, A. Hussain, in *Proceedings of the International Workshop on Computational Intelligence in Security for Information Systems CISIS'08*. Score information decision fusion using support vector machine for a correlation filter based speaker authentication system (Springer, Berlin, Germany, 2009), pp. 235–242. https://doi.org/10.1007/978-3-540-88181-0_30
70. E. Alpaydin, Introduction to machine learning. (MIT press, 2020)
71. C. Leaper, R.D. Robnett, Women are more likely than men to use tentative language, aren't they? A meta-analysis testing for gender differences and moderators. Psychol. Women Q. **35**(1), 129–142 (2011)
72. T. Holtgraves, Social desirability and self-reports: testing models of socially desirable responding. Personal. Soc. Psychol. Bull. **30**(2), 161–172 (2004)
73. M.W. Baldwin, B. Fehr, On the instability of attachment style ratings. Pers. Relatsh. **2**(3), 247–261 (1995)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.