Brainprint: Assessing the uniqueness, collectability, and permanence of a novel method for ERP biometrics

Blair C. Armstrong (a), Maria Ruiz-Blondet (b), Negin Khalifian (b), Kenneth J. Kurtz (b), Zhanpeng Jin(c, d), and Sarah Laszlo(b,e)

(a) Basque Center on Brain, Cognition, and Language
(b) Department of Psychology, Binghamton University
(c) Department of Electrical and Computer Engineering, Binghamton University
(d) Department of Biomedical Engineering, Binghamton University
(e) Program in Linguistics, Binghamton University

*Maria Ruiz-Blondet: mruizbl1@binghamton.edu, 607.777.2326
Blair C. Armstrong: blair.c.armstrong@gmail.com
Negin Khalifian: nkhalif1@binghamton.edu
Kenneth J. Kurtz: kkurtz@binghamton.edu
Zhanpeng Jin: zjin@binghamton.edu
Sarah Laszlo: slaszlo@binghamton.edu

Binghamton University
4400 Vestal Parkway East
Binghamton, NY, 13902, USA

Basque Center on Brain, Cognition, and Language
Paseo Mikeletegi
20009 San Sebastian
Donostia, SPAIN

RUNNING HEAD: Brainprint

ABSTRACT

The human brain continually generates electrical potentials representing neural communication. These potentials can be measured at the scalp, and constitute the electroencephalogram (EEG). When the EEG is time-locked to stimulation-- such as the presentation of a word--, and averaged over many such presentations, the Event-Related Potential (ERP) is obtained. The functional characteristics of components of the ERP are well understood, and some components represent processing that may differ uniquely from individual to individual-- such as the N400 component, which represents access to the semantic network. We applied several pattern classifiers to ERPs representing the response of individuals to a stream of text designed to be idiosyncratically familiar to different individuals. Results indicate that there are robustly identifiable features of the ERP that enable labeling of ERPs as belonging to individuals with accuracy reliably above chance (in the range of 82-97%). Further, these features are stable over time, as indicated by continued accurate identification of individuals from ERPs after a lag of up to six months. Even better, the high degree of labeling accuracy achieved in all cases was achieved with the use of only 3 electrodes on the scalp-- the minimal possible number that can acquire clean data.

1. Introduction

The electroencephalogram (EEG) is a measure of post-synaptic brain activity that has recently received substantial attention as a potential biometric (e.g., Palaniappan & Mandic, 2005; 2007; review in Campisi & La Rocca, 2014). Here, we will extend this work by examining the feasibility of cognitive components of the Event-Related Potential (ERP) as a biometric measure. Event-Related Potentials are obtained when the EEG is averaged time-locked to some stimulation of interest (e.g., presentation of a word). Individual *components* of the ERP have well understood functional characteristics (see Luck & Kappenman, 2011), that correspond to specific cognitive events. When compared with the background EEG, this has the advantage that it is possible to make an analysis of the desired cognitive state of a user, and design a biometric challenge protocol that can tap that cognitive state. Here, in particular, we will investigate a protocol that taps access to *semantic memory*.

Semantic memory can be thought of as the network of concepts and connections between them that all individuals possess. We argue that semantic memory is a system that, although generally similar across individuals, is likely to be highly individualized when examined in detail, and therefore likely to be able to provide a distinctive biometric. To see why this is the case, consider the concepts [bee] and [anaphylaxis]. Even when considering only these concepts, it is easy to imagine a number of plausible semantic networks including them across individuals. For example, some individuals might be allergic to bees and therefore link these concepts strongly; some individuals might realize that bee allergies can cause anaphylaxis and therefore have a link between them-- but a weaker link than that possessed by a person with a bee allergy--, and some individuals might not know what anaphylaxis is and therefore not represent it in memory. Of course, there are many more concepts represented in semantic memory than just [bee] and [anaphylaxis], and the more concepts that are represented, the more opportunities that arise for there to be differences in how they are represented across people. For example, as the pool of concepts grows from even just [bees, anaphylaxis] to [bees, anaphylaxis, clowns, cilantro, sharks, prawns, spiders], many more plausible combinations of mappings between concepts become possible, and it effectively becomes impossible, from a statistical perspective, that any two individuals will have an identical network.

While there are likely many neuro-cognitive networks that might differ between individuals besides the semantic network, semantic memory is an especially viable target for ERP biometrics because access to the semantic network is known to produce a robust deflection in the ERP, known as the N400 component. The N400 is a negative wave that peaks around 400 ms post stimulus onset and has a centro-parietal maximum (see review in Kutas & Federmeier, 2011). The N400 is known to represent the language comprehension system's automatic attempt to access the semantic network (Laszlo & Federmeier, 2007; 2008; 2009; 2011; 2014). One characteristic of the N400 that is central here is that N400s elicited in response to items that individuals are unfamiliar with differ from N400s elicited in response to items that individuals are familiar with (Laszlo & Federmeier, 2007, 2011, 2014). This is a useful characteristic of a potential biometric, because it means that when individuals are familiar with different subsets of items, those individuals will elicit different profiles of N400s. Here, we will make use of this feature by presenting participants with a stream of text that includes a large number of acronyms. In previous work, we have demonstrated that it is extremely unlikely for any two individuals to be familiar with exactly the same subset of these items (Ruiz-Blondet et al., 2014). Consequently, we expect the profile of N400s elicited by individuals exposed to these items to be different.

In the preceding paragraphs, we introduced the concept of semantic memory, described a theoretical account under which semantic memory is likely to differ across individuals, and linked that account with a biomarker-- the N400 ERP component. This process highlights the potential advantage of the use of ERPs as biometrics over the background EEG. The background EEG is elicited regardless of what a participant is doing, meaning that there is reduced (or no) experimental control over the resultant data ERP biometrics, in contrast, have the potential to begin with principled theories about why a particular protocol should produce unique ERPs, and enable a focused analysis centered only on the component or components most likely to produce identifiable responses

To guide the present work, we consider a theoretical framework for biometrics that requires the demonstration of four characteristics: universality, collectability, uniqueness, and permanence (Jain, Ross, & Prabhakar 2004). Universality refers to the necessity that, if a characteristic is to be used as a biometric, every person must possess that characteristic. This is is already established for EEG, as the lack of EEG is a clinical indicator of brain death (Wijdicks, 1995). As ERPs represent the averaged EEG, universality is therefore established for ERPs as well. Collectability refers to the requirement that a biometric measure must be possible (and, ideally, easy and comfortable) to collect. One of the principal issues that decreases the collectability of EEG (and, by extension ERP) biometrics is that many studies of EEG biometrics have demonstrated a need for data acquisition from a large array of electrodes in order to acheive > 90% identification accuracy (see review in Palaniappan & Mandic, 2007, see also La Rocca 2012). To address this, here we will perform biometric identification on the basis of only 3 electrodes: a reference, a ground for common mode rejection, and a single active sensor. This is the minimum number of electrodes with which clean EEG/ERP data can be required, and thus maximizes the collectibility of this protocol on this metric.

Permanence refers to the requirement that a biometric must be stable over time (see Brigham & Kumar, 2010, and Campisi & La Rocca, 2014, for review of the issue of permanence in EEG biometrics). Here, we will explore this issue by asking participants to provide ERPs in three sessions with a gap of from between one week and six months between the first and final session (see also Ruiz-Blondet, Laszlo, & Jin, under review).

Distinctiveness refers to the requirement that a biometric be different in each individual, and is seemingly the most difficult of the four requirements to assess. Distinctiveness is an unexplored topic in terms of the ERPs associated with semantic memory access, although it is widely accepted that there are quantifiable *individual differences* in brain organization or activity in response to tasks of this sort in general (e.g., Raz, Lindenberger, Rodrigue, Kennedy, & Head, et al., 2005; La Rocca et al., 2014; see also Khalifian & Laszlo, In Revision). Here, we will assess distinctiveness by applying several pattern classifiers (more details below) to individuals' ERP data, to determine whether ERPs are robustly identifiable via machine learning.

