Title: On the problem of spike-wave discharge detection by humans and machines.

Abbreviated Title: Rigorous classification of SWDs.

Authors and Affiliations: Jesse A. Pfammatter[1], Rama K. Maganti[2], and Mathew V. Jones[1]

[1]Department of Neuroscience, University of Wisconsin, 1111 Highland Avenue, Madison, WI, 53705

[2]Department of Neurology, University of Wisconsin, 1685 Highland Avenue, Madison, WI 53705

Author Contributions: JP designed research, analyzed data, and wrote the paper; RM analyzed data and edited the paper; MJ designed research, analyzed data and edited the paper.

Corresponding Author:

Jesse Pfammatter
5505 WIMR II
1111 Highland Ave.
Madison WI, 53705
email: pfammatter@wisc.edu

Number of Figures: 5
Number of Tables: 1
Number of Multimedia: 1
Number of Words for Abstract: 247
Number of Words for Significance Statement: 110
Number of Words for Introduction: 750
Number of Words for Discussion: 956

Acknowledgements: We would like to thank Kile P. Mangan for scoring SWDs, Eli Wallace for scoring sleep, and Rebecca Willet for helpful conversations regarding SVMs. Antoine Madar and Eli Wallace provided invaluable feedback during the development of this work. The data analyzed were collected by KPM and Aaron Nelson in the lab of Chiara Cirelli.

Conflicts of Interest: Authors report no conflicts of interest.

Funding Sources: RO1 NS075366 'Tonic Inhibition in Thalamocortical Network Function and Absence Epilepsy', DOD PR161864 'Reduced GABAergic Tonic Inhibition as a Shared Mechanism of Post-Traumatic Sleep Disorders and Epilepsy'

Abstract

The detection of events from EEG records has traditionally been done manually even though manual detection is time intensive, costly and subject to inconsistencies and biases. The need for automated methods for seizure-related event detection is widely recognized and indeed development of these methods has been abundant. However, clinicians and researchers continue to rely on manual event detection because current automated algorithms either fail to produce accurate results, to meet users' expectations or are difficult to employ. Thus, there is a need for new solutions to the problems inherent in EEG event detection. Here we develop a support vector machine (SVM)-based algorithm for the automated detection of spike-wave discharges (SWDs) in a mouse model of absence epilepsy. The algorithm first identifies putative SWD events using frequency- and amplitude-based peak detection. Then, using predictors calculated from the Morlet wavelet transform of each event, it discretely classifies (SWD/nonSWD) and assigns a confidence score to each event. The result of this algorithm is scoring of SWDs along a 'confidence' continuum that is highly correlated with human confidence and that allows us to more effectively characterize ambiguous events in order to gain a better understanding of 'ground truth' in SWD detection. Finally, we validate the algorithm by showing that the detected events are temporally correlated, on the scale of seconds, with abrupt changes in spectral power bands relevant to normal behavioral states including sleep. Our results demonstrate the value of viewing detection of SWDs as a continuous, rather than discrete, classification problem.

67 **Significance Statement**

68 Manual detection of epilepsy-related events from EEG is time consuming,
69 expensive, and fraught with problems associated with inter- and intra-rater consistency.
70 Both clinicians and researchers would benefit from an automated method of absence
71 seizure detection that also provides a framework for the description of seizure events
72 and how they relate to other electrographic events (e.g. sleep characteristics). We
73 present an algorithm for the fully automated, consistent, and rapid scoring of SWDs and
74 demonstrate the use of this algorithm. We anticipate that this algorithm, in conjunction
75 with other automated EEG analysis tools, can be used to individualize treatments and to
76 identify mechanisms that may be common across different types of epilepsy.

Introduction

The electroencephalogram (EEG) is the gold standard for epilepsy diagnosis, treatment, and research. Scoring of electrographic events (e.g., convulsive seizures, absence seizures, interictal spikes (IISs), sleep stages, etc.) has traditionally been done manually; a practice that presents a major bottleneck in clinical and research workflows.

One major problem with manual scoring of EEG is that it is intensely time-consuming and costly. A skilled human requires 2-3 hours to score a 24-hour record. In practice, this means that much data goes unanalyzed. In epilepsy monitoring units, it is common practice to scan EEG to obtain a qualitative appraisal that guides diagnoses and treatment. Unfortunately, this technique may not be sufficient to treat all patients (Boly and Maganti, 2014). In research, groups often use limited segments of EEG for manual event detection and analysis.

Another problem with manual scoring of EEG is that agreement between expert scorers is often low. Agreement on sleep scoring is often no greater than ~80% (Danker Hopfe et al., 2009). The detection of nonconvulsive epileptiform events (e.g. spike-wave discharges (SWDs), IISs, etc.) is even more challenging. In fact, Webber et al. (1993) found 52% agreement between humans scoring 3-5 minute EEGs for spikes, demonstrating that establishing objective criteria to ensure accurate, consistent and unbiased scoring is difficult and should be a high priority.

Nonconvulsive epileptiform events are difficult to identify for numerous reasons. First, the definition of ‘epileptiform’ is often vague (Fisher et al., 2014; Koutroumanidis et al., 2017a; 2017b). For example, Adjouadi et al. (2005) suggests that IISs should have a duration between 20-70 ms whereas Polivannyi et al. (2017) suggests that the range

should be 5-50 ms. Adjouadi et al. (2005) also suggests that the rising and falling phases of a spike must be ‘fast’. Such criteria are somewhat subjective and subject to disagreement. Second, clinicians and researchers have traditionally restricted themselves to categorizing epilepsy-related events into discrete groups, whereas events probably exist on a continuum (as we illustrate). Unfortunately, manual scoring of continuous events is not practical and may not even be possible. Clinicians and researchers end up using subjective criteria, such as that an event “stands out of the background” (Webber and Lesser, 2017), to force events into binary categories. Finally, many electrographic features appear hidden in the time domain (Lüttjohann, 2017). As we show below, events can appear ambiguous to the same expert scorer such that, on repeated presentations, a scorer changes their mind about an event’s categorization. Such ambiguity is reminiscent of human perception of optical illusions such as the Necker Cube (Kornmeier and Bach, 2005) and suggests that ‘ground truth’ in identifying electrographic events may need to be considered in terms of confidence rather than certainty.

There has been significant development of algorithms for automated detection of convulsive seizures (Bergstrom et al., 2013), IISs (Wilson and Emerson, 2002; Scheuer et al., 2017), SWDs (Ovchinnikov et al., 2010; Xanthopoulos et al., 2010; Bauquier et al., 2015; Richard et al., 2015), sleep (Sunagawa et al., 2013) and artifacts (Nolan et al., 2010). However, despite progress, most clinicians and researchers continue to rely on manual scoring methods. These algorithms may not be widely used because installation, tuning, and proper implementation often requires significant programing and analysis experience. Even users that do have such skills may lack confidence in the output, either

because it does not match their impression of how events 'should' be detected or because the algorithm does not perform as advertised on novel data. Finally, many proprietary software packages that tout quality classification are expensive and will never be tried by the majority of users.

There is a critical need for development of tools to *rapidly*, *accurately and consistently* score long-duration EEG records for epileptiform events and sleep stages. These tools must aid in the establishment of quantitative 'ground truth' for epileptiform events, even for the case of ostensibly stereotyped events such as SWDs. Development of such software is imperative a) to speed up (and make less expensive) both clinical diagnoses and research, b) to make transparent and quantitative the consensus criteria used by experts, and c) to understand the mechanistic links between epilepsy and sleep.

We present a support vector machine (SVM)-based algorithm for the detection of SWDs in mice expressing a $GABA_A$ receptor mutation (γ2R43Q) that causes absence epilepsy (Tan et al., 2007; Petrou and Reid, 2012). This algorithm is rapid, scoring a 24-hour record in ~3 minutes, 100% consistent, mimics characteristics of human uncertainty, and creates a pathway for better understanding of the 'ground truth' for SWDs.

Methods

All use of animals was approved by the University of Wisconsin - Madison Institutional Animal Care and Use Committee (IACUC).

Animal colonies. Animals were bred from a colony maintained at the University of Wisconsin, Madison. In a paired breeding scheme, female wild-type (RR) mice of a

146 C57BL/6J-OlaHsD background (formerly Harlan, currently Envigo, Madison, WI) were
147 bred with male heterozygous mice expressing the γ2R43Q knock-in mutation in the
148 same background. Heterozygous knock-in mice (RQ) experience SWDs concurrent with
149 behavioral arrests (Tan et al., 2007) whereas homozygous mutants were almost never
150 born. The γ2R43Q strain was provided by Dr. Steven Petrou.

