Schizophrenia Research: Cognition 25 (2021) 100196

Contents lists available at ScienceDirect

Schizophrenia Research: Cognition

journal homepage: www.elsevier.com/locate/scog

Research Paper

Computational methods for integrative evaluation of confidence, accuracy, and reaction time in facial affect recognition in schizophrenia

Varsha D. Badal [a,b], Colin A. Depp [a,b,c,*], Peter F. Hitchcock [d], David L. Penn [e,f], Philip D. Harvey [g,h], Amy E. Pinkham [i,j]

[a] Department of Psychiatry, University of California San Diego, San Diego, CA, United States of America
[b] Sam and Rose Stein Institute for Research on Aging, University of California San Diego, San Diego, CA, United States of America
[c] VA San Diego Healthcare System, La Jolla, CA, United States of America
[d] Brown University, Providence, RI, United States of America
[e] Department of Psychology, University of North Carolina, Chapel Hill, NC, United States of America
[f] School of Psychology, Australian Catholic University, Melbourne, VIC, Australia
[g] Department of Psychiatry and Behavioral Sciences, University of Miami Miller School of Medicine, Miami, FL, United States of America
[h] Research Service, Miami VA Healthcare System, United States of America
[i] School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX, United States of America
[j] Department of Psychiatry, University of Texas Southwestern Medical School, Dallas, TX, United States of America

ARTICLE INFO

Keywords:
Machine learning
Neural networks
Social cognition
Psychosis

ABSTRACT

People with schizophrenia (SZ) process emotions less accurately than do healthy comparators (HC), and emotion recognition have expanded beyond accuracy to performance variables like reaction time (RT) and confidence. These domains are typically evaluated independently, but complex inter-relationships can be evaluated through machine learning at an item-by-item level. Using a mix of ranking and machine learning tools, we investigated item-by-item discrimination of facial affect with two emotion recognition tests (BLERT and ER-40) between SZ and HC. The best performing multi-domain model for ER40 had a large effect size in differentiating SZ and HC ($d = 1.24$) compared to a standard comparison of accuracy alone ($d = 0.48$); smaller increments in effect sizes were evident for the BLERT ($d = 0.87$ vs. $d = 0.58$). Almost half of the selected items were confidence ratings. Within SZ, machine learning models with ER40 (generally accuracy and reaction time) items predicted severity of depression and overconfidence in social cognitive ability, but not psychotic symptoms. Pending independent replication, the results support machine learning, and the inclusion of confidence ratings, in characterizing the social cognitive deficits in SZ. This moderate-sized study ($n = 372$) included subjects with schizophrenia (SZ, $n = 218$) and healthy controls (HC, $n = 154$).

1. Introduction

The ability to recognize emotions in faces is key to healthy social function. Diminished abilities in facial affect recognition have been associated with SZ since Kraepelin and Bleuler's early clinical observations in the late 19th century (Maatz et al., 2015; Kerr and Neale, 1993; Schneider et al., 1995; Mandal et al., 1998; Hooker and Park, 2002; Gur et al., 2007). Several emotion recognition and discrimination tasks have been designed to assess facial affect recognition with a variety of emotions represented (Gur et al., 2002; Bell et al., 1997; Kerr and Neale, 1993). Social COgnition Psychometric Evaluation SCOPE (Pinkham et al., 2017), a comprehensive psychometric study of social cognitive instruments, evaluated effect sizes for the accuracy of emotion recognition alongside other social cognition measures. Significant differences between SZ and HC with medium effect sizes were seen comparing accuracy Bell Lysaker Emotion Recognition Test (BLERT) and Emotion Recognition 40 (ER-40). However, differences in accuracy scores across individual emotions were more variable, with the majority of emotions failing to significantly differentiate groups (Pinkham et al., 2019). The parameters of reaction times while making emotion recognition judgements was different across the groups, but to a lesser extent than accuracy (Cornacchio et al., 2017). Although confidence ratings were

* Corresponding author at: Department of Psychiatry, San Diego VA, Stein Institute for Research on Aging, Department of Psychiatry (0664), University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0664, United States of America.
E-mail address: cdepp@ucsd.edu (C.A. Depp).

https://doi.org/10.1016/j.scog.2021.100196
Received 14 October 2020; Received in revised form 6 March 2021; Accepted 10 March 2021
Available online 22 April 2021

somewhat less able to differentiate HC and SZ compared to accuracy, correlations between confidence and real world outcomes shared 12–16% of the variance (compared to 1% for accuracy). Therefore, accuracy can be integrated with other concurrent performance metrics, including reaction time (RT) and confidence ratings (CR), in order to gain a deeper understanding of social cognitive problems.

Confidence ratings (CR) may serve as a particularly informative index for understanding emotion processing difficulties in schizophrenia; a subset of SZ show overconfidence on performance (Jones et al., 2019), yet as an index with non-linear effects (Silberstein et al., 2018). In a previous publication from the SCOPE study (Jones et al., 2019), CR in facial affect recognition was found to be diminished on average in SZ, as was accuracy. However, a sizeable proportion of SZ had extremely exaggerated confidence ratings, with 13% of the participants stating that they believed that they were 100% confident that they were correct on every item (compared to 1.4% HC). These patients were also the least accurate within the sample.