1.1 Difference from Past Approaches

In utilizing the averaged ERP and the N400 component more specifically as the biometric measure, our approach to electrophysiological biometrics differs from prominent and successful biometric protocols that utilize the EEG (e.g., Palanappian & Mandic, 2005; 2007; La Rocca et al, 2014; Abdullah et al, 2010; La Rocca et al., 2012). As already discussed, we have taken this approach primarily because of what is known about the N400 component, but it is an approach that possibly has other benefits as well. First, the ERP is less sensitive to background noise than is the ongoing EEG, which is a critical characteristic for applied use. This is because any applied use of EEG/ERP biometrics is likely to occur in environments where there are many likely sources of electrical noise, for instance, from other electronic devices or lights. In the ERP, electrical noise that is not time-locked to the signal of interest are likely to be at least partially reduced during averaging---a process that does not occur for analysis of the EEG. While EEG can be digitally filtered to remove background noise in some particular frequency band, this is a process that uses the information of neighbor data points instead of adding in new data. In contrast, the averaging performed in ERP analysis removes the noise from all frequency bands by adding new data where the meaningful information will be enhanced while random noise will tend to zero. This means that the ERP may be more robust to noise, relative to EEG.

Another issue with EEG is that to be stable enough for biometric analysis, it is necessary to record it for a relatively long duration to obtain sufficient data to first train and then test a classifier. For example, Marcel and Millan (2007) used data collected from 12 4-minute sessions, during which the task was altered every 15 seconds (for a review of other related studies, see Campisi & La Rocca, 2014). Even in the best work, with respect to overall recording duration, La Rocca, et al., (2014) report one of the shortest duration recordings for EEG data: a 10-second recording as the critical test after 50 seconds of prior recording used for training. Individual ERPs, in contrast, typically are only 1 second long, and semantic memory effects in particular are detectable from as early as 250 milliseconds (e.g., Laszlo & Federmeier 2014). Even when 50-100 trials are averaged, this still means no more than approximately 1.5 minutes of data recording. Here, for example, we will analyze averages of 50 trials during training and testing, which corresponds to 50 seconds of data in each case (i.e., the training and testing classification is based on ERPs aggregated over several trials --- we do not attempt single-trial recognition, although future work will be needed to determine how small a number of trials can still yield accurate recognition). The data included for analysis, then, is therefore similar in quantity to that collected by the field-leading work of La Rocca et al. (2014) and vastly less than that reported in Marcel and Millan (2014). This difference in collectability and data set size is therefore a potential advantage of ERPs over EEG for biometric use.[1]

[1] For transparency, it should be noted that this description bears on the total duration of the recordings only (i.e., information content as a function of recording time). Of course, such vastly different approaches differ on other characteristics as well, some of which may be for principled reasons and others not. For example, the inter-trial interval in the present study, which was included for similarity to past ERP work on ERP components related to isolated word reading --- adds a small amount of total time to the experiment itself, even if no data from this “filler” period is included in the recording. The duration of this “filler” period in obtaining clean classification results, however, has yet to be investigated to determine if it is necessary or not for the present aims. Other methods, such as those employed by La Rocca et al. (2014) can record continuously, which offers efficiency over the present method in that respect. However, those recordings also employed a 64 channel montage, which, if factoring in setup time over our approach, which requires only a single active electrode, would substantially increase the total duration of the experiment. Additional work is clearly needed to equate “clock time” across these and other approaches.

1.2 Pattern Classification

As a benchmark classifier, we employed support vector machines (SVM; Schölkopf & Smola, 2002; Milenova et al., 2005). SVMs are known to be excellent pattern classifiers in many respects; however, SVMs were originally developed for the purpose of binary classification and suffer from practical challenges when extended to multi-class problems, as in the present case (for discussion, see Hsu & Lin, 2002). Here, in order to extract classifier performance from SVM, we transform the more difficult biometric *identification* problem (of labeling a token as belonging to one of a large number of users) to easier *verification* problem (of deciding whether a token belongs to one particular user or not).

As a second benchmark, we will use a simple linear discriminant based on cross-correlation. Cross-correlation is a known method for quantifying the similarity between pairs of electrophysiological waveforms when classification is required (e.g., Chandaka, Chaterjee, & Munshee, 2009). When used as the basis function for a linear discriminant, it is also a highly efficient algorithm when compared to either SVM or neural network classifiers (such as those we will describe next), because it does not require training in order to perform classifications and consequently requires little computational overhead.

Finally, we will use two neural network classifiers. The first is the Divergent Autoencoder (DIVA; Kurtz, 2007). DIVA was originally developed as a cognitive model of human category learning, but has additional potential as a more general-purpose classifier for machine learning. A Divergent Autoencoder is similar to a standard multi-layer autoencoder, except that there is a separate output layer ("channel") for each category in the classification problem. The key design principle is training the autoencoder to reconstruct the members of each category with the constraint that each output channel shares a common hidden layer. Classification outcomes are a function of reconstruction error on each output channel: an item is a good member of a class to the extent it can be recoded and decoded along the appropriate channel with minimal distortion. The output channel that reconstructs an input with the least error provides its label to that input.

The final method for classification we will implement is Naive Discriminant Learning. NDL is also a learning neural network approach, but for present purposes its primary advantage over DIVA (or other neural network approaches) is that it does not use back-propagation to learn its weights. Instead, it relies on the Danks equilibrium equations (Danks, 2003) to discover optimal weights for input classification in one step. This characteristic of NDL allows it to retain the advantages of a learning classifier (e.g., the ability to emphasize some portions of the data over others) without one of the major pitfalls of large learning classifiers for applied use-- namely, lengthy training time due to iterative learning.

The prior literature would seem to favor SVM and cross-correlation as the methods most likely to produce high accuracy results, as these two are gold-standard methods for classification (in the case of SVM) and comparison of electrical waveforms (in the case of cross-correlation). However, the ability of the neural networks to learn may provide some advantages over cross-correlation, particularly if the relevant input-output structure is nonlinear, and the ability of DIVA and NDL to natively handle multi-way classification problems may provide them with an advantage over SVM.

2. Methods

The general schematic for data collection was as follows. Participants came to the lab and read text silently to themselves while ERPs were collected. A subset of these returned to the lab multiple times with intervening lags of one week to six months and performed the same task, with the same stimuli, in order to address permanence. After all data were collected, classifiers were applied to quantify the distinctiveness of the resultant ERPs. For the three non-SVM classifiers, the outputs of each classifier were transformed to rankings where possible labels for an ERP token was ranked from 0 to N-1 (where N = the number of participants). The label ranked 0 was each classifier's best guess as to which participant each token belonged to. This method was applied identically to NDL, DIVA, and cross-correlation. For each token presented to each classifier, then, we computed a *identification accuracy*, defined as (1 - [Rank of correct label / Number of participants]). This method of quantifying accuracy reflects the idea that classifiers should be given more credit for ranking the correct label highly, even if the correct label is not given the top rank (e.g., Mitchell, Shinkareva, Carlson, Chang, Malve, et al., 2008). For SVM we instead used a two-class, verification scenario. Here, each ERP could be classified either as authorized user or not. SVM's output was considered "correct" when it either 1) verified an authorized user's token as a match or 2) rejected an un-authorized user's token as unauthorized.