151 *Computer Programing and Code Accessibility.* All analyses, data processing, and
152 event classification was done in Matlab (Mathworks, Natick, MA) using a combination of
153 home-written and publicly available scripts. The final version of this algorithm was
154 executed with Matlab 2017b and all code is available at the GitHub repository:
155 https://github.com/jessePfammatter/detectSWDs.

156 *EEG implantations, recording, and data.* EEG electrodes were implanted as
157 described in Nelson et al. (2013). Briefly, male heterozygous knock-in (RQ) and wild-
158 type (RR) mice were implanted at p24, under isoflurane anesthesia (1-2% in 100% O2).
159 Each animal was implanted with gold plated miniature screw electrodes over the right
160 and left frontal and parietal cortices, and one over the cerebellum as reference. Two
161 vinyl-coated braided stainless-steel wire electrodes were placed in the nuchal muscle
162 for electromyogram (EMG) recording. The EEG and EMG leads were wired to a head
163 cap, which was affixed to the skull with dental acrylic.

164 After recovery from surgery (2-3 days), animals were hooked up to a
165 multichannel neurophysiology recording system (Tucker-Davis Technologies, TDT,
166 Alachua, FL, USA) to continually sample EEG and EMG signals at 256 Hz (digitally
167 bandpass filtered between 0.1 and 100 Hz) for up to fourteen days. Data were notch
168 filtered at 60 Hz using a Chebyshev Type II digital filter and high-pass (>2 Hz) filtered

with a Chebyshev Type I digital filter. EEG signals were then normalized to have a mean of zero and a standard deviation of 1 using the following method (herein referred to as Gaussian normalization): First, we fit a Gaussian distribution to the all-points histogram of each 24-hour record using least-squares minimization and Nelder-Mead Simplex optimization (Nelder and Mead, 1965). Data points equal to zero were excluded. We then normalized the 24-hour record by subtracting the mean from each point and dividing by the standard deviation of the model fit. All analyses here used a frontal EEG channel (usually front-left).

Automated Event Detection. We used a two-stage algorithm for the detection and classification of SWDs. In Stage 1, we automatically detected putative SWD events using a frequency and amplitude threshold-based approach. The parameters we selected for this first stage are tunable and we chose parameters to be ‘all inclusive’ of potential SWD events at the cost of also finding false positives, which were filtered out in Stage 2. First, we detected all peaks above 3 standard deviations of the mean of the normalized EEG signal, excluding peaks within a critical distance (here, 86 ms) from a previous peak. We then calculated the normalized (Gaussian fit method above) first derivative of the signal and identified all peaks above 3 standard deviations, excluding peaks within a critical distance (here, 83 ms) from a previous peak. We retained all peaks that coincided between the normalized EEG signal and the first derivative, where the criterion for coincidence was that a peak in the first derivative occurred within 59 ms prior to the corresponding peak in the normalized signal. Finally, we grouped the peaks into ‘possible events’ if the peaks were nearer than 316 ms. This procedure selected events that had peaks with both a fast rising phase and a prominent amplitude and that

192 fell within previously published guidelines for the frequency of SWDs in mice (Mangan,
193 2013; Tan et al., 2007) and rats (Sitnikova and van Luijtelaar, 2007; Pinault et al.,
194 2001).

195 *Manual Event Classification.* We presented 2500 putative SWD events (i.e.,
196 identified by the Stage 1 detection) to each of four human scorers for manual
197 classification. Of these, 2050 were randomly selected unique events from 10 24-hour
198 EEG records from 5 animals (3 RQ and 2 RR). We also randomly selected 50 events to
199 be repeated 10 times each. Each of the 2500 events was shown to the human scorers
200 in random order via a graphical user interface (GUI). Each event was presented as an 8
201 second length of normalized single-channel EEG with the putative event highlighted in
202 yellow. Scorers were asked to classify each event as either a SWD or nonSWD and
203 were given as much time as they needed to view/review/reclassify events. All four
204 human scorers were already skilled in the identification of SWDs but were blind to the
205 genotypes, and two of the scorers were not aware of the presence of repeated events.
206 None of the scorers were aware of which specific events were repeated because the
207 order of presented events was random. This manual scoring of a small subset of data
208 (much less than 1%) required about 1-2 hours per scorer.

209 EEG/EMG signals were also manually scored for sleep stages (wake, NREM, or
210 REM) (see Fig. 3) in 4 second epochs with Serenia Sleep Pro software (Pinnacle
211 Technology, Lawrence, KS, USA).

212 *Event Predictor Variables:* From the putative events detected by Stage 1, we
213 extracted 12 predictor variables for use in Stage 2 of the automated algorithm: event
214 classification with a Support Vector Machine (SVM). These variables were calculated

from four sets of scale ranges from the wavelet transform using the cwt function in Matlab (Morlet wavelet) of each event. These ranges were picked to target the characteristic ~6Hz frequency of SWDs and several of its harmonic components. For each of the 4 sets of scales (49-40 ≈ 4.4-8.2 Hz, 39-30 ≈ 8.8-16.4 Hz, 29-20 ≈ 17.6-32.8 Hz, and 19-10 ≈ 35.1-65.5 Hz) we summed the absolute value of the wavelet transform scalogram amplitude across the scale set and across the duration of the event and then calculated the mean, standard deviation and maximum (4 scale ranges * 3 statistics = 12 predictors).

Automated Event Classification. For Stage 2 of the algorithm we developed an SVM-based automated method to classify SWD events identified in Stage 1 using the 12 event predictors. We trained the SVM using the labels provided by the human scorers (10000 labels, 2500 per scorer for 2050 unique events) where the algorithm created a weighted response for each unique event based on the average human label across all four scorers. In our preliminary testing of the algorithm and upon visual inspection of the predictor variables for these events, we choose a Gaussian kernel for the SVM. In order to choose the proper hyperparameters (kernel scale, box constraint, and cost matrix) for the SVM we employed a grid-search based optimization process. We trained the SVM under each unique combination of the following parameters; kernel scale = [1, 5, 10, 15, 20], box constraint = [5, 10, 15, 20, 25], and a cost matrix of [0, 1; x, 0] where x = [1, 1.5, 2, 2.5, 3]. We then chose hyperparameters with the following guidelines: A) Each scorer should contribute to the model as equally as possible. This was achieved by minimizing the standard deviation of the agreement between human scorers and the SVM. B) To avoid overfitting to subsets of data, the model should

perform well in 5-fold cross-validation. We quantified this by calculating the mean 5-fold cross-validation accuracy between the SVM and the training labels for each of 10 iterations of model construction (there is variability as this is a stochastic process) and taking the standard deviation of this value. Minimizing this quantity allowed us to select a model that performs similarly on numerous randomly chosen subsets of data. C) The model should match human uncertainty as much as possible. We calculated this agreement by calculating the R^2 between the model scores (distance from each event to the nearest support vector) and the weighted human responses for each of the 50 repeated events within the data set (i.e. intra-rater reliability). Maximizing this quantity produces a model with high fit to the human uncertainty. Although A, B, and C were the primary parameters for optimization, we considered several other quantities including: i) agreement between the SVM and the events that all humans agreed upon (human consensus data or HC), ii) average agreement between the human scorers and the SVM and iii) average of the 5-fold cross-validation process. While these parameters are important, their optimal set point is unclear given that variability between human scorers suggests that ground truth labels may not exist for these data. Here, we present data as classified by an algorithm with the following hyperparameters: kernel scale = 10, box constraint = 10, cost matrix = [0 1, 1.5 0]. A nice summary of using SVM for EEG event detection is provided by Nandan et al. (2010).

In order to highlight the relative predictive importance of the variables used in the SVM we performed a LASSO regression with the SVM classification score as a response variable and the 12 variables calculated from the wavelet transform of each event as predictor variables. We allowed the LASSO regression 100 iterations with

261 cross validated (10 fold) fits and selected the final coefficients from the model with a
262 lambda value (indicating lowest cross-validation error) that was one standard deviation
263 larger than the model with the lowest lambda value.

264 *Application of the algorithm and preliminary testing of the SWD output scores.*
265 We then applied the algorithm to classify SWDs in several 24-hour EEG records that
266 were not used for algorithm development. Using the output of the algorithm we
267 calculated the event-triggered average of 50 random examples of strong SWDs
268 (distance to support vector of >1 to 2), moderate SWDs (>0.5 to 1), weak SWDs (>0 to
269 0.5), weak nonSWDs (<0 to -0.5), moderate nonSWDs (<-0.5 to -1), and strong
270 nonSWDs (<-1 to -2), and 50 random locations not identified by our algorithm to delta
271 (0.5-4 Hz) , theta (6-9 Hz), sigma (10-14 Hz), and gamma (25-100 Hz) power
272 surrounding each event during lights on and lights off periods.