Confidence also appears to be related to RT and clinical symptoms as well. For example, Jones et al. reported a positive correlation between increased confidence and faster RT in HC (Jones et al., 2019), and RT, when used as a proxy for effort, converged with the difficulty of the items for HCs only. Clinical factors, such as presence of psychotic symptoms and delusion may increase reported confidence (Hoven et al., 2019) while anxiety and paranoia may reduce it (Hoven et al., 2019; Jones et al., 2019). Separate analyses also indicated that reduced confidence was associated with higher levels of depression in both people with SZ and the HC sample in this study were found to be associated with reduced confidence (Oliveri et al., 2019) and, in people with SZ, reduced confidence rates were linked to more accurate self-assessments of their everyday functioning (Harvey et al., 2019). Taken together, confidence, reaction time, and accuracy are related to one another in non-linear and complex ways in terms of both differentiating SZ from HC and in association with symptoms, with both underestimation and overestimation of abilities correlating with lower accuracy (Harvey et al., 2019). However, in these studies, confidence ratings were averaged across different emotions, which may obscure other important processes.

Computational approaches may prove useful for disentangling complex inter-related constructs in facial affect recognition, including accuracy, reaction time, and confidence at the item/discrete emotion level. Machine learning (ML) is a subfield of artificial intelligence (AI) and is a set of generalized software tools that enable learning of patterns in the data, with those patterns subsequently used to predict or classify new data. ML optimizes predictions by considering all features simultaneously, such as item-by-item data on confidence, RT and accuracy, and exploits learned patterns among them. It has some similarities to statistical regression but differs in its ability to handle high-dimensional, non-linear relationships. To date, we are aware of no studies that have attempted to apply ML to disentangling reaction time, confidence, and accuracy of facial affect recognition in SZ.

In this study, we extended the results of (Jones et al., 2019) by applying ML techniques to the ER40 and BLERT facial recognition tasks in the SCOPE study ($n = 372$; 218 SZ, 154 HC) using item-by-item level data across different emotions on accuracy, reaction time, and confidence. Our overall goal was to examine whether simultaneous evaluation of accuracy, confidence, and reaction time could shed light on the relative impact of these metrics in discriminating SZ from HC. Specifically, we a) identified the optimal set of variables based on each of ER40 and BLERT items that distinguished SZ from HC considering accuracy, reaction time, and confidence across the different emotions, b) evaluated the performance of the resulting feature set to discriminate SZ from HC, and c) repeated these steps within SZ sample alone to assess the ability of features to distinguish the severity of positive and negative symptoms, severity of depression, and self- and informant-rated social cognitive ability. We hypothesized that the resulting ML models with selected features would be more sensitive in discrimination of HC and SZ than traditional statistical comparison of each dimension alone (i.e., accuracy, CR, RT).

2. Method

2.1. Sample characteristics

The dataset was derived from the Social Cognition Psychometric Evaluation study, final Validation Study (SCOPE-5; (Pinkham et al., 2018a), which included patients with diagnoses of SZ or schizoaffective disorder ($n = 218$) and healthy controls ($n = 154$). Data were collected data from three sites: The University of Texas at Dallas (UTD), The University of Miami Miller School of Medicine (UM), and The University of North Carolina at Chapel Hill (UNC) and the study was approved by the Institutional Review Board at each site. All participants provided written informed consent. Details on recruitment and inclusion/exclusion criteria are available elsewhere (Pinkham et al., 2018a).

2.2. Facial affect recognition measures

We limited our analysis to two of the social cognition measures covered by the SCOPE-5 study as each measure was modified to include item-by-item accuracy, reaction time, and confidence dimensions. Participants were instructed to respond as rapidly as possible to the items while maintaining accuracy. We also limited analyses to visit one of the SCOPE study, as participants completed these same measures a second time for the purpose of evaluating test-retest reliability. For each item on these measures, participants responded as rapidly as possible, provided an emotion recognition response, and generated a rating of confidence in the correctness of that response on a scale from 0 (not confident at all) to 100 (extremely confident). Participants were not provided performance feedback.

2.2.1. Penn Emotion Recognition Task (ER-40)

The ER40 asks participants to assign an emotion to 40 photographs depicting expressions of one of 5 emotions (happiness, sadness, anger, fear and neutral), 4 each of low and high intensity (Gur et al., 2002).

2.2.2. Bell Lysaker Emotion Recognition Task (BLERT)

The task involves identification of 7 emotions, 3 each of (happiness, sadness, fear, disgust, surprise, anger, or neutral) from 21 audio-visual clips of an actor (Bell et al., 1997). Thus, there were 5 overlapping emotions and 2 additional emotions in the BLERT that were not in the ER-40. Evidence suggests that surprise and its initial perception may be negative (Noordewier and Breugelmans, 2013), hence it is considered being of negative polarity We included neutral emotion in the positive set or the absence of negative.

2.3. Psychiatric symptoms and social cognition

2.3.1. Positive and Negative Syndrome Scale (PANSS)

The PANSS assesses the severity of 7 positive symptoms (including hallucinations and delusions), 7 negative symptoms (loss off normal functions) and 16 items of general psychopathology (Kay et al., 1987). We used the PANSS positive and negative syndrome subscales and, for classification models, we used a cut-off of 15, the defined boundary between mild to moderate symptom severity on the scale. We also evaluated the PANSS Reduced Emotional Experience factor (Emotional Withdrawal (N2), Passive-apathetic Social Withdrawal (N4) and Active social avoidance (G16)) were correlated to functional deficits in SZ (Strassnig et al., 2018; Harvey et al., 2017) as an outcome, with any of the factors (N2, N4 or G16) equal to or surpassing a cut off of 4.

2.3.2. Beck Depression Inventory (BDI) – 2

BDI-2 was used to assess depressive symptom severity (Beck et al., 1996). BDI-2 contains 21 items with 0–4 scale rating. For classification models, we used scores of 9, 15, 19 and 29 corresponding to minimal,

mild, moderate and severe levels respectively (Chemerinski et al., 2008).