2.1 Data Acquisition (Event-Related Potentials)

Data were acquired following the methods of past studies that demonstrate differences on the N400 on the basis of individual acronym knowledge (Laszlo & Federmeier, 2007; 2011; 2014; Laszlo, Stites, & Federmeier, 2012). ERPs were recorded from 45 adult participants (11 female, age range 18-25, mean age 19.12). Of these 45, 30 participated only once, 15 (10 female, age range 18-23, mean age 20.71) returned to the lab for a second session between 5 days and 40 days after the first session (mean 12.73 days), and 9 (5 females, age range 18-23, mean age 20.22) returned for a third session, between 134 and 188 days after the first session (mean 156 days). The EEG was digitized at 6 midline electrode sites in addition to the reference, ground, and EOG channels; these corresponded roughly to fPz, Cz, Pz, Oz, O1 and O2 in the international 10-20 system. Only data from O2 was analyzed, as pilot work indicated this was the most robust channel (Ruiz-Blondet et al., 2013). Ag/AgCl active amplification electrodes were used; interelectrode impedance was maintained at < 50 Kohm (Laszlo, Ruiz-Blondet, Khalifian, Chu, & Jin, 2014). Data were acquired with a hardware high pass filter (.016 Hz) to reduce the influence of DC shifts. Participants viewed 75 acronyms intermixed with fillers from other lexical types. For more details about the items and the task, see Laszlo & Federmeier (2007; 2011; 2014). Acronyms were repeated once at a lag of 0, 2 or 3 intervening items. This repetition allows for relatively homogenous (though not identical, due to repetition effects; Laszlo & Federmeier, 2007; 2011; Laszlo & Armstrong, 2014) but non-overlapping segmentation of the data into train and test corpora for machine learning: first responses to acronyms were used for training, and second responses were used for testing. ERPs were computed by averaging the data at each electrode, time-locked to the onset of each acronym, on each of the two presentations. Data were digitized at 500 Hz and contain a 100 ms pre-stimulus baseline; thus each 1.1 second long ERP includes 550 samples. We did not apply software filters or artifact rejection prior to pattern classification because pilot work demonstrated that neither of these measures positively impacted classifier accuracy and added time to the overall procedure.

2.2 Pattern Classifiers

During recording, a small number of trials were lost from some participants (e.g., due to movement artifact), but all participants retained at least 70 trials, so 70 random trials were selected from all participants to keep the size of each participant's data set uniform. The neural networks require multiple examples from each participant in order to learn input-output mappings robustly, so it was not sufficient to simply create 1 ERP from each participant for network training. Instead, a bootstrapping procedure was used, where 100 ERPs were generated for each participant with a random 50 of that participant's 70 trials selected each time. After bootstrapping, 100 ERPs were available from each participant, for a total of 3000 (30 participants x 100 random averages). Bootstrapping was applied to both the train and test data, meaning that 3000 averages were available for training, and a completely non-overlapping 3000 averages were available for testing.

2.2.1 Cross-Correlation

To classify by cross-correlation, we first computed the maximum absolute value of the cross-correlation between pairs of waveforms. Each of the 100 random averages for each participant was cross-correlated with both 1) another average from that same participant (a self-self pair) and a random average from every other participant (a self-other pair), for a total of 30 pairs per average. The cross-correlations between pairs were then normalized to reduce variability caused by scalp thickness and other cognitive-unrelated events. This operation was performed for each of the 100 averages of each of the 30 participants (i.e. 3000 times). Each time, the highest cross-correlation value received a rank of zero and the lowest value received a rank of 29. Then, identification accuracy for each of the 3000 test cases was 1 - [rank of correct pair / number of pairs (zero indexed)]. Thus, a “correct” response on each case would be for the self-self pair to be given a rank of 0, and all the self-other pairs be given a higher rank. The mean identification accuracy was the mean accuracy across the 3000 trials. 95% confidence intervals on this mean were computed on the basis of the t-distribution.

2.2.2 Divergent Autoencoder (DIVA)

The DIVA network was a 550:200:550[30] feedforward autoencoder. The 550 input units correspond to the 550 samples in each waveform. A 200 unit hidden layer was used based on pilot simulations that determined that this was the smallest sized hidden layer that enabled near perfect (99% accuracy) learning of the training set. The [30] signifies that, instead of having only one output layer, as in a standard autoencoder, there were 30 output layers, one for each participant.

During learning, hidden to output weights were adjusted only along the correct category channel as a function of the mean squared error across that channel’s output. The network was trained through 1000 iterations of the 3000 training examples; this was determined to be a level that allowed excellent performance (>99% classification accuracy) on the training data without overfitting. After these 1000 iterations, weights in the model were fixed.

At test, the model was presented with each of the 3000 test examples, and activation was allowed to propagate forward through the network. Reconstruction error was measured on each channel. The channels were then assigned ranks based on their output error. Identification accuracy was computed as described above, with the accuracy for each of the 3000 test cases being 1 - [rank of correct channel] / [number of channels] (ranks and number of channels zero indexed); mean identification accuracy was given as the mean accuracy across the 3000 trials. 95% confidence intervals on this mean were computed on the basis of the t-distribution. Figure 1 displays an example of an empirically derived ERP along with its best and worst DIVA reconstructions.

2.2.3 Naive Discriminant Learning (NDL)

The NDL classifier was trained by providing the 3000 training patterns as input across a 550 unit input layer, and requiring the network to indicate its classification by activating one of 30 units in an output layer. To speed classification, rather than rely on online (trial-by-trial) or batch (groups of trials) iterative learning methods, the Danks (2003) equations were employed to estimate the end-state of iterative learning using the classic Rescorla-Wagner (1972) discriminative learning algorithm, but in a single iteration. Estimated weights at equilibrium were obtained using the implementation of NDL provided by Shaoul, Arppe, Hendrix, Milin, and Baayen (2014).   After training, NDL correctly classified all of the participants on the basis of the input ERPs using a winner-take-all evaluation method.

To test the classification performance of NDL, the trained NDL network's parameters were fixed and the 3000 patterns of testing data were presented. Network output was then normalized and transformed into a ranked classification. Identification accuracy and confidence intervals were computed identically as for DIVA.

2.2.4 Support Vector Machines (SVM)

Separate training and testing data sets were specifically designed for SVM. Here, a verification scenario was used (instead of the identification scenario above). For this, 87 tokens were selected from a participant (the authorized user), and compared with 3 tokens from each of the other 29 participants (87 total, the intruders). The SVM was implemented with an RBF kernel and $\sigma=1000$ as indicated by pilot work. Across the 87 tokens, the SVM's output was considered correct if it verified an authorized user as being authorized (placed it in the authorized class), or indicated an intruder was unauthorized (placed it in the unauthorized class). Mean accuracy was given as the number of correct classifications divided by the total number of classifications. 95% confidence intervals on the mean were computed on the t-distribution.

3. Results

3.1 First session

Figure 2 presents mean identification accuracies for cross-correlation, DIVA, and NDL, along with verification accuracy for SVM. As visualized in the figure, mean identification accuracy for cross-correlation was .92 (95% confidence interval: 0.92-0.93). Mean identification accuracy for DIVA was .89 (95% confidence interval: 0.89-0.90). Mean identification accuracy for NDL was .82 (95% confidence interval: 0.81-0.83). Finally, SVM's mean verification accuracy was .83 (95% confidence interval: 0.77-0.88). The confidence intervals indicate that, with 95% confidence, cross-correlation was reliably the most accurate classifier, followed by DIVA, followed by NDL, with SVM's accuracy overlapping that of NDL.

The null hypothesis for identification accuracy for the identification scenario is that the classifiers are assigning ranks to the correct class by chance, consequently the expected random accuracy is 50%. In the case of SVM, the null hypothesis for verification accuracy is also 50%, since each trial can be randomly classified as either authorized user or impostor. Clearly, all classifiers performed substantially better than chance. To quantify this statistically in the 30 class classifiers, we computed the distribution of accuracies across 50 000 random permutations of the ranking matrix. We then assigned p-values to each observed accuracy by determining the proportion of random accuracies that were higher than the observed accuracy for each classifier (an approximate randomization test). The null hypothesis was rejected for all classifiers with $p < .0001$. To quantify SVM's verification accuracy, we compared the observed mean accuracy (.83) against a binomial distribution with a success probability of .5, which indicated that the observed mean would be observed by chance with $p < .0001$.

