Psychopathology

Research Article

Psychopathology
DOI: 10.1159/000534811

Received: September 6, 2022
Accepted: October 16, 2023
Published online: November 27, 2023

Machine Learning Facial Emotion Classifiers in Psychotherapy Research: A Proof-of-Concept Study

Martin Steppan[a, b] Ronan Zimmermann[a, b] Lukas Fürer[b]
Matthew Southward[c] Julian Koenig[d, e] Michael Kaess[d, f]
Johann Roland Kleinbub[g] Volker Roth[h] Klaus Schmeck[b]

[a]Faculty of Psychology, University of Basel, Basel, Switzerland; [b]Psychiatric University Hospital, Basel, Switzerland; [c]Department of Psychology, University of Kentucky, Lexington, KY, USA; [d]University Hospital of Child and Adolescent Psychiatry and Psychotherapy, University of Bern, Bern, Switzerland; [e]Section for Experimental Child and Adolescent Psychiatry, Department of Child and Adolescent Psychiatry, Centre for Psychosocial Medicine, University of Heidelberg, Heidelberg, Germany; [f]Section for Translational Psychobiology in Child and Adolescent Psychiatry, Department of Child and Adolescent Psychiatry, Centre for Psychosocial Medicine, University of Heidelberg, Heidelberg, Germany; [g]Department of Philosophy, Sociology, Education and Applied Psychology, University of Padova, Padova, Italy; [h]Department of Mathematics and Informatics, University of Basel, Basel, Switzerland

Keywords

Adolescents · Borderline personality disorder · Emotions · Facial expressions classifiers · Psychotherapy

Abstract

Background: New advances in the field of machine learning make it possible to track facial emotional expression with high resolution, including micro-expressions. These advances have promising applications for psychotherapy research, since manual coding (e.g., the Facial Action Coding System), is time-consuming. ***Purpose:*** We tested whether this technology can reliably identify in-session emotional expression in a naturalistic treatment setting, and how these measures relate to the outcome of psychotherapy. ***Method:*** We applied a machine learning emotion classifier to video material from 389 psychotherapy sessions of 23 patients with borderline personality pathology. We validated the findings with human ratings according to the Clients Emotional Arousal Scale (CEAS) and explored associations with treatment outcomes. ***Results:*** Overall, machine learning ratings showed significant agreement with human ratings. Machine learning emotion classifiers, particularly the display of positive emotions (smiling and happiness), showed medium effect size on median-split treatment outcome (d = 0.3) as well as continuous improvement (r = 0.49, p < 0.05). Patients who dropped out form psychotherapy, showed significantly more neutral expressions, and generally less social smiling, particularly at the beginning of psychotherapeutic sessions. ***Conclusions:*** Machine learning classifiers are a highly promising resource for research in psychotherapy. The results highlight differential associations of displayed positive and negative feelings with treatment outcomes. Machine learning emotion recognition may be used for the early identification of drop-out risks and clinically relevant interactions in psychotherapy.

karger@karger.com
www.karger.com/psp

Karger
OPEN ACCESS

Correspondence to:
Martin Steppan, mhsteppan@gmail.com

Introduction

Facial expressions of emotion are evolved signals, showing homology not only in humans and primates but also in more distant species [1]. Despite this similarity, the assessment of specific emotions in psychiatric research and psychotherapy is complex and time-consuming. In psychotherapeutic settings, questionnaires have been developed to describe emotional states, although these questionnaires are usually completed after the emotional experience, potentially leading to recall bias [2]. To address this limitation, Ekman developed the Facial Action Coding System (FACS), which allows for a standardized and more precise quantification of facial expressions of emotions from images and videos [3]. FACS has been applied to psychotherapy research at an early stage [4], but it is almost impossible to manually rate the millions of frames that comprise videos of a full course of treatment from 1 patient, let alone larger samples. Hence, in psychiatric research, mainly indirect measures (e.g., vocal tone [5], heart rate variability [6], skin conductance [7], hormonal status [8], or neuroimaging [9]) have been used to get objective markers of individuals' emotional arousal.

Whereas previously, only indirect measures of emotional arousal (e.g., skin conductance) were available, new technologies make it possible to compare facial expressions with large libraries of prototypical emotions (e.g., happiness, disgust, surprise) in almost real-time [10], often referred to as facial emotion recognition (FER). This possibility offers a whole new area of research, in which displayed emotions can be tracked both within sessions and over the full course of treatment without requiring the same time investment from human raters. These algorithms have shown substantial agreement with Ekman's FACS, and commercial software products (e.g., FaceReader) have become available [11]. Researchers have demonstrated the utility of such classifiers to assess facial expressions of emotion in psychotherapeutic interventions among patients with borderline personality disorder (BPD) [12]. However, this study was based on only 29 psychotherapeutic sessions from 12 patients. Due to the novelty of this technology as opposed to verbal emotion expression [13], facial emotion expression is still largely understudied in psychiatric research.

Despite the early stage of this research, several recent studies have investigated the usefulness of emotion classifiers for diagnostic purposes in a psychiatric setting: Two independent studies used facial expression video analysis to detect depression in patients [14, 15]. Considering the vast amount of other data that might be available (medical records, biological markers, social variables) the use of artificial intelligence (AI) and deep learning is heavily discussed as a promising tool in psychiatry and psychiatric treatment [16]. Also within the diagnostic process and patient interactions, there are novel applications. Feedback from emotion classifiers has been used to train medical students to improve their interview technique [17]. Indeed, one study found good technological acceptance among counselors and general comfort among students emphasizing also the practical potential of this technology [18].

When moving from a monadic to a dyadic level, the analysis of body language, movement, and facial expression can go further, particularly for psychotherapeutic interventions. Considering that both patients' and therapists' signals can be analyzed at the same time, there is an increasing interest in the synchrony of these signals and the development of the rapport between patient and therapist. Previous research has focused on movement synchrony between patient and therapist, which was positively linked to therapeutic success [19], but also to the quality of the relationship [20]. Hence, first studies have investigated facial expressions as another sign of nonverbal synchrony and found positive associations with therapeutic alliance [21].

