# Classifier for the Rapid Simultaneous Determination of Sleep-Wake States and Seizures in Mice

Brandon J. Harvey[1,2], Viktor J. Olah[3], Lauren M. Aiani[4], Lucie I. Rosenberg[2], and Nigel P. Pedersen[2,4,5*]

[1] Graduate Program in Neuroscience, Emory University, 201 Dowman Drive, Atlanta, GA 30322, USA.

[2] Department of Neurology, University of California, Davis, 1515 Newton Court, Davis, CA 95618, USA.

[3] Department of Cell Biology, Emory University, 615 Michael St., Atlanta, GA 30322, USA.

[4] Department of Genetics, Emory University School of Medicine, 201 Dowman Drive, Atlanta, GA 30322, USA.

[5] Centers for Neuro-Engineering and Medicine, Center for Neuroscience, University of California, Davis, CA 95618, USA.

(*) Authors to whom correspondence should be addressed:

Nigel P. Pedersen, MD
Department of Neurology
Center for Neuroengineering in Medicine,
1515 Newton Court
Davis, CA 95618
Email: nppedersen@ucdavis.edu

**Abstract**

Independent automated scoring of sleep-wake and seizures have recently been achieved; however, the combined scoring of both states has yet to be reported. Mouse models of epilepsy typically demonstrate an abnormal electroencephalographic (EEG) background with significant variability between mice, making combined scoring a more difficult classification problem for manual and automated scoring. Given the extensive EEG variability between epileptic mice, large group sizes are needed for most studies. As large datasets are unwieldy and impractical to score manually, automatic seizure and sleep-wake classification are warranted. To this end, we developed an accurate automated classifier of sleep-wake states, seizures, and the post-ictal state. Our benchmark was a classification accuracy at or above the 93% level of human inter-rater agreement. Given the failure of parametric scoring in the setting of altered baseline EEGs, we adopted a machine-learning approach. We created several multi-layer neural network architectures that were trained on human-scored training data from an extensive repository of continuous recordings of electrocorticogram (ECoG), left and right hippocampal local field potential (HPC-L and HPC-R), and electromyogram (EMG) in the murine intra-amygdala kainic acid model of medial temporal lobe epilepsy. We then compared different network models, finding a bidirectional long short-term memory (BiLSTM) design to show the best performance with validation and test portions of the dataset. The SWISC (sleep-wake and the ictal state classifier) achieved >93% scoring accuracy in all categories for epileptic and non-epileptic mice. Classification performance was principally dependent on hippocampal signals and performed well without EMG. Additionally, performance is within desirable limits for recording montages featuring only ECoG channels, expanding its potential scope. This accurate classifier will allow for rapid combined sleep-wake and seizure scoring in mouse models of epilepsy and other neurologic diseases with varying EEG abnormalities, thereby facilitating rigorous experiments with larger numbers of mice.

**Keywords**

1. Epilepsy
2. Sleep
3. Seizures
4. Machine Learning
5. Sleep-wake

## *1. Introduction*

While the relationship between sleep and seizures has been widely appreciated for centuries (Crespel et al., 1998; Janz, 1962; Shouse and Sterman, 1982), mechanisms remain obscure. As a prelude to rodent studies examining this relationship, we sought a means to comprehensively label (score) sleep, wake, and seizure-related activity in large datasets from continuous chronic mouse electroencephalographic (EEG) recordings.

Several factors hinder the meticulous investigation of sleep in rodent models of epilepsy: Firstly, there are variable changes in the EEG background and prominent interictal EEG abnormalities (Pitkänen et al., 2017). Second, these abnormalities are not static, typically evolving from seizure induction throughout the recording period. Lastly, there is substantial variability in the number, severity (Almeida Silva et al., 2016), and electrophysiological morphology (Henshall et al., 2000) of seizures and interictal findings, necessitating larger cohorts than might be needed for studies of sleep-wake alone. These large datasets often involve thousands of hours of multi-channel recording. Manual scoring of this data for both seizures and sleep-wake is often impractical; a trained expert scorer may take up to 25 minutes to score sleep-wake for 12 hours of data (Kloefkorn et al., 2020), and even longer when EEGs are abnormal. Conventional algorithmic and parametric sleep-wake scoring methods are not applicable as these approaches typically depend on the ratio of EEG power in the theta and delta bands (Costa-Miserachs et al., 2003), and perform poorly with variably abnormal EEGs. Therefore, we considered using a machine learning approach to address this need.