3.2 Improving performance with a combination of models.

The separate examination of each of the models indicates that accurate classifications can be obtained by each one. Perfect classification is, however, the ultimate aim for applied use of an ERP biometric. To this end, we examined whether a combination of models could be used to improve overall classification performance. Assuming that each model accounts for a random subset of the total variability in the signal, and knowing that all of the models are at approximately 90% accuracy, the likelihood of two models failing on a given trial is only 1%. Combining three models could further reduce the failure rate to 0.1%. However, it is also possible that all of the models make errors on some same subset of difficult to classify data (i.e., that errors are distributed non-randomly). If this is the case, examination of those difficult to classify data may reveal characteristics in the signal, and/or common aspects of the individual models, that can guide future work.

To gain initial insight into these issues, we examined the mean identification accuracy for DIVA, NDL, and cross-correlation on the testing data at the trial level (SVM was excluded for the reasons outlined above). In the first analysis, we calculated the maximum possible identification accuracy across all three models for each individual trial---the accuracy that is achieved if each trial is identified only by the model that provides the best classification of any of the models for that trial. This establishes the upper bound for classification accuracy that could be established, in principle, by combining the different models (although this does not mean that a classifier could actually learn this optimal classification, an issue we visit next). The results showed that mean maximum accuracy increased to 97.6% --- a substantial increase over the performance of the best single model (cross correlation, at 92%), but still below the

99.9% expected if the three models were each tapping a random portion of the overall variability.

To provide a more realistic assessment of how the results from the different models could be combined, we developed a simple meta-classifier that integrated the results of the NDL, DIVA, and cross-correlation classifiers. Given that cross-correlation performed the best of the three overall during testing, the meta-classifier's default response when the first-ranked classification was not the same across the three algorithms was that of the cross-correlation classifier. However, the cross-correlation classifier could be overridden if the two other classifiers both agreed on a different response. This voting scheme aimed to capitalize on the fact that the errors committed by each algorithm were partially independent from one another. Per this meta-classifier, identification accuracy increased to 93.7% (95% confidence interval: 93.6-93.7).

3.3 Participant Level Classification-- Insights for Future Work

One possibility as to why the meta-classifier did not perform to the theoretical ceiling limit is that there is a subset of trials or a subset of participants that are particularly difficult to classify, across all models. A plot of maximum accuracy across the models for each participant clearly shows that this is the case (Figure 3). Whereas for most participants, maximum accuracy was at ceiling, for a small number of participants---and participant 21, in particular---classification accuracy was, noticeably lower, meaning that no model could correctly classify this participant with 100% accuracy. Inspection of the data for that participant indicates a possible explanation for this failure (see Figure 4). It appears that participant 21 produced an ocular artifact in response to first presentations of items that was not re-produced in response to second presentations, which would make the second presentation waveforms difficult to match to this participant's first presentation waveforms for all classifiers. Fortunately, such ocular artifacts also have a clear neural signature that is distinct from N400 activity, and which can be automatically detected and removed with a variety of algorithms (e.g., Gratton, Coles, and Donchin, 1983; Jung, Makeig, Humphries, Lee, Mckeown, et al., 2000). Consequently, one avenue of future work would be to include a provision in the classification algorithm to attempt ocular artifact correction for individuals classified poorly.

3.4 Initial assessment of permanence

Cross-correlation was shown to be the most accurate method for identification in the analysis above and is the least computationally expensive of the methods. For these reasons, cross-correlation was the only method used to provide some initial insights into the permanence of our biometric. Note that because of the smaller sample size of these assessments of repeatability, these data serve primarily to provide some basic validity that permanence may be achievable with this approach; additional more extensive work targeting this issue is clearly needed. Nevertheless, these initial insights do provide reason to be optimistic in this respect: The mean identification accuracy when test tokens from a participant's second ERP session were compared to training tokens from a participants' first ERP session was 89% (95% confidence interval: 0.88 to 0.90). These results are better than chance at $p < .0001$ (the null hypothesis here being the same as for the analogous analyses in the single-session data). The mean identification accuracy when test tokens from a participant's 3rd ERP session were compared to training tokens from the first ERP session was 93% without one outlier classified with 7% accuracy. (95% confidence interval: 0.87 to 0.99). Again, these results are better than chance at p < .0001. Further, the mean accuracy for the third session is within the confidence interval of accuracy for the first session, indicating that classification accuracy did not decline significantly over time. Figure 5 displays cross correlation identification accuracy over time; Figure 2 places these results side-by-side with the results of the single-session analysis. Inspection of the data from the individual participants who participated in a third session also showed that several (4 out of 9) are still being classified with perfect accuracy after this extended period of time.

4. Discussion

We set out to address the collectibility, permanence, and uniqueness of a novel method for biometrics that makes use of the averaged Event-Related Potential. We were motivated in this exploration by an interest in making use of a vast literature from cognitive neuroscience that provides understanding of the cognitive events that elicit the ERP-- in particular, access to semantic memory-- to inform biometric design. We reasoned that semantic memory was a cognitive system likely to vary uniquely across individuals, and that designing a challenge protocol to tap semantic memory directly could result in a highly accurate biometric.

To address uniqueness, we applied several classifiers to the ERP biometrics. All were able to classify the data with accuracy far above chance. The cross-correlation classifier was numerically the strongest, with an identification accuracy of 92%, and the meta-classifier was able to reach an accuracy of 97%. Worthy of particular emphasis, this high accuracy was achieved in a relatively collectable protocol, that is, one where data from only 50 seconds of recording from only one active sensor was used for training and testing of the classifiers. This is at least comparable and in a number of cases substantially higher accuracy than has been achieved with a single sensor in many past EEG biometric applications. For instance, Palanippan & Mandic (2007) report only 13% accuracy when using only one electrode. More recently, Abdullah et al. (2010) reported classification accuracies in the 70% - 87% range with only one channel and needed data from four channels to obtain comparable accuracy to that reported here. Similarly, Riera, Soria-Frisch, Caparrini, Grau, & Ruffini (2007; see also Su et al., 2014) achieved comparable accuracy to ours (error rates of 3.4%) with EEG recordings lasting 2-4 minutes (see Campisi & La Rocca, 2014, for a review of 16 other recent articles, of which only those with more electrodes show considerably better performance, even ignoring the brevity of our recording time). Thus, the ERP biometric explored here seems to be at least on par with field-leading work in EEG biometrics in terms of both uniqueness and collectability.

To address permanence, we asked a subset of participants to return to the lab between a week and six months after their first session. Figure 5 displays classification accuracy over time; in fact, classification accuracy for some participants remained as high as 100% even after 178 days. These results are consistent with predictions from the semantic memory literature, which suggest that this particular type of memory should be relatively stable over time, not sensitive to strong interference from new knowledge on old knowledge, and degrade gracefully when memories are lost (McClelland et al., 1995; O'Reilly, 2011; Rogers & McClelland, 2003).

4.1 Future Work

Of course, even the 97% accuracy attained by the meta-classifier is below the optimally 100% accuracy that would be desirable for applied biometric use. Additionally, the one active sensor we analyzed data from, at O2, is located over the back of the head, and therefore sits atop hair on most people; this requires the application of electrolytic gel for adequate signal quality. Both the collectability and the uniqueness of the ERP biometric could therefore still bear improvement.

Regarding collectability, it would be beneficial to be able to record the ERP biometric from a site that does not typically have hair on it (e.g., above the eyes, see Riera et al., 2007; Su et al., 2014), so that there is no need for electrolytic gel. To this end, we are currently conducting an investigation of ERP biometrics using a higher density electrode montage than that used here, in order to see how ERP biometric accuracy varies across different sites on the scalp.