Like elsewhere, also in psychiatry AI is an emerging area of research [22]. In recent years, AI approaches have been used for the early diagnosis of psychiatric and neurodegenerative disorders (e.g., dementia) [23]. AI approaches have also been used in to personalize psychiatric treatments and promote "high performance medicine" [24] or "precision psychiatry" [25]. A prototypical example of this is an AI monitor suggested by Jan et al. which can predict Beck Depression Inventory II (BDI-II) scores from vocal and visual expressions [26]. AI approaches demonstrate at least three clear benefits to psychotherapy research. First, they offer comprehensive standardization. AI algorithms produce a standardized form of measurement that can incorporate more relevant inputs and dimensions than human raters can effectively process [27]. Second, AI approaches save time and boost resolution. Given that AI can accurately measure emotional change, it could also be used to complement or replace highly laborious human ratings with higher resolution information of emotions displayed by patients and therapists (for instance traditional FACS coding requires the manual scoring of about 30 frames of video for each second of therapy). Third, AI approaches can offer enhanced reliability. However, it is worth noting that not all AI methods are necessarily deterministic. While consistency (i.e., reliability) in outcomes is possible, it does not guarantee the validity of results.

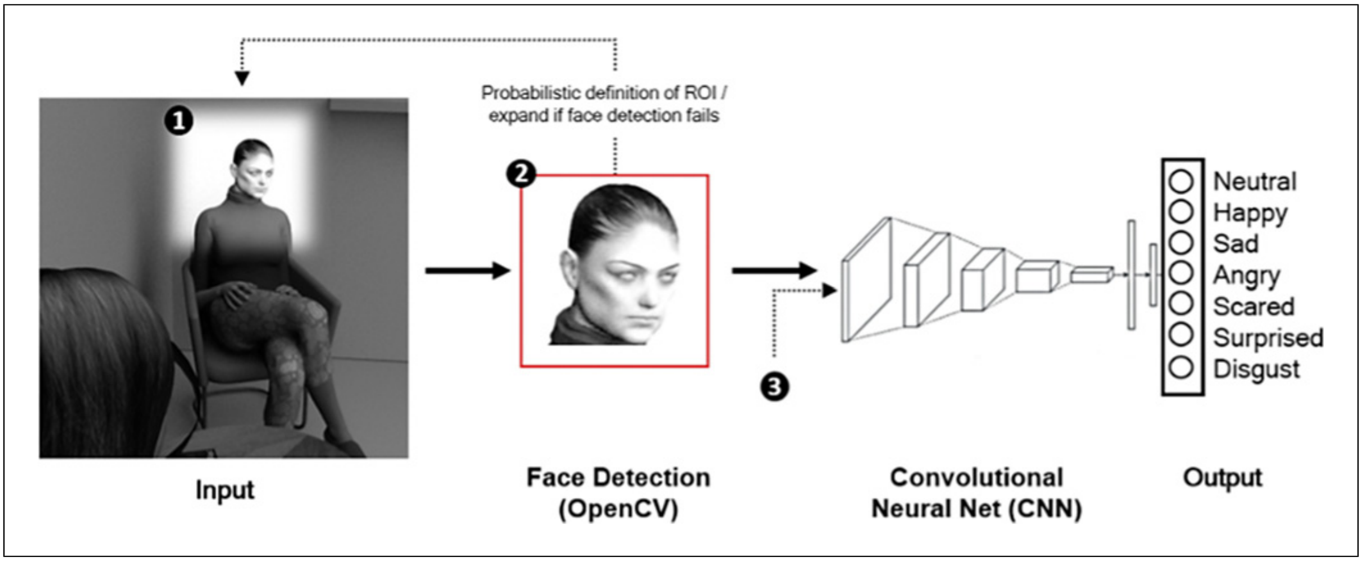
Probabilistic definition of ROI / expand if face detection fails
1
2
3
Neutral
Happy
Sad
Angry
Scared
Surprised
Disgust
Input
Face Detection (OpenCV)
Convolutional Neural Net (CNN)
Output

Fig. 1. Design for image processing. 1 = to reduce error, a region of interest (ROI) can be defined where a face is detected with high likelihood which is expanded if face detection fails; 2 = faces are detected and transformed into 48 × 48 pixels grayscale pictures; 3 = the convolutional neural net had been trained using the FER-2013 dataset.

Therefore, to enhance the practical applicability of this research, researchers need to demonstrate the external and criterion validity of AI, e.g., its utility for naturalistic treatment data. Thus, the aim of this study is to test (a) how emotion classifiers applied to videos of psychotherapy sessions compare to "gold standard" expert ratings of emotional arousal in patients and (b) whether emotion classifiers show clinically relevant associations with treatment outcomes.

Methods

Ethical Approval

This study is part of the multi-centre study "Evaluation of Adolescent Identity Treatment" that has been registered at clinicaltrials.gov (NCT02518906) [28, 29]. The current analyses are based on the entire available data collected at one participating center (Psychiatric University Hospitals, Basel). Ethical approval was obtained from the Local Ethics Committee. All adolescents, their parents, and the therapists provided written informed consent for participation.

Sample

The sample consists of 23 adolescent patients with borderline personality pathology, defined as (1) meeting DSM-IV criteria for BPD using the Structured Clinical Interview for DSM-IV Axis II Personality Disorders; SCID-II [30] and (2) presenting with identity diffusion according to the Assessment of Identity Development in Adolescence (AIDA; total T score >60) [31, 32]. The mean age of the patients was 16.2 years (SD = 1.6), 21 (91.3%) out of 23 patients were female. The original material consisted of footage of 423 therapy sessions (duration approx. 50 min each). Thirty-four videos were excluded mainly due to errors in face detection (e.g., the algorithm consistently detected a background object as a face) leading to a final sample of 389 analyzed psychotherapy sessions.

