ISSN: 2638-9231

Gratis

Research Article

Annals of Behavioral Neuroscience

Comparison of Diagnostic Accuracy Using Bootstrapping Methods versus a Machine Learning Algorithm Using the Complex Trial Protocol (CTP)

Ward AC*, Rosenfeld JP, Kelley J and McCann D

Department of Psychology, Northwestern University, Evanston IL, United States

***Correspondence:** Anne Cable Ward, Department of Psychology, Scientist & Subject Matter Expert at Brainwave Science Inc, Northwestern University, Evanston IL, United States, E-mail: anneward2013@u.northwestern.edu

Received date: August 08, 2019; Accepted date: January 31, 2020; Published date: February 13, 2020

Abstract

Because Machine Learning (ML) algorithms are objective and can find patterns in large data sets that might go unnoticed by the naked eye, ML often improves clinical diagnosis of disease. Since the P300-based Concealed Information Test (CIT) also uses biomedical data to make a "diagnosis" about whether someone is knowledgeable about crime-related details, its accuracy might also improve through the use of ML. Using the countermeasure-vulnerable "three stimulus protocol" (3SP), others have shown that ML outperformed the traditional bootstrapping and cross-correlational methods for diagnosis. Here, we expand on their work by applying ML techniques to data acquired using the countermeasure-resistant Complex Trial Protocol (CTP) version of the P300-based CIT, with the goal of comparing diagnostic accuracy between this approach and our currently used bootstrapping method. As expected, grand averaged ERPs showed the CIT effect (i.e., a larger probe compared to irrelevant response) in the guilty group but not the innocent group. While results comparing diagnostic methods revealed a seemingly higher hit rate (80% vs. 73%) and area under the ROC curve (AUC) (.872 vs. .712) using bootstrapping compared to ML, standardized *z*-scores did not provide evidence suggesting the superiority of one approach over the other ($Z = 1.06$, $p>.2$, two-tailed). These findings offer no evidence that the ML algorithm used here increases diagnostic accuracy of the CTP, and thus do not support that the ML method should replace the currently used bootstrapping method. Limitations that may pertain to our findings are discussed, along with future directions.

Introduction

The concealed information test (CIT)

The Concealed Information Test [1] is a method for detecting recognition of details that only someone related to a crime would know. During a CIT, subjects are presented with three types of stimuli. The first-the probe-is the item investigators and involved suspects know to be related to the crime (e.g., the murder weapon). Items in the next stimulus category are selected to be from the same domain as the probe and are called "irrelevants" because, although they are similar to the probe and typically require the same behavioral response as the probe, they are not related to the crime under investigation. For example, if a murder were committed with a gun, irrelevant items might include images of a knife, bat, etc. When shown an image of the murder gun (i.e., the probe) presented among images of other weapons unrelated to the crime (i.e., irrelevants), those who are unfamiliar with the crime's details should respond similarly to all items. However, because the murder gun is distinguishable amongst the other items for knowledgeable individuals due to its familiarity and salience, individuals involved in the crime are expected show a differential physiological response to the probe (i.e., the "CIT effect). The third class of stimuli requires a different behavioral response than

the probe and irrelevants, and is called "targets" because they are included to force task engagement/attention. Because its unique button response makes the target meaningful, elevated physiological target responses are expected for both knowledgeable and unknowledgeable subjects.

The CIT has a firm theoretical basis, is more empirically established than alternative methods for detecting recognition, and has been researched using a variety of physiologically measures, including Skin Conductance Rate (SCR), Respiration Line Length (RLL), and Heart Rate (HR) [2,3]. While these Autonomic Nervous System (ANS) measures can effectively detect recognition (with Cohen's *d* effect sizes of *d* = 1.55, 1.11, and 0.89 for SCR, RLL, and HR, respectively), the most robust CIT indicator of recognition is P300 amplitude (*d* = 1.89;) [4]. Because this endogenous ERP occurs involuntarily when a rarely presented meaningful item-like a murder weapon-is recognized, it can be used to determine who is privy to a crime's details by comparing P300 amplitude in response to probes and irrelevant items.

Concealed information detection using bootstrapping

Much of the original work using P300 to detect concealed information used t-tests to look for intra-individual differences between the means obtained from single sweep probe and irrelevant ERPs [5,6]. Because single sweeps are very noisy and several sweeps are typically needed to detect signals unless they are very pronounced, most analyses in ERP physiological work has examined group differences by comparing mean amplitudes of grand averages, which include the averages of all participants (Ps) in a condition. For each individual to have several probe and irrelevant averages, they would need to complete the test many times, which is not feasible (due to time constraints, expense, habituation, the risk of irrelevants becoming relevant through repeated exposure, etc.). As a result of comparing noisy single sweeps, individual diagnostic accuracy rates in these original studies were low (< 80%).

Farwell and Donchin (1991) solved the problem of comparing single sweep probe and irrlevants and made accurate individual diagnosis possible by applying the bootstrapping method [7,8] to P300-based concealed information detection to produce multiple probe and irrelevant averages for each individual. Using bootstrapping as we currently do, the ultimate goal is to establish that a CIT effect is present in a given individual (e.g., determine with at least a 90% confidence that the probe P300 amplitude is truly larger than the irrelevant P300 value). To do this, distributions of probe and the combined irrelevant (Iall) averages are needed for each P. As detailed below (see "Bootstrapped Amplitude Difference Method" section), distributions are created by randomly sampling with replacement from sets of single probe and single irrelevant waves, creating average probe and irrelevant ERPs based on these sampled sweeps, finding the average probe-irrelevant difference and adding it to a distribution (consisting of 100 difference values after this process is repeated 100 times), and then drawing the lower bound of the confidence interval at the point in the distribution that makes the confidence interval > 0 (e.g., a cutoff -1.29 SDs corresponds to a 90% confidence that the probe and irrelevants differ). While guilt is often inferred if the probe-Iall bootstrapped differences are > 0 in at least 90 of 100 iterations, this criterion can be adjusted to achieve the desired sensitivity/specificity (based on the severity of the consequences of false positives or false negatives).