Regarding uniqueness, the use of acronyms as challenge stimuli here was largely motivated by our own prior work, and, as is necessarily the case for a first step in any investigation, may not constitute the optimal set of stimuli for eliciting individuating brain responses. Consequently, our ongoing work examines whether other categories of stimuli may be able to elicit more unique responses than those elicited by acronyms. Relatedly: here items were presented multiple times to each participant. This was done partly to ensure compatibility with past work, and partly to create similar, but non-overlapping, train and test data sets. However, the repetition of stimuli raises two important questions related to long-term uniqueness of challenge protocols of this type: will participants' neural responses be experience dependent, and actually be changed over time by repeated exposure in this protocol and others like it? And, how robust will the response to any particular item be over a time frame longer than the 6 months tested here, given that each participant's experience outside the lab is unpredictable?

As pertains to the first question, the neural mechanism of N400 repetition effects is well enough understood to be instantiated in an explicit computational model (Laszlo & Armstrong, 2014), and seems to be the result of short-term resource depletion, rather than substantive changes to long term memory. This suggests that N400 responses to repetitions of items over multiple sessions should not change dramatically (as is also suggested by the permanence data reported here). In agreement with this analysis, N400 repetition effects for 2 repetitions are not typically shown to be different than N400 repetition effects for 3 or more repetitions (e.g., Young and Rugg, 1992) even within a session, let alone across recording sessions when cellular resources are able to replenish. Thus, there are good theoretical and some empirical reasons to believe that multiple presentations of challenge items will not be deleterious to ERP biometric accuracy, but this is ultimately still an empirical question which we plan to address in future work through examination of ERP biometric identification accuracy at even more remote time points and with even more repetitions of items.

The question of whether an individual's experience outside the lab might deleteriously affect N400 biometric accuracy is also an important one. As already discussed, numerous investigations in the field of formal semantics suggest that the semantic network is: relatively stable over time, not sensitive to strong interference from new knowledge on old knowledge, and degrades gracefully when memories are lost (McClelland et al., 1995; O'Reilly, 2011; Rogers & McClelland, 2003). However, we nonetheless aim in future work to investigate the use of ERP biometrics based on components of the ERP that are likely to be even less sensitive to experience, such as early visual components representing the encoding of line segments (which are known to be very stable over time, Hubel & Weisel, 1968). These components are known to be sourced in the early visual cortices, and it is established that there are substantial individual differences in the cortical folding of these areas (e.g., Dougherty, Koch, Brewer, Fischer, Modersitzki, & Wandell, 2003; Shwarzkopf, Song, & Rees, 2010). This anatomical variability should be expected to increase the uniqueness of ERP biometrics based on early visual components, while being entirely unrelated to experience. Once again, the process of selection of early visual components as possible ERP biometrics highlights an advantage of ERPs over EEG biometrics: where it was desirable to identify a measure that would be relatively stable over time regardless of experience, visual processing was a strong candidate and, knowing the correlates of early visual processing in the ERP, it is possible to select both stimuli and spatial and temporal regions of interest for analysis in a principled manner.

5. Conclusions

Here, we investigated, for the first time in the literature, the use of an ERP biometric based on the uniqueness of individual's semantic networks and resultant N400 effects. We demonstrated identification accuracy that was robustly above chance even when only 1 active sensor and 50 seconds of data were used for classification, and we further demonstrated that some individuals could still be identified with perfect accuracy even after as long as six months. This work thus constitutes an encouraging proof-of-concept for the use of ERP biometrics and has yielded a number of targeted directions for further refinement.

Acknowledgements

The authors wish to acknowledge M. Monk for insightful discussion throughout the course of this project. BCA acknowledges support from Marie Curie IIF PIIF-GA-2013-627784. SL acknowledges support from NSF CAREER-1252975. ZJ and SL received supports from NSF TWC SBE-1422417, the Binghamton University Interdisciplinary Collaborative Grants program, and the Binghamton University Health Sciences Transdisciplinary Area of Excellence.

References

Abdullah, M. K., Subari, K. S., Loong, J. L. C., & Ahmad, N. N. (2010, November). Analysis of effective channel placement for an EEG-based biometric system. In Biomedical Engineering and Sciences (IECBES), 2010 IEEE EMBS Conference on (pp. 303-306). IEEE.

Almehmadi, Abdulaziz, and Khalil El-Khatib. "The state of the art in electroencephalogram and access control." In Communications and Information Technology (ICCIT), 2013 Third International Conference on, pp. 49-54. IEEE, 2013.

“BBC News: Malaysia car thieves steal finger”, last modified March 31, 2005, http://news.bbc.co.uk/2/hi/asia-pacific/4396831.stm

Brammer, M. J., Bullmore, E. T., Simmons, A., Williams, S. C. R., Grasby, P. M., Howard, R. J., ... & Rabe-Hesketh, S. (1997). Generic brain activation mapping in functional magnetic resonance imaging: a nonparametric approach.*Magnetic resonance imaging*, *15*(7), 763-770.

Brigham, K. & Kumar, B. V. (2010). Subject identification from electroencephalogram (EEG) signals during imagined speech. *Proceedings of the IEEE 4th International Conference on Biometrics: Theory Applications and Systems*, 1-8.

Cavanagh, J. F., & Allen, J. J. (2008). Multiple aspects of the stress response under social evaluative threat: An electrophysiological investigation. *Psychoneuroendocrinology*, *33*(1), 41-53.

Campisi, P., & La Rocca, D., (2014). Brain waves for automatic biometric-based user recognition. *Information Forensics and Security, IEEE Transactions , 9*(5), 782-800.

Cree, G. S., & Armstrong, B. (2012). Computational models of semantic memory. In Spivey, M., McRae, K. & Joanisse, M. (Eds.), *Cambridge Handbook of Psycholinguistics* (pp. 259-282). New York: Cambridge University Press.

Chandaka, S., Chatterjee, A., & Munshi, S. (2009). Cross-correlation aided support vector machine classifier for classification of EEG signals. *Expert Systems with Applications*, *36*(2), 1329-1336.

Danks, D. (2003). Equilibria of the Rescorla-Wagner model. Journal of Mathematical Psychology, 47, 109-121.

Dougherty, R. F., Koch, V. M., Brewer, A. A., Fischer, B., Modersitzki, J., & Wandell, B. A. (2003). Visual field representations and locations of visual areas V1/2/3 in human visual cortex. *Journal of Vision*, *3*(10), 1.

Gratton, G., Coles, M. G., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and clinical neurophysiology*,*55*(4), 468-484.

Hsu, C-W., & Lin, C-J. (2002). A comparison of methods for support vector machines. *Neural Networks*, 13(2), 415-425.

Jain, A. K., Ross, A., & Prabhakar, S. (2004). “An introduction to biometric recognition”. Circuits and Systems for Video Technology, IEEE Transactions on, 14(1), 4-20.

Jung, T. P., Makeig, S., Humphries, C., Lee, T. W., Mckeown, M. J., Iragui, V., & Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source separation. Psychophysiology, 37(02), 163-178.

Khalifian, N. (2013) Life of ERPLE: Developing a silent reading task for use with children and adults. SUNY Binghamton, Binghamton, New York.

Khalifian, N., & Laszlo, S. (Minor Revision). Don’t Judge a Reader by Their Scores: Event-Related Potentials Demonstrate a Stronger Relationship with Reading-Related Report Card Achievement than Behavior Alone. *Developmental Science*.

Khalifian, N., & Laszlo, S. (2014). Predicting Individual Scholastic Reading Performance with Event-Related Potentials: Results of Year Two of the Binghamton

Reading Brain Project. Presented at the 54th Annual Meeting of the Society for Psychophysiological Research, Atlanta, Georgia.