273 **Results**

274 We found that SWDs in γ2R43Q animals have a characteristically strong and
275 continuous power in the ~6 Hz frequency range during the entire discharge event (Fig.
276 1) consistent with previous descriptions of SWDs in these animals (Tan et al., 2007).
277 Additionally, SWDs appear to have a very reliable 'fingerprint' in the wavelet transform
278 time-frequency decomposition (Fig. 1).

279 *Manual Scoring and Inter- and Intra-Rater Reliability.* Overall, human scorers
280 were quite variable in their scoring. Out of the 2050 unique events, the four human
281 scorers were in complete agreement (i.e., consensus labels) for only 57% of the events
282 (1159 events, 940 nonSWD and 219 SWD). Figure 2 shows a characterization of
283 human scoring of the 2500 events presented as 2D cross-sections in predictor space for

6 of the predictor variables used in the SVM. Humans tended to agree about nonSWD events (blue) whereas disagreement (green) and agreement about SWDs (red) tended to occupy similar regions of feature space. Further characterization of the inter- and intra-rater reliability are shown in Figure 3, which demonstrates the difference in rating trends of 3 of the human scorers. Figure 3B shows 8 example events pulled from the 50 events presented to each human scorer 10 times. These 8 events are ranked in order of agreement where events in dark red were fully agreed upon SWDs (0:40, nonSWD:SWD), blue were fully agreed upon nonSWDs (40:0) and intermediate colors represent ambiguous responses. For each of the 2500 events presented to human scorers, agreed upon SWD events have high values for both the maximum values in 4.4-8.2 Hz and 17.6-32.8 Hz ranges whereas agreed upon nonSWD events have a low maximum value for these bands (Fig. 3B). Intra- and inter-rater reliability data are shown in Figure 3C-E. Inter-rater reliability on the 50 repeated events (RE) data agreed with the trend in distribution of human agreement for all of the 2500 events presented (low-resolution data). Average intra-rater reliability across all scorers was 100% for 13 of 50 repeated events. 13 of 50 had an average intra-rater agreement of 90% or less across all scorers. For individual scorers, intra-rater agreement was as low as 50% for the most ambiguous events. Average Intra-rater reliability was 94% across all scorers and all 50 repeated events. Inter-rater agreement values for the human scorers are shown in Table 1, along with agreement between humans and the SVM (below).

Performance of Automated Detection Compared to Manual Scoring. Traditional clustering methods such as K-means failed to partition the data similar to the human scorers (notice the ‘fuzziness’ in SWD/nonSWD scoring in Figure 3A) but rather drew a

307 hard line between SWD/nonSWD events (Fig. 4A). Figure 4B shows the output of the
308 SVM that was trained using the human scoring labels and demonstrates and displays a
309 similar 'fuzziness' in SWD/nonSWD scoring as human scorers. In fact, the SVM
310 classification scoring associated with each event (e.g. distance to the nearest support
311 vector) is strongly correlated with human inter-rater reliability (Fig. 4C). Overall, the
312 sensitivity of our final algorithm was 0.98 and the specificity was 0.95 as compared to
313 the human consensus labels. Figure 4D shows the coefficients for the LASSO
314 regression. The absolute value of these coefficients indicates the relative importance of
315 each variable whereas the sign indicates whether a large versus small value of that
316 variable, relative to the others, was most predictive of the SVM classification scores.
317 Importantly, negative coefficients do not indicate that the magnitude for those variables
318 were lower for SWDs than nonSWDs on average.
319 We present graphical results for the optimization of our tunable SVM parameters
320 in Figure 5.
321 *Application of the Algorithm.* We applied the algorithm to analyze SWDs in a set
322 of 60 24-hour EEG recordings from RR and RQ animals, 50 of which were not used for
323 SVM development. The algorithm analyzed these records in ~40 minutes, or ~1.5
324 minutes per 24-hour record (compare with 1-3 hours required for a human to score
325 SWDs in a 24-hour record). We found an average of 737.51 ± 482.28 events per 24-
326 hour RQ record (n = 8 animals, 45 24-hour records) where 194.09 ± 169.37 of these
327 events are considered SWD events (> 0 SVM score). Per 24-hour record in RR animals
328 (n = 3 animals, 14 24-hour records), we find 409.64 ± 207.28 possible events where
329 26.79 ± 21.18 of these events are considered SWD events. Figure 6 shows the

application of our algorithm on a representative 24-hour record from a RQ animal. Using the output of the algorithm (a list of event times and distance to nearest support vector) on an RQ animal, we demonstrated the relatedness of SWD and nonSWD events (as labeled by the SVM) to sleep staging. We found strong and abrupt relationships between SWDs and each power band that sometimes preceded the SWD by several minutes (Fig. 6 D-K).

As previously mentioned, the detection algorithm found 26.79 ± 21.18 events considered to be SWD in RR animals. This value might be considered by some as the 'false positive' rate for our detection algorithm in wild-type animals. However, some proportion of these events have very strong SWD features and show a temporal relationship to the delta, theta, sigma, and gamma power bands as events in RQ animals (data not shown). Figure 7 shows six events pulled from normal animals that have strong visual SWD characteristics.

Discussion

We have developed a fully automated, SVM-based algorithm for the detection of SWDs that leverages human agreement to develop a continuous, confidence-based score in addition to a binary classification (SWD/nonSWD). The algorithm was trained using a dataset with 2500 putative events (2050 unique) and labels (SWD or nonSWD) from four expert human scorers, using five-fold internal cross-validation. The resultant SVM agreed with humans as well as they agreed with each other (~70%) and, unlike the other classification techniques we investigated such as K-means clustering (Fig. 4A) but like humans, did not draw sharp boundaries in predictor space between regions containing SWDs vs nonSWDs.

353 We optimized the algorithm to minimize the variance between human scorers
354 and the SVM, thus reducing biases introduced by any individual human scorer.
355 Importantly, as evident from the LASSO regression and similar to human scorers, the
356 algorithm keyed in on an ensemble of predictor variables that allows for a maximization
357 of the signal to noise ratio in SWD events. After training, fully automated scoring of 24-
358 hour EEG records was conducted, requiring ~3 minutes per 25-hour record. Thus, the
359 algorithm reproduced features of the human decision-making process while adding
360 important aspects that humans lack, including consistent, unbiased and rapid scoring of
361 events.
362 There are, of course, some caveats. Although the method works quite well on the
363 γ2R43Q model of absence epilepsy, we have yet to fully test it in other models of
364 absence epilepsy. Continued development will include testing and expansion to other
365 animal models as well as to human EEG. We anticipate that Stage 1 (event detection)
366 should work quite well for most types of absence seizures exhibiting an ~6 Hz character
367 (as occurs in rodents). However, for Stage 2 (the SVM classification) we anticipate that
368 users would need to retrain the algorithm with human labels that are representative of
369 the different types of animals/treatments being analyzed. The time/cost of such
370 retraining may be as low as a few hours per human scorer. It is possible that the current
371 SVM could be tuned to perform broadly across additional models of epilepsy without re-
372 training by a tuning of the SVM hyperparameters. In this current version, we selected a
373 set of optimization parameters that best fit the characteristics of human scoring,
374 however other hyperparameters produced results that we believe would be more
375 applicable as a generalized algorithm for SWD detection. Indeed, selection and tuning

of model hyperparameters can be challenging and result in selection bias that influence model performance as much as the selection of different predictor variables (Cawley and Talbot, 2010). Continued development of this algorithm may also include components shown to be important for other SWD detection algorithms such as using phase information from the wavelet transform such as was done in (Richard et al., 2015).

Despite its limitations, we have used this algorithm to highlight a few important aspects of SWDs that would have been difficult or impossible to identify solely with human classification. Our data show that events along the spectrum of the SVM score (and thus SWD confidence) show temporal correlations with abrupt changes in EEG power bands. While these correlations are most prominent in SWDs with the highest SVM scores, events with lower scores (and even those below threshold) continue to show correlations, albeit with smaller effects. Qualitatively similar temporal patterns were observed in all of the R43Q animals analyzed in this study (data not shown). Thus, the SVM reveals physiologically relevant changes in brain activity that are commonly used in the scoring of sleep and shows that these changes are temporally linked to, and could potentially be used to predict, the occurrence of SWDs. Importantly, our findings align with previous work showing shifts in phase-amplitude coupling up to a minute prior to SWDs (Maheshwari et al., 2017), and that SWDs are related to sleep timing (Halasz et al., 2002). Our future work will continue to expand upon the view of SWD ‘ground truth’ as presented in this manuscript.