Instead of diagnosing concealed information based on the number of bootstrapped iterations where the probe is larger than Iall, others have instead relied on cross-correlation coefficients [7]. Essentially, this method creates bootstrapped average ERPs for the probe, Iall, and the target, which are then used to determine whether the probe P300 is more similar in shape to the target or to Iall. Single sweeps from the probe are compared to those of the target and Iall, and these values are used to make two distributions: probe vs. Iall and probe vs. target. This approach assumes that a greater cross-correlation between the probe and Iall (A) indicates innocence, because it suggests that the probe is not meaningful for the P, while a greater cross-correlation between the probe and target (B) indicates knowledge recognition (guilt), because it suggests that the probe is meaningful like the target. In this case, concealed information is typically indicated if at least 90 of 100 bootstrapped correlation subtractions (B-A) are > 0 (positive).

The bootstrapped cross-correlation approach has limitations however, including that it uses target P300s as templates for probe P300s, and targets and probes are inherently different; probe P300s occur because an item is meaningful, while target P300s are evoked due to task-relevance. This can result in latency, phase, and other morphology differences [9,10], these differences make the target an inadequate index for probe comparison, and so unlike the cross-correlation approach, the more

commonly used bootstrapped amplitude difference method instead simply compares probe and irrelevant amplitudes.

Research comparing the bootstrapped peak-to-peak (p-p) amplitude difference method to the bootstrapped cross-correlation approach has supported the superiority of the former [9,11,12]. However, these traditional time-domain based approaches are not ideal since they fail to tease apart functionally separate, yet co-occurring neural events, the summed effects of which are reflected in ERPs. Considering the limitations associated with restricting analyses to a time domain, one weakness of the bootstrapped p-p amplitude difference method is that it only looks at part of the wave, meaning that any activity occurring outside a given time frame that could improve diagnostic accuracy gets ignored. When a frequency domain has instead been used, components in different frequency ranges have demonstrated their association with discrete neural events. For example, some frequencies have shown stronger associations with information processing, and others with behavioral events [13]. However, while Fourier transformation can help begin to relate frequency to functionally separate neural events, it fails to take time into account.

Concealed information detection using machine learning (ML)

To improve the identification of separate functional components, EEG/ERP signals must be considered in dimensions that account for both frequency and time, which can be accomplished using computer algorithms. One such algorithm that incorporates both domains-wavelet transformation-can establish when and to what extent distinct events occur within a waveform (allowing multiple underlying neural events to be represented), as well as when separate frequencies occur, and how this changes over the course of the wave [14]. Because Machine Learning (ML) incorporates both time and frequency domains, ML methods can be applied to recognize patterns in complex data sets, through the use of a statistical classifier and wavelet features.

To illustrate how features are selected, consider how one can tell a circle and a square apart; these objects can be broken down into discrete features, like number of sides, angles within shape, and curvature measures within shape, etc., which discriminative ML algorithms (e.g., logistic regression cost minimization) can use to find a plane or line in the feature vector space that best delineates squares from circles, based on the predefined features. Put differently, features or attributes selected because they are thought to be associated with a given output (e.g., shape type) are extracted, and statistical classifiers (i.e., functions) input this data and assign it a label output (e.g., circle vs. square, probe vs. irrelevant). A major benefit of ML is that it allows one to incorporate features or attributes which are not apparent to the naked eye, allowing the classifier to customize a unique and potentially non-linear delineation based upon the data, instead of on preconceived notions about how the data is expected to behave, which could ultimately lead to more objectivity and diagnostic accuracy.

Applied to concealed information detection, probe and irrelevant responses can be broken down into different features or attributes, which a ML algorithm can use to find a division between the classes “guilty” and “innocent,” based on quantitative feature measures. In an attempt to do this using P300,[11]compared their ML pattern recognition method to the bootstrapped amplitude difference method and the cross-correlation approach and found evidence favoring either ML or the bootstrapped amplitude difference method (depending upon the chosen Receiver Operator Characteristic[ROC] criterion), and that the bootstrapped cross-correlation method performed consistently (though slightly) worse. Extending upon this research, [15] further refined their initial pattern recognition system by adding new morphological and frequency features to the previously used wavelet features. In addition to outperforming their previous attempt, their results also exceeded the bootstrapped p-p amplitude difference and cross-correlation approaches.

Applying ml to complex trial protocol (CTP) data

While the findings of [11,15] suggest that ML approaches may be preferable to the traditionally used bootstrapped p-p amplitude difference method, it is important to note that the comparisons between the methods for diagnosis discussed above were drawn from data acquired using the suboptimal “Three Stimulus Protocol” [10]. One shortcomings of this approach is that it combines implicit probe/irrelevant and explicit target/nontarget discrimination tasks; presenting probe, irrelevant, and target stimuli at random requires a simultaneous implicit probe/irrelevant categorization and explicit target/nontarget discrimination on every trial (since the target requires a unique response). While the probe’s amplitude typically exceeds that of the irrelevants in knowledgeable individuals, the secondary target/nontarget task unnecessarily consumes cognitive resources and diverts

attention away from the primary probe/irrelevant categorization, reducing probe P300 amplitude [16,17]. In addition to this reduction of the CIT effect-the difference between probe and combined irrelevant responses, upon which a diagnosis is determined-another weakness of the 3SP is its vulnerability to "countermeasures," or deliberate attempts to beat the test. Specifically, making irrelevant stimuli meaningful by applying countermeasures when they are presented increases their P300, thus also reducing the CIT effect [9,12].

The Complex Trial Protocol (CTP) (Figure 1) [18] was developed to address the shortcomings of the 3SP and separates probe/irrelevant and target/nontarget responses in three ways: 1) temporally, by about a second, 2) spatially, by assigning left-hand probe/irrelevant responses and right-hand target/nontarget responses, and 3) by domain, using numbers as targets instead of simply making an irrelevant a target by assigning it significance.

