UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Lukács, G., & Steyrl, D. (2022). Machine Learning Mega-Analysis Applied to the Response Time Concealed Information Test: No Evidence for Advantage of Model-Based Predictors Over Baseline. *Collabra: Psychology*, *8*(1). https://doi.org/10.1525/collabra.32661

Cognitive Psychology

Machine Learning Mega-Analysis Applied to the Response Time Concealed Information Test: No Evidence for Advantage of Model-Based Predictors Over Baseline

Gáspár Lukács[1] [a], David Steyrl[2]

[1] Department of Cognition, Emotion, and Methods in Psychology, University of Vienna, Vienna, Austria; Department of Philosophy, University of Vienna, Vienna, Austria; JSPS International Research Fellow at the Department of Psychology, Aoyama Gakuin University, Tokyo, Japan, [2] Department of Cognition, Emotion, and Methods in Psychology, University of Vienna, Vienna, Austria; Department of Psychiatry, Psychotherapy and Psychosomatics, Psychiatric University Hospital, University of Zurich, Switzerland

Keywords: binary classification, machine learning, response time, concealed information test, deception

https://doi.org/10.1525/collabra.32661

Collabra: Psychology

Vol. 8, Issue 1, 2022

The response time Concealed Information Test (RT-CIT) can help to reveal whether a person is concealing the knowledge of a certain information detail. During the RT-CIT, the examinee is repeatedly presented with a *probe*, the detail in question (e.g., murder weapon), and several *irrelevants*, other details that are similar to the probe (e.g., other weapons). These items all require the same keypress response, while one further item, the *target*, requires a different keypress response. Examinees tend to respond to the probe slower than to irrelevants, when they recognize the former as the relevant detail. To classify examinees as having or not having recognized the probe, RT-CIT studies have almost always used the averaged difference between probe and irrelevant RTs as the single predictor variable. In the present study, we tested whether we can improve classification accuracy (recognized the probe: yes or no) by incorporating the average RTs, the accuracy rates, and the SDs of each item type (probe, irrelevant, and target). Using the data from 1,871 individual tests and incorporating various combinations of the additional variables, we built logistic regression, linear discriminant analysis, and extra trees machine learning models (altogether 26), and we compared the classification accuracy of each of the model-based predictors to that of the sole probe-irrelevant RT difference predictor as baseline. None of the models provided significant improvement over the baseline. Nominal gains in classification accuracy ranged between −1.5% and 3.1%. In each of the models, machine learning captured the probe-irrelevant RT difference as the most important contributor to successful predictions, or, when included separately, the probe RT and the irrelevant RT as the first and second most important contributors, respectively.

Introduction

In the recent decades, machine learning (ML) led to groundbreaking advances in applied sciences and engineering, its ubiquitous use ranging, for example, from medical diagnostics (Benjamens et al., 2020) and brain imaging analysis (Lemm et al., 2011) to computer applications that are by nowadays commonplace (e.g., Deng & Li, 2013) and various forensic tools (Carriquiry et al., 2019). Within forensics, one very practical ML application is to improve the prediction accuracy of computerized deception detection methods. There is increasing interest in this area, with most ML studies utilizing complex large data sets such as those from EEG and fMRI testings (e.g., Bablani et al., 2019; Davatzikos et al., 2005; Derakhshan et al., 2020; Dodia et al., 2019). However, ML can also be applied to comparatively simple data such as behavioral response times (RTs). In the present study, we assess the potential benefit of combining variables from RT-based Concealed Information Tests (CITs), in a machine learning mega-analysis using data from 1,871 individual tests. We do this by creating several ML models based on the combinations of up to 12 variables, and test whether the classification accuracy in predicting deception is better when using any of these models, than when using the baseline predictor that is conventional for this purpose in the RT-CIT.

RT-CIT and Feature Selection

The RT-CIT aims to disclose whether the tested person

[a] Corresponding author: Gáspár Lukács, Aoyama Gakuin University, 4-4-25 Shibuya, Shibuya-ku, Tokyo, 150-8366, Japan. E-mail: lkcsgaspar@gmail.com

recognizes certain relevant items ("probes"), such as a weapon used in a recent homicide, among a set of other objects ("irrelevants"), based on comparatively slower responding to the recognized probes (Meijer et al., 2016; Suchotzki et al., 2017). During the RT-CIT, examinees have to categorize items that are presented on a computer screen by pressing one of two keys (e.g., either "E" or "I" on a regular keyboard). They are asked to press one of those keys (e.g., "E") when they see the probe or one of four irrelevants, and they are asked to press the other key (e.g., "I") when they see a certain target item (an additional irrelevant item designated for this purpose). It is assumed that knowledgeable examinees recognize the probe as the relevant item in respect of the deception detection scenario, and that they generally respond slower to the probe as compared to irrelevants. Thereby, based on probe-irrelevant RT differences, "guilty" (knowledgeable) examinees can be distinguished from "innocent" (naive) ones. Since RT-CIT studies almost unanimously use the probe-irrelevant difference as the sole (or at least primary) predictor variable, this may here be declared as the *baseline* to which any modified or model-based predictors may be compared in view of diagnostic accuracy (regarding the importance of baseline, see e.g., DeMasi et al., 2017).

Properly built and cross-validated ML models represent a robust way of multivariate data analysis, which can incorporate and control for a large number of input variables (known as *features, or independent variables*) while avoiding overfitting (Cawley & Talbot, 2010; Hastie et al., 2009). Nonetheless, including too many redundant or inconsequential features reduces the accuracy of ML-based diagnostics: Given excessive possibilities of feature combinations, the model may capture random fluctuations in the training set (on which it is built) and, when cross-validating it on the test data (previously "unseen"), it will prove less accurate (Guyon & Elisseeff, 2003; Theodoridis & Koutroumbas, 2009). Therefore, features should be carefully selected based on previous knowledge and hypotheses in the given domain.