We also identified SWD-like events in wild-type mice, consistent with the hypothesis that SWD events represent a corruption of otherwise normal brain processes

399 (Kelly 2004). If so, finding examples of SWDs with lower intensity (i.e. smaller distance
400 to nearest SV) is to be expected. Additionally, Letts et al. (2014) found that 8 of 27
401 strains of wild type mice had SWDs including several from the C57-related
402 backgrounds. It is important to have a rigorous method that is capable of detecting
403 these lower intensity events because, similar to higher intensity SWD events, many
404 show a temporal relationship with several bands of EEG power. These low intensity
405 events would mostly likely be missed by human scorers since they either have lower
406 amplitude or a less consistent pattern in the time domain (i.e. not easily identifiable by
407 human scorers). Nonetheless, these events show temporal relationship with several
408 EEG power bands that coincide with confident SWD events (data not shown).
409 Although other groups have developed algorithms for the detection of SWDs
410 (Ovchinnikov et al., 2010; Xanthopoulos et al., 2010; Bauquier et al., 2015; Richard et
411 al., 2015), to the best of our knowledge, no other group has designed an algorithm to
412 mirror human confidence in scoring. We used the confidence-based scores generated
413 from our algorithm and based on human scoring characteristics as a starting point to
414 help establish ‘ground truth’ for SWD events and here we present some data, namely
415 that the relationship between event-triggered averages and EEG power spectra change
416 in a continuous trend with decreasing likelihood of SWD, that indicates that variability in
417 inter- and intra-rater reliability are extremely important in helping us learn more about
418 ‘ground truth’ in SWDs. However, it is important to note that this is only a first step
419 towards a clearer understanding of SWDs. While we have learned a lot about ground
420 truth, or the lack thereof, from the characteristics of human and machine scoring, we

421 believe that an algorithm that truly targets ‘ground truth’ and targets human decision

422 making, will be valuable going forward.

References

Adjouadi M, Cabrerizo M, Ayala M, Sanchez D, Yaylali I, Jayakar P, Barreto A (2005) Detection of interictal spikes and artifactual data through orthogonal transformations. J Clin Neurophysiol 22:53–64.

Bauquier SH, Lai A, Jiang JL, Sui Y, Cook MJ (2015) Evaluation of an automated spike-and-wave complex detection algorithm in the EEG from a rat model of absence epilepsy. Neurosci Bull 31:601–610.

Bergstrom RA, Choi JH, Manduca A, Shin H-S, Worrell GA, Howe CL (2013) Automated identification of multiple seizure-related and interictal epileptiform event types in the EEG of mice. Sci Rep 3:1483.

Cawley GC, Talbot NLC (2010) On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res 11:2079–2107.

Fisher RS, Scharfman HE, deCurtis M (2014) How can we identify ictal and interictal abnormal activity? Adv Exp Med Biol 813:3–23.

Halasz P, Terzano MG, Parrino L (2002) Spike-wave discharge and the microstructure of sleep-wake continuum in idiopathic generalised epilepsy. Clin Neurophysiol 32:38–53.

Kelly KM (2004) Spike-wave discharges: Absence or not, a common finding in common laboratory rats. Epilepsy Curr 4:176-177.

Kornmeier J, Bach M (2005) The Necker cube - an ambiguous figure disambiguated in early visual processing. Vision Res 45:955–960.

Koutroumanidis M, Arzimanoglou A, Caraballo R, Goyal S, Kaminska A, Laoprasert P, Oguni H, Rubboli G, Tatum W, Thomas P, Trinka E, Vignatelli L, Moshé SL (2017a)

The role of EEG in the diagnosis and classification of the epilepsy syndromes: a tool for clinical practice by the ILAE Neurophysiology Task Force (Part 1). Epileptic Disord 19:233–298.

Koutroumanidis M, Arzimanoglou A, Caraballo R, Goyal S, Kaminska A, Laoprasert P, Oguni H, Rubboli G, Tatum W, Thomas P, Trinka E, Vignatelli L, Moshé SL (2017b) The role of EEG in the diagnosis and classification of the epilepsy syndromes: a tool for clinical practice by the ILAE Neurophysiology Task Force (Part 2). Epileptic Disord 19:385–437.

Letts VA, Beyer BJ, Frankel WN (2014) Hidden in plain sight: spike-wave discharges in mouse inbred strains. Genes Brain Behav 13:519–526.

Lüttjohann A (2017) Disclosing hidden information in the electroencephalogram using advanced signal analytical techniques. J Physiol (Lond) 595:7021–7022.

Maheshwari A, Akbar A, Wang M, Marks RL, Yu K, Park S, Foster BL (2017) Persistent aberrant cortical phase-amplitude coupling following seizure treatment in absence epilepsy models. J Physiol (Lond) 10:387–7260.

Mangan KP (2013) GABAergic tonic inhibition, thalamocortical function, and absence epilepsy. PhD dissertation, University of Wisconsin, Madison. 160 p.

Nandan M, Talathi SS, Myers S, Ditto WL, Khargonekar PP, Carney PR (2010) Support vector machines for seizure detection in an animal model of chronic epilepsy. J Neural Eng 7.

Nelder JA, Mead R (1965) A Simplex Method for Function Minimization. Comput J 7:308–313.

468 Nelson AB, Faraguna U, Zoltan JT, Tononi G, Cirelli C (2013) Sleep patterns and
469 homeostatic mechanisms in adolescent mice. Brain Sci 3:318–343.
470 Nolan H, Whelan R, Reilly RB (2010) FASTER: Fully automated statistical thresholding
471 for EEG artifact Rejection. J Neurosci Meth 192:152–162.
472 Ovchinnikov A, Lüttjohann A, Hramov A, van Luijtelaar G (2010) An algorithm for real-
473 time detection of spike-wave discharges in rodents. J Neurosci Meth 194:172–178.
474 Petrou S, Reid CA (2012) The GABAAγ2(R43Q) mouse model of human genetic
475 epilepsy. In: Jespers Basic Mechanisms of the Epilepsies, 4 ed. Jasper's Basic
476 Mechanisms of the Epilepsies. Bethesda (MD): National Center for Biotechnology
477 Information (US).
478 Pinault D, Vergnes M, Marescaux C (2001) Medium-voltage 5-9-Hz oscillations give rise
479 to spike-and-wave discharges in a genetic model of absence epilepsy: In vivo dual
480 extracellular recording of thalamic relay and reticular neurons. Neuroscience
481 105:181–201.
482 Polivannyi F, Igasaki T, Murayama N, Neshige R (2017) Wavelet Transform Based
483 Algorithm for Automatic Detection of Patient- Specific Single Spike-and-Wave
484 Discharges in Simulated Real-Time Conditions. Neuroscience and Biomedical
485 Engineering 4:263–272.
486 Richard CD, Tanenbaum A, Audit B, Arneodo A, Khalil A, Frankel WN (2015)
487 SWDreader: A wavelet-based algorithm using spectral phase to characterize spike-
488 wave morphological variation in genetic models of absence epilepsy. J Neurosci
489 Meth 242:127–140.

Scheuer ML, Bagic A, Wilson SB (2017) Spike detection: Inter-reader agreement and a statistical Turing test on a large data set. Clin Neurophysiol 128:243–250.

Sitnikova E, van Luijtelaar G (2007) Electroencephalographic characterization of spike-wave discharges in cortex and thalamus in WAG/Rij rats. Epilepsia 48:2296–2311.

Sunagawa GA, Séi H, Shimba S, Urade Y, Ueda HR (2013) FASTER: an unsupervised fully automated sleep staging method for mice. Genes Cells 18:502–518.

Tan HO, Reid CA, Single FN, Davies PJ, Chiu C, Murphy S, Clarke AL, Dibbens L, Krestel H, Mulley JC, Jones MV, Seeburg PH, Sakmann B, Berkovic SF, Sprengel R, Petrou S (2007) Reduced cortical inhibition in a mouse model of familial childhood absence epilepsy. Proc Natl Acad Sci USA 104:17536–17541.

Webber WRS, Lesser RP (2017) Automated spike detection in EEG. Clin Neurophysiol 128:241–242.

Wilson SB, Emerson R (2002) Spike detection: a review and comparison of algorithms. Clin Neurophysiol 113:1873–1881.