Kirschbaum, C., Pirke, K. M., & Hellhammer, D. H. (1993). The ‘Trier Social Stress Test’–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, *28*(1-2), 76-81.

Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin & Review*, *14*(4), 560-576.

Kutas, M., & Federmeier, K.D. (2011).Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). Annual Review of Psychology.

La Rocca, D., Campisi, P. , & Scarano, G. (2012). EEG biometrics for individual recognition in resting state with closed eyes. *Proceedings of the International Conference of the Biometrics Special Interest Group*, 1-12.

La Rocca, D., Campisi, P., Vegso, B., Cserti, P., Kozmann, G.; Babiloni, F., & De Vico Fallani, F. (2014). Human Brain Distinctiveness Based on EEG Spectral Coherence Connectivity. *IEEE Transactions on Biomedical Engineering, 61*(9), 2406-2412.

Laszlo, S., & Federmeier, K. D. (2007). Better the DVL You Know Acronyms Reveal the Contribution of Familiarity to Single-Word Reading. *Psychological Science*, *18*(2), 122-126.

Laszlo, S., & Federmeier, K.D. (2008). Minding the PS, queues, and PXQs: Uniformity of semantic processing across multiple stimulus types. *Psychophysiology,* 45, 458-466.

Laszlo, S., & Federmeier, K.D. (2009). A beautiful day in the neighborhood: An event-related potential study of lexical relationships and prediction in context. *Journal of Memory and Language*, 61, 326-338.

Laszlo, S., & Federmeier, K.D. (2011). The N400 as a snapshot of interactive processing: evidence from regression analyses of orthographic neighbor and lexical associate effects. *Psychophysiology,* 48, 176-186.

Laszlo, S., Stites, M., & Federmeier, K.D. (2012). Won't Get Fooled Again: An Event-Related Potential Study of Task and Repetition Effects on the Semantic Processing of Items without Semantics. Language and Cognitive Processes, 27, 257-274.

Laszlo, S., Ruiz-Blondet, M., Khalifian, N., Chu, F., & Jin, Z. (2014). A Direct Comparison of Active and Passive Amplification Electrodes in the Same Amplifier System. *Journal of neuroscience methods,* 235, 298-307.

Laszlo, S., & Federmeier, K. D. (2014). Never seem to find the time: evaluating the physiological time course of visual word recognition with regression analysis of single-item event-related potentials. Language, Cognition and Neuroscience, 29, 642-661.

Laszlo, S., & Armstrong, B.C. (2014). PSPs and ERPs: Applying the dynamics of post-synaptic potentials to individual units in simulation of ERP reading data. Brain and Language, 132, 22-27.

Marcel, S. & Millan, J. D. R. (2006). Person authentication using brainwaves (EEG) maximum a posteriori model adaptation. *IEEE Transactions in Pattern Analysis and Machine Intelligence, 29*(4), 743-748.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102*(3), 419-457.

McGonagle, K. A., & Kessler, R. C. (1990). Chronic stress, acute stress, and depressive symptoms. *American Journal of Community Psychology*, *18*(5), 681-706.

Milenova, B.L., Yarmus, J.S., Campos, M.M., "SVM in Oracle Database 10g: Removing the Barriers to Widespread Adoption of Support Vector Machines", Proceedings of the 31st VLDB Conference, Trondheim, Norway, 2005.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K. M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *science*, *320*(5880), 1191-1195.

O'Reilly, R.C., Bhattacharyya, R., Howard, M.D. & Ketz, N. (2011). Complementary Learning Systems. *Cognitive Science, 38*(6), 1229-1248.

Palaniappan, R., & Mandic, D. P. (2005). Energy of brain potentials evoked during visual stimulus: A new biometric?. In *Artificial Neural Networks: Formal Models and Their Applications–ICANN 2005* (pp. 735-740). Springer Berlin Heidelberg.

Palaniappan, R., & Mandic, D. P. (2007). Biometrics from brain electrical activity: A machine learning approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(4), 738-742.

Raz, N., Lindenberger, U., Rodrigue, K. M., Kennedy, K. M., Head, D., et al. (2005). Regional brain changes in aging healthy adults: general trends, individual differences and modifiers. *Cerebral Cortex,* 15(11), 1676-1689.

Riera, A., Soria-Frisch, A., Caparrini, M., Grau, C., & Ruffini, G. (2008). Unobtrusive biometric system based on electroencephalogram analysis. *EURASIP Journal on Advances in Signal Processing, 2008.*

Ruiz-Blondet, M., Laszlo, S., & Jin, Z. (Under Review). Assessment of Permanence of Non-Volietional EEG Brainwaves as a biometric. *IEEE International Conference on Identity, Security, and Behavior Analysis (ISBA, 2015).*

Luck, S. J., & Kappenman, E. S. (Eds.). (2011). *The Oxford handbook of event-related potential components*. Oxford university press.

McClelland, J. L., and Rogers, T. T. (2003). The Parallel Distributed Processing approach to semantic cognition. *Nature Reviews Neuroscience, 4*(4), 310-322.

Ruiz-Blondet, M., Khalifian, N., Armstrong, B.C., Jin, Z., Kurtz, K.J., & Laszlo, S. (2014). Brainprint: Identifying Unique Features of Neural Activity with Machine Learning. Proceedings of the 36th Annual Conference of the Cognitive Science Society, Mahwah, NH: Lawrence Erlbaum Associates.

Schölkopf, B., & Smola, A. J. (2002). Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press.

Schwarzkopf, D. S., Song, C., & Rees, G. (2010). The surface area of human V1 predicts the subjective experience of object size. *Nature neuroscience*,*14*(1), 28-30.

Su, F., Zhou, H., Feng, Z., & Ma, J.. (2012). A biometric-based covert warning system using EEG. *Proceedings of the 5th IARP International Conference on Biometrics*, 342-347.

Wijdicks, E. F. (1995). Determining brain death in adults. *Neurology*, *45*(5), 1003-1011.

Young, M. P., & Rugg, M. D. (1992). Word frequency and multiple repetition as determinants of the modulation of event□related potentials in a semantic classification task. *Psychophysiology*, *29*(6), 664-676.

Figure Captions

Figure 1: Sample data and DIVA reconstructions. Left: A true ERP elicited by Participant 0. Center: The best DIVA reconstruction of that ERP. Right: The worst DIVA reconstruction of that ERP. The best DIVA reconstruction appears as a slightly filtered version of the true ERP, with early component activity emphasized (grey box).