Separate semi-automatic sleep-wake and seizure scoring has been successful. Varying methodologies for sleep scoring include machine learning approaches utilizing Support Vector Machines (SVMs) (Lampert et al. 2015) and machine learning-based approaches. These algorithms include several classifiers driven by convolutional feature extraction. AccuSleep (Barger et al., 2019) uses a linear classification layer, and MC-SleepNet (Yamabe et al., 2019) uses a BiLSTM with a Dense layer for classification. Previously, we used techniques including convolutional feature extraction to score sleep and cataplexy in cataplexic mice (Exarchos et al., 2020). However, these existing classifiers are trained on non-epileptic mice with a non-pathological EEG background. Several effective seizure detection approaches have been published recently, with methodologies including parametric (Tieng et al., 2017), machine-learning (Wei et al., 2020), and deep-learning-based (Jang and Cho, 2019) algorithms. These previous works are also adequate for identifying seizures across various mouse models of epilepsy (Wei et al., 2021). However, like the existing published sleep scoring classifiers, none combine sleep-wake and seizure classification into one tool.

We aimed to create an automated sleep-wake and seizure scoring method that could batch-process files from control mice and mice with varying degrees of epilepsy-related EEG background abnormalities. We focused on machine learning methods that are either theoretically appropriate or empirically suited to classifying EEG time series data. As a benchmark, we choose the classification accuracy seen with human scoring as determined by inter-rater agreement. The agreement between scorers in our laboratory is high (>93%) (Kloefkorn et al., 2020), in accord with other reports of 92% (Rytkönen et al., 2011). Here we describe a highly accurate method for simultaneous automated sleep-wake and seizure classification.

## *2. Methods and Materials*

***2.1. Mice*** - Mice used in the study (11-30 weeks old, N=60) of either sex with a C57BL/6J background were housed and used in accordance with the Emory Institutional Animal Care and Use Committee. Mice in this dataset include transgenic VGAT-CRE C57BL/6J (N=46) strains obtained from Jackson Labs (Stock No. 028862), as well as wild-type C57BL/6J (N=14, Stock No. 000664), bred in our animal facility with the oversight of the Department of Animal Resources. Breeding procedures included backcrossing every five generations. DNA samples were obtained via ear punch before surgery to determine the genotype using polymerase chain reaction per the Jackson Labs protocol. Mice were provided with food and water *ad libitum* and maintained on a 12-hour light-dark cycle (lights on 7 am-7 pm). Cages were changed weekly and in the same session for all mice.

***2.2. Surgery*** - Surgical procedures have been described previously (Zhu et al., 2020). Briefly, mice were induced with ketamine (100 mg/kg i.p.) and xylazine (10 mg/kg i.p.), followed by meloxicam (5 mg/kg s.c.) in 1 cc saline, and anesthesia was maintained with isoflurane (0-1.5%). Four holes are drilled for head plate screws, two for depth electrodes, and one for a guide cannula targeting the basolateral amygdala. Bilateral hippocampal depth electrodes were placed in the perforant path region, immediately overlying the dorsal blade of the dentate gyrus (±2.00 mm ML, -2.53 mm AP, -1.80 mm DV from the brain surface). Screw electrodes for electrocorticography were placed in the frontal (-1.30 mm ML, +1.00 mm AP) and contralateral parietal bones (+2.80 mm ML, -1.50 mm AP), as well as over the midline cerebellum (reference; 0 mm ML, -6.00 mm AP) and opposite frontal bone (ground; +1.30 mm ML, +1.00 mm AP). Finally, the guide cannula (5mm, Plastics1, Roanoke, VA) was implanted with the tip 1.75 mm dorsal to the basolateral amygdala target

(-2.70 or -3.40 mm ML, -1.34 mm AP, -3.80 mm DV from the brain surface). Electromyogram (EMG) paddles (Plastics1, Roanoke, VA) were inserted subcutaneously above the posterior neck muscles, and instrumentation was then secured with cyanoacrylate adhesive and dental cement. The mice were then allowed time to recover from anesthesia and regain their righting reflex before being singly housed in seven-inch diameter clear acrylic recording barrels with food and water available *ad libitum,* nesting material, and a 12-hour light-dark cycle.

***2.3. Mouse Recording*** **-** Mice recovered from surgery for three to four days and were then connected to the tether for two days of habituation before seven days of baseline recording. Video, EEG (electrocorticogram [ECoG] and bilateral hippocampal field potential [HPC-L/R]), and EMG were recorded continuously throughout the experiment. A one-megapixel day-night camera was used for video recording (ELP, Shenzhen, China). Pre-amplifying headsets (8406-SE31 M, 100x gain, Pinnacle Technologies, KS) were used with a commutator (model 8408, Pinnacle Technologies, KS) and analog breakout boxes (8443-PWR, Pinnacle Technologies, KS) with additional gain and digitization (Power 1401, CED, Cambridge, UK). Synchronized video and EEG/EMG files were saved every 12 hours and automatically restarted with Spike2 (v9.10, CED, Cambridge, UK).