Xanthopoulos P, Rebennack S, Liu C-C, Zhang J, Holmes GL, Uthman BM, Pardalos PM (2010) A Novel Wavelet Based Algorithm for Spike and Wave Detection in Absence Epilepsy. In, pp 14–19. IEEE.

507 **Legends**

508 ***Figure 1:*** A) An EEG trace from the frontal left channel of a mutant
509 γ2R43Q mouse showing three putative SWD events detected by Phase 1. B) The
510 wavelet transform scalogram of the trace in A. We use the maximum, average and
511 standard deviation of the scalogram amplitude from four bands (~4.4-8.2 Hz, ~8.8-16.4
512 Hz, ~17.6-32.8 Hz, and, ~35.1-65.5 Hz shown as bands with white frequency labels) to
513 generate 12 predictors to train the SVM-based classification algorithm.

514 ***Figure 2:*** Each of the 2500 human-scored events labeled by agreed SWD (red),
515 consensus agreed nonSWD (blue), and consensus disagreed (green) shown as a 2D
516 cross-sections in predictor space for 6 of the 12 predictor variables used to characterize
517 a SWD event. The diagonal shows the histograms for each category of label for each of
518 the 6 predictor variables. Notice how disagreed upon events tend to align with SWD
519 rather than nonSWD events.

520 ***Figure 3:*** A characterization of the variability of human classification of SWD events.
521 Panel A shows each of the 2500 events as labeled by human scorers 1-3. Note the
522 dramatic variability in labeling between each of the scorers. Panel B shows examples of
523 eight events identified as 'putative' SWD events by the first stage of the algorithm. The
524 colored bar below each trace represents the start and end of each event, where color
525 shows the degree of agreement between human scorers. This color scale matches that
526 above Panel C. Panel C shows each of the 2500 putative SWD events presented to
527 each human scorer, plotted by the maximum value of the wavelet transform from ~ 4.4-
528 8.2 Hz versus the maximum value of the wavelet transform from ~ 17.6-32.8 Hz. The
529 numbers mark locations the 8 example events from panel B. Panel D shows the

average ***inter-rater*** agreement for each of the 50 repeated events (RE, large circles) plotted over the same data. Note that the agreement between raters approaches a coin flip (green) near the intersection of the strong SWD and nonSWD events. Panel D E shows the average ***intra-rater*** reliability for the same 50 repeated events as shown in panel D.

Figure 4: SVM-based approach to SWD classification. Panels A-B show each of the 2500 potential SWD events displayed by the maximum value of the wavelet transform from ~ 4.4-8.2 Hz versus the maximum value of the wavelet transform from ~ 17.6-32.8 Hz. Panel A shows events as clustered by K-means clustering using the 12 wavelet transform predictor variables. Panel B shows events as labeled by a SVM trained with the human scoring labels and fit with a Gaussian kernel. The numbers 1-8 are overlaid to show the SVM's classification of the 8 example events from Figure 2. Panel C shows the relationship between the distance to nearest support vector (x axis, 2 is more likely SWD) and the average intra-rater reliability (from Fig. 2C) for the 50 repeated events. Panel D shows the coefficients from a LASSO regression which highlights the relative importance (absolute value of coefficients) and relative magnitude (sign of coefficients) of the variables used to predict the SVM score.

Table 1: The agreement matrix for classification of 2500 potential SWD events as classified by four human scorers (S1-S4), a SVM trained with the KM labels, and the human consensus data (HC, the events on which all human scorers agreed). Note that the agreement between all of the humans (bold, green) was comparable to that between the SVM classification and the humans (italic, blue) although the agreement

552 between the SVM and the humans shows much lower variability than the agreement
553 between the humans alone.
554 ***Figure 5*:** Parameter optimization for the SVM. We used a grid search method to vary
555 box constraint, kernel scale, and the cost function parameters of the SVM. A-F show
556 unique pair-wise combinations of our output parameters: 1) the standard deviation of
557 the 5-fold cross fold validation, 2), the agreement between the humans and the SVM 3)
558 the standard deviation of the agreement between the humans and the SVM, and 4) the
559 R^2 between the SVM classification score and the repeated event data set.
560 ***Figure 6:*** Implementation of the SWD detection algorithm on a 24-hour EEG recording
561 from a mutant γ2R43Q mouse with absence epilepsy and a demonstration of the high-
562 resolution data provided the 'distance to SV' measure associated with each identified
563 event. A) A hypnogram showing WAKE (blue), nREM (green) and REM (magenta)
564 epochs (4 s) manually scored by RM. B) A visual representation of the events identified
565 by the SVM-based algorithm. The vertical height of each bar represents the distance to
566 nearest support vector where values > 0 are likely to be SWDs whereas events < 0 are
567 likely nonSWD events. The color at the end of each vertical bar shows the sleep stage
568 associated with the identified event. The histogram on the right of panel C shows the
569 distribution of event scores and colors are representative of the categories of events
570 used in panels D-K. Panel C shows the average number of SWDs (score > 0) per hour
571 for all sleep stages (red, larger circles), WAKE, nREM, and REM sleep stages. Notice
572 the pattern of increased SWD events during lights on periods (07:00 - 18:00, normal
573 sleep time for mice) as compared to lights off. Panels D-K show the event-triggered
574 average of delta (D, H), theta (E, I), sigma (F, J), and γ (G, K) power (calculated in 4 s

epochs) for events at several sets of distances from the nearest support vector where redder colors are more likely to be SWDs (red: >1 to 2, yellow: >0.5 to 1, lime: 0 to 0.5, green: 0 to -0.5, cyan: <-0.5 to -1, blue: <-1 to -2). Note that the color scale in this figure matches the color scale in panel B of Figure 3. The black trace in each of these panels shows the null-hypothesis: the event-triggered average of each power value for 100 randomly selected epochs in the EEG record. Events with the largest positive distance to SV (red) show a strong relationship to all power bands during lights on (white panels) and lights off (grey panels) and this relationship deteriorates as event distances decrease. Note that all events identified by the algorithm show some sort of temporal relationship with theta and sigma power as compared to randomly selected EEG locations. Further investigation is needed to identify the relationship of each group of these events to absence epilepsy.

Figure 7: Six SWD-like events from a 24-hour record of a wild type, nonepileptic animal as identified by the SWD detection algorithm.

Figure 1:

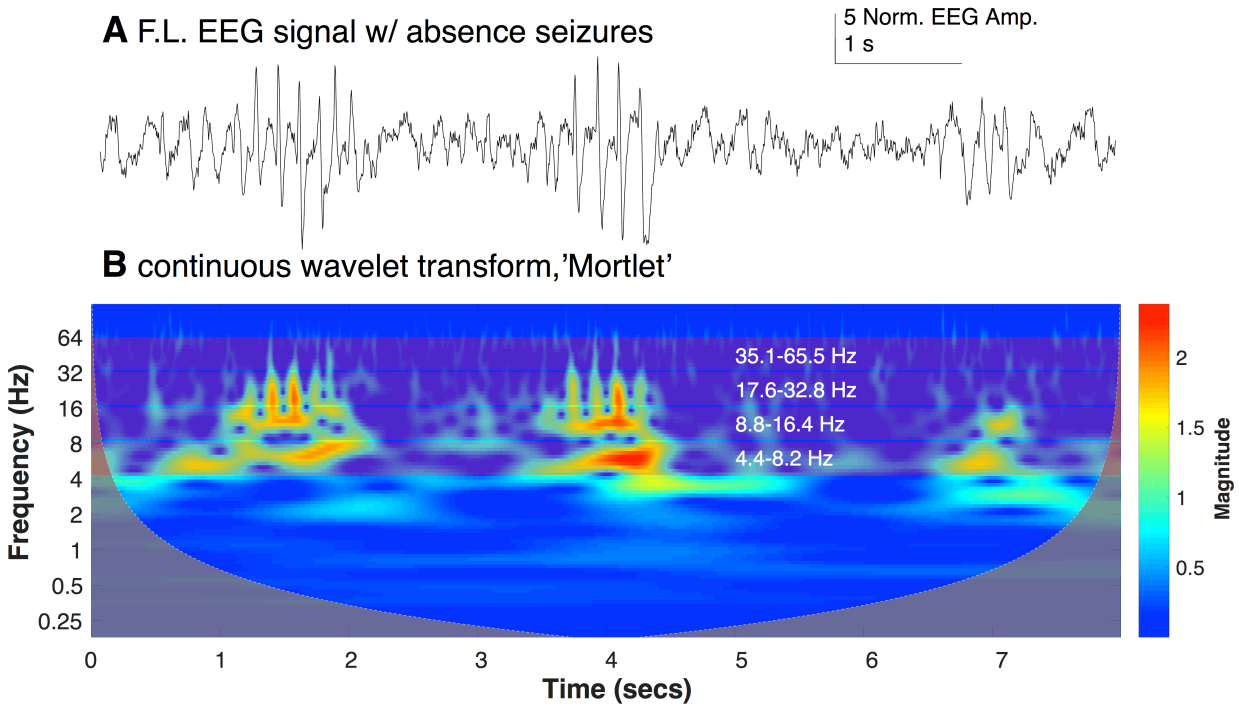

A F.L. EEG signal w/ absence seizures
5 Norm. EEG Amp.
1 s
B continuous wavelet transform,'Mortlet'
35.1-65.5 Hz
17.6-32.8 Hz
8.8-16.4 Hz
4.4-8.2 Hz
Frequency (Hz)
64
32
16
8
4
2
1
0.5
0.25
0
1
2
3
4
5
6
7
Time (secs)
Magnitude
2
1.5
1
0.5