Video Processing and Machine Learning

Figure 1 illustrates the video processing design. OpenCV was used to detect faces [33]. Since faces were relatively stationary throughout the therapy sessions, face detection was optimized by dynamically defining an area of high likelihood of facial presence. If no face was detected for longer than thousand milliseconds, the region of interest was extended to ensure the accuracy of face detection. To validate this method, we saved an image of the selected region of interest every 100 frames for manual control of the procedure. To detect emotions, we implemented a pretrained model based on convolutional neural networks [34]. This model achieved an accuracy of 66% in correctly classifying emotions in the FER-2013 dataset. This accuracy rate is almost identical with human classification, which had been 65% ± 5% [35]. Although even more accurate models have been developed [36], we applied the algorithm by Arriaga et al. [34] which was designed for real-time use which is more suited for future applications in psychotherapy research and was therefore also computationally lighter than other algorithms.

This network was trained using the "FER-2013 dataset" (https://bit.ly/3gQLm9T), a collection of 35,685 facial images designed to develop machine learning algorithms for FER [35]. Similar to previously used algorithms (e.g., the software FaceReader [11]), probabilities of the occurrence of six basic emotions (happiness, surprise, anger, disgust, sadness, and fear), plus a percentage of neutral expression were extracted. Because we analyzed more than 1 month of total video material, and to reduce computing time, calculations were performed at sciCORE (http://scicore.unibas.ch/), a high-performance cluster computer and scientific computing center at the University of Basel.

Client Emotional Arousal Scale

Three observers were trained to code videos using the Client Emotional Arousal Scale (CEAS). For each 1 min interval, raters code (1) the intensity of the patient's arousal on a 1 (*low/no arousal*) to 7 (*full arousal*) ordinal Likert scale and (2) identify the predominant discrete emotion (e.g., joy, sadness, fear, anger, disgust, love). To note, a discrete emotion was not identified for intervals rated low in emotional arousal (<3). Raters first completed a training phase, in which 7 sessions were independently rated followed by a consensus rating. A total of 232 sessions were rated (on average 9.8 sessions per patient). Altogether 11,960 min-intervals (i.e., 186.5 h) of video material were rated manually. Inter-rater reliability of CEAS has been reported from 0.75 to 0.81 in previous studies [37, 38].

External Criteria for Reliability and Validity

To verify the validity of the FER results, we cross validated them with human-rated CEAS scores [39, 40]. To test for clinical validity (i.e., the importance of FER for therapeutic outcome), we regressed results of FER on pre- and post-treatment measures of several psychological questionnaires. These included five subscales of the Youth Outcome Questionnaire (Y-OQ) [41]; Intrapersonal Distress, Somatic, Interpersonal Relationships, Social Problems, Behavioral Dysfunction; four subscales of the Level of Personality Functioning Questionnaire (LoPFQ 12–18) [31]; Self Direction, Empathy, Identity, and Intimacy; as well as the Zanarini Rating Scale for BPD (ZAN-BPD) [42]. Six individuals dropped out of treatment, leaving 17 patients with treatment outcome data.

Statistical Analysis

We first tested the degree of agreement between FER and CEAS human ratings on unspecified emotional arousal using Pearson's correlations between CEAS ratings and the probability of a neutral facial expression according to FER ("1-neutral"). We controlled for multilevel effects by including a random intercept to allow for different baseline levels in arousal for different patient-therapist dyads, using the R package *lme4* [43]. To provide information regarding objectivity of FER, we report the correlation of FER nonspecific arousal ("1-neutral") for three different human raters. Fisher's *z* test was performed to test whether individual rater correlations differed from the overall trend. We then showcased the temporal agreement of nonspecific CEAS-arousal ratings and FER nonspecific arousal over the course of one session as an exemplar case.

For more specific human ratings using CEAS, we applied multilevel logistic regressions to determine how well FER predicted human classifications of emotions using *lme4* [43]. In addition to nonspecific CEAS ratings, human raters also labeled situations with specific emotions: anger, disgust, fear, joy, sadness, surprise, and love. We fit multilevel logistic regressions for each emotion separately including a random intercept for each patient. The presence of an emotion was used as the dependent binary variable, and seven FER scores were used as predictors. Sensitivity and specificity were calculated by adding up individual classification tables of each logistic regression model.

Regarding therapy outcome, we calculated a mean rank of every individual on all subscales of the questionnaires (YOQ, LoPF, ZAN-BPD; scored so higher scores indicate greater severity) before and after treatment. Change on these measures was calculated as the difference between these mean ranks, Δ. Using median-split, we transformed this de facto continuous delta also into two groups of "good outcome" (≥median) and "poor outcome" (<median). To illustrate the association of FER with therapy outcome, we plotted FER against each minute (1–50) in all therapy sessions and calculated independent samples *t*-tests, as well as Cohen's *d* for each comparison between the good outcome, poor outcome, and dropout group. Furthermore, to reduce complexity in the signals, we applied a principal component analysis (PCA) to all seven FER categories. The first two principal components (eigenvalues >1) were plotted for the different outcome groups (good, poor, dropout) to illustrate the relative localization of these groups within the higher order space of emotionality. Finally, we report a correlation matrix of FER categories with binary and continuous therapy outcome, using FER aggregated at the highest level (i.e., the level of patient-therapist dyads).

Except for the between-patient correlations, we analyzed relations involving FER by averaging FER scores across blocks of minutes. This choice allowed us to directly compare FER ratings with human CEAS ratings which were made across minute intervals. All analyses and visualization were done in R Studio using R (Version 3.5.3) [44]. We used the R package ggplot2 for all visualizations [45].

Results

Agreement between Human Raters and FER

Figure 2 illustrates the agreement of human raters and machine learning FER classifiers. Video sequences, which were rated as relatively highly emotionally arousing by humans (CEAS>3), were also classified as less neutral by FER. Figure 2a shows an example of one session with comparatively high agreement between human ratings and FER ($r = 0.50$). Figure 2b illustrates the overall correlation between FER and human CEAS ratings. The overall correlation (irrespective of rater) was $r = 0.08$, $p < 0.001$. Figure 2b also shows the agreement of FER with human ratings, broken down by three human raters who evaluated the video material. For all three human raters, the association was positive and significant (see also online suppl. Fig. 1; for all online suppl. material, see https://doi.org/10.1159/000534811). When applying a random effects model, controlling for baseline differences between raters and patients, the effect remained highly significant ($p < 0.001$). When comparing each human rater to the overall agreement between CEAS and FER (Fisher's *z*), all three raters showed essentially similar slopes (see Fig. 2b).