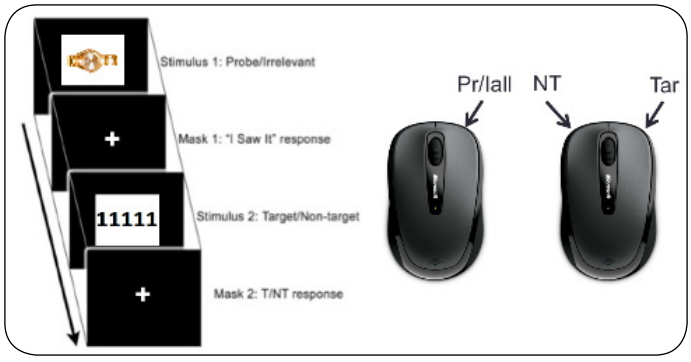

Figure1: Complex Trial Protocol trial design. "Pr" = probe; "Iall" = combined irrelevants; "NT" = nontarget; "Tar/T" = target.

These modifications increase target discriminability, which frees processing resources and allows for increased attention toward the probe, resulting in greater probe P300 amplitudes and a more dramatic CIT effect. Given the sparsity of research testing ML diagnosis methods on P300 CIT data and the notable differences between the 3SP and CTP, it unclear if ML methods might optimize concealed information diagnosis using the more sophisticated CTP version of the P300-based CIT. To that end, the goal of the current effort is to compare a discriminative ML algorithm to the currently used bootstrapped amplitude difference approach, using data acquired during the CTP.

Methods

Participants

The ML and traditional bootstrapped amplitude difference methods were tested on a dataset with 30 Ps from the Northwestern University's Introductory Psychology pool. All Ps had normal or corrected to normal vision, reported no history of neurological or psychological abnormalities, consented to participation, and received class credit.

Procedure

Ps were randomly assigned to either a simple guilty (SG; n = 16) or an innocent control group (n = 14). After being told the premise of the CIT they would later take (see Appendix A "Pre-Task Instructions"), Ps completed either a mock-crime or an innocent control task, followed by a P300-based CIT.

Mock-crime/Control task

Ps read along as their participation and task, which involved entering an office to steal an item from an envelope (SG) or to simply look inside an envelope (innocents), and then returning to the lab, were explained by the experimenter (see Appendix A). Half of Ps in the SG group stole a watch while remaining Ps stole a bracelet.

P300-based CIT

After the goal of the CIT and instructions on how to respond behaviorally to stimuli during the task were described (see Appendix B), all Ps completed the CTP version of the P300-based CIT. Ps were seated three feet from a monitor, upon which images of the probe (watch or bracelet), six irrelevants (comparable jewelry items), the target (i.e., "11111"), and four nontargets (i.e., "22222…55555"), appeared. Each CTP trial (Figure 1) began with a fixation cross "+" in the middle of the screen (200ms), which was replaced with probe or irrelevant (300ms), followed by 1100 ms of observation, and finally a randomly varying inter stimulus interval of 50-200 ms(with fixation "+"). This was repeated in the second part of the trial, except with target and nontarget stimuli. The target and four nontargets were each presented 20% of the time, resulting in a 20%-80% target to nontarget ratio. We also used a symmetric protocol–meaning that the target and each of the nontarget stimuli followed each irrelevant and the probe with the same frequency–in an attempt to prevent Ps from focusing their attention on detecting patterns to the target occurrences.

As described above, probe and irrelevant stimuli

required the same left-hand button response, simply acknowledging that the item was seen, while right-handed response were made for target (right mouse button) and nontarget (left mouse button) stimuli. The CIT consisted of 210 trials and lasted about 15 minutes. To help force attention to the stimuli, Ps were stopped unpredictably about every 50 trials and asked to name the previous item they had seen (as they were forewarned during their instructions).

Data acquisition

EEG was collected using Ag/AgCl electrodes attached midline to the scalp at Fz, Cz, and Pz, sites. Electro-oculogram (EOG) was recorded referentially with an electrode placed above the left eye to collect eye movements and blinks, and all electrodes were referenced to linked mastoids, and Ps were grounded via a forehead electrode. Impedance remained under 5 kΩ, eye blink artifacts were corrected using the Semlitsch method [19], and trials containing artifacts over 90 μV were dropped. The 19 channel Mitsar amplifier (Model 201) used a 30 Hz low pass and a 0.16 Hz high pass filter setting, and output passed through a 16-bit Mitsar A/D converter sampling at 500 Hz. For display and analyses, a Kaiser (alpha = 1.8) filtering algorithm (with the digital filter set to pass frequencies from 0 to 6 Hz), was used off-line to remove higher frequencies in single sweeps and averages.

Data Analysis Plan

ERP measurement

P300 was analyzed at Pz, where it is typically largest [20], using the p-p method, which has been shown to be 25% more accurate at detecting knowledge than the base-to-peak method [21]. To find p-p P300 values, the algorithm searched from 300-700 ms for the maximally positive 100ms segment, the midpoint of which is P300 latency. Searching from this P300 latency to 1400 ms, the algorithm also establishes a maximally negative 100 ms segment average, and then subtracts it from the maximally positive average to find the p-p amplitude value. As recommended by [22], our search windows were chosen based on a grand average including all Ps, across conditions. Individual ERPs were also examined, and search windows were minimally expanded for the eleven Ps withP300 peaks outside the default windows.

Since we expected to see the CIT effect only in the SG group, we used analysis of variance (ANOVA) methods to assess for differences in probe and irrelevant responses at the Pz site, and between the SG and innocent groups. Partial eta squared (ηp^2) is also provided as an effect size estimate [23]. Bayes Factors (JZS BFs, scaled r = .707) [24] calculated at http://pcl.missouri.edu/bayesfactor) are reported as directly interpretable odds ratios, and test whether a true difference between group means exists (favoring the alternative hypothesis), or not (favoring the null hypothesis). For example, if JZS BF = 1.0, the null and alternative hypotheses are equally likely, and this is an indeterminate outcome. BFs > 3 provide evidence for the alternative (+3) or the null (-3) hypothesis.