***2.4. Kainic Acid Injection*** **-** After continuous baseline recording, either kainic acid (N=51) or normal saline vehicle (N=9) was injected into the basolateral amygdala 5-7 hours after lights on (0.3 μg in 200 nL over 5 minutes) via the internal cannula after removal of the stylet. Video-EEG recordings were continued throughout the KA infusion. Mice were injected with diazepam (5 mg/kg, i.p.) to terminate status epilepticus 40 minutes after the end of the KA infusion. The recording then continued for an additional three weeks.

***2.5. Sleep and Seizure Scoring*** - Scoring was performed by three authors (LR, LA, NP) and then confirmed by those with the most experience (LA, NP). Sleep scoring was performed manually for mouse sleep using modified American Academy of Sleep Medicine (AASM) and Rechtschaffen and Kales criteria (Moser et al., 2009; Rechtschaffen and Kales, 1968). Epochs lasted for 20 seconds and were labeled based on the presence of more than 50% of a state except for the following rules: Rapid-eye-movement sleep (REM) can only follow non-REM (NREM) and, in the case of both NREM and wake, there must be two or more consecutive epochs to score the respective states (see Figure 1). The post-ictal state

is not characteristic of sleep or wake and was characterized by initial behavioral obtundation and post-ictal electrographic suppression. The resumption of behavior and EEG normalization delimited the end of the post-ictal period.

Seizure scoring was performed via visual identification of epochs. The start of a seizure is marked by the emergence of unihemispheric ictal spiking in the hippocampus, which then evolves into high-amplitude and high-frequency spiking that can be detected in both hippocampal channels. See Figure 1 for more detail and examples.