2.3.3. Observable Social Cognition (OSCARS)

OSCARS is an 8 item self or informant-rated instrument addressing social cognitive abilities (Healey et al., 2015). Both self- and informant based assessments were to the SZ patients in SCOPE, for both sources we employed the suggested optimal cutoff based upon the Youden Index of 17 (Healey et al., 2015). Informants, as described by (Pinkham et al., 2018a) included high contact clinicians, family members or other close associates.

2.4. Analysis

For comparison to machine learning methods, we computed traditional, univariate statistical measures: *t*-tests and measures of effect size (Cohen's d and AUC). We compared these traditional approaches to machine learning approaches. The key steps of this procedure are to select a set of best predictors (feature selection) and evaluate how well these features distinguish (classify) SZ from HC.

For feature selection, we used the Gini index (Kantardzic, 2011) to each of the 40 items in the ER40 (120 total features, 40 of each of response accuracy, reaction time and confidence score and the 21 items in the BLERT (63 features for BLERT, 21 each for response accuracy, reaction time and confidence score). The choice of Gini index was based on it being one of the fast and recommended filtering techniques (Bommert et al., 2020) compared to information gain (and several others). Additionally, it is not adversely affected by imbalance in sample sizes (Dubey et al., 2014). The calculation of the Gini index involves subtracting sum of squared probabilities of each class from one. Gini index for dataset (S) may be defined as (Kantardzic, 2011):

$$Gini(S) = 1 - \sum_{i=0}^{c-1} p_i^2$$

where

c = number of predefined classes {here: SZ, HC}, C_i are classes for i = 1, …, c-1, s_i = number of samples belonging to class C_i,

$p_i = s_i/S$ is a relative frequency of class C_i in the set {here: SZ, HC}.

From the ranked the set of all features we selected the subset that produced the best ML results empirically (features were incrementally included, highest to the lowest, until a decline in performance was observed). We contrasted the distribution of selected features with the distribution in the superset (e.g., 40 each of confidence, accuracy, and RT for the ER40; 24 negative emotions and 16 neutral/happy emotions).

Next, we submitted the subset of top Gini-ranked features utilizing several ML algorithms (see Appendix A) to evaluation discrimination between SZ and HC. Feature ranking, selection (GINI index) and ML implementations (several models), are available in Orange (Demsar et al., 2013). ML algorithms applied were Neural Network (with various activation functions such as ReLu, Tanh, Logistic), SVM (with RBF, Linear and Polynomial kernels), KNN (k-nearest neighbor), AdaBoost, decision tree, Naïve Bayes, Random Forest and stacked models. The data was split 80%–20% randomly, where 80% was used for training while the 20% was used for testing. We used a grid search, a simple approach for hyperparameter estimation (when dealing with a small enumerable set) in ML (Claesen and De Moor, 2015). We identified the approximate number of features, in multiples of 10, that worked best and subsequently refined the value by trying all the values in the neighborhood. We used Orange (Demsar et al., 2013) for test and performance results as well (best of 5 trial runs are reported). To estimate performance for ML models, we used the F1 score, AUC and equivalent Cohen's *d* using conversion tables (Salgado, 2018) for comparison to traditional statistics. The performance measures shown are average over classes and computed as documented (Demsar et al., 2013; Pedregosa et al., 2011; Scikit-Learn, n.d.).

Finally, the same set of features and identical methodology were extended to distinguish (or predict targets) based on cutoffs scores on PANSS, OSCARS Self-Rated, OSCARS Informant and derived measures such as OSCARS assessment inaccuracy and overconfidence. OSCARS assessment inaccuracy is defined as the difference between OSCARS Informant and OCARS Self-Rated. For overconfidence, we limit our analysis to SZ subjects with OSCARS Self-Rated >OSCARS Informant.

3. Results

3.1. Traditional statistical comparisons across ER40 and BLERT emotion categories (Tables 1, 2)

t-tests indicated that the Cohen's d for the BLERT overall total was 0.58 and for the ER40 0.48. Effect sizes for accuracy, RT and confidence ratings (CR) comparing emotions for each group (HC, SZ) were generally in the very small to small range. For the ER40 effect size ranges for accuracy (within individual emotions) were d = 0.07 to 0.20, for RT d = 0.09 to 0.2, and for confidence d = 0.02 to 0.15. For the BLERT, effect size ranges for accuracy (d = 0.10 to 0.40), RT (d = 0.07 to 0.16) and confidence (d = 0.19 to 0.26) were also generally in the small range.

3.2. Feature selection for ER40 and discrimination of SZ from HC

For the ER40, 25 out of 120 features were selected using GINI variable selection procedure. Of the 25 selected features, almost half (48%) were confidence ratings compared to the two other performance variables (36% RT, 16% accuracy), χ^2 (for homogeneity) = 3.73 and ρ = 0.15. In addition, sadness was disproportionately represented (sad:36%, fear:24%, angry:20%; happy:12% and neutral:8%), χ^2 (for homogeneity) = 7.57 and ρ = 0.10. These 25 features were included in ML models and the best performing model achieved a Cohen's d of 1.24 (and AUC of 0.81 and F1 score of 0.78), considerably higher than traditional comparisons (d = 0.48) (Table 3). Supplementary Table 2 section details parameters used in ML models. Supplementary Table 3 shows the performance of various ML models on classification tasks and associated

Table 1
Effect sizes: Cohen's d, T-value and p-value over 5 emotions, and Combined for ER40: A) RT B) CR and C) Correct (Corr) across the groups (HC & SZ).