Bootstrapped amplitude difference method

As described above, the p-p bootstrapped amplitude difference method was used to determine if the probe's P300 response exceeded that of the irrelevants, for each individual. In this procedure, 30 probe waves are first randomly sampled (with replacement) from the original set of 30 single probe sweeps. These selected single sweeps are then used to create a probe ERP, from which an average bootstrapped probe P300 value is computed (where P300 is defined as the largest 100 ms segment average in the 300-700 ms following the stimulus, minus the maximally negative 100 ms segment occurring between the P300 latency and 1400 ms, for example). The same process is then carried out with irrelevant stimuli, and 30 of the original 180 irrelevant single sweeps are sampled and used to create an ERP, from which an average bootstrapped irrelevant value is computed. The difference between these two values-the bootstrapped probe average minus the bootstrapped irrelevant average-is then calculated. This process of: a) randomly selecting 30 probe and 30 irrelevant sweeps, b) creating probe and irrelevant ERPs, and c) comparing their amplitudes, is typically carried out 100 times. The resulting 100 probe-irrelevant difference values are then used to create a distribution, from which a diagnosis is made, based on the number of iterations where the probe>Iall. For example, if the criterion for a "knowledgeable" determination is 90%+probe>Iall iterations, a cutoff is drawn 1.29 standard deviations below the mean, making the lower bound of the 90% confidence interval >0 (since negative values in the distribution represent the iterations where Iall>probe).

Bootstrapping offers an alternative to gauging the CIT effect based on measures reported directly in voltages(i.e., probe-irrelevant amplitude differences).To determine if the CIT effect is present (i.e., to diagnose concealed knowledge), we typically use a criterion of 90 out of 100 iterations where the probe P300 is larger than Iall P300, measured p-p. However, this 90% criterion is arbitrary, and studies testing for episodic information

(e.g., stolen jewelry, as used here), which is less salient and thus typically produces smaller P300s [25] than semantic information (e.g., name, hometown), often use a less stringent criterion of 85% or 80%. A final and related measure of the CIT effect is the difference between bootstrapped probe and irrelevant mean amplitudes. During the bootstrapping procedure, bootstrapped P300 amplitude averages are produced for both probes and irrelevants during each iteration, and the average of these sample means can be used to estimate a probe and irrelevant population mean difference.

ML method

The discriminative ML algorithm tested here was modeled from [15]. However, instead of using linear discriminant analysis, we used logistic regression. For a binary classification-such as knowledgeable vs. unknowledgeable-logistic discriminant analysis predicts a normal density function for each classification feature, establishing a linear boundary where they meet. However, unlike linear discriminant analysis, logistic regression instead creates a class boundary by predicting the log-odd function between the classes, which may be preferable since it does not make a-priori assumptions about distributions of predictor variables. Additionally, while [15] employed discrete wavelet transforms, we did not. Finally, while [15] used genetic algorithms to pick the classifier features, we instead selected a subset of the features they employed. After testing combinations to see which best fit the data, the following were included in our set our features:

Amplitude: The maximum signal value.

Positive area: The sum of the positive signal values.

Negative area: The sum of the negative signal values.

Total area: The sum of positive and negative signal values.

Total absolute area: The sum of the positive signal values and the absolute value of the negative signal values.

Zero crossing: The number of times the signal value equaled zero, in the peak-to-peak time window.

Slope sign alterations: the number of slope sign alterations of a pair of points adjacent on the ERP signal.

Using these features, the statistical classifier was trained on a random subset (SG: n = 8; innocent n = 7) of the sample, and then applied to the remaining half (SG: n = 8; innocent n = 7) for diagnosis. After the chosen features were extracted from a given probe EEG single sweep, the algorithm determined whether that sweep more closely resembled a probe or an irrelevant response. For purposes of classification (and in an attempt to reduce variance across Ps), cutoffs were established for each P based on the assumption that probe responses mirror target responses. Since target responses should be elevated in all Ps while elevated probe responses should be restricted to the SG group, a knowledgeable diagnosis was made if the classifier identified probe sweeps (in the first part of the trial) as such≥ the percentage of target sweeps classified as a probe (in the second part of the trial). For instance, if a P's target response was classified as a probe 60% of the time, that P would be determined guilty/knowledgeable if probe sweeps were classified as such at a rate ≥ 60%.

Comparing bootstrapping and ml

Because diagnostic accuracy rates for each method are based on different wave characteristics (e.g., strictly amplitude for bootstrapping and several features for ML) and are produced using different criteria (ML: % target sweeps classified as a probe; bootstrapping: number of iterations out of 100 where the probe>Iall), hit rates are not an ideal index for making methodological comparisons. To best compare ML algorithm to the traditional bootstrapping method, areas under the ROC curve (AUCs), which reflect both sensitivity (i.e., hits) and specificity (i.e., correct rejections) at each possible decision criterion, were obtained through Signal Detection Theoretical analysis [26]. To determine if one method outperformed the other, we compared the ML and bootstrapping AUCs using Z-tests, as suggested by [27]. Since z-scores are standardized (i.e., they indicate how many standard deviations away from the mean a particular score lies), they are useful for comparing AUCs derived using different sample sizes and classification criteria, as is the case here. We expected that, although both methods would have a high AUC (.9+), the ML algorithm might improve upon our traditional bootstrapping diagnostic technique, as determined by z-score.

Results

ERP qualitative analysis

After excluding the <10% of Ps who failed to follow instructions (e.g., incorrect recall during more than one pop quiz) or who had excessive artifacts, 30 data sets remained. The mean number of probe trials across all subjects was 28, and all Ps had at least 20 probe sweeps. Figure 2 shows the grand average ERPs at Pz for the SG and innocent groups, which demonstrate the expected CIT effect: elevated P300 responses to the probe when compared to Iall in the SG group, but not the innocent group.