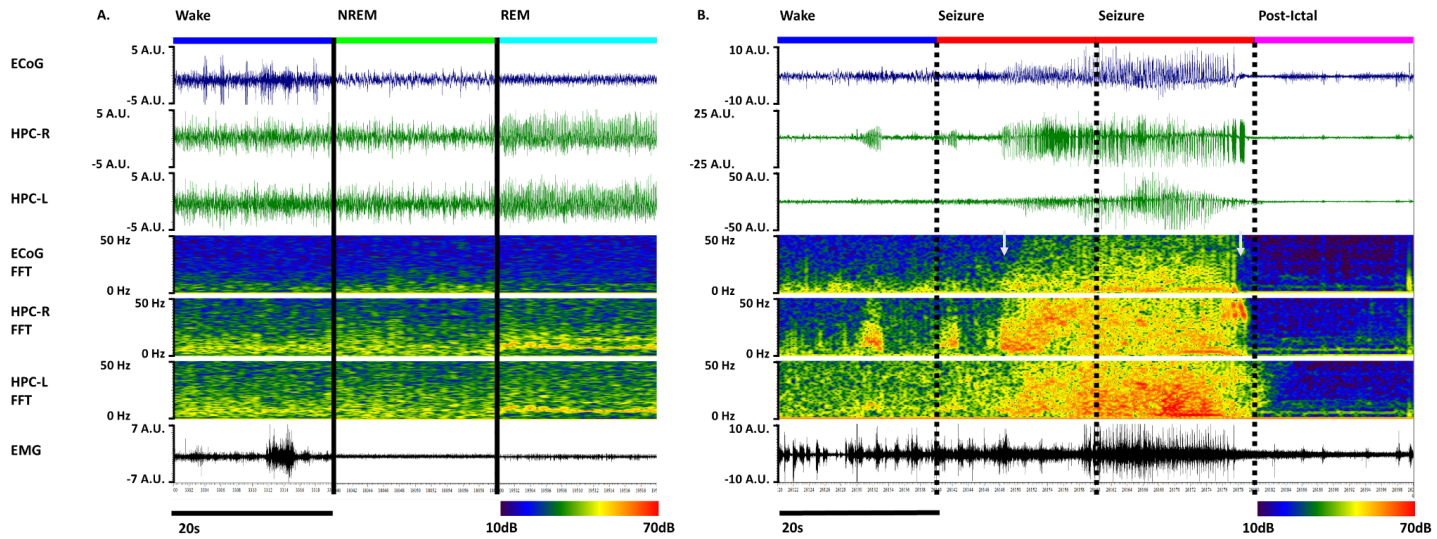



**Figure 1. Scoring of mouse recordings.** A. Sleep-wake categorization on three non-consecutive twenty-second epochs drawn from ECoG, left and right hippocampus (HPC-L and HPC-R), and electromyogram (EMG) channels. Sleep states are discriminated from wake by a reduction in EMG amplitude and a decrease in high-frequency activity in the EEG channels. REM discrimination from NREM is marked by a decrease in delta frequencies and an increase in theta activity, best seen in the hippocampal electrodes. Y-axes are presented in arbitrary units (a.u.) to reflect potential voltage range differences between mouse cohorts. The use of normalization in the pre-processing pipeline addresses concerns about using specified units on inputs. B. Same channel order as A. A spontaneous seizure occurs from wake in the IAKA mouse model, divided into twenty-second consecutive epochs, followed by post-ictal obtundation. A spike train begins in the left hippocampus during the epoch labeled "wake," progresses into a full seizure before the midpoint of the second epoch (left arrow), and ends just before the third epoch ends (right arrow), resulting in a label of "seizure" for the second and third epochs. The fourth epoch shows the suppressed electrographic signal characteristic of a post-ictal state.

***2.6. Preprocessing -*** A custom script exported files for each mouse from Spike2. A Python script then downsampled and extracted features (Figure 2). Files were imported from the .mat format using HDF5Storage (version 0.1.18), then filtered using a first-order Butterworth high-pass filter at 1 Hz, using Scipy (version 1.7.3). Scipy's signal package was then used

to decimate the data from 2 kHz to 200 Hz, including a second-order infinite impulse response anti-aliasing zero-phase filter for ECoG, and an eighth-order version of the same filter for the HPC-L, HPC-R, and EMG channels. Numpy (version 1.21.6), was used to order the data into an array of 20-second epochs and save the data in the .npy format. Feature extraction was then performed via the pyFFTW Fourier Transform library (version 0.13.0 (Frigo, 1999; Gomersall, 2016)) and the Scipy Signal Welch Power Spectral Density function, as described below.

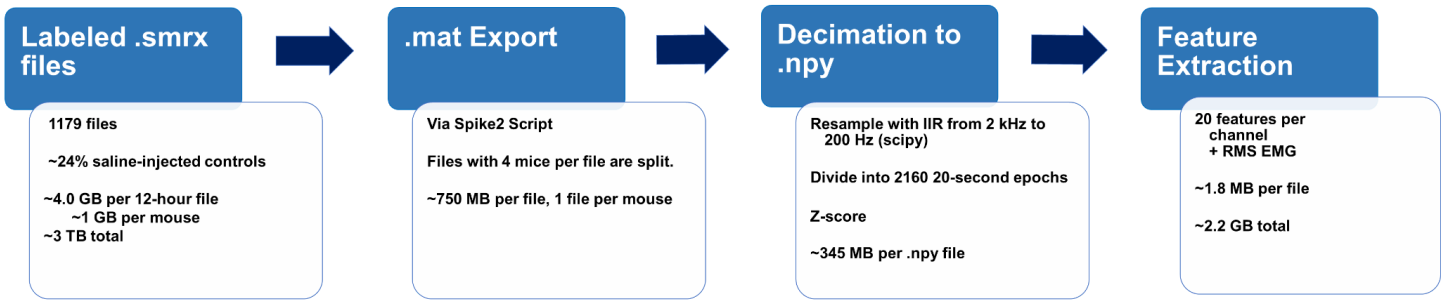



**Figure 2. Data Pre-Preprocessing Pipeline in Python.** Recorded data as in Figure 1 is stored as .smrx, an EEG-specific file format. These files are very large and not easily read into Python, they must first be exported with 20s epochs to MATLAB .mat format via a script built in Spike2, the recording and analysis software. The .mat files are then downsampled by a factor of 10, including a zero-phase IIR anti-aliasing filter implemented in Scipy, then z-scored for normalization and exported to .npy to further save disk space. Finally, feature extraction is performed, described further in Feature Extraction.

***2.7. Feature Extraction -*** Six statistical features for each channel were extracted for each epoch: mean, median, standard deviation, amplitude variance, skewness, and kurtosis. Spectral features consist of the absolute magnitudes of Fast Fourier transforms (FFT) and Welch Power Spectral Density (PSD) at each frequency bin of interest. The six conventional frequency bins analyzed include delta (δ, 2-4 Hz), low theta (θ, 