Emotion	T value	*p* value	AUC	Cohen's d
A) Effect size for RT				
Combined[a]	−2.99	0.003	0.59	0.32
Neutral	2.38	0.017	0.52	0.09
Sad	4.69	<0.001	0.54	0.17
Happy	5.74	<0.001	0.55	0.20
Angry	2.33	0.019	0.52	0.09
Fearful	4.65	<0.001	0.54	0.17
B) Effect size for CR				
Combined[a]	0.776	0.44	0.52	0.08
Neutral	0.60	0.54	0.50	0.02
Sad	3.95	<0.001	0.54	0.14
Happy	4.29	<0.001	0.54	0.15
Angry	1.01	0.31	0.51	0.04
Fearful	1.62	0.11	0.51	0.06
C) Effect size for Accuracy				
Combined[a]	4.69	<0.001	0.63	0.48
Neutral	2.15	0.031	0.52	0.08
Sad	4.78	<0.001	0.54	0.17
Happy	4.50	<0.001	0.54	0.16
Angry	2.01	0.045	0.52	0.07
Fearful	5.57	<0.001	0.55	0.20

n1 = total number of HC participants (i.e. 154) * number of times of each emotion (i.e. 8 of each emotion out of 40) = 1232 and **n2** = total number of SZ participants (i.e. 218) * number of times of each emotion (i.e. 8 of each emotion out of 40) = 1744[+]. n1, n2 is used to calculate degree of freedom for t-test. Approximate associated AUC is also shown (Salgado, 2018). [+] One response for neutral emotion for SZ was not available.

[a] Table 8 from (Pinkham et al., 2018a).

Table 2
Effect sizes: Cohen's d, T-value and *p*-value over 7 emotions, and Combined for BLERT: A) RT B) CR and C) Correct (Corr) across the groups (HC & SZ).

Emotion	T value	p value	AUC	Cohen's d
A) Effect size for RT				
Combined[a]	−1.54	0.124	0.54	0.16
Neutral	2.65	0.008	0.54	0.15
Sad	1.89	0.058	0.53	0.11
Happy	1.25	0.209	0.52	0.07
Angry	2.71	0.006	0.54	0.16
Fearful	1.37	0.168	0.52	0.08
Surprise	2.01	0.044	0.53	0.12
Disgust	2.40	0.016	0.54	0.14
B) Effect size for CR				
Combined[a]	3.20	0.001	0.59	0.32
Neutral	4.02	<0.001	0.56	0.23
Sad	4.54	<0.001	0.57	0.26
Happy	3.89	<0.001	0.56	0.23
Angry	4.20	<0.001	0.56	0.24
Fearful	3.17	0.002	0.55	0.19
Surprise	3.52	<0.001	0.55	0.20
Disgust	4.12	<0.001	0.56	0.24
C) Effect size for Accuracy				
Combined[a]	5.70	<0.001	0.66	0.58
Neutral	6.85	<0.001	0.61	0.40
Sad	5.72	<0.001	0.59	0.34
Happy	2.27	0.02	0.54	0.14
Angry	2.59	0.009	0.54	0.16
Fearful	3.28	0.001	0.55	0.20
Surprise	3.40	<0.001	0.55	0.20
Disgust	1.57	0.116	0.52	0.10

n1 = total number of HC participants (i.e. 154) * number of times of each emotion (i.e. 3 of each emotion out of 21) = 462[+] and **n2** = total number of SZ participants (i.e. 218) * number of times of each emotion (i.e. 3 of each emotion out of 21) = 654[+]. n1, n2 is used to calculate degree of freedom for t-test. Approximate associated AUC is also shown (Salgado, 2018). [+]One response for disgust emotion in SZ and one response for neutral in HC was not available.
[a] Table 8 from (Pinkham et al., 2018a).

ROC curves are depicted in Supplementary Fig. 1.

3.3. Feature selection for BLERT and discrimination of SZ from HC

A total of 33 out of 63 features were selected for the BLERT. Identical to the ER40, almost half were confidence ratings (48%) compared to the two performance variables (18.1% RT, 33.3% accuracy), χ^2 (for homogeneity) = 9.54 and ρ = 0.008. Also similar to ER40, BLERT selected features included a higher representation of sadness (21.2%), as well as was anger (21.2%). Other emotions were fear (15.1%), surprise (12.1%), disgust (9.0%), vs. neutral (15.1%), and happy (6.0%) categories, χ^2 (for homogeneity) = 9.54 and ρ = 0.14. The 33 selected features submitted to ML for the BLERT resulted in a best performing model with a Cohen's d of 0.87 (and an F1 score of 0.71 and AUC of 0.73) using Neural Network with ReLu as activation function, slightly lower than that of the ER40 but considerably higher than that in traditional statistics (d = 0.58) (Table 3). Overall, these converging results from the ER40 and BLERT suggest ML models resulted in increased discriminability compared to uni-dimensional comparison of total scores, and confidence ratings and sadness items were most relevant to distinguishing SZ from HC.

3.4. Feature selection and discrimination levels of self and informant rated social cognition (OSCAR self and informant reports) and overconfidence

Within the SZ subsample, there were 80 selected features for the OSCARS Self-Rated Score and 40 for OSCARS Informant Subscale. For OSCARS Self-Rated Scale, features comprised 42.5% confidence, 42.5% RT and 15% accuracy and the component emotions were 22.5% fear, 17.5% sad, 21.25% angry, 23.75% neutral and 15% happy. For OSCARS Informant, features comprised 67.5% confidence, 15% RT and 17.5% accuracy and the component emotions were 25% fear, 25% sad, 22.5% angry, 17.5% neutral and 10% happy. In general, models predicted OSCARS Self-Rated (d = 0.91, F1 score = 0.74, AUC = 0.74) and OSCARS Informant (d = 0.87, F1 score = 0.81, AUC = 0.73) performed well (Table 4). Our results indicate strong accuracy in ML models predicting OSCARS Self-Rated and OSCARS Informant.