Figure 2:

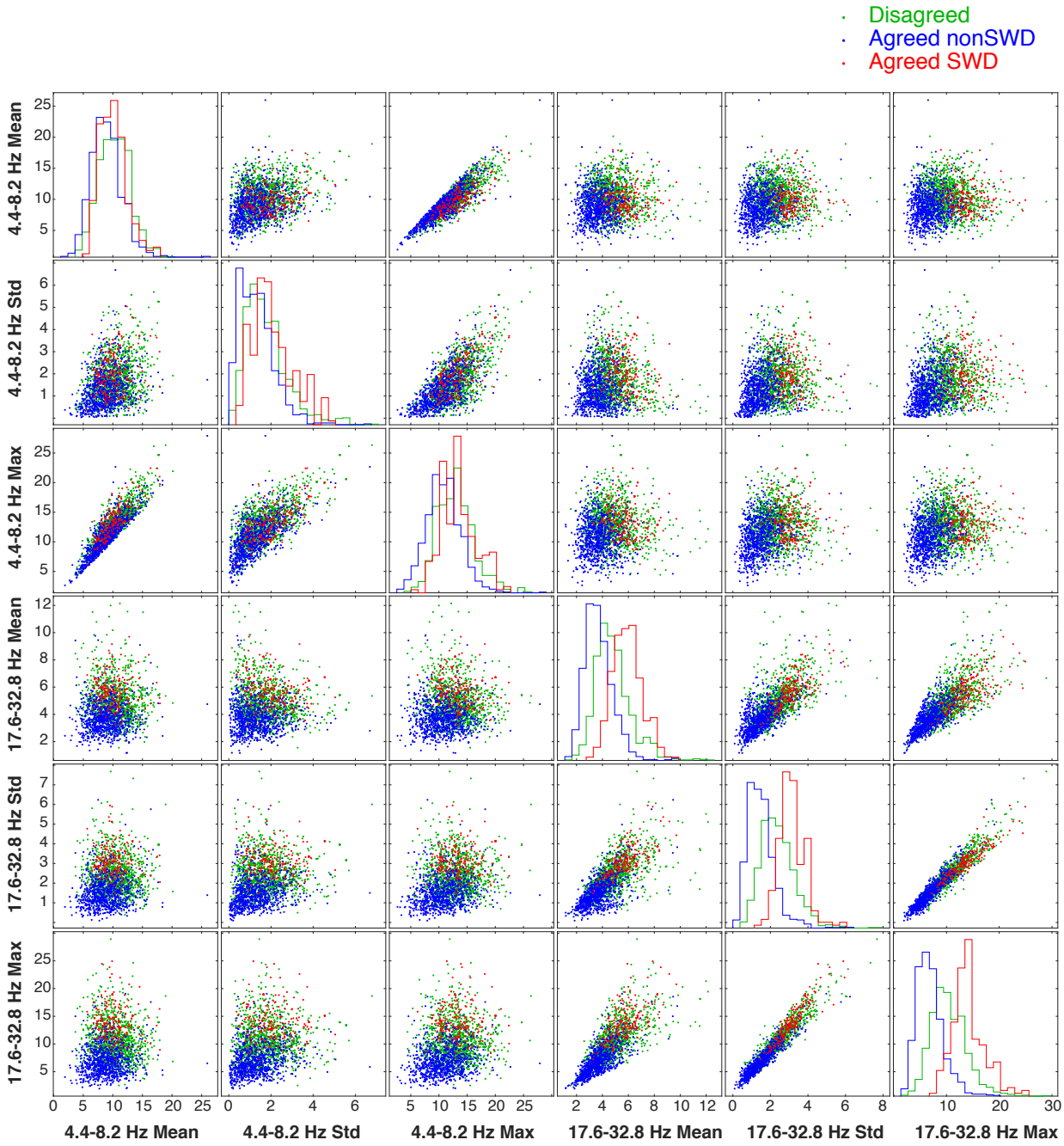

Disagreed
Agreed nonSWD
Agreed SWD
4.4-8.2 Hz Mean
4.4-8.2 Hz Std
4.4-8.2 Hz Max
17.6-32.8 Hz Mean
17.6-32.8 Hz Std
17.6-32.8 Hz Max

Figure 3:

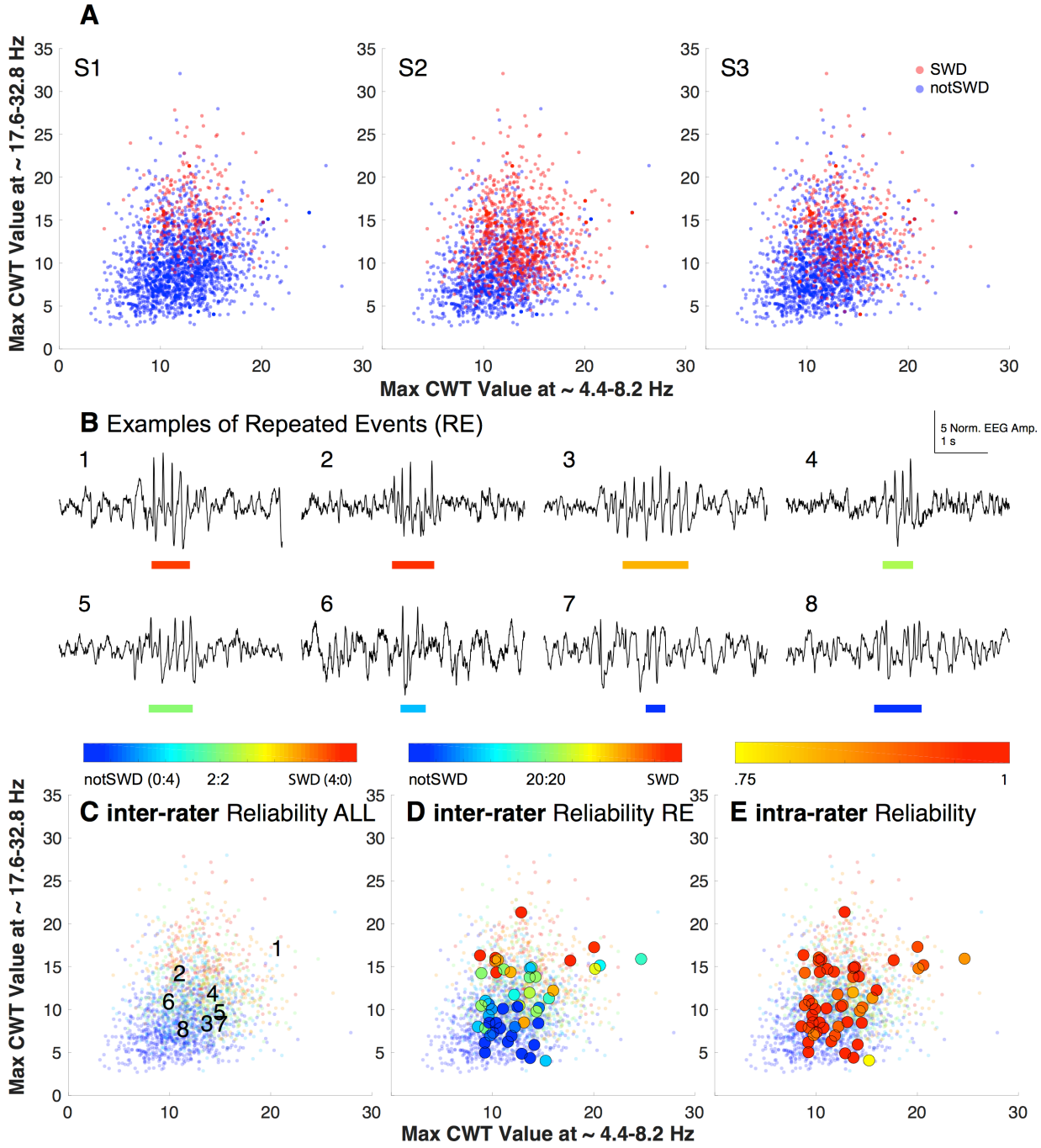

A
S1
S2
S3
SWD
notSWD
Max CWT Value at ~ 17.6-32.8 Hz
Max CWT Value at ~ 4.4-8.2 Hz
B Examples of Repeated Events (RE)
5 Norm. EEG Amp.
1 s
1
2
3
4
5
6
7
8
notSWD (0:4)
2:2
SWD (4:0)
notSWD
20:20
SWD
.75
1
C inter-rater Reliability ALL
D inter-rater Reliability RE
E intra-rater Reliability