Table 1 illustrates the relative utility of each of the seven FER emotions for classifying specific emotions as labeled by human raters. On average, the correct classification was 76.6% for all human classifications. This percentage is relatively stable for all emotions, except for "disgust" (99.8%) and "surprise" (95.9%). In comparison to the other labels, these two emotions were rarely labeled, i.e., this high accuracy is mainly due to high specificity,

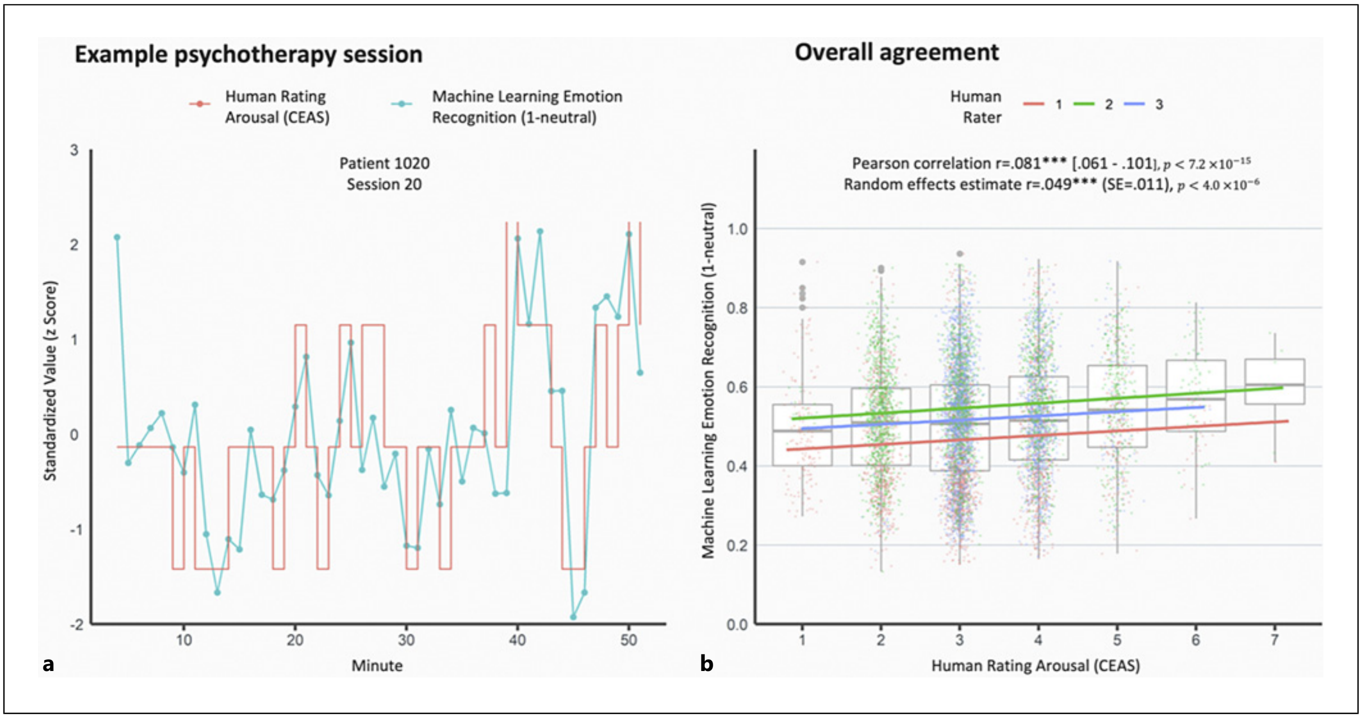

Fig. 2. Agreement between human ratings of emotional arousal (CEAS) and machine learning facial emotion recognition (FER). Nonspecific emotional arousal (1 – neutral). **a** Example plot of one session with high agreement. **b** Overall agreement between CEAS and FER for all cases and human raters. Colors indicate different human raters.

indicating overfitting for these two emotions. The average sensitivity was 0.70, and the average specificity was 0.78, indicating a relatively robust, however imperfect prediction of human ratings based on machine learning FER.

Association between FER and the Outcome of Psychotherapy

Figure 3a–g illustrates the display of emotions within the course of psychotherapeutic sessions, stratified by three outcome groups (good outcome, poor outcome, and dropout). All three groups show relatively distinct patterns, as well as some similar features regarding the display of basic emotions. For example, at the beginning of each session, all three groups show a relatively high level of happiness, which decreases after about 5 min (Fig. 3e). Only the dropout group shows almost no increase in happiness in the first 5 min of the interaction. Conversely, sadness and fear are relatively lower at the beginning of each session, and only reach their average level (horizontal lines) after 15–20 min, indicating psychotherapeutic work triggering also negative emotions.

We found distinct emotional expression in our good outcome, poor outcome, and the dropout group. For instance, all FER emotional categories except surprise, demonstrated significant between-group differences ($d > 0.10$). The most pronounced effect can be seen for happiness, which is significantly higher in patients with a good therapy outcome, compared to those with a poor outcome, $d = 0.30$, $p < 0.01$, and those who dropped out ($d = 0.32$, $p < 0.01$). In contrast, the poor outcome group demonstrated consistently and significantly more sadness, anger, disgust, and fear compared to the good outcome group. This corresponds with a PCA (Fig. 3h). Based on component loadings (see online suppl. Material Table S1), the first two principal components can be described as positive affectivity (positive loadings on happiness and nonspecific arousal) and negative affectivity (positive loadings on anger, sadness, fear, and nonspecific arousal). When plotting the factor scores of every minute analyzed using FER, the three outcome groups show relatively distinct, albeit overlapping areas of higher presence. Whereas the good outcome group (green dots) is more often located in the bottom right quadrant (positive affectivity), the poor outcome group shows a higher density in the upper two quadrants (negative affectivity). Those in the dropout group also demonstrated

Table 1. Multilevel logistic regressions to predict expert classifications from machine learning facial emotion recognition (FER)

Human classification	Total correct (%)	Sensitivity	Specificity	Labeled minutes
«Anger»	71.3	0.597	0.745	1,691
«Disgust»	99.8	0.750	0.998	8
«Fear»	71.6	0.655	0.733	1,718
«Joy»	75.1	0.711	0.756	795
«Sadness»	72.0	0.609	0.750	1,641
«Surprise»	95.9	0.714	0.960	14
«Love»	55.3	0.941	0.546	136
Neutral (CEAS <3)	72.1	0.637	0.749	1,948

Experts labeled the emotional state of patients during psychotherapy from video (minute-wise). Multilevel logistic regressions were applied to predict each classification from MLER taking into account random intercepts for $N = 23$ individuals and fixed effects for each MLER emotion (averaged per minute; max. 30 frames/second). Prediction was based on the maximum area under the curve (ROC, Youden-Index).