Table 3
Performance of ML on ER40 and BLERT datasets compared to performance inferred from t-tests in SCOPE study. (Classification target: SZ vs. HC). For ER401, 120 features (cognition: 40 RT, 40 Corr and meta-cognition: 40 CR) were considered while for BLERT 63 features (cognition: 21 RT, 21 Corr and meta-cognition: 21 CR) were considered as input.

Dataset	Features	Ranked feature/total[b,c]	Count of selected negative emotion features[b]	Count of (CR) features[b,c]	Count of negative CR features[b]	Best performing model	AUC	F1	Equivalent Cohen's d
ER40	GINI filtered (RT + CR + Accuracy)	25/120	20	12	11	Stack[a]	0.81	0.78	1.24[e]
	t-Test Accuracy	–	–	–	–	–	0.63[!]	–	0.48[d]
	t-Test Reaction Time	–	–	–	–	–	0.59[!]	–	0.32[d]
	t-Test Confidence	–	–	–	–	–	0.52[!]	–	0.08[d]
BLERT	GINI filtered (RT + CR + Accuracy)	33/63	26	16	13	Neural Network, ReLu	0.73	0.71	0.87[e]
	t-test Accuracy	–	–	–	–	–	0.66[!]	–	0.58[d]
	t-test Reaction Time	–	–	–	–	–	0.54[!]	–	0.16[d]
	t-Test Confidence	–	–	–	–	–	0.59[!]	–	0.32[d]

[a] Stack method implies stacking of methods (Naïve Bayes, Neural Network (ReLU), Random Forest, Tree).
[b] Negative features considered for ER40 are S, A and F while negative features considered for BLERT are S, A, F, D and SU.
[c] Positive features considered for ER40 and BLERT are N, H.
[d] Table 8 from (Pinkham et al., 2018a).
[e] Table 2 from (Salgado, 2018).

Table 4
Performance of ML with OSCAR, SLOF as targets for SZ group. Gini was used to rank top features. Input variables considered were 120 ER40(cognitive: 40 RT, 40 corr and meta-cognitive: 40 CR). Target variable considered were categorical (0,1) while all others were numeric. Here threshold refers to the value of target used for categorization. Performance is shown for the best ranked features and model.

Target	Threshold	Method	Number of ranked features[a,b]	Count of (CR)[a,b] features	Count of negative features[a]	Count of negative (CR)[a] features	F1	AUC	Equivalent Cohen's d[c]
Oscars Self Report	0–17 as 0, 18 and above as 1	Stochastic Gradient Descent (SGD)	80	34	49	22	0.74	0.74	0.91
Oscars Informant Report	0–17 as 0, 18 and above as 1	AdaBoost	40	27	29	18	0.81	0.73	0.87

[a] Negative features considered for ER40 are S, A and F.
[b] Positive features considered for ER40 are N, H.
[c] Table 2 from (Salgado, 2018).

Additionally, overconfidence (only positive valued difference between OSCARS Self-Rated and OSCARS Informant, 55 subjects in SZ met the criteria) was strongly predictable on a continuous scale using ER40 features ($R^2 = 0.53$, MAE = 1.79, top 42 features, Neural Network with logistic activation).

3.5. Feature selection and discrimination levels of positive and negative symptom severity (PANSS)

Within the SZ subsample, there were 24 selected features for the PANSS Positive Syndrome Subscale and 35 for PANSS Negative Subscale. The emotions for PANSS positive comprised 25% fear, 20.8% sad, 8.3% angry, 20.8% neutral and 25% happy emotions. For PANSS Negative, the feature comprised 20% fear, 25.7% sad, 8.6% angry, 20% neutral and 25.7% happy emotions. For machine learning models predicting cut points (15) on positive and negative symptoms the overall separation was acceptable as evidenced by the F1 score (Table 5). In general, models predict PANSS Positive Syndrome Scale (d = 0.70 (Salgado, 2018), F1 score = 0.69, AUC = 0.69) with acceptable performance and Negative syndrome scale (d = 0.25 (Salgado, 2018), F1 score = 0.65, AUC = 0.57) with reduced performance. Our results indicate lowered accuracy in ML models predicting severity of positive and general negative symptoms.

We further explored whether the prediction of the PANSS Reduced Emotional Experience factor (Emotional Withdrawal (N2), Passive-apathetic Social Withdrawal (N4) and Active social avoidance (G16)) have been found to be correlated with social deficits in SZ making them worthy clinical targets (Kalin et al., 2015). A classification of subjects based on any of the three factors greater than 4 can be predicted using ER40 features (Table 5) (F1-score = 0.86, AUC = 0.717 and equivalent Cohen's d = 0.811).

3.6. Feature selection and discrimination levels of depression severity (BDI)

For BDI scores corresponding to minimal (cutoff = 9), mild (cutoff = 15), moderate (cutoff = 19) and severe (cutoff = 29) levels, best performance was achieved using top ranked 25, 25, 20 and 20 features respectively. The classification performance of subjects based upon BDI cutoffs increased with increasing severity cutoffs (F1 scores increased from 0.72 to 0.91 and AUCs of 0.69 to 0.90 respectively). Table 5 shows the confidence ratings that are common for various BDI cutoffs and their overlap with ER40 confidence ratings. Although sad and angry emotions dominated across the specified BDI thresholds, with increasing BDI cutoff thresholds, confidence ratings increasingly include neutral and positive emotion (happiness).

Table 5
Performance of ML with targets PANSS, BDI on the ER40 (only SZ group was considered). Gini was used to rank top features. Input variables considered were 120 (cognitive: 40 RT, 40 corr and meta-cognitive: 40 CR). Target variable considered were categorical (0, 1) while all others were numeric. Here threshold refers to the value of target used for categorization. Thresholds for categorization for BDI was taken from (Chemerinski et al., 2008) while for PANSS_pos and PANSS_neg was based on the number closer to the mean of the respective distributions. Performance is shown for the best ranked features and model.