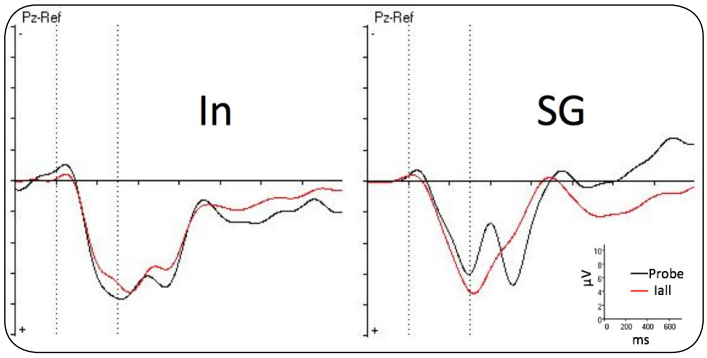

Figure 2: Grand average innocent (In; *n* = 14) and guilty (SG; *n* = 16) event related potentials, in response to probe and combined irrelevant ("Iall") stimuli.

ERP quantitative analysis

A 2 (group: SG vs. innocent) X 2(stimulus: probe vs. Iall) mixed ANOVA was run on p-p P300 amplitudes to confirm the CIT effect illustrated above. Average probe and Iall amplitudes (in micro volts [μV]) were probe = 10.68, Iall = 9.92 and probe = 14.59, Iall = 9.97 for the innocent and SG groups, respectively. The main effect of group did not reach significance, $F(1,28) = 1.318$, $p = .261$, possibly due to the interaction seen in (Figure 3). The ηp^2approached medium at .045, but the JZS BF favored the null hypothesis at 2.83. As expected, the main effect of stimulus type was large, $F(1,28) = 14.87$, $p = .001$, $\eta p^2 = .347$, and the JZS BF clearly favored the alternative at 45.754. The interaction was also significant, $F(1,28) = 7.64$, $p = .01$, $\eta p^2 = .214$, with JZS BF favoring the alternative at 5.494. Figure 3 supports that the source of this interaction is the larger probe in the SG group compared to the innocent group, and illustrates the larger probe-Iall difference (i.e., the CIT effect) in the SG group. Post-hoc t-tests provided some support for this, with $t(1,28) = 1.919$, $p = .065$, $\eta p^2 = .116$, and JZS BF suggesting the equal likelihood of the null and alternative hypothesis (1.332, favoring the alternative)for probes, and $t(1,28) = .033$, $p = .974$, $\eta p^2 < .001$, with JZS BF favoring the null at 2.898for Iall.

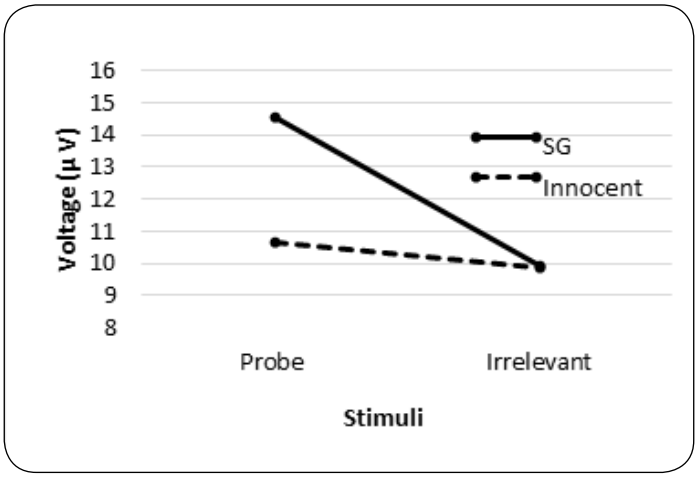

Figure 3: Computed grand average peak-to-peak P300 values for probes and combined irrelevants as a function of condition (Guilty [SG] vs. Innocent) and stimulus type (probe vs. irrelevant).

Bootstrapping

Further evidence supporting that the CIT effect is present and limited to the SG group is offered by results derived using the bootstrapping procedure: A t-test was conducted to assess group differences in the number of bootstrapped iterations where the probe amplitude exceeded that of the irrelevants, and revealed that the probe was larger than Iall more frequently in the SG group (SG mean = 83.0; Innocent mean = 49.8; $t[1,28] = 4.475$, $p<.001$, $\eta p^2 = .070$,JZS BF = 184.22, favoring the alternative). A final t-test on the average bootstrapped mean p-p probe-Iall differences (SG mean = 4.19; Innocent mean = 0.22) showed that this index of the CIT effect was also larger in the SG compared to the innocent group ($t[1,28] = 3.214$, $p<.01$, $\eta p^2 = .060$, JZS BF = 11.952, favoring the alternative).

Outcomes of both the bootstrapping and ML procedures are displayed in table 1, by group (SG vs. innocent). Using the criterion of 85 out of 100 iterations where the probe>Iall for a knowledgeable diagnosis, the bootstrapping procedure accurately classified 24 of 30 (80%) individuals (SG: 11/16 = 68.75%; Innocent: 13/14 = 92.86%). Loosening this criterion to 80 of 100 iterations did not change any determinations, and investigation of b-p bootstrapped results showed that measuring P300 b-p caused diagnostic accuracy rates to suffer by >15% (and thus formal b-p analyses were not conducted).

ML

Values produced by the ML method-which represent the percentage of stimuli in the first part of the trial that were labeled as a probe (i.e., another irrelevant item coded as the probe for analysis purposes for innocent Ps), were higher in the SG group compared to the innocent group, but not significantly so (SG mean = .466; Innocent mean = .283; t[1,13] = 1.804, p = .094, ηp^2 = .200, JZS BF = 1.144, favoring the alterative).Using a participant-specific criterion of % target sweeps classified as probes, the ML algorithm produced an overall diagnostic accuracy of 73.33% (SG: 5/8 = 62.5%; Innocent: 6/7 = 85.71%).

ROC analysis

ROC curves derived from the bootstrapping and ML methods are displayed in figure 4. While results comparing diagnostic methods showed a higher AUC when the bootstrapping approach was used (.872 vs. .712 for ML), standardized z-values revealed that diagnostic accuracies did not differ significantly (Z = 1.06, p>.2, two-tailed).