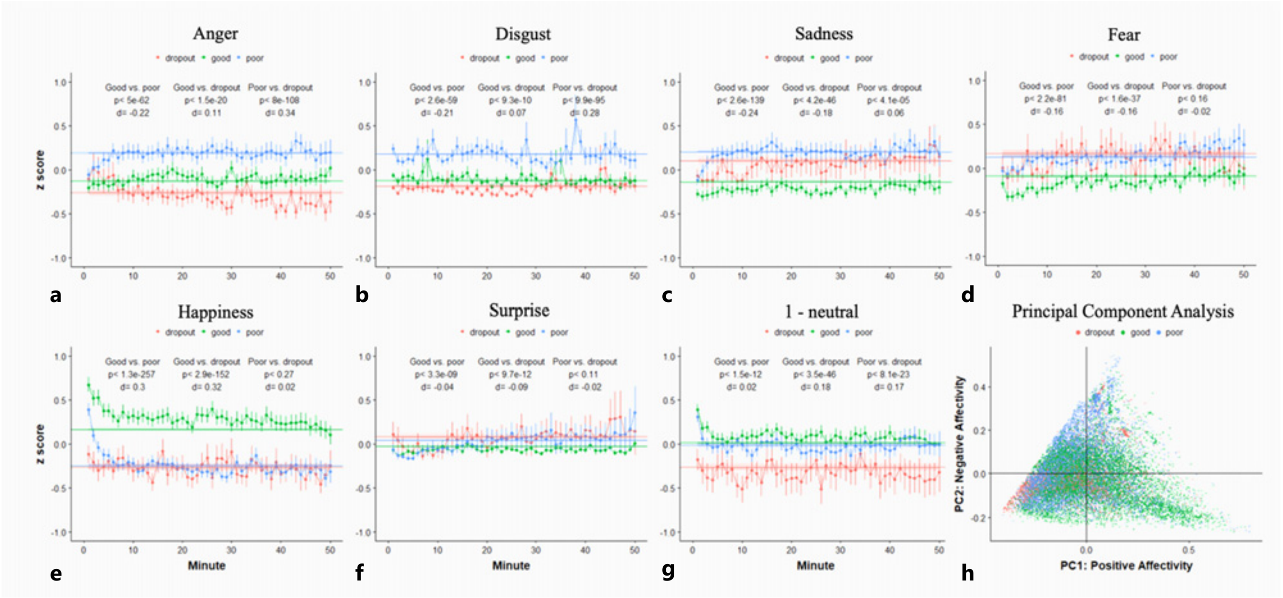

Fig. 3. a–g Emotional expression assessed with machine learning facial emotion recognition (FER) within all therapy sessions (per minute) by therapy outcome group (green = good, blue = poor, red = dropout; vertical lines indicate standard errors; standardized values). Test statistics are shown for each group comparison. **h** Scatter plot showing the results of a PCA on all minute intervals where FER scores were available.

scores in an area of higher density in the lower left quadrant, which can be described as neither positive nor negative affectivity. This is also in line with a generally reduced overall nonspecific arousal of the dropout group compared to both other groups (vs. good outcome: $d = 0.18$, $p < 0.01$; vs. poor outcome: $d = 0.17$, $p < 0.01$).

Table 2 shows a conservative estimate of potential associations of FER with pre- and post-treatment symptomatology, dropout, as well as the continuous improvement over therapy, Δ. Aggregated at the highest level (individuals), the display of happiness was significantly (negatively) correlated, $r = -0.53$, $p < 0.01$, with post-treatment symptomatology, but not pre-treatment

Table 2. Correlation matrix of machine learning facial emotion recognition (FER) and the outcome of psychotherapy

Emotion classifier	Pre	Post	Delta	Dropout
Angry	−0.12	0.32	0.29	−0.27
Disgust	−0.33	−0.05	−0.23	−0.33
Scared	0.26	0.26	−0.05	0.15
Happy	0.06	**−0.53***	**0.49***	−0.14
Sad	0.00	0.33	−0.42	0.02
Surprised	0.31	0.37	−0.13	0.11
Neutral	−0.07	0.15	−0.15	0.21

Pearson correlations. MLER aggregated on the level of individuals. Delta = pre – post symptomatology rank; unspecific arousal = "1-neutral." *p < 0.05.

symptomatology. Similarly, individual improvement, Δ, was significantly and positively correlated, $r = 0.49$, $p < 0.02$, with the display of happiness in treatment. For several other cells, numerically relevant, but not significant association are reported here, suggesting a general tendency for negative affectivity to be associated with less improvement (negative correlations for sadness, anger, disgust, and fear), as well as with lower nonspecific arousal (negative correlation with neutral).

Discussion

In a real-world examination of 389 psychotherapy sessions, we evaluated the efficacy of machine learning in emotion recognition for psychotherapy research. Our findings provide significant support for the technology's utility, further underscored by its external and ecological validity, as evidenced by notable correlations with human evaluations and tangible associations with therapy outcomes and patient dropout rates.