Regardings the RT-CIT, there have been suggestions for certain combinations of features. Firstly, since targets are similar to probes in certain important respects (rare, task-relevant items in the test), and both are, in these respects, opposed to irrelevants (frequent, less relevant items). Therefore, it has been hypothesized that those who do not respond to targets (much) slower than to irrelevants may also not respond to probes (much) slower than to irrelevants, despite recognizing the probe as the relevant item (Noordraven & Verschuere, 2013). Thereby, the prediction based on probe-irrelevant difference may be adjusted: In case of smaller target-irrelevant difference, one may predict guilt (i.e., probe having been recognized) in case of relatively smaller probe-irrelevant difference. Any such advantage can be captured by ML models, if the target (or target-irrelevant difference) is included as a feature.

Secondly, probe-irrelevant accuracy rates (ARs; ratio of correct responses per all responses) have repeatedly been found to also have some degree of predictive power (although generally much lower than probe-irrelevant RT differences): Guilty participants tend to have lower accuracy to probes than to irrelevants (e.g., Hu et al., 2013; Noordraven & Verschuere, 2013). Therefore, if the RT and AR differences do not strongly correlate, the AR may contribute to better predictions (Lukács, Gula, et al., 2017): For example, when the RT measure yields ambiguous values (in a "grey area" between typical guilty and typical innocent values), the AR measure may be given more weight in the decision. Again, such possibilities could be captured via ML models.

Thirdly and lastly, probe-irrelevant RT differences have often been divided by the irrelevant *SD* to obtain a standardized measure (Noordraven & Verschuere, 2013; Verschuere et al., 2015), although this in itself does not seem to improve diagnostic accuracy (Lukács & Specker, 2020). Nonetheless, the *SD*s of RTs could still contribute to better predictions in some other specific interaction with the rest of the included features, which would then be captured with ML models (see also Elaad & Ben-Shakhar, 1997; Hu et al., 2013).

Therefore, altogether, it seemed reasonable to include each main item type (probe, irrelevant, target), and each measured via RT, ER, and *SD*. The ML models were built using these features in specific combinations (see Methods), and the models were systematically compared to the baseline. With this, based on the considerations described above, we hoped and expected to improve the diagnostic accuracy of the RT-CIT. Of course, at the same time, we were aware of the possibility of no improvement. Notwithstanding its vast impact, ML is not beneficial in every case it is applied to: Instances of success are often widely publicized and well-known, but, unsurprisingly, instances where ML does not prove useful do not usually garner attention (perhaps with the exception of the notorious failures of economic prediction algorithms; López de Prado, 2018; but see also, e.g., Matsuda et al., 2019). In the academic world specifically, this discrepancy is aggravated by publication bias favoring positive outcomes and related conceptual and statistical issues (which we committedly aim to avoid; Riley, 2019).

Study 1

In this first study, we focused on RTs and ARs, in combinations of probe, target, and irrelevant items, as suggested by some previous papers, including probe-irrelevant as well as target-irrelevants differences (Lukács, Gula, et al., 2017; Noordraven & Verschuere, 2013), but also each item type's RT mean and AR separately (see below for details).

Method

All analyses were preregistered (Study 1: https://osf.io/gsk89; Study 2: https://osf.io/cph5s) except where we state otherwise. We conducted no pilot or any other prior studies or analyses not reported in the manuscript; we report all examined measures and conditions. We had no prior knowledge (e.g., Van den Akker et al., 2019) regarding the characterics of the data relevant to our ML analysis, except for the sporadic findings reported in previous individual studies as described above in the Introduction.

Data

For all our analysis we used, as a convenience sample, a recently published extensive database with trial-level results from CIT studies, available via https://osf.io/sfbkt/ (Lukács & Specker, 2020). This data includes 12 different datasets with different experimental designs (each including data from guilty as well as innocent participants), from seven different papers (Geven et al., 2018; Kleinberg & Verschuere, 2015, 2016; Lukács, Kleinberg, et al., 2017; Noordraven & Verschuere, 2013; Verschuere et al., 2015; for detailed database description, see Lukács & Specker, 2020, pp. 5–6).

We excluded (following the criteria by Lukács & Specker, 2020) all participants with accuracy rate not higher than 75% for the main items (probes and irrelevants merged), or accuracy rate not higher than 50% for (a) target items or (b) target-side fillers or (c) nontarget-side fillers. ("Fillers" are additional items presented throughout the task, used in one of the included experimental designs – Lukács, Kleinberg, et al., 2017 – but are otherwise not relevant to the present paper.) This left 1,871 participants in the data set: 752 innocent (age = 31.7±11.1 [21 unknown]; 357 male, 375 female [20 unknown]), and 1,119 guilty (age = 29.4±10.9 [18 unknown]; 493 male, 609 female [17 unknown]). For all further calculations, responses below 150 ms were excluded. For RT measures, only correctly answered trials were used (i.e., we excluded incorrect and too slow trials). ARs were calculated as the ratio of correct responses to correct, incorrect, and too slow responses.

Procedure

The ML-based multivariable classification analyses were performed in Python (v3.8.3, scikit-learn library v0.23.1; Pedregosa et al., 2011) to predict "guilt" or "innocence" from the features of the RT-CIT and to identify the features relevant for a successful classification. Logistic regression (LR; Hastie et al., 2009), linear discriminant analysis (LDA; Hastie et al., 2009), and Extra Trees (ET; Geurts et al., 2006) were used as methods for machine learning models. LR is a generalized linear model that wraps a logistic function around a linear regression. LDA is a linear, computationally highly efficient classification algorithm that uses linear hyperplanes to distinguish between classes. ET is a computationally efficient and highly accurate nonlinear classifier. ETs implement an ensemble of "extremely randomized trees". Ensemble methods improve the performance of base predictors, such as decision trees, by accumulating the predictions of those base predictors via, for example, majority voting. However, to obtain diverse predictions from the same base predictors, processes that introduce randomness when building the base predictors are applied. Hence the name "randomized trees." In general, decision trees can capture linear and non-linear relationships between features and the prediction targets (Hastie et al., 2009), in this case guilt or innocence.