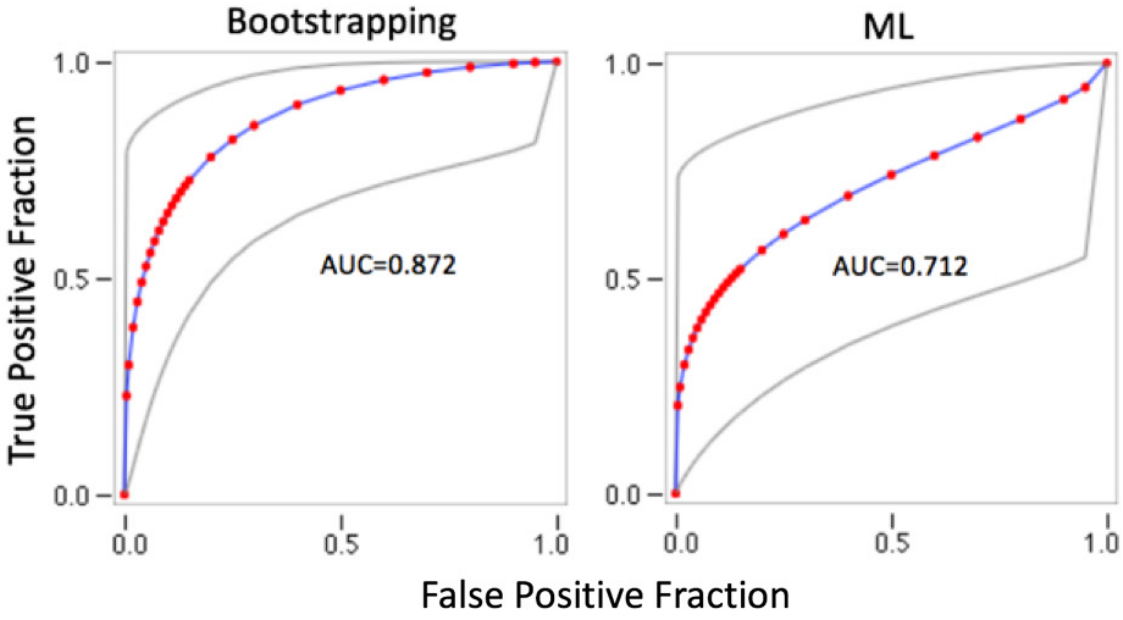

Figure 4: Receiver operator characteristic (ROC) curves (blue lines with red dots) and areas under the curve (AUCs) for the bootstrapping and machine learning (ML) methods. Gray curves show the 95% CI.

Table 1: Values obtained by bootstrapping and machine learning (ML) procedures, used for receiver operator characteristic (ROC) analysis, by group (simple guilty [SG] and Innocent [In]).

Bootstrapping	SG	100, 100, 99, 98, 98, 98, 94, 92, 92, 91, 85, 75, 58, 56, 52, 40
	In	23, 30, 31, 34, 34, 36, 39, 56, 59, 60, 61, 63, 79, 92
ML	SG	0.65, 0.59, 0.56, 0.55, 0.52, 0.48, 0.38, 0.0
	In	0.1, 0.16, 0.16, 0.17, 0.32, 0.49, 0.58

Discussion

Our goal was to compare the traditionally used bootstrapped amplitude difference method with a novel ML approach designed to diagnose concealed knowledge, using the most empirically validated and countermeasure-resistant version of the P300-based CIT: the CTP. While we expected both diagnostic approaches to yield high detection rates for both the SG and innocent groups, it was less clear whether the theoretical benefits of the ML approach, and the promising results of previous research using ML algorithms for P300-baseddiagnosis of concealed information, would be supported. To investigate this, we applied both methods to a single dataset which included both an innocent and a SG condition, and then compared standardized z-scores to determine if one approach outperformed the other.

Visual inspection of the SG and innocent ERPs (Figure 2) shows an obvious CIT effect in the former but not the

later, as expected. This was supported by data suggesting that the elevated P300 response in the SG compared to the innocent condition was specific to the probe, and group differences in the CIT effect, as measured by: probe-Iall amplitudes, number of bootstrapped iterations where the probe>Iall, and average bootstrapped mean probe-Iall differences. This effect is also suggested by the ML outcome data showing the % probe sweeps classified as a probe, which was marginally larger (p<.1)in the SG group.

However, while these expected group differences in the CIT effect emerged, individual hit rates using the well-vetted bootstrapping method were lower than typically observed. A few limitations may explain the relatively low hit rates. The first-limited exposure to the probe-is a result of probing for episodic information (i.e., jewelry stolen during a mock crime). Because semantic information is well-rehearsed and salient, it oftenproducesP300 responses large enough to obscure differences in accuracies between the bootstrapping and ML approaches. In attempt to prevent these ceiling effects, we used episodic information, which may not have been adequately encoded during the very brief exposure to the stolen item during the mock-crime.

Other factors that may have contributed to suboptimal accuracy rates concern stimuli selection and presentation. For example, because CTP stimuli presentation occurs so quickly, guilty Ps may not have recognized the item they stole, reducing its P300. Alternatively, guilty and innocent Ps may only be sure or unsure about what a few items in the stimuli set are, which is problematic for multiple reasons. First, this uncertainty might consume cognitive resources, for example, if a P focuses on the item in an attempt to identify it, oris worried the experimenter will quiz them on an ambiguous item. In guilty Ps, increased cognitive load may decrease the probe's amplitude [16] and reduce the expected CIT effect (probe-Iall difference). When unsure of what items are, Ps may also make unintended associations between them. For example, Ps might think the cufflinks are earrings (which appear as another irrelevant), or a guilty P might think the irrelevant image of a ring is also a bracelet (their probe item). In these cases, irrelevants could generate an oddball response that would harm diagnostic accuracy.

To reduce item confusion in the future, possible solutions include showing all Ps the images they will see on the subsequent CIT (with labels), or using words (e.g., "bracelet") as stimuli. The false positives seen in the SG group also highlight why it is important to test all stimuli sets on a known innocent group before tests are administered to suspects, so that irrelevant items that produce an elevated response can be removed.