When benchmarking FER against the human ratings "gold standard" (CEAS) for patients' facial emotional expressions during psychotherapy [38], we noted consistent yet numerically modest correlations for specific emotions and nonspecific emotional arousal. This aligns with prior research demonstrating congruence between FER and human emotion ratings [11, 46], bolstering the argument for FER's satisfactory criterion validity, that is, its ability to accurately measure its intended target [47]. In analyzing the alignment of FER with varied human raters, we observed uniformly positive trends, with no significant deviation in individual rater trends from the general association. This suggests that FER showcases inter-rater agreement, meaning its measurements are largely consistent regardless of the observer it is benchmarked against. Nonetheless, it is worth highlighting the subdued correlations between FER and human CEAS ratings. This disparity seems rooted in differing evaluation scopes and timeframes. Human raters assessed emotions over minute-long intervals, whereas FER offered a granular view, capturing up to 1,800 frames (or 30 frames/second) within that span. Moreover, human assessors incorporated verbal emotion expressions in their evaluations. This difference in both timeframe and evaluation modality between FER and CEAS likely accounts for the modest yet statistically significant correlations observed.

Research comparing our findings is limited, given that this is the first study that applies FER to psychotherapeutic video materials of such duration. Our results echo previous studies that employed FACS in treatment sessions, emphasizing the significance of smiling and positive feelings in psychotherapy [4]. Notably, the display of happiness and smiling emerged as the most potent predictor of therapy outcomes, as well as patient dropout rates. The first 5 min of each psychotherapy session appear especially indicative (see Fig. 3e). Typically, interactions commence with a pronounced display of happiness, which diminishes rapidly within these initial minutes. This pattern may be attributed to the widespread social convention of beginning interactions with courtesy smiles or "social smiling." Intriguingly, patients who eventually drop out (represented by the red line) seem less inclined to adhere to this convention compared to the other groups. Clinically speaking, a patient's omission of these courtesy smiles at a session's onset is not merely indicative of "poor etiquette." It might instead signify a diminished outcome expectation from the patient's end, a factor we have identified as crucial in the treatment of BPD [48]. Future studies should explore the evolution of such courteous behaviors over time and their transformation during psychotherapeutic treatment.

Contrastingly, Arango et al. (2019) observed that, over the course of a session among 29 therapy sessions of 12 patients with BPD in Mexico, expressions of happiness and fear intensified while those of sadness diminished [12]. While this study involved a comparable patient population with BPD, the observed discrepancies might hint at cultural variations in emotional expression [49]. This research also resonates with the foundational concept of the "affective circumplex," which delineates the "arousal" and "valence" emotional dimensions [49]. The initial two components pinpointed through PCA align well with this theoretical framework. Both PC1 (indicative of positive emotionality) and PC2 (denoting negative emotionality) map onto the “valence” dimension, whereas “arousal” amplifies along the diagonal depicted in Figure 3h. While our study was not tailored explicitly to emulate this taxonomy, the findings are congruent with the bidimensional structure associated with facial expressions [50].

To assess the ecological validity of emotional expressions, we examined the correlations between FER and treatment outcomes. We identified small- to medium-sized associations between FER-rated expressions of all emotions except for surprise, and the shifts in outcomes throughout treatment. Notably, our findings somewhat correspond with prior studies and meta-analyses, which indicate that emotion suppression adversely affects psychotherapy outcomes [13, 51]. While we did not detect associations with pre-treatment symptomatology, we observed consistent group differences concerning post-treatment symptomatology, dropout rates, and individual progress (referred to as “delta”). These findings hint that emotional expressions might possess predictive value concerning treatment results. This should not be interpreted to suggest that specific emotions are direct causative factors. Rather, the intensity of these emotional expressions could signify potential causative elements, like treatment discomfort or the consistency of patient-therapist interactions. Supporting this, a past meta-analysis revealed that the quality of the relationship between the patient and therapist is pivotal for treatment success [52], which might be manifested in the more frequent display of positive emotions observed in this study.

Several strengths and limitations of our approach warrant mention: (1) computational load versus accuracy: the machine learning algorithms we used are open-source and designed for real-time emotion analysis. While these were chosen for their potential future accessibility to practitioners, more specialized algorithms, such as those in advanced robotics, might have produced more potent results [53]. Consequently, our applied FER might not fully capture the potential of this methodology; (2) generalizability versus individual tailoring: the algorithm was not customized to individual patients. Although tailoring the algorithm to each patient might yield stronger associations, we aimed to prevent overfitting and ensure broader applicability of the results; (3) sample specificity: our sample focused on adolescents with borderline personality pathology. Results may vary with other groups. Exploring emotional expression using FER across diverse patient populations and treatments seems a promising avenue for future work; (4) statistical consideration: our analysis involves data nested within 23 individuals. While we identified solid associations with post-treatment symptom changes, there were none with pre-treatment symptomatology. This hints that positive emotional displays might be more linked to the treatment process than to innate personality traits. However, with only 23 participants, our study was adequately powered only for strong correlations ($r >$ 0.55). Larger studies may unveil more nuanced relationships between facial expressions and therapy outcomes; (5) computational intensity: despite FER’s speed advantage over human coding, it remains computationally demanding. Utilizing this method in real-time during therapy sessions would necessitate further technological advancements; (6) dichotomization of outcomes: we dichotomized treatment outcomes mainly for visual clarity. While dichotomization can occasionally produce misleading results [52], our findings align with continuous psychotherapy outcomes; (7) data depth versus breadth: while our study furnishes rich longitudinal data, we concentrated on aggregate measures to predict broad symptom changes. Future investigations might delve into the longitudinal correlations between expressed emotions and specific therapeutic methods.

This pilot study advances the integration of machine learning emotion recognition in psychiatric and psychological research. We have delivered evidence highlighting FER’s alignment with core quality benchmarks such as reliability, objectivity, and prognostic validity. This suggests that FER could either complement or potentially surpass traditional emotion recognition methods reliant on physiological metrics like skin conductance and heart rate variability. Clinically, FER holds promise in pinpointing pivotal emotional sequences. With ongoing advancements minimizing computational demands, FER could serve in individual-focused studies and assist mental health professionals in their supervision roles. There is potential for this technique to function as a clinical asset, spotlighting significant emotional junctures that might otherwise be overlooked, thereby enriching the therapeutic experience and possibly bolstering treatment outcomes. Given our study’s promising findings, we advocate for a collaborative international approach. Analogous to efforts in other research arenas, aggregating anonymized FER data could amplify statistical robustness and shed more light on the technique’s applicability across diverse disorders and age brackets. We welcome engagement from researchers possessing video footage of psychotherapeutic sessions.