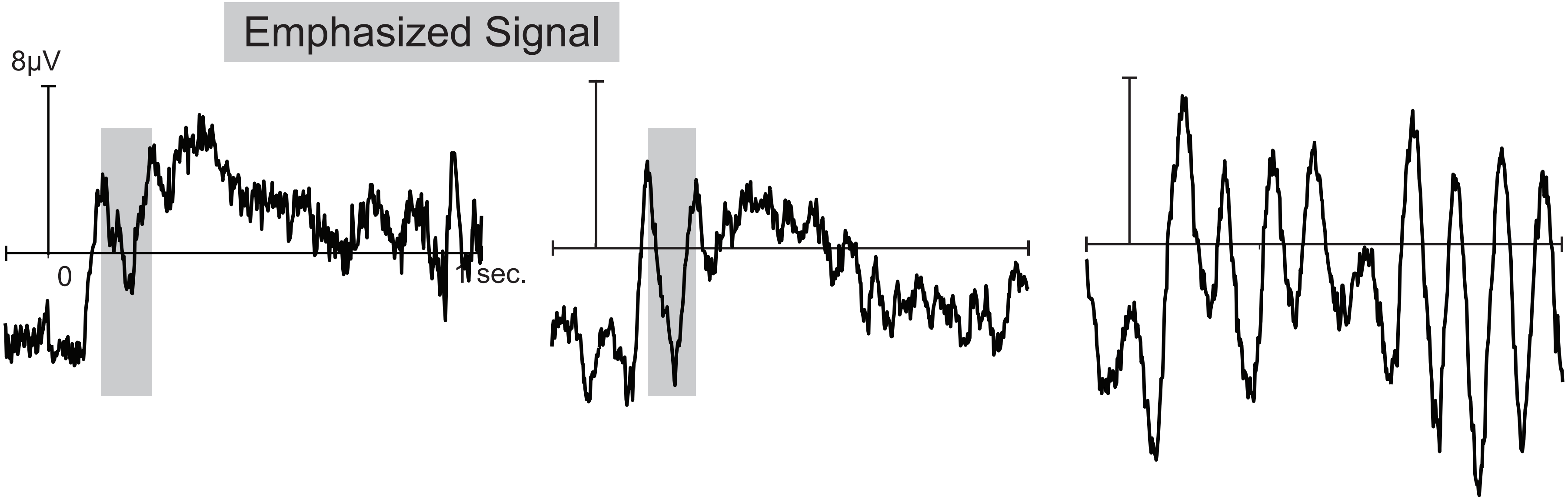
Figure 2: Accuracy values for the four different classifiers, the meta-classifier and the permanence data for 2nd and 3rd sessions. Error bars denote 95% confidence intervals on the means.

Figure 3. Maximum classification accuracy across classifiers, by participant. Error bars denote the standard error of the mean. The participant with the lowest classification accuracy is colored in black. Note that many participants are classified with 100% accuracy across models.

Figure 4. Sample train and test data for a well classified participant (participant 4, left) and the most poorly classified participant in the study (participant 21, right). It is clear that Participant 21 is difficult to classify due to the presence of an ocular artifact (dashed box) present in only the train data. Participant 4 is classified with 100% accuracy across trials.

Figure 5. Cross-correlation identification accuracy over time. Each dot represents a single participant. The majority of participants are still very accurately classified with a delay between first and last session of as much as 178 days.

Figure 1



Participant 0: True ERP

Participant 0: DIVA best reconstruction [Rank 1]

Participant 0: DIVA worst reconstruction [Rank 32]

Figure 3

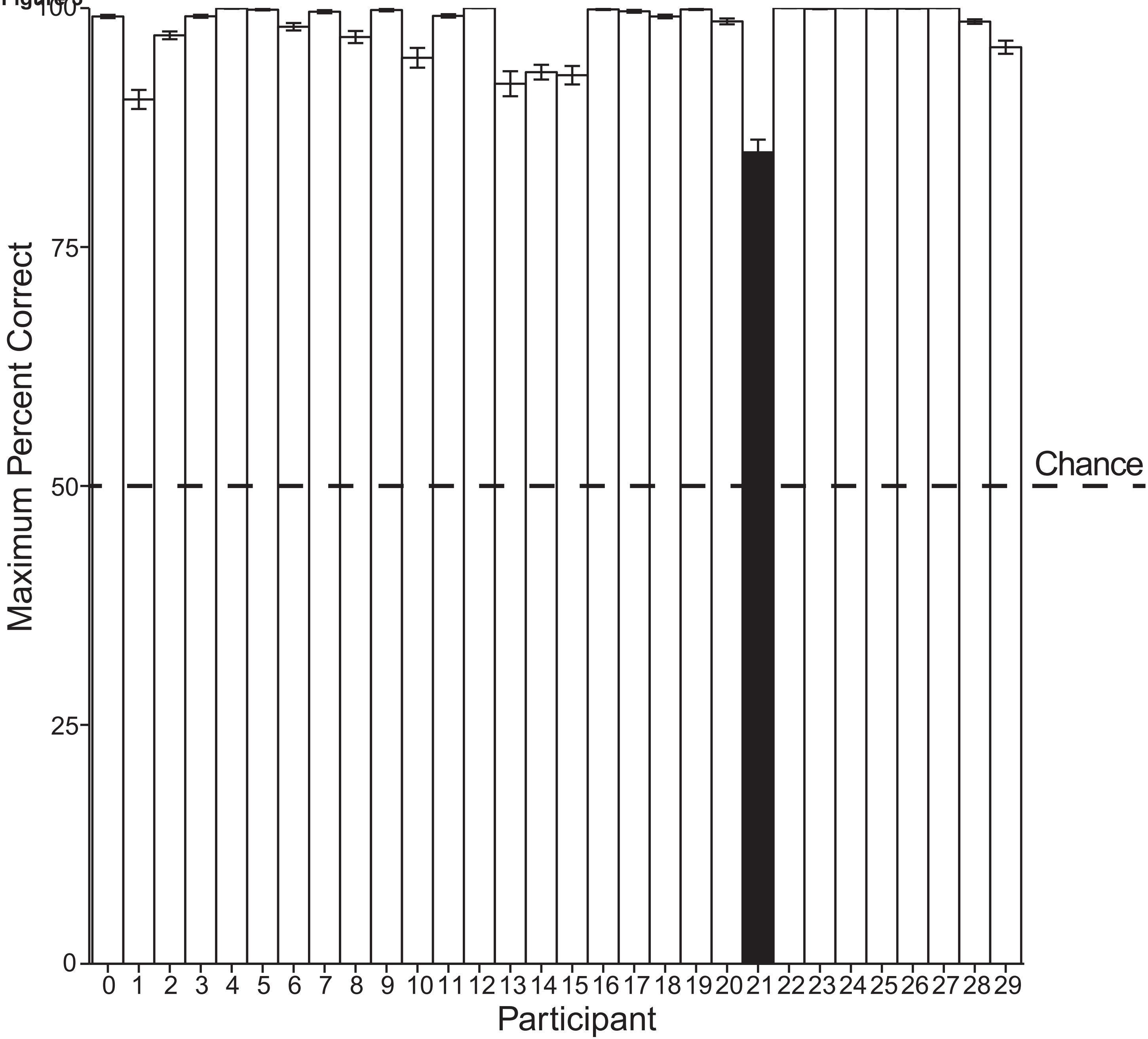
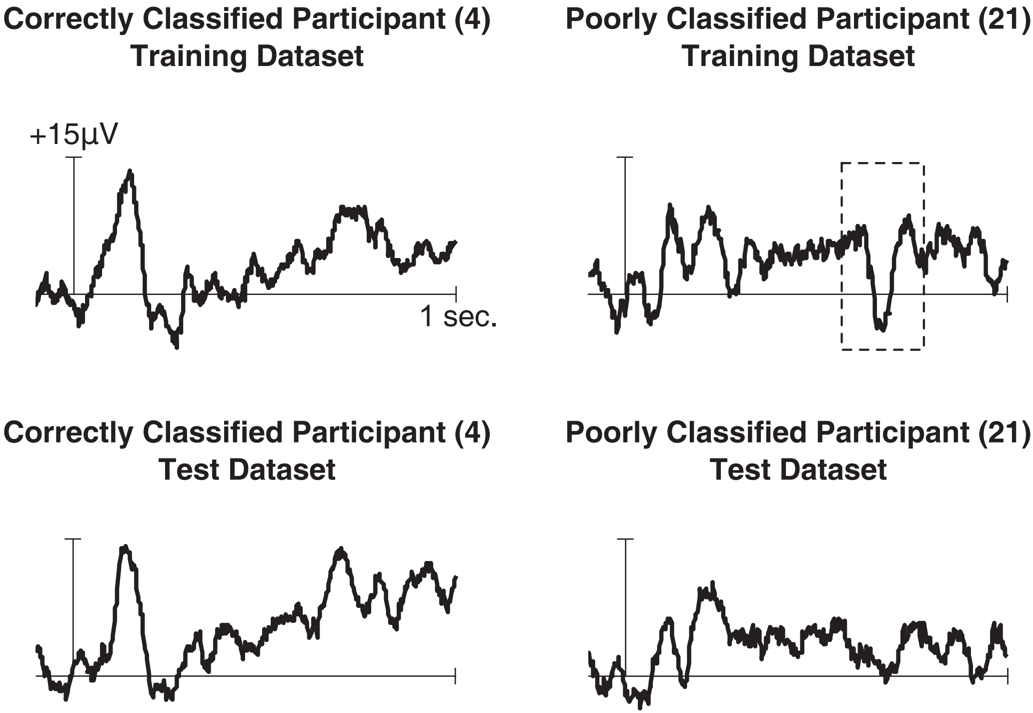