Target	Threshold	Method	Number of ranked features[a,b]	Count of (CR)[a,b] features	Count of negative features[a]	Count of negative CR[a] features	F1	AUC	Equivalent Cohen's d[c]
PANSS1_neg	15	Neural Network (ReLu)	35	6	19	3	0.65	0.57	0.25
PANSS1_pos	15	Tree	24	10	13	6	0.69	0.69	0.70
PANSS Reduced Emotional Experience[d]	See below[e]	Stack[f]	50	17	33	13	0.86	0.78	0.81
BDI	9	Neural Network (tanh)	25	6	16	4	0.72	0.69	0.70
BDI	15	SVM (linear)	25	7	15	5	0.81	0.77	1.04
BDI	19	Random Forest	20	6	13	3	0.78	0.83	1.35
BDI	29	Neural Network (ReLu)	20	5	9	2	0.91	0.90	1.81

[a] Negative features considered for ER40 are S, A and F.
[b] Positive features considered for ER40 are N, H.
[c] Table 2 from (Salgado, 2018).
[d] The items in the *PANSS Reduced Emotional Experience* factor are: Emotional Withdrawal (N2), Passive-apathetic Social Withdrawal (N4) and Active social avoidance (G16).
[e] 1 if any factor is >4 else 0.
[f] Comprising Naïve Bayes, Neural Network (ReLu), Random Forest, Tree.

4. Discussion

Our study applied computational methods to evaluate relationships among facial affect recognition accuracy, reaction time and confidence ratings at an item level, in order to find relationships that may not be obvious from standard statistical comparisons. Although ML techniques are inherently exploratory and require independent replication, there were several potentially important findings. For one, confidence ratings were disproportionately represented in selected features discriminating HC from SZ, comprising an identical 48% of selected features for the BLERT and ER40. Moreover, among emotions, sadness was also disproportionally selected among features for both BLERT and ER40. Secondly, machine learning models with selected variables separating HC from SZ achieved very large (ER40) or large (BLERT) effect sizes, compared to medium effect sizes in standard statistical comparison with unselected and averaged features. Third, confidence ratings were the majority of selected items in models predicting informant rated social cognitive ability and PANSS emotional expression (even more so than self-ratings of social cognition), and resulting models incorporating these confidence ratings were accurate in predicting informant ratings cognition as judged by the model performance metrics. Fourth, selected features were poor predictors of levels of severity of psychotic symptoms but were effective in separating levels of severity of depression. Notably, confidence ratings were not over-represented in any of the models predicting symptoms with SZ. Overall, these findings indicate that computational methods identified subsets of commonly used emotion recognition tasks that better discriminated SZ from healthy comparators than standard comparisons with independent dimensions of accuracy, reaction time, and confidence. Moreover, confidence judgements were particularly influential in the selected models differentiating HC and SZ and for social cognitive ability as rated by informants and the self. Therefore, impairments in confidence judgements, which were not evident in standard mean comparisons, may be important for understanding aberrant social cognition in SZ. Also important is that the selected emotions for discriminating of SZ and HC and for symptom variation were primarily negative.

Although in need of replication in an independent sample, our cross validated ML models for ER40 and BLERT resulted in superior predictive ability in discriminating HC and SZ. Effect sizes for the best performing ML models were $d = 1.24$, compared to 0.48 for standard comparison of ER40 Total Score in the same sample. The gap between BLERT-based ML models and traditional comparison was smaller, with $d = 0.87$ for the best performing ML model and 0.58 for the BLERT total score via t-test. One possible reason for the lower performance on the BLERT is that fewer inputs were involved (120 for the ER40 and 63 for the BLERT), and the extra negative emotions included, disgust and surprise, do not induce changes in confidence as might fear and anger, which as a proportion are lower. We controlled for this effect by including how many of them may be expected. Nevertheless, these results indicate adding RT and confidence ratings to standard tasks of emotion recognition can improve discrimination between healthy and clinically affected groups. One reason why confidence ratings and RT may add to the discriminations between is SZ and HC is that other factors, beyond diagnosis, may influence accuracy. For example, emotion recognition is impacted by male gender and older age (Sasson et al., 2010) which may obfuscate differences between HC and SZ. Additionally, overconfidence may be a central deficit in schizophrenia (Jones et al., 2019) and RT is associated with confidence. Inclusion of confidence and RT, which together are sensitive to clinical symptoms, with the standard tasks could help to provide additional ability to discriminate diagnostic or other groupings, as well as to potentially lead to creation of briefer tasks that preserve discriminative performance.

Another implication is that confidence, particularly for judging negative emotions, is an important dimension distinguishing SZ from HCs as negative emotions on these tasks are more numerous, as well as in discriminating levels of observer rated social cognitive ability. Nearly 50% of selected items in discriminating SZ from HC were confidence ratings for both the BLERT and ER40. A total of 91% of selected confidence ratings were for negative emotions on the BLERT. This stands in contrast to the effect sizes for standard comparison for total averaged confidence across items, which were very small (ER40 $d = 0.08$) or small-medium (BLERT $d = 0.38$) Due to the nature of ML, it is challenging if not impossible to determine if systematic overconfidence or under confidence drive this discrimination. However, convergence with other findings within the SZ sample provide clues about what might and might not drive aberrant confidence. Although depressive symptoms might be associated with negatively biased confidence judgements, discrimination models with the ER40 items across levels of depression were not disproportionally inclusive of confidence features. Interestingly, ML models were more accurate at higher levels of severity of depression, and yet the rate of confidence ratings selected was stable across levels of severity depression. Additionally, subjective and more global judgements about social cognitive ability on the self-rated OSCARS indeed were slightly less accurate in association with the ER40 in ML models, and a higher proportion of confidence ratings were evident in feature selection for the OSCARS Informant Report than the Self Report administration. Taken together, the findings may implicate generalized inaccuracy in awareness of performance of judging negative emotion as a discriminating factor in SZ from HCs. Aberrant confidence ratings of negative facial affect may not be well explained by a systematic negative bias diminishing confidence as might be associated with depression. Greater awareness is associated in some models with greater accuracy and depression (e.g. depressive realism (Alloy and Abramson, 1979, Bortolotti and Antrobus, 2015).