Acknowledgments

Calculations were performed at sciCORE (http://scicore.unibas.ch/) scientific computing center at University of Basel.

Statement of Ethics

This study is part of the multi-centre study "Evaluation of Adolescent Identity Treatment" registered at clinicaltrials.gov (NCT02518906). Ethical approval has been obtained from the Local Ethics Committee (Ethikkommission Nordwest-und Zentralschweiz: Nr.: 2015–230). All adolescents, their parents, and the therapists provided written informed consent for participation. Post-trial care is ensured by the local clinical teams.

Conflict of Interest Statement

The authors have no conflicts of interest to disclose.

Funding Sources

The study was funded by the University of Basel (https://www.unibas.ch/en.html) and the Psychiatric University Clinics of the University of Basel (https://www.upk.ch/en/home.html). Additionally, we would like to thank the "Thalmann-Stiftung" (https://www.fundraiso.ch/sponsor/thalmann-stiftung) and the "Stiftung zur Förderung von Psychiatrie und Psychotherapie" (https://stiftung-fpp.ch) for the financial support. Finally, we want to thank the open access funding provided by the University of Basel (https://www.oa-fund.unibas.ch/de/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

Martin Steppan: design, data analysis, writing; Ronan Zimmermann and Lukas Fürer: design, writing; Matthew Southward, Johann Roland Kleinbub, and Volker Roth : writing; Julian Koenig, Michael Kaess, and Klaus Schmeck: design, data collection, writing.

Data Availability Statement

Due to ethical and privacy reasons and an unknown risk how unique the flow of emotional expressions in psychotherapy is (it might be as unique as a fingerprint), the data used in this study are currently not made publicly available. A preprint preliminary version of this manuscript is available on PsyAxiv [54].

References

1 Waller BM, Julle-Daniere E, Micheletta J. Measuring the evolution of facial "expression"using multi-species FACS. Neurosci Biobehav Rev. 2020;113:1–11.

2 Schiepek G, Stöger-Schmidinger B, Kronberger H, Aichhorn W, Kratzer L, Heinz P, et al. The Therapy Process Questionnaire-Factor analysis and psychometric properties of a multidimensional self-rating scale for high-frequency monitoring of psychotherapeutic processes. Clin Psychol Psychother. 2019;26(5):586–602.

3 Ekman P, Friesen WV. Facial action coding systems. Consulting Psychologists Press; 1978.

4 Bänninger-Huber E. Prototypical affective microsequences in psychotherapeutic interaction. In: Ekman P, Rosenberg E, editors. What the face reveals Oxford University Press; 1997. p. 414–33.

5 Soma CS, Baucom BR, Xiao B, Butner JE, Hilpert P, Narayanan S, et al. Coregulation of therapist and client emotion during psychotherapy. Psychother Res. 2020;30(5):591–603.

6 Dikecligil GN, Mujica-Parodi LR. Ambulatory and challenge-associated heart rate variability measures predict cardiac responses to real-world acute emotional stress. Biol Psychiatry. 2010;67(12):1185–90.

7 Kleinbub JR. State of the art of interpersonal physiology in psychotherapy: a systematic review. Front Psychol. 2017;8:2053.

8 Schumacher S, Niemeyer H, Engel S, Cwik JC, Knaevelsrud C. Psychotherapeutic treatment and HPA axis regulation in posttraumatic stress disorder: a systematic review and meta-analysis. Psychoneuroendocrinology. 2018;98:186–201.

9 Oathes DJ, Ray WJ, Yamasaki AS, Borkovec TD, Castonguay LG, Newman MG, et al. Worry, generalized anxiety disorder, and emotion: evidence from the EEG gamma band. Biol Psychol. 2008;79(2):165–70.

10 Jain N, Kumar S, Kumar A, Shamsolmoali P, Zareapoor M. Hybrid deep neural networks for face emotion recognition. Pattern Recognit Lett. 2018;115:101–6.

11 Skiendziel T, Rösch AG, Schultheiss OC. Assessing the convergent validity between the automated emotion recognition software noldus FaceReader 7 and facial action coding system scoring. PloS One. 2019;14(10):e0223905.

12 Arango I, Miranda E, Sánchez Ferrer JC, Fresán A, Reyes Ortega MA, Vargas AN, et al. Changes in facial emotion expression during a psychotherapeutic intervention for patients with borderline personality disorder. J Psychiatr Res. 2019;114:126–32.

13 Peluso PR, Freund RR. Therapist and client emotional expression and psychotherapy outcomes: a meta-analysis. Psychotherapy. 2018;55(4):461–72.

14 Wang Q, Yang H, Yu Y. Facial expression video analysis for depression detection in Chinese patients. J Vis Commun Image Represent. 2018;57:228–33.

15 de Melo WC, Granger E, Hadid A. A deep multiscale spatiotemporal network for assessing depression from facial dynamics. IEEE Trans Affect Comput. 2022;13(3):1581–92.

16 Su C, Xu Z, Pathak J, Wang F. Deep learning in mental health outcome research: a scoping review. Transl Psychiatry. 2020;10(1):116.

17 Yatera D, Nishiya K, Karouji Y, Kojiri T. Facial expression visualization system for medical interview practice support. Proced Comput Sci. 2019;159:1986–94.

18 Gom-os DFK, Yong KY. An empirical study on the use of a facial emotion recognition system in guidance counseling utilizing the technology acceptance model and the general comfort questionnaire. Appl Comput Inform. 2022.

19 Altmann U, Schoenherr D, Paulick J, Deisenhofer A-K, Schwartz B, Rubel JA, et al. Associations between movement synchrony and outcome in patients with social anxiety disorder: evidence for treatment specific effects. Psychother Res. 2020;30(5):574–90.

20 Ramseyer F, Tschacher W. Nonverbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome. J Consult Clin Psychol. 2011;79(3):284–95.

21 Yokotani K, Takagi G, Wakashima K. Nonverbal synchrony of facial movements and expressions predict therapeutic alliance during a structured psychotherapeutic interview. J Nonverbal Behav. 2020;44(1):85–116.