The model's classification performances (weighted classification accuracy) were estimated using a nested cross-validation procedure (Cawley & Talbot, 2010). Cross-validation allows to assess the performance of the model that can be expected on new, unseen data, hence, the generalizability of the model. Cross-validation implements repeated train-test splits of the data, where a separate model is trained and tested in each cross-validation repetition. In our main cross-validation loop, a shuffle-split data partitioning with 10% of the participants in the testing-set was repeated 100 times, resulting in 100 models each. In each repetition, the training-set was used for data scaling (z-scoring and model complexity optimization. Complexity optimization is necessary if models allow different regularization strengths or different model sizes (in our case LR and ET models) to avoid overfitting the data. Complexity optimization was implemented in an inner (nested) cross-validation procedure. Hence, a separate cross-validation was carried out for each repetition of the outer cross-validation loop. The inner cross-validation loop again used a shuffle-split partitioning scheme (with 10% of the participants of the outer training-set in the inner testing set), but in these cases with 50 repetitions only, to save computation time. To control model complexity, we tuned the regularization parameter C and the L1 ratio parameter in the LR classifier, as well as the maximum number of leaf nodes per tree in the ET classifier. The candidates for the parameters were randomly drawn (randomized search procedure, 50 random draws, via RandomizedSearchCV, scikit-learn, v0.23.1). The parameters that led to the highest weighted classification accuracy were subsequently used in the outer cross-validation loop.

The obtained models were then tested on the respective hold-out set of the main cross-validation loop. The hold-out set (10% of the participants) were not used in the inner cross-validation loop. In each repetition of the main cross-validation loop, model prediction accuracy was computed. To counter unbalanced classes (more samples in one class than in the other, e.g., more guilty than innocent participants) weighted accuracy was used (Hastie et al., 2009). For prediction accuracy, higher values indicate a better model fit. Classification accuracy values lie between zero (a model that classifies every participant incorrectly) and one (perfect model that gets all classifications right). The theoretical chance level is 0.5. However, in the case of unbalanced classes and limited number of repetitions, the chance level varies substantially and needs to be assessed. For that, an additional model was trained and tested on a shuffled version of the data in each cross-validation loop.

The contributions of single features to the models' performance were also assessed. For non-linear classification models (e.g., ET classifier), this is not as straightforward as for linear models. One procedure that can be applied to all kinds of models is permutation feature importance testing, which works as follows. First, a baseline accuracy score is computed by passing a testing-set through the model. Second, the values of a single factor are permuted and the testing-set is passed again through the model. Third, the accuracy score is recomputed. Fourth, the importance of a factor is the difference between the baseline accuracy and the accuracy score after permutation (Molnar, 2019). The permutation thus disentangles the relationship between a factor and the prediction, that is, the drop in the model score (classification accuracy) is indicative of how much the model depends on that factor (Molnar, 2019). The statistical significance of the features' importances were assessed using a shuffle (randomization) testing procedure (Edgington & Onghena, 2007; Ojala & Garriga, 2010). The null hypothesis was that there is no relationship between the features (independent variables) and the prediction target. Therefore, following the evaluation of the features' importances, the model was refitted using the same training-data, but with shuffled targets (to fulfill the null hypothesis) and the features' importances were assessed again. The shuffling was repeated 64 times in each main cross-validation loop (hence, 6,400 times in total). To reject the null hypothesis at 95% confidence, no more than 5% of the features' importances obtained with shuffle data should be more extreme than the features' importances obtained by the original data (Edgington & Onghena, 2007; Ojala & Garriga, 2010). This percentage of more extreme values is equivalent to the *p* value that our results occurred under the null hypothesis. Subsequently, the obtained *p* values were Bonferroni-corrected for multiple comparisons. This whole analysis (computing the models and the importance of features) was carried out for each method and each feature set separately.

In Study 1, we ran five separate analyses with different sets of features (independent variables):

1. Baseline models with the conventional "probe-minus-irrelevant RT mean difference" predictor only.
2. Two features included in the models: Probe-minus-irrelevant RT mean differences; probe-minus-irrelevant AR differences.
3. Five features included in the models: Probe-minus-irrelevant RT mean differences; target-minus-irrelevant RT mean differences; probe RT mean; irrelevant RT mean; target RT mean.
4. Twelve features included in the models: Probe-minus-irrelevant RT mean differences; probe-minus-irrelevant AR differences; target-minus-irrelevant RT mean differences; target-minus-irrelevant AR differences; probe RT mean; irrelevant RT mean; target RT mean; probe AR; irrelevant AR; target AR; age; gender. (For this analysis, 39 participants with unknown age or gender were removed from the data for this analysis, leaving 1,832.)
5. Models with all six significant features of the 4th models: Probe-minus-irrelevant RT mean differences; probe-minus-irrelevant AR differences; target-minus-irrelevant AR differences; probe RT mean; irrelevant RT mean; age. (Again, those with unknown age were removed from the data, leaving 1,832.)

In total we computed a total number of 1,500 analyses (100 repetition, 3 models, 5 sets of features) in Study 1.