To test the diagnostic efficacy of each approach, our main comparison of interest was between the AUC values derived from each method. While AUCs were computed by the bootstrapping procedure >15% higher than the value found for the ML approach, standardized *z*-scores did not suggest the superiority of one approach over another. However, the findings presented here are admittedly limited by their sample size, and while *z*-scores allow for standardization of unequal sample sizes so AUCs can be compared, our findings would have benefited from data from more participants. As such, we plan compare the ML and bootstrapping methods again, incorporating the data of a recently run study that included an innocent condition (to allow for ROC analysis).

While findings presented here do not support that the ML algorithm should replace the currently used bootstrapping method for diagnosis using the CTP, further refinement of the algorithm and inclusion of additional features might improve its accuracy. For example, since probe P300 latencies and behavioral reaction times, and target/nontarget error rates are often elevated in guilty individuals, algorithms might benefit from incorporating such non-neural features. Indeed, combining ERP and behavioral measures has been shown to outperform single outcome diagnosis [28], and might offer a promising avenue for future investigation. Another pragmatic consideration that supports the use of the bootstrapping method is that it can be applied on the individual, single subject level and does not require a training data set as ML does. While the ML algorithm did not improve CIT diagnostic accuracy, it may be useful in clinical contexts where P300 is used to make diagnoses based on subjective clinician judgment, which may benefit more from adopting a ML approach than the already objective bootstrapping method we used for comparison here.

Reference

1. Lykken DT. The GSR in the detection of guilt. *Journal of Applied Psychology*. 1959; 43(6):385-388. DOI: https://doi.org/10.1037/h0046060
2. Verschuere B, Ben-Shakhar G, Meijer E. Memory detection: Theory and application of the concealed information test. *Cambridge University Press*. 2011. Availble from: https://psycnet.apa.org/

record/2011-07195-000

3. Ben-Shakhar G. Current research and potential applications of the concealed information test: an overview. *Front Psychol*. 2012; 3. DOI: https://doi.org/10.3389/fpsyg.2012.00342
4. Meijer EH, klein Selle N, Elber L, Ben-Shakhar G. Memory detection with the Concealed Information Test: A meta analysis of skin conductance, respiration, heart rate, and P300 data. *Psychophysiology*. 2014; 51(9):879-904. DOI: https://doi.org/10.1111/psyp.12239
5. Rosenfeld JP, Nasman VT, Whalen R, Cantwell B, Mazzeri L. Late vertex positivity in event-related potentials as a guilty knowledge indicator: a new method of life detection. *Int J Neurosci*. 1987; 34(1-2):125-129. DOI: https://doi.org/10.3109/00207458708985947
6. Rosenfeld JP, Cantwell B, Nasman VT, Wojdac V, Ivanov S, Mazzeri L. A modified, event-related potential-based guilty knowledge test. *International Journal of Neuroscience*. 1988; 42(1-2):157-161. DOI: https://doi.org/10.3109/00207458808985770
7. Farwell LA, Donchin E. The truth will out: interrogative polygraphy ("Lie detection") with event-related brain potentials. *Psychophysiology*. 1991; 28(5):531-547. DOI: https://doi.org/10.1111/j.1469-8986.1991.tb01990.x
8. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. *CRC Press.* 1994; Available from: https://www.crcpress.com/An-Introduction-to-the-Bootstrap/Efron-Tibshirani/p/book/9780412042317
9. Rosenfeld JP, Soskins M, Bosh G, Ryan A. Simple, effective countermeasures to P300-based tests of detection of concealed information. *Psychophysiology*. 2004; 41(2):205-219. DOI: https://doi.org/10.1111/j.1469-8986.2004.00158.x
10. Rosenfeld JP. P300 in detecting concealed information. Memory Detection: Theory and application of the concealed information test. 2011; 63-89.
11. Abootalebi V, Moradi MH, Khalilzadeh MA. A comparison of methods for ERP assessment in a P300-based GKT. *Int J Psychophysiol*. 2006;62(2):309-320. DOI: https://doi.org/10.1016/j.ijpsycho.2006.05.009
12. Mertens R, Allen JJB. The role of psychophysiology in forensic assessments: Deception detection, ERPs, and virtual reality mock crime scenarios. *Psychophysiology*. 2008; 45(2):286-298. DOI: https://doi.org/10.1111/j.1469-8986.2007.00615.x
13. Başar E, Başar-Eroglu C, Karakaş S, Schürmann M. Gamma, alpha, delta, and theta oscillations govern cognitive processes. *Int J Psychophysiol*. 2001; 39(2):241-248. DOI: https://doi.org/10.1016/s0167-8760(00)00145-8
14. Unser M, Aldroubi A. A review of wavelets in biomedical applications. *Proc IEEE*. 1996; 84(4):626-638. DOI: https://doi.org/10.1109/5.488704
15. Abootalebi V, Moradi MH, Khalilzadeh MA. A new approach for EEG feature extraction in P300-based lie detection. *Comput Methods Programs Biomed*. 2009; 94(1):48-57. DOI: https://doi.org/10.1016/j.cmpb.2008.10.001
16. Donchin E, Kramer A, Wickens C. Applications of event-related brain potentials to problems in engineering psychology. *Psychology Faculty Publications.* 1986; Availble from: https://scholarcommons.usf.edu/psy_facpub/280
17. Donchin E, Porges SW, Coles MGH. Psychophysiology: systems, processes, and applications. *Psychology Faculty Publications.* 1986; pp:702-710. Available from: https://scholarcommons.usf.edu/psy_facpub/167/
18. Rosenfeld JP, Labkovsky E, Winograd M, Lui MA, Vandenboom C, Chedid E. The Complex Trial Protocol (Ctp): a new, countermeasure-resistant, accurate, P300-based method for detection of concealed information. *Psychophysiology*. 2008; 45(6):906-919. DOI: https://doi.org/10.1111/j.1469-8986.2008.00708.x
19. Semlitsch HV, Anderer P, Schuster P, Presslich O. A solution for reliable and valid reduction of ocular artifacts, applied to the P300 ERP. *Psychophysiology*. 1986; 23(6):695-703. DOI: https://doi.org/10.1111/j.1469-8986.1986.tb00696.x
20. Johnson R. On the neural generators of the P300 component of the event-related potential. *Psychophysiology*. 1993; 30(1):90-97. DOI: https://doi.org/10.1111/j.1469-8986.1993.tb03208.x
21. Soskins M, Rosenfeld JP, Niendam T. Peak-to-peak measurement of P300 recorded at 0.3 Hz high pass filter settings in intraindividual diagnosis: complex vs. simple paradigms. *Int J Psychophysiol*. 2001; 40(2):173-180. DOI: https://doi.org/10.1016/s0167-8760(00)00154-9
22. Keil A, Debener S, Gratton G, et al. Committee report: publication guidelines and recommendations for studies using electroencephalography and magnetoencephalography. *Psychophysiology*. 2014; 51(1):1-21. DOI: https://doi.org/10.1111/psyp.12147
23. Cohen J. Statistical Power Analysis for the Behavioral