22 Durstewitz D, Koppe G, Meyer-Lindenberg A. Deep neural networks in psychiatry. Mol Psychiatry. 2019;24(11):1583–98.

23 Egger K, Rijntjes M. Big data and artificial intelligence for diagnostic decision support in atypical dementia. Nervenarzt. 2018;89(8):875–84.

24 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56.

25 Bzdok D, Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. Biol Psychiatry Cogn Neurosci Neuroimaging. 2018;3(3):223–30.

26 Jan A, Meng H, Gaus YFBA, Zhang F. Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. IEEE Trans Cogn Dev Syst. 2018;10(3):668–80.

27 Dawes TR, Eden-Green B, Rosten C, Giles J, Governo R, Marcelline F, et al. Objectively measuring pain using facial expression: is the technology finally ready? Pain Manag. 2018;8(2):105–13.

28 Schmeck K, Pick OG, Milidou M, Schenk N, Schlüter-Müller S, Zimmermann R. Früherkennung von persönlichkeitsstörungen. Persönlichkeitsstörungen Theor Ther. 2018;22(3):179–85.

29 Zimmermann R, Fürer L, Schenk N, Koenig J, Roth V, Schlüter-Müller S, et al. Silence in the psychotherapy of adolescents with borderline personality pathology. Personal Disord. 2021;12(2):160–70.

30 First MB, Spitzer RL, Gibbon M, Williams JB. The structured clinical interview for DSM-III-R personality disorders (SCID-II). Part I: description. J Pers Disord. 1995;9(2):83–91.

31 Goth K, Foelsch P, Schlüter-Müller S, Birkhölzer M, Jung E, Pick O, et al. Assessment of identity development and identity diffusion in adolescence-Theoretical basis and psychometric properties of the self-report questionnaire AIDA. Child Adolesc Psychiatry Ment Health. 2012;6:27–16.

32 Lind M, Vanwoerden S, Penner F, Sharp C. Inpatient adolescents with borderline personality disorder features: identity diffusion and narrative incoherence. Personal Disord. 2019;10(4):389–93.

33 Bradski G, Kaehler A. Learning OpenCV: computer vision with the OpenCV library. O'Reilly Media, Inc.; 2008.

34 Arriaga O, Valdenegro-Toro M, Plöger P. Real-time convolutional neural networks for emotion and gender classification. 2017. ArXiv171007557 [Preprint].

35 Goodfellow IJ, Erhan D, Carrier PL, Courville A, Mirza M, Hamner B, et al. Challenges in representation learning: a report on three machine learning contests. Neural information processing: 20th international conference, ICONIP 2013, daegu, korea, november 3-7, 2013. Proceedings, Part III 20 Springer; 2013. p. 117–24.

36 Abdullah SMS, Abdulazeez AM. Facial expression recognition based on deep learning convolution neural network: a review. J Soft Comput Data Min. 2021;2(1):53–65.

37 Missirlian TM, Toukmanian SG, Warwar SH, Greenberg LS. Emotional arousal, client perceptual processing, and the working alliance in experiential psychotherapy for depression. J Consult Clin Psychol. 2005;73(5):861–71.

38 Greenberg L, Auszra L, Herrmann I. The relationship among emotional productivity, emotional arousal and outcome in experiential therapy of depression. Psychother Res. 2007;17(6):737–93.

39 Rosner R. The relationship between emotional expression, treatment and outcome in psychotherapy: an empirical study. Peter lang gmbh. Internationaler Verlag Der Wissenschaften; 1996.

40 Machado P. Clients emotional arousal in therapy: development of a rating scale. Unpubl Manuscr. 1992.

41 Burlingame GM, Wells MG, Lambert MJ, Cox JC. Youth outcome questionnaire (Y-OQ). In: Maruish ME, editor. The use of psychological testing for treatment planning and outcomes assessment Routledge; 2014. p. 235–73.

42 Zanarini MC, Vujanovic AA, Parachini EA, Boulanger JL, Frankenburg FR, Hennen J. Zanarini rating scale for borderline personality disorder (ZAN-BPD): a continuous measure of DSM-IV borderline psychopathology. J Pers Disord. 2003;17(3):233–42.

43 Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. 2014. ArXiv14065823 [Preprint].

44 Allaire J. RStudio: integrated development environment for R. Boston MA; 2012. Vol. 770(394). p. 165–71.

45 Wickham H. ggplot2. WIREs Comput Stats. 2011;3(2):180–5.

46 Haines N, Southward MW, Cheavens JS, Beauchaine T, Ahn W-Y. Using computer-vision and machine learning to automate facial coding of positive and negative affect intensity. PLoS One. 2019;14(2):e0211735.

47 Bornovalova MA, Choate AM, Fatimah H, Petersen KJ, Wiernik BM. Appropriate use of bifactor analysis in psychopathology research: appreciating benefits and limitations. Biol Psychiatry. 2020;88(1):18–27.

48 Bäumer A-V, Fürer L, Birkenberger C, Wyssen A, Steppan M, Zimmermann R, et al. The impact of outcome expectancy on therapy outcome in adolescents with borderline personality disorder. Borderline Personal Disord Emot Dysregul. 2022;9(1):30–11.

49 Russell JA. A circumplex model of affect. J Pers Soc Psychol. 1980;39(6):1161–78.

50 Calder AJ, Burton AM, Miller P, Young AW, Akamatsu S. A principal component analysis of facial expressions. Vis Res. 2001;41(9):1179–208.

51 Scherer A, Boecker M, Pawelzik M, Gauggel S, Forkmann T. Emotion suppression, not reappraisal, predicts psychotherapy outcome. Psychother Res. 2017;27(2):143–53.

52 Maxwell SE, Delaney HD. Bivariate median splits and spurious statistical significance. Psychol Bull. 1993;113(1):181–90.

53 Ruiz-Garcia A, Elshaw M, Altahhan A, Palade V. A hybrid deep learning neural approach for emotion recognition from facial expressions for socially assistive robots. Neural Comput Appl. 2018;29(7):359–73.

54 Steppan M, Zimmermann R, Fürer L, Schenk N, Schmeck K. Machine learning facial emotion recognition in psychotherapy research. A useful approach? PsyArxiv. 2020.