Model Comparison. Having obtained the results (100 classification accuracy scores per method and set of features) of the 100 repetitions of the main (outer) cross-validation loop, the last step was to assess whether any of our four multi-factor models (2 - 5) performed better than the single-factor baseline (model 1). In our preregistration, we specified a bootstrap procedure (Efron, 1992) and rank-based Bayes factors (BFs; Van Doorn et al., 2020) as statistical significance tests. However, this was a mistake: These tests are actually not applicable in this context (notwithstanding their widespread use in other papers, which are also erroneous), because the scores obtained in the ML procedure are based on the same underlying data, and therefore the assumption of independence is violated, and the comparisons would result in a very high Type I error rate (false positives). How to solve this problem is not straightforward, and no well-established procedure exists. It has been suggested to just report 95% confidence intervals (CIs) of the scores of each model (Mitchell, 2013) – which is what we preregistered for Study 2 (see below). This however provides no formal assessment of whether there is a statistically significant difference. One solution has been offered by Nadeau and Bengio (2003), which is a variance correction applied to Student's *t*-test. This yields a conservative *p* value (low chance of false positives, but relatively higher chance of false negative) that takes into account the reduced variability due to the choice of the training sets and the choice of the test sets. This Nadeau-Bengio correction has since been suggested by several other authors for machine learning procedures such as ours (Bouckaert & Frank, 2004; Witten et al., 2011).

We therefore report the means and 95% CIs of the scores for each model method and feature set, and, exploratorily, Nadeau-Bengio corrected *p* values for each comparison to the baseline. For completeness however, we do report the outcomes of the pre registered tests in an online Appendix (available via https://osf.io/zhmb2/), although these are, as explained above, not to be relied on.

Secondary analysis. To demonstrate how the outcomes may differ in case of using single experimental designs, we repeat the exact same model preparation and analysis (1,500 analyses), but add, for each method and feature set, experimental design (datasets 1-12) as a factorial feature. This may show whether there are any predictive features specific to experimental designs, although such predictors would not necessarily be generalizable to other or new experiments or designs. For the rest of this paper, we denote this feature as *Experiment*, although, technically, some of the experimental designs were originally reported within as a single "experiment" (or "study") using two or three different RT-CIT designs (e.g., Verschuere et al., 2015; again, for details, see Lukács & Specker, 2020).

Results

All statistical results are reported in Table 1[1] for the ML without Experiment as feature, and in Table 2 for the ML with Experiment as feature. None of the model comparisons yielded statistically significant results. The nominal gain using multiple features, as compared to baseline, was also

[1] The achieved classification accuracies reported here may seem relatively low, but this is largely unrelated to our ML procedure (i.e., this is simply the accuracy generally provided by the RT-CIT, see, e.g., table 2 in Lukács & Specker, 2020), and it is still superior to human judgments alone, which are generally around chance-level accuracy (e.g., Hartwig & Bond, 2011). Furthermore, recently developed improved RT-CIT designs can provide substantially higher accuracies (Lukács, Kleinberg, et al., 2017; Lukács & Ansorge, 2021; Wojciechowski & Lukács, 2022).

Table 1. Model comparisons (without Experiment as feature)

	Mean	95% CI	$p_{\text{N-B}}$
LR baseline	69.9	[69.3, 70.4]	
LR 2 features	70.9	[70.4, 71.5]	.457
LR 5 features	70.2	[69.6, 70.7]	.846
LR 12 features	70.9	[70.3, 71.6]	.500
LR 6 sign. features	71.0	[70.3, 71.6]	.470
LDA baseline	67.9	[67.3, 68.5]	
LDA 2 features	69.3	[68.6, 69.9]	.365
LDA 5 features	68.0	[67.4, 68.6]	.938
LDA 12 features	68.3	[67.7, 68.9]	.770
LDA 6 sign. features	69.1	[68.5, 69.8]	.432
ET baseline	69.2	[68.6, 69.9]	
ET 2 features	72.3	[71.7, 72.9]	.056
ET 5 features	69.1	[68.5, 69.7]	.929
ET 12 features	71.2	[70.6, 71.9]	.211
ET 6 sign. features	71.6	[71.1, 72.2]	.113

Note. The means and 95% CIs are reported for each model. The Nadeau-Bengio-corrected *p* values ($p_{\text{N-B}}$) refer to the comparison of the given multi-feature model to the baseline of the same ML method (LR, LDA, or ET). The baseline always refers to the model with the sole probe-irrelevant RT mean difference feature. The multi-feature models include the features described under Procedure, in the same order.

Table 2. Model comparisons with Experiment as feature

	Mean	95% CI	$p_{\text{N-B}}$
LR baseline + Exp.	75.4	[74.8, 76.0]	
LR 2 features + Exp.	76.6	[76.1, 77.2]	.402
LR 5 features + Exp.	75.2	[74.6, 75.9]	.919
LR 12 features + Exp.	75.9	[75.3, 76.5]	.740
LR 6 sign. features + Exp.	76.0	[75.4, 76.7]	.686
LDA baseline + Exp.	72.4	[71.7, 73.1]	
LDA 2 features + Exp.	74.6	[74.0, 75.1]	.181
LDA 5 features + Exp.	73.7	[73.2, 74.3]	.409
LDA 12 features + Exp.	74.5	[73.8, 75.2]	.230
LDA 6 sign. features + Exp.	74.0	[73.4, 74.7]	.349
ET baseline (+ Exp.	74.1	[73.4, 74.7]	
ET 2 features + Exp.	75.4	[74.8, 75.9]	.382
ET 5 features + Exp.	74.1	[73.6, 74.7]	.970
ET 12 features + Exp.	75.6	[75.0, 76.2]	.319
ET 6 sign. features + Exp.	74.5	[73.9, 75.1]	.799

Note. Similarly as in Table 1, the means and 95% CIs are reported for each model. The Nadeau-Bengio-corrected *p* values ($p_{\text{N-B}}$) refer to the comparison of the given multi-feature model to the baseline of the same ML method (LR, LDA, or ET). The baseline always refers to the model with the probe-irrelevant RT mean difference as the only feature apart from the Experiment factor. The multi-feature models include the features described under Procedure (in the same order), and the Experiment as feature.

fairly small in all cases. In the case of models without Experiment as a feature, the average gain was 1.1%, 95% CI [0.8, 1.4] (minimum: −1.5%, maximum: 3.1%). In the case of the models with Experiment as a feature, the average gain was practically the same, 1.1%, 95% CI [0.6, 1.5] (minimum: −0.2, maximum: 2.2). This means that our ML methods did not find any features that could substantially contribute to better predictions either overall (generalizable to all included experiments, and hence likely to the RT-CIT in general), or specific to given experiments.