Maximum Percent Correct
100
75
50
25
0
Chance
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
Participant

Figure 4



Correctly Classified Participant (4)
Training Dataset

Poorly Classified Participant (21)
Training Dataset

Correctly Classified Participant (4)
Test Dataset

Poorly Classified Participant (21)
Test Dataset

Figure 2

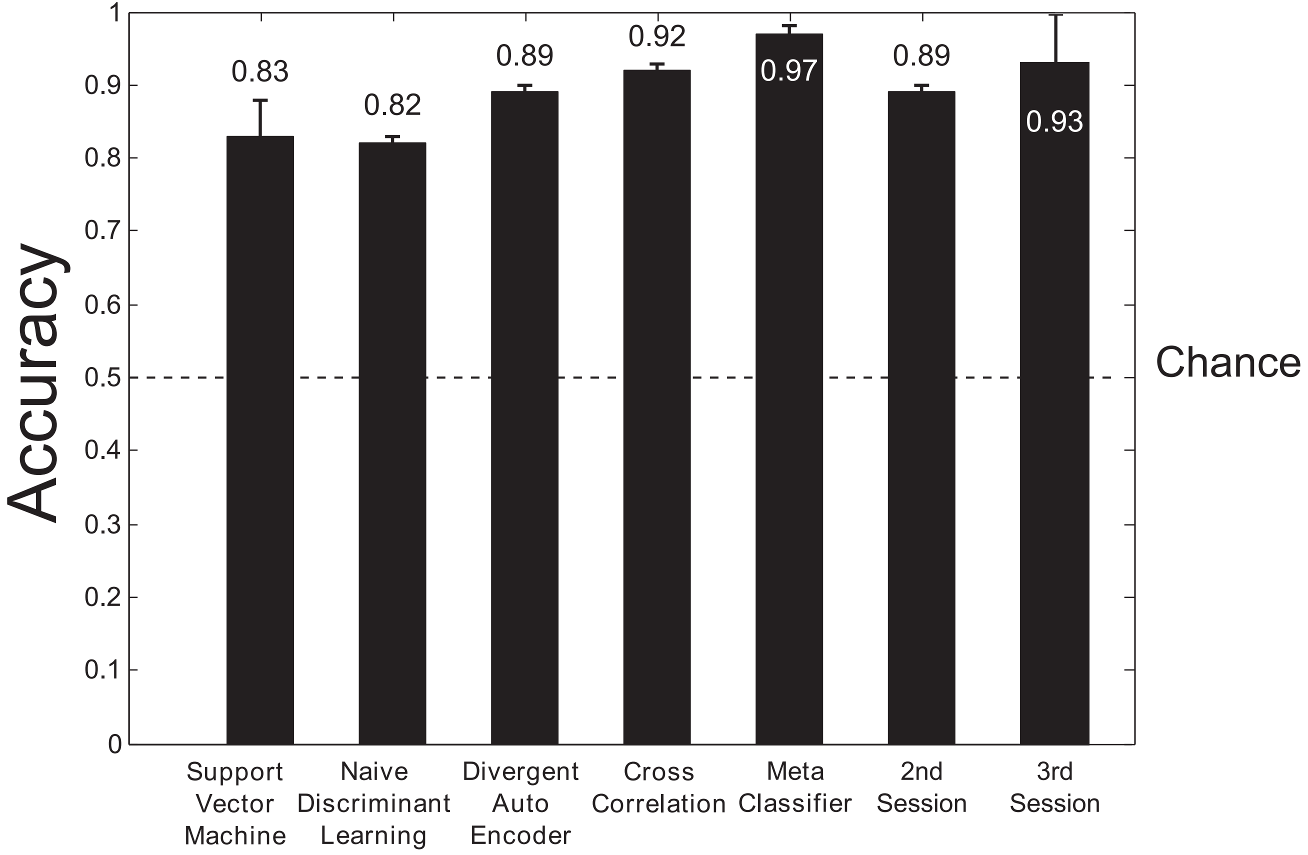
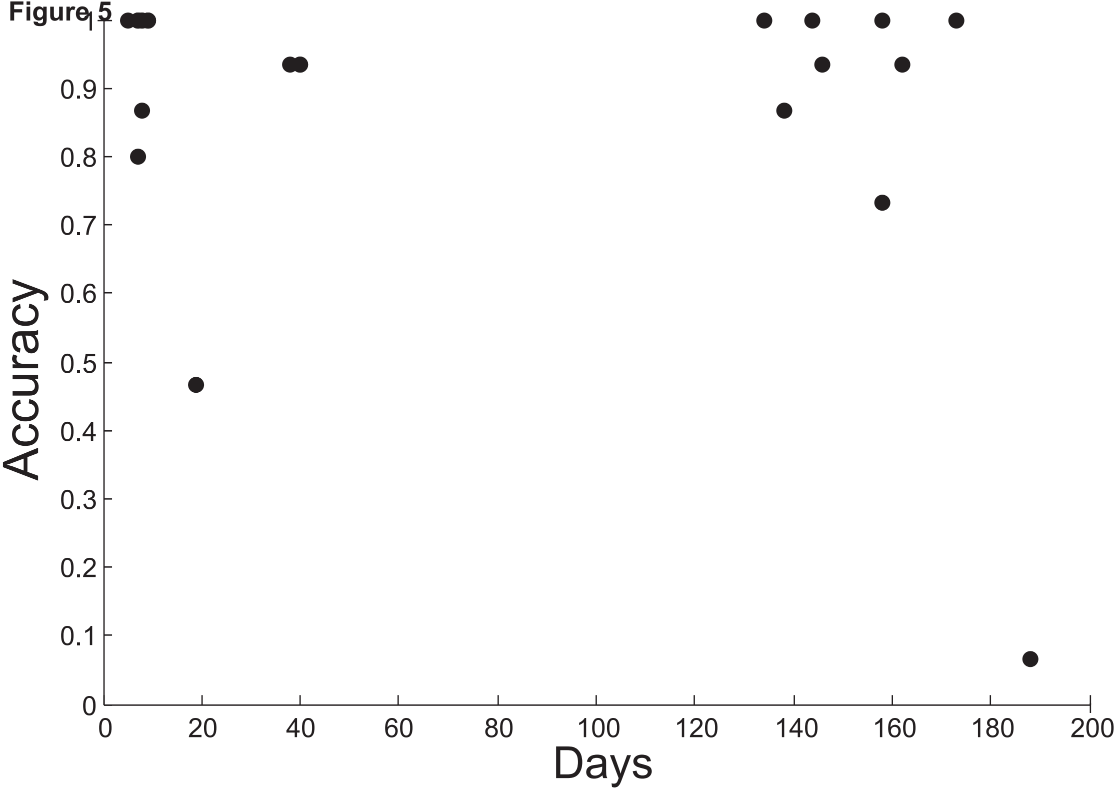

Accuracy
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.83
0.82
0.89
0.92
0.97
0.89
0.93
Chance
Support Vector Machine
Naive Discriminant Learning
Divergent Auto Encoder
Cross Correlation
Meta Classifier
2nd Session
3rd Session

Figure 5


Accuracy
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
20
40
60
80
100
120
140
160
180
200
Days