Table 1:

	S1	**S2**	**S3**	**S4**	**K-means**	**SVM**	**HC**
S1	1	**70.32**	**79.44**	**54.32**	*66.76*	*72.84*	1
S2		1	**75.68**	**76.88**	*73.72*	*79.36*	1
S3			1	**64.96**	*67.76*	*75.28*	1
S4				1	*75.12*	*74.64*	1
K-means					1	86.88	89.56
SVM						1	95.60
HC							1

Avg. Agreement, Humans: 70.27 ± 9.38

Avg. Agreement, K-means: 70.84 ± 4.19

Avg. Agreement, SVM to Humans: 75.53 ± 2.75

Figure 4:

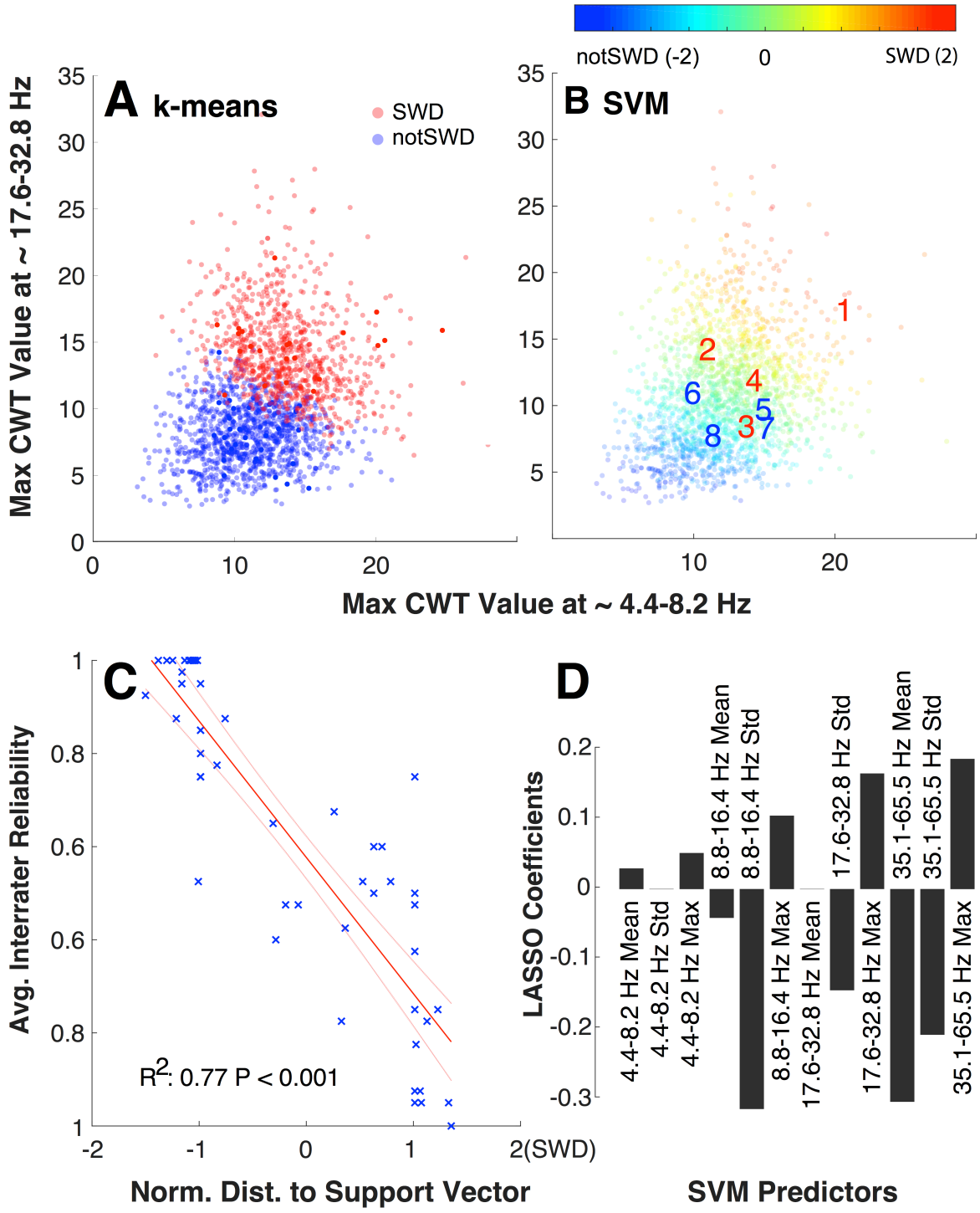

notSWD (-2)
0
SWD (2)
A k-means
SWD
notSWD
B SVM
Max CWT Value at ~ 17.6-32.8 Hz
Max CWT Value at ~ 4.4-8.2 Hz
C
Avg. Interrater Reliability
R2: 0.77 P < 0.001
Norm. Dist. to Support Vector
2(SWD)
D
LASSO Coefficients
4.4-8.2 Hz Mean
4.4-8.2 Hz Std
4.4-8.2 Hz Max
8.8-16.4 Hz Mean
8.8-16.4 Hz Std
8.8-16.4 Hz Max
17.6-32.8 Hz Mean
17.6-32.8 Hz Std
17.6-32.8 Hz Max
35.1-65.5 Hz Mean
35.1-65.5 Hz Std
35.1-65.5 Hz Max
SVM Predictors

Figure 5:

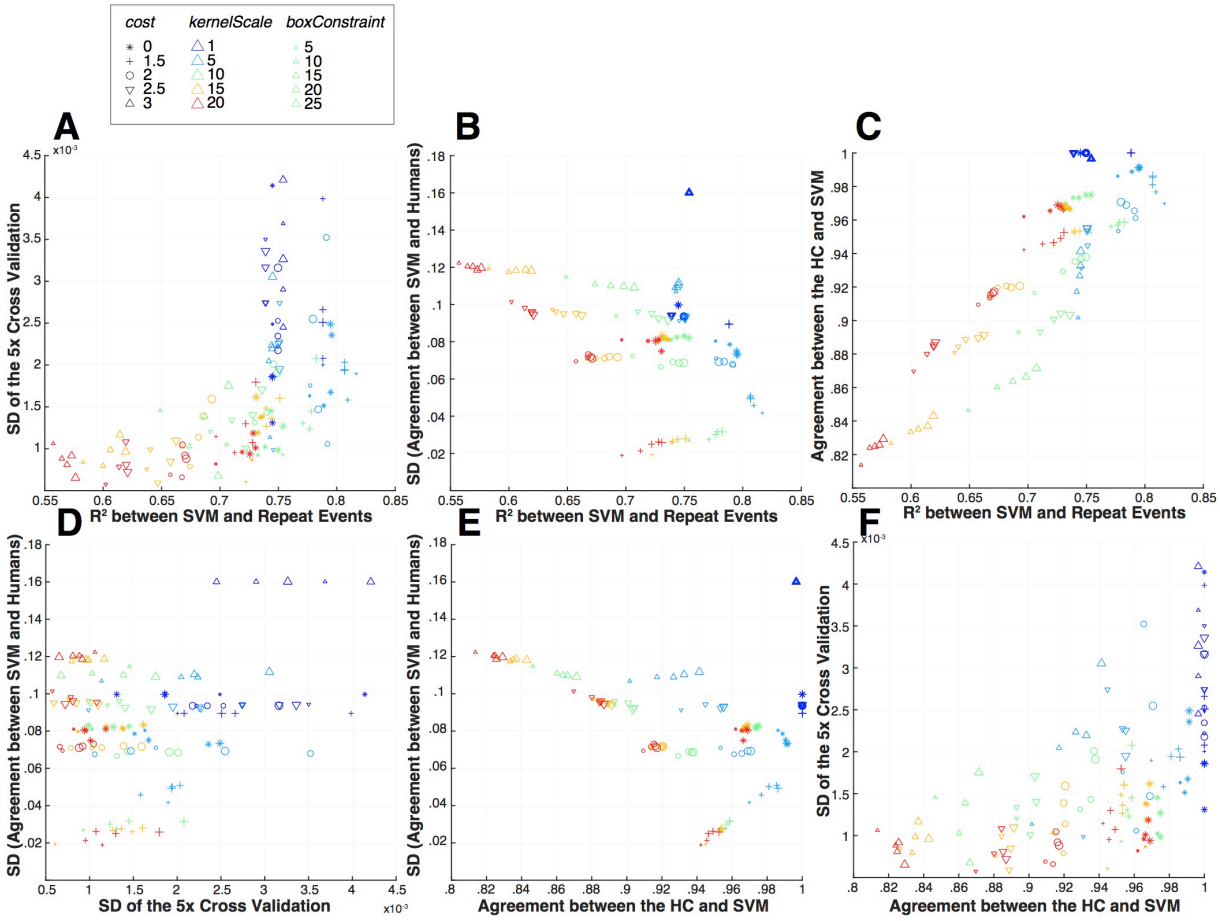
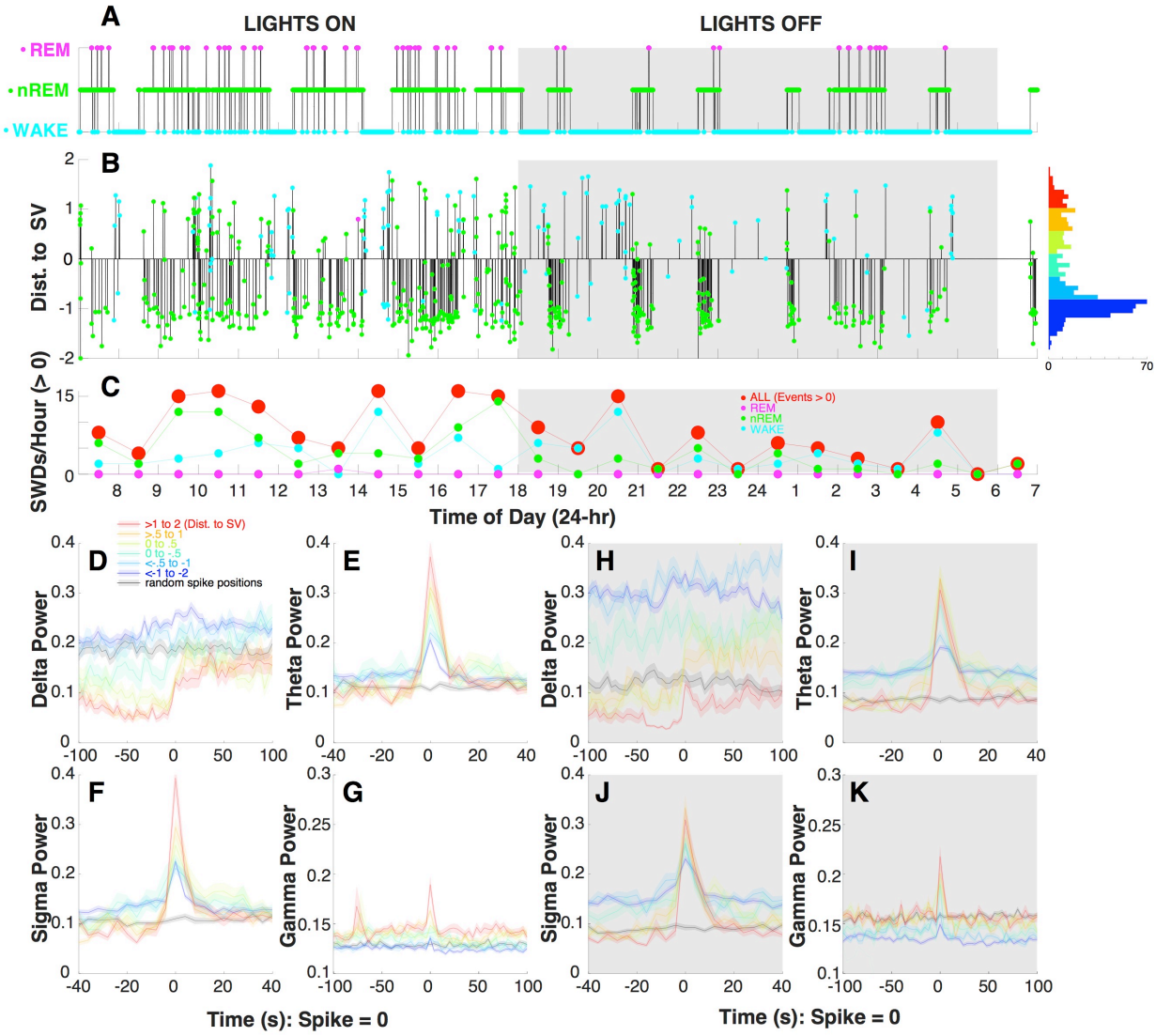
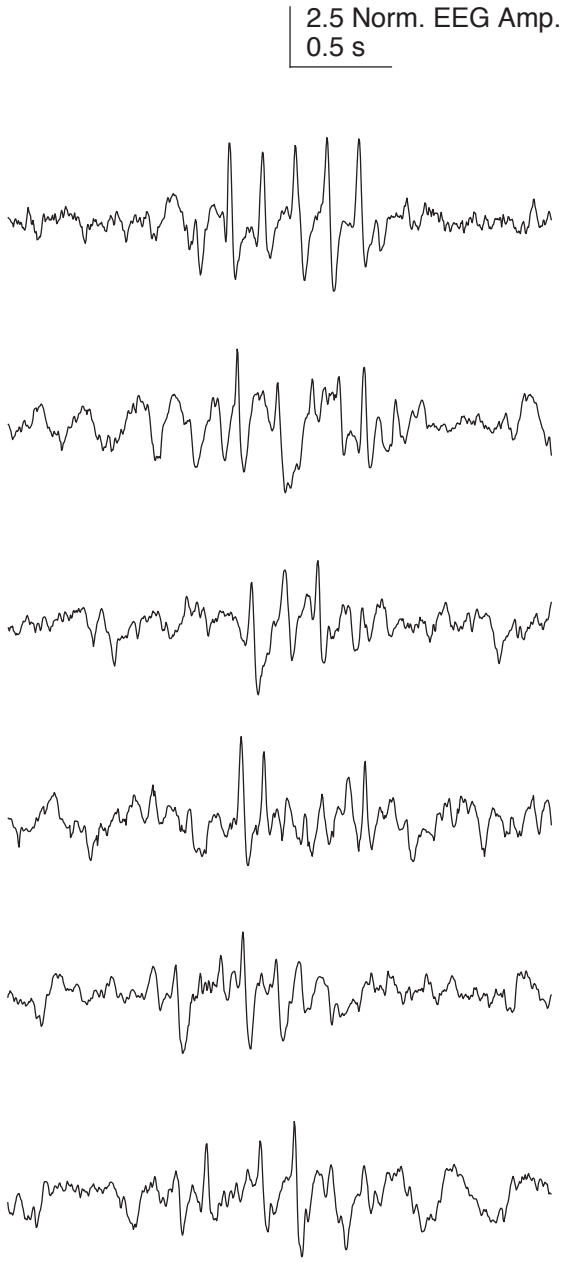

cost
0
1.5
2
2.5
3
kernelScale
1
5
10
15
20
boxConstraint
5
10
15
20
25
A
B
C
D
E
F
SD of the 5x Cross Validation
SD (Agreement between SVM and Humans)
Agreement between the HC and SVM
R² between SVM and Repeat Events
x10⁻³

Figure 6:


A
LIGHTS ON
LIGHTS OFF
REM
nREM
WAKE
B
Dist. to SV
C
SWDs/Hour (> 0)
ALL (Events > 0)
REM
nREM
WAKE
Time of Day (24-hr)
>1 to 2 (Dist. to SV)
>.5 to 1
0 to .5
0 to -.5
<-.5 to -1
<-1 to -2
random spike positions
D
E
H
I
F
G
J
K
Delta Power
Theta Power
Sigma Power
Gamma Power
Time (s): Spike = 0
Time (s): Spike = 0

603 Figure 7:



604