In every model, the probe-irrelevant RT mean difference was found as the most important contributor to correct predictions. Figures depicting the relative importance of each feature, in each model, can be found in the online Appendix.

Study 2

In Study 2, to empirically verify the theory-based "probe-irrelevant RT difference" feature, we separately included all lower-level features (e.g., probe mean RT, probe ER, irrelevant RT, irrelevant ER, etc.), with the expectation that the ML too would find probe mean RT and irrelevant mean RT as the most important contributors, and that, once again, the model based on these lower-level features will not outperform the baseline of the plain probe-irrelevant RT difference predictor. Furthermore, here we also included *SD*s per item type as potentially beneficial additional features (Elaad & Ben-Shakhar, 1997; Hu et al., 2013; Noordraven & Verschuere, 2013).

Method

Data

The analysis was performed on the same data as in Study 1.

Procedure

The procedure was largely the same as in the previous analysis, except for the included features. Here, apart from the baseline (probe-minus-irrelevant RT mean difference), we created only one other model, including the following eleven features: probe RT mean; irrelevant RT mean; target RT mean; probe RT *SD*; irrelevant RT *SD*; target RT *SD*; probe AR; irrelevant AR; target AR; age; gender.

The ML analyses were again performed in Python (v3.9.2, scikit-learn library v0.23.2; Pedregosa et al., 2011). Here we only used LR and ET. Classification performances were again assessed using a nested cross-validation procedure. A shuffle-split scheme with 128 repetitions (20% of the participants in the testing-set, 80% in the training-set) was applied in the main (outer) cross-validation loop. In each repetition, the training-set was used for data scaling (min-max-scaling) and model complexity optimization. Model complexity optimization was carried out in a nested (inner) cross-validation procedure using a sequential bayesian optimization procedure in combination with a shuffle-split scheme (20% testing, 80% training, 64 repetition) to find the best complexity parameters (BayesSearchCV, scikit-optimize, v0.8.1, LR: C 1e-4 to 1e4, l1_ratio 0.011 to 1, ET: min_samples_leaf 1 to 32 log-prior, max_features 1 to total number of features). The complexity parameters that lead to the lowest squared error in the training-set were subsequently used along with the following constant parameters – LR: penalty=elasticnet, tol=1e-4, solver=saga, max_iter=1e4, warmstart=True, and ET: n_estimators=512, criterion=friedman_mse – to train an LR and ET classifier model in the main cross-validation loop, respectively. The models were subsequently tested on the respective testing-set of the main cross-validation loop. The testing-set was not used in the inner cross-validation loop. As in Study 1, to counter unbalanced classes (more samples in one class than in the other), weighted accuracy was used. The statistical analyses were the same as in Study 1.

Model Comparison. Same as in Study 1, here in Study 2 we report the means and 95% CIs of the scores for each model (Mitchell, 2013), and, exploratorily, Nadeau-Bengio corrected *p* values for each comparison to the baseline (Bouckaert & Frank, 2004; Nadeau & Bengio, 2003; Witten et al., 2011).

Results

For the LR modelling, the baseline mean score was 70.0%, 95% CI [69.7, 70.3], while the multi-feature model's mean score was 70.9%, 95% CI [70.5, 71.3], and the Nadeau-Bengio-corrected *p* value showed no statistically significant difference ($p_{\text{N-B}}$ = .363). For the ET modelling, the baseline mean score was 70.0%, 95% CI [69.6, 70.4], while the multi-feature model's mean score was 68.5%, 95% CI [68.1, 68.9], hence the latter actually gave nominally lower accuracy (but again with no statistically significant difference, $p_{\text{N-B}}$ = .148).

The importance of each feature in achieving the given classification accuracies in the multi-feature models (70.9% for LR, 68.5% for ET) is shown in Figure 1 (the underlying data is available via https://osf.io/zhmb2/). In both models, the most important feature is the probe RT mean, and the second most important feature is the irrelevant RT mean. In the LR model in particular, where learning procedure is relatively straightforward, the contribution of the other variables is negligible: The ML no doubt found the probe-irrelevant RT difference as the main and only important feature combination. In the case of the ET model, whose learning procedure is more complex, other features contribute to some degree as well: However, as reflected in the nominally decreased classification accuracy, these serve more as confounds rather than beneficial contributors.

Lastly, following the suggestion of an anonymous reviewer, we exploratorily performed the same analysis using a Random Forest classifier (RF; Biau & Scornet, 2016; Breiman, 2001; Geurts et al., 2006; with the relevant parameters identical to those for ET). The RF is an ensemble method very similar to ET. While the ET is more computationally efficient, the RF may provide better results under certain conditions. However, in our experience, the difference in the obtained results is typically very small and even negligible. In the present case too, the RF led to no noteworthy improvement: The baseline mean score of RF was 68.3%, 95% CI [67.9, 68.8], while the multi-feature model's mean score was 69.0%, 95% CI [68.7, 69.4] (and the difference between the two scores was not statistically significant: $p_{\text{N-B}}$ = .536).

General Discussion

Considering all the potential information in the outcome of an individual RT-CIT might leave one perplexed at why, from all this data, it is only the probe-irrelevant RT mean difference that is used for diagnostics – and, in turn, it may be very tempting to try incorporating some additional characteristics. A well-established way to do this is via ML-based multivariable models and proper cross-validation (e.g., Kleinberg et al., 2018), as we did in the present study. However, we found no substantial diagnostic advantage of incorporating any of 15 additional variables as potential predictive features.