There are several important limitations to this work. Although we evaluated two separate instruments as a form of validation, and evaluated performance on a held-out sample, ML is inherently exploratory and as noted before, our findings require replication in an independent sample. Although multinomial ML prediction models are possible, we had simplified outcomes to binary based on established cutoff without a priori consideration of performance. Psychotic symptoms and mood may have important influence on confidence. Further, the strength of emotions depicted introduce a complexity to the analysis and is affected by both age and sex (Sasson et al., 2010), and is inversely related in its ability to discriminate. We have not included these factors in the current analysis to limit the scope, but future research could apply these same methods to evaluate the variation within schizophrenia. The sample population was stable outpatients and so these results may not apply to more symptomatic or hospitalized people with SZ. Arguably, a stable sample may have attenuated our ability to discriminate SZ from HC.

Key next steps for this work, in addition to replication in an independent sample with the same paradigm, would include evaluating whether the item-by-item approach incorporating confidence is consistent with other social cognitive and non-social cognitive domains. It may be that item-by-item analyses with multiple dimensions may reveal different patterns of deficits in SZ than previously recognized in traditional summative and independent analyses. Additional work with ML could incorporate the time dimension, such that confidence at the item-by-item level may be influenced and updated by prior responses on the same measure. Alternative paradigms, such as those that involve performance feedback, may be useful to further disentangle biases in confidence from inaccuracy. Finally, if aberrant confidence is an important aspect of SZ and relates to social cognitive ability, then mechanistic approaches employing neuroimaging, including of key brain regions involved in introspection such as the insula and right rostrolateral prefrontal cortex (Pinkham et al., 2018b), may yield important information.

Role of the funding source

The study was supported by the National Institute of Mental Health (grant numbers R01 MH093432 to P.H., D.P. & A.P, R01 MH112620 to A.P., MH116902 to C.D. and T32 MH019934).

CRediT authorship contribution statement

Varsha D. Badal: Helped design and implement the study, analyzed results, edited and prepared the manuscript.

Peter Hitchcock: Edited and contributed to the manuscript.

David L. **Penn**: Edited and contributed to the manuscript.

Philip D. Harvey: Edited and contributed to the manuscript.

Amy E. Pinkham: Edited and contributed to the manuscript.

Colin A. Depp: Helped design and implement the study, oversaw the study, analyzed results, edited and prepared the manuscript.

Declaration of competing interest

Dr. Harvey has received consulting fees or travel reimbursements from Allergan, Alkermes, Akili, Biogen, Boehringer Ingelheim, Forum Pharma, Genentech (Roche Pharma), Intra-Cellular Therapies, Jazz Pharma, Lundbeck Pharma, Minerva Pharma, Otsuka America (Otsuka Digital Health), Sanofi Pharma, Sunovion Pharma, Takeda Pharma, and Teva. He receives royalties from the Brief Assessment of Cognition in Schizophrenia and the MATRICS Consensus Battery. He is chief scientific officer of i-Function, Inc. He has a research grant from Takeda and from the Stanley Medical Research Foundation. None of the other authors have commercial interests to report.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scog.2021.100196.

References

Alloy, L.B., Abramson, L.Y., 1979. Judgment of contingency in depressed and nondepressed students: sadder but wiser? J. Exp. Psychol. Gen. 108, 441.

Beck, A.T., Steer, R.A., Brown, G.K., 1996. Beck Depression Inventory-II. San Antonio, 78, pp. 490–498.

Bell, M., Bryson, G., Lysaker, P., 1997. Positive and negative affect recognition in schizophrenia: a comparison with substance abuse and normal control subjects. Psychiatry Res. 73, 73–82.

Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., Lang, M., 2020. Benchmark for filter methods for feature selection in high-dimensional classification data. Comput. Stat. Data Anal. 143, 106839.

Bortolotti, L., Antrobus, M., 2015. Costs and benefits of realism and optimism. Curr. Opin. Psychiatry 28, 194.

Chemerinski, E., Bowie, C., Anderson, H., Harvey, P.D., 2008. Depression in schizophrenia: methodological artifact or distinct feature of the illness? J. Neuropsychiatry Clin. Neurosci. 20, 431–440.

Claesen, M., De Moor, B., 2015. Hyperparameter search in machine learning. arXiv preprint arXiv:1502.02127. https://arxiv.org/abs/1502.02127.

Cornacchio, D., Pinkham, A.E., Penn, D.L., Harvey, P.D., 2017. Self-assessment of social cognitive ability in individuals with schizophrenia: appraising task difficulty and allocation of effort. Schizophr. Res. 179, 85–90.

Demsar, J.C.T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., Mozina, M., Polajnar, M., Toplak, M., Staric, A., Stajdohar, M., Umek, L., Zagar, L., Zbontar, J., Zitnik, M., Zupan, B., 2013. Orange: data mining toolbox in python. J. Mach. Learn. Res. 14 (Aug), 2349–2353.

Dubey, R., Zhou, J., Wang, Y., Thompson, P.M., Ye, J., Initiative, A.S.D.N., 2014. Analysis of sampling techniques for imbalanced data: an n= 648 ADNI study. NeuroImage 87, 220–241.