Sciences. New York: Academic Press; 1969.

24. Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon Bull Rev*. 2009; 16(2):225-237. DOI: https://doi.org/10.3758/PBR.16.2.225

25. Rosenfeld JP, Shue E, Singer E. Single versus multiple probe blocks of P300-based concealed information tests for self-referring versus incidentally obtained information. *Biol Psychol*. 2007; 74(3):396-404. DOI: https://doi.org/10.1016/j.biopsycho.2006.10.002

26. Green DM, Swets JA. Signal Detection Theory and Psychophysics. *New York : Wiley*. 1966; Available from: https://trove.nla.gov.au/version/12407339

27. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*. 1983; 148(3):839-843. DOI: https://doi.org/10.1148/radiology.148.3.6878708

28. Hu X, Rosenfeld JP. Combining the P300-complex trial-based concealed information test and the reaction time-based autobiographical implicit association test in concealed memory detection. *Psychophysiology*. 2012; 49(8):1090-1100. DOI: https://doi.org/10.1111/j.1469-8986.2012.01389.x

Appendices

Appendix A:

Pre-Task Instructions: "The project in which you are about to participate focuses on lie-detection tests. This is of major interest in the field of applied psychophysiology. Criminals who are questioned about a certain crime or crime scene show brain wave responses when they see test item stimuli which they recognize from the crime." TO SG GROUP ONLY: "In the next time period you will commit a laboratory crime, and afterwards, you will take our brainwave test."

Mock-Crime Instructions: "Go into room 203E, the last door on your left nearest the window as you enter the lab. In room 203E, as you enter, there will be a set of 8 drawers on your left. In the topmost drawer on the left, you will find a padded envelope with a valuable item inside. Open the envelope, take it out, hold it in your hand and rotate it as you examine it from all angles, then put it in your pocket and exit the room without letting the experimenter know which item it is. Return to the room where you signed the consent form. Any questions? Now, exit the room and perform the robbery."

Innocent Instructions: "Go into room 203E, the last door on your left nearest the window as you enter the lab. In room 203E, as you enter, there will be a set of 8 drawers on your left. In the topmost drawer on the left, you will find a padded envelope. Open the envelope, look inside, and then return to the room where you signed the consent form. Any questions? Now, exit the room and complete your task."

Appendix B
CTP Instructions

Pretest Instructions: In the present scenario, you are suspected of having committed a crime. In order to find out if you are guilty or innocent of the crime, you will take a brain wave lie-detection test during which we will measure your brain wave responses. For this reason I have attached the electrodes to your head. In the test, several items will be presented on this screen to you, including the image corresponding to the item you are suspected of taking. The test is based on the theory that your brain wave responses get bigger when you recognize an item related to the crime.

Test Instructions: These tests will be administered on the computer and I will monitor them myself. In this experiment, electrodes will be placed on your head and behind your ears. A harmless conductive paste is used to apply the electrodes; it can be easily cleaned off. A sink, water, and towels are available in the lab if you wish to clean yourself immediately afterwards. Do NOT wear contact lenses. If you need glasses, you MUST be sure to wear them, not contact lenses. You will be seated in a comfortable chair while a series of stimuli are presented on a screen in front of you. The stimuli will be presented in pairs. Each pair includes an item – that would be the first stimulus in a pair and it will be followed by a string of numbers 11111, or, 22222, or, 33333, or, 44444, or, 55555) – that would be the second stimulus in the same pair. There will be 210 pairs presented in a run.

THE FIRST STIMULUS: When you see the first stimulus in a pair, you have to immediately press the RIGHT button using your index finger on the LEFT-HAND mouse to indicate that you have seen the item. Some images might be familiar or relevant to you, but others will have no relevance to you. No matter if an item is relevant or irrelevant to you, you should respond to all of them exactly the same way (with a right button press on the left mouse). Remember, no matter which stimulus is presented, you always want to press the button as soon as possible.

THE SECOND STIMULUS: which you see a bit later in the trial, will be a string of numbers and it will appear soon after the first stimulus. The 11111 string is your target. When you see the target string 11111, respond immediately by pressing the RIGHT ("YES") button on the RIGHT-HAND mouse. If the second stimulus is 22222, or 33333, or 44444, or 55555, to press the LEFT ("NO") button on the RIGHT-HAND mouse. PAY ATTENTION TO ALL STIMULI! At a few random times during testing, the experimenter will pause the stimulus presentation and ask you to identify the last stimulus you saw on the screen. More than one incorrect identification will result in unusable data and early termination of your participation. It is crucial that you control your blinking, trying to reduce it as much as you can. If needed, you can blink when you see numbers (not when you see an item).

Remember: during the course of a trial, keep your eyes focused on the center of the screen where the stimuli appear. Please try to sit still, relax all your facial and body muscles and try to minimize blinking.

Thank you!