A part of the explanation is that the baseline probe-irrelevant RT difference predictor is already a fairly good predictor, generally speaking, and provides reasonable accuracies in almost all RT-CIT studies. Thereby, even if some other features are to a small degree predictive of guilt, they are simply dwarfed by the probe-irrelevant RT difference and are therefore redundant: Although their contribution to the model-based prediction may be significant in some cases, their benefit does not sufficiently outweigh the disadvantage of decreasing the contribution of the strongest predictor, the probe-irrelevant RT difference. Relatedly, there is some extent of correlation between probe-irrelevant RT differences and other variables that indicate guilt: For example, the (Pearson) correlation between probe-irrelevant RT differences and probe-irrelevant ER differences is .298, 95% CI [.243, .350]. Consequently, probe-irrelevant ER difference may predict guilt in many cases, but the probe-irrelevant RT difference also already predicts it in these cases (along with cases where the former does not). Therefore, the addition of probe-irrelevant ER differences is mostly redundant.

The ML approaches that we employed (LR, LDA, ET) are state-of-the-art and among the best established and most widely used binary classifiers, covering all essential feature interaction possibilities (e.g., Hastie et al., 2009). Therefore, it is unlikely that other ML methods on the same or similar variables would fare much better – rather, trying many more options could misleadingly indicate better results in some cases merely by chance (Cawley & Talbot, 2010): Even with a dataset as large as in the present study, the potential issues of multiple-testing and data dredging have to be kept in mind (Wicherts et al., 2016).

On the other hand, future ML research could incorporate other and more sophisticated data characteristics as features, such as those based on the modelling of individual trial-level data (e.g., De Boeck & Jeon, 2019). Cross-validating ML models built on the top of a large number of individual-level models, each built on raw data, would require extreme computational power – but it seems a worthwhile topic for exploration. The only study (Strahm, 2017) to our knowledge that so far explored such modelling, though presented an inspiring introduction and discussion of the topic, found no substantial improvement in diagnostic accuracy when using distribution models or explanatory process models, as compared to the conventional means and *SD*s (represented by the Gaussian model). Nonetheless, the study involved only a small sample (*N* = 94), and, also, that specific dataset's very high baseline accuracy (97.7% for the Gaussian model) may have created a ceiling effect. However, whether or not diagnostic improvement can be achieved, modelling may still be interesting to help uncover the cognitive processes underlying the RT-CIT effect, which, in turn, may also have very practical uses: For example, Reich et al. (2018) suggested that drift diffusion models (Ratcliff, 2002) may help uncovering attempts at faking in the RT-CIT. Such analysis may however be hindered by that the RT-CIT results typically contain no or only very few incorrect responses for probes and controls (see supplementary figure at https://osf.io/unhzx/ for the numbers of participants in our study per error RT count, per item type), such that would be desirable for the process model. Future RT-CIT experiments could attempt to increase the proportion of error RTs by, for example, decreasing the time allowed for responding (which may force participants to respond faster and, thereby, more likely incorrectly) – this would be particularly interesting in view of speed-accuracy tradoff models (let alone the potential practical effects in terms of probe-irrelevant differences and related diagnostic accuracy; Lubczyk et al., 2022). However, the very contrary, removing or greatly lengthening the response time limit, may on the other hand allow better fitting for various process models (Strahm, 2017, p. 32). What is more, this latter option might even yield a larger number of error RTs too, if it may be the case that a substantial portion of the too-slow responses in case of a strict time limit, which effectively prevents executing these response at all, would be executed as incorrect responses in case of no or a very long time limit.

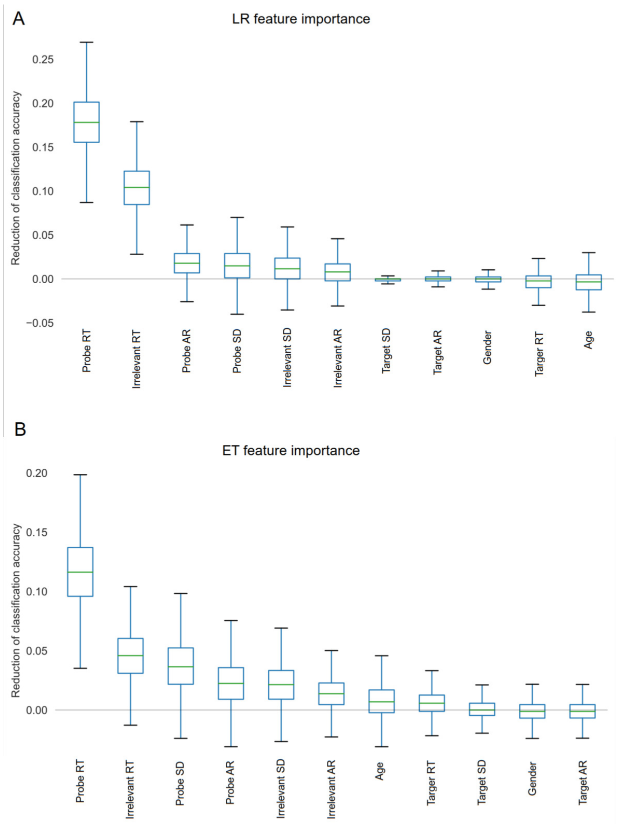

Figure 1. Feature importance per model

Note. The box plots per feature show the given feature's importance in the given model, calculated as the reduction in classification accuracy (ratio of correct classifications) in case the given feature's values are shuffled (see details in the Procedure section). To note, the chance level is 50%, and, therefore, a reduction of 17.8%, as in case of the Probe RT in the LR model, would almost completely obliterate the original classification accuracy (reducing it from 70.9% to 53.1%).

Finally, although neither statistically significant nor robust in magnitude, we did find nominal improvements up to 2-3% in some cases. It is possible that with more data available, this improvement can be proven statistically significant – especially if the data is from RT-CITs using similar task designs, in which case the relation between the variables may be more consistent. If so, the few percent improvement could still be of practical use: If the RT-CIT is eventually applied widely in real life, a 3% improvement would essentially mean 3 more correct guilty or innocent classifications out of every 100 tests – where each classification may have important consequences in investigations or court cases.

Regardless of whether improvements are still possible in any of the ways described above, our findings show very strong support for the probe-irrelevant RT difference as the primary and by far the most robust predictor in the RT-CIT. Since the ML did not capture any of the other features as having substantial predictive power in the included 12 experiments, it is unlikely that any of them is of great interest as a predictor of guilt (i.e., probe recognition) on either practical or theoretical grounds that is generalizable to the RT-CIT.

Contributions

Concept and data curation by GL, design and statistical analysis by DS and GL, manuscript by GL and DS.

Funding information

Gáspár Lukács has been funded by the OeAW Post-DocTrack Program and the JSPS Postdoctoral Fellowship for Research in Japan. These funding sources had no role or involvement related to this study.

Competing interests

No competing interests exist.

Data accessibility statement

All data, analysis scripts, supplementary material, and preregistrations are available via https://osf.io/zhmb2/.

Submitted: July 15, 2021 PST, Accepted: February 15, 2022 PST

REFERENCES

Bablani, A., Edla, D. R., Tripathi, D., & Kuppili, V. (2019). An efficient Concealed Information Test: EEG feature extraction and ensemble classification for lie identification. *Machine Vision and Applications*, *30*(5), 813–832. https://doi.org/10.1007/s00138-018-0950-y

Benjamens, S., Dhunnoo, P., & Meskó, B. (2020). The state of artificial intelligence-based FDA-approved medical devices and algorithms: An online database. *Npj Digital Medicine*, *3*(1), 118. https://doi.org/10.1038/s41746-020-00324-0

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Bouckaert, R. R., & Frank, E. (2004). Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms. In H. Dai, R. Srikant, & C. Zhang (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 3–12). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-24775-3_3

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Carriquiry, A., Hofmann, H., Tai, X. H., & VanderPlas, S. (2019). Machine learning in forensic applications. *Significance*, *16*(2), 29–35. https://doi.org/10.1111/j.1740-9713.2019.01252.x

Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, *11*(70), 2079–2107.

Davatzikos, C., Ruparel, K., Fan, Y., Shen, D. G., Acharyya, M., Loughead, J. W., Gur, R. C., & Langleben, D. D. (2005). Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection. *NeuroImage*, *28*(3), 663–668. https://doi.org/10.1016/j.neuroimage.2005.08.009

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology*, *10*, 102. https://doi.org/10.3389/fpsyg.2019.00102

DeMasi, O., Kording, K., & Recht, B. (2017). Meaningless comparisons lead to false optimism in medical machine learning. *PLOS ONE*, *12*(9), e0184604. https://doi.org/10.1371/journal.pone.0184604

Deng, L., & Li, X. (2013). Machine Learning Paradigms for Speech Recognition: An Overview. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(5), 1060–1089. https://doi.org/10.1109/tasl.2013.2244083

Derakhshan, A., Mikaeili, M., Gedeon, T., & Nasrabadi, A. M. (2020). Identifying the Optimal Features in Multimodal Deception Detection. *Multimodal Technologies and Interaction*, *4*(2), 25. https://doi.org/10.3390/mti4020025

Dodia, S., Edla, D. R., Bablani, A., Ramesh, D., & Kuppili, V. (2019). An efficient EEG based deceit identification test using wavelet packet transform and linear discriminant analysis. *Journal of Neuroscience Methods*, *314*, 31–40. https://doi.org/10.1016/j.jneumeth.2019.01.007

Edgington, E. S., & Onghena, P. (2007). *Randomization tests* (4th ed.). Chapman & Hall/CRC.

Efron, B. (1992). Bootstrap methods: Another look at the jackknife. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in Statistics* (pp. 569–593). Springer New York. https://doi.org/10.1007/978-1-4612-4380-9_41

Elaad, E., & Ben-Shakhar, G. (1997). Effects of item repetitions and variations on the efficiency of the guilty knowledge test. *Psychophysiology*, *34*(5), 587–596. https://doi.org/10.1111/j.1469-8986.1997.tb01745.x

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

Geven, L. M., Ben-Shakhar, G., Kindt, M., & Verschuere, B. (2018). Memory-based deception detection: Extending the cognitive signature of lying from instructed to self-initiated cheating. *Topics in Cognitive Science*, *12*(2), 608–631. https://doi.org/10.1111/tops.12353

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Hartwig, M., & Bond, C. F. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin*, *137*(4), 643–659. https://doi.org/10.1037/a0023589

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Hu, X., Evans, A., Wu, H., Lee, K., & Fu, G. (2013). An interfering dot-probe task facilitates the detection of mock crime memory in a reaction time (RT)-based concealed information test. *Acta Psychologica*, *142*(2), 278–285. https://doi.org/10.1016/j.actpsy.2012.12.006

Kleinberg, B., van der Toolen, Y., Vrij, A., Arntz, A., & Verschuere, B. (2018). Automated verbal credibility assessment of intentions: The model statement technique and predictive modeling. *Applied Cognitive Psychology*, *32*(3), 354–366. https://doi.org/10.1002/acp.3407

Kleinberg, B., & Verschuere, B. (2015). Memory detection 2.0: The first web-based memory detection test. *PLOS ONE*, *10*(4), e0118715. https://doi.org/10.1371/journal.pone.0118715

Kleinberg, B., & Verschuere, B. (2016). The role of motivation to avoid detection in reaction time-based concealed information detection. *Journal of Applied Research in Memory and Cognition*, *5*(1), 43–51. https://doi.org/10.1037/h0101804

Lemm, S., Blankertz, B., Dickhaus, T., & Müller, K.-R. (2011). Introduction to machine learning for brain imaging. *NeuroImage*, *56*(2), 387–399. https://doi.org/10.1016/j.neuroimage.2010.11.004

López de Prado, M. M. (2018). *Advances in financial machine learning*. Wiley.

Lubczyk, T., Lukács, G., & Ansorge, U. (2022). Speed versus accuracy instructions in the response time concealed information test. *Cognitive Research: Principles and Implications*, *7*(1), 3. https://doi.org/10.1186/s41235-021-00352-8

Lukács, G., & Ansorge, U. (2021). The mechanism of filler items in the response time concealed information test. *Psychological Research*, *85*(7), 2808–2828. https://doi.org/10.1007/s00426-020-01432-y

Lukács, G., Gula, B., Szegedi-Hallgató, E., & Csifcsák, G. (2017). Association-based Concealed Information Test: A novel reaction time-based deception detection method. *Journal of Applied Research in Memory and Cognition*, *6*(3), 283–294. https://doi.org/10.1037/h0101811

Lukács, G., Kleinberg, B., & Verschuere, B. (2017). Familiarity-related fillers improve the validity of reaction time-based memory detection. *Journal of Applied Research in Memory and Cognition*, *6*(3), 295–305. https://doi.org/10.1016/j.jarmac.2017.01.013

Lukács, G., & Specker, E. (2020). Dispersion matters: Diagnostics and control data computer simulation in Concealed Information Test studies. *PLOS ONE*, *15*(10), e0240259. https://doi.org/10.1371/journal.pone.0240259

Matsuda, I., Ogawa, T., & Tsuneoka, M. (2019). Broadening the Use of the Concealed Information Test in the Field. *Frontiers in Psychiatry*, *10*, 24. https://doi.org/10.3389/fpsyt.2019.00024

Meijer, E. H., Verschuere, B., Gamer, M., Merckelbach, H., & Ben-Shakhar, G. (2016). Deception detection with behavioral, autonomic, and neural measures: Conceptual and methodological considerations that warrant modesty. *Psychophysiology*, *53*(5), 593–604. https://doi.org/10.1111/psyp.12609

Mitchell, T. M. (2013). *Machine learning* (Nachdr.). McGraw-Hill.

Molnar, C. (2019). *Interpretable machine learning. A guide for making black box models explainable*. https://christophm.github.io/interpretable-ml-book/

Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. *Machine Learning*, *52*(3), 239–281. https://doi.org/10.1023/a:1024068626366

Noordraven, E., & Verschuere, B. (2013). Predicting the sensitivity of the reaction time-based Concealed Information Test. *Applied Cognitive Psychology*, *27*(3), 328–335. https://doi.org/10.1002/acp.2910

Ojala, M., & Garriga, G. C. (2010). Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, *11*(62), 1833–1863.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review*, *9*(2), 278–291. https://doi.org/10.3758/bf03196283

Reich, L., Gula, B., & Alexandrowicz, R. W. (2018, September 27). *Detection of deception in a RT-CIT mock crime paradigm with the Drift Diffusion Mode* [Conference presentation abstract]. 13th Alps-Adria Psychology Conference (AAPC18). https://arts.units.it/retrieve/handle/11368/2944232/271042/AA_2018_abstracts.pdf

Riley, P. (2019). Three pitfalls to avoid in machine learning. *Nature*, *572*(7767), 27–29. https://doi.org/10.1038/d41586-019-02307-y

Strahm, S. (2017). *Reaction time models applied to the RT-CIT: Improvement in fit and classification accuracy through accounting for skewness and inclusion of descriptive parameters*. https://www.researchgate.net/publication/322250401

Suchotzki, K., Verschuere, B., Van Bockstaele, B., Ben-Shakhar, G., & Crombez, G. (2017). Lying takes time: A meta-analysis on reaction time measures of deception. *Psychological Bulletin*, *143*(4), 428–453. https://doi.org/10.1037/bul0000087

Theodoridis, S., & Koutroumbas, K. (2009). *Pattern recognition* (4th ed.). Elsevier Acad. Press.

Van den Akker, O., Weston, S. J., Campbell, L., Chopik, W. J., Damian, R. I., Davis-Kean, P., Hall, A. N., Kosie, J. E., Kruse, E. T., Olsen, J., Ritchie, S. J., Valentine, K. D., van 't Veer, A. E., & Bakker, M. (2019). *Preregistration of secondary data analysis: A template and tutorial*. https://doi.org/10.31234/osf.io/hvfmr

Van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E.-J. (2020). Bayesian rank-based hypothesis testing for the rank sum test, the signed rank test, and Spearman's ρ. *Journal of Applied Statistics*, *47*(16), 2984–3006. https://doi.org/10.1080/02664763.2019.1709053

Verschuere, B., Kleinberg, B., & Theocharidou, K. (2015). RT-based memory detection: Item saliency effects in the single-probe and the multiple-probe protocol. *Journal of Applied Research in Memory and Cognition*, *4*(1), 59–65. https://doi.org/10.1016/j.jarmac.2015.01.001

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01832

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). Morgan Kaufmann.

Wojciechowski, J., & Lukács, G. (2022). Importance‐related fillers improve the classification accuracy of the response time concealed information test in a crime scenario. *Legal and Criminological Psychology*, *27*(1), 82–100. https://doi.org/10.1111/lcrp.12198

SUPPLEMENTARY MATERIALS

Peer Review History

Download: https://collabra.scholasticahq.com/article/32661-machine-learning-mega-analysis-applied-to-the-response-time-concealed-information-test-no-evidence-for-advantage-of-model-based-predictors-over-basel/attachment/82658.docx?auth_token=1mGrZ-AibRGilOBqhQmM