Gur, R.C., Sara, R., Hagendoorn, M., Marom, O., Hughett, P., Macy, L., Turner, T., Bajcsy, R., Posner, A., Gur, R.E., 2002. A method for obtaining 3-dimensional facial expressions and its standardization for use in neurocognitive studies. J. Neurosci. Methods 115, 137–143.

Gur, R.E., Calkins, M.E., Gur, R.C., Horan, W.P., Nuechterlein, K.H., Seidman, L.J., Stone, W.S., 2007. The consortium on the genetics of schizophrenia: neurocognitive endophenotypes. Schizophr. Bull. 33, 49–68.

Harvey, P.D., Khan, A., Keefe, R.S.E., 2017. Using the positive and negative syndrome scale (PANSS) to define different domains of negative symptoms: prediction of everyday functioning by impairments in emotional expression and emotional experience. Innov Clin Neurosci 14, 18–22.

Harvey, P.D., Deckler, E., Jones, M.T., Jarskog, L.F., Penn, D.L., Pinkham, A.E., 2019. Autism symptoms, depression, and active social avoidance in schizophrenia: association with self-reports and informant assessments of everyday functioning. J. Psychiatr. Res. 115, 36–42.

Healey, K.M., Combs, D.R., Gibson, C.M., Keefe, R.S., Roberts, D.L., Penn, D.L., 2015. Observable Social Cognition—a Rating Scale: an interview-based assessment for schizophrenia. Cogn. Neuropsychiatry 20, 198–221.

Hooker, C., Park, S., 2002. Emotion processing and its relationship to social functioning in schizophrenia patients. Psychiatry Res. 112, 41–50.

Hoven, M., Lebreton, M., Engelmann, J.B., Denys, D., Luigjes, J., VAN Holst, R.J., 2019. Abnormalities of confidence in psychiatry: an overview and future perspectives. Transl. Psychiatry 9, 1–18.

Jones, M.T., Deckler, E., Laurrari, C., Jarskog, L.F., Penn, D.L., Pinkham, A.E., Harvey, P. D., 2019. Confidence, performance, and accuracy of self-assessment of social cognition: a comparison of schizophrenia patients and healthy controls. Schizophrenia Research: Cognition.

Kalin, M., Kaplan, S., Gould, F., Pinkham, A.E., Penn, D.L., Harvey, P.D., 2015. Social cognition, social competence, negative symptoms and social outcomes: inter-relationships in people with schizophrenia. J. Psychiatr. Res. 68, 254–260.

Kantardzic, M., 2011. Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons.

Kay, S.R., Fiszbein, A., Opler, L.A., 1987. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr. Bull. 13, 261–276.

Kerr, S.L., Neale, J.M., 1993. Emotion perception in schizophrenia: specific deficit or further evidence of generalized poor performance? J. Abnorm. Psychol. 102, 312.

Maatz, A., Hoff, P., Angst, J., 2015. Eugen Bleuler's schizophrenia—a modern perspective. Dialogues Clin. Neurosci. 17, 43.

Mandal, M.K., Pandey, R., Prasad, A.B., 1998. Facial expressions of emotions and schizophrenia: a review. Schizophr. Bull. 24, 399–412.

Noordewier, M.K., Breugelmans, S.M., 2013. On the valence of surprise. Cognit. Emot. 27, 1326–1334.

Oliveri, L.N., Awerbuch, A.W., Jarskog, L.F., Penn, D.L., Pinkham, A., Harvey, P.D., 2019. Depression predicts self assessment of social function in both patients with schizophrenia and healthy people. Psychiatry Res. 112681.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.

Pinkham, A.E., Harvey, P.D., Penn, D.L., 2017. Social cognition psychometric evaluation: results of the final validation study. Schizophr. Bull. 44, 737–748.

Pinkham, A.E., Harvey, P.D., Penn, D.L., 2018a. Social cognition psychometric evaluation: results of the final validation study. Schizophr. Bull. 44, 737–748.

Pinkham, A.E., Klein, H.S., Hardaway, G.B., Kemp, K.C., Harvey, P.D., 2018b. Neural correlates of social cognitive introspective accuracy in schizophrenia. Schizophr. Res. 202, 166–172.

Pinkham, A.E., Morrison, K.E., Penn, D.L., Harvey, P.D., Kelsven, S., Ludwig, K., Sasson, N.J., 2019. Comprehensive comparison of social cognitive performance in autism spectrum disorder and schizophrenia. Psychol. Med. 1–9.

Salgado, J.F., 2018. Transforming the area under the normal curve (AUC) into Cohen'sd, Pearson's rpb, odds-ratio, and natural log odds-ratio: two conversion tables. J. Exp. Psychol. Gen. 10, 35–47.

Sasson, N.J., Pinkham, A.E., Richard, J., Hughett, P., Gur, R.E., Gur, R.C., 2010. Controlling for response biases clarifies sex and age differences in facial affect recognition. J. Nonverbal Behav. 34, 207–221.

Schneider, F., Gur, R.C., Gur, R.E., Shtasel, D.L., 1995. Emotional processing in schizophrenia: neurobehavioral probes in relation to psychopathology. Schizophr. Res. 17, 67–75.

Scikit-Learn. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html [Online]. [Accessed December 2019].

Silberstein, J.M., Pinkham, A.E., Penn, D.L., Harvey, P.D., 2018. Self-assessment of social cognitive ability in schizophrenia: association with social cognitive test performance, informant assessments of social cognitive ability, and everyday outcomes. Schizophr. Res. 199, 75–82.

Strassnig, M., Bowie, C., Pinkham, A.E., Penn, D., Twamley, E.W., Patterson, T.L., Harvey, P.D., 2018. Which levels of cognitive impairments and negative symptoms are related to functional deficits in schizophrenia? J. Psychiatr. Res. 104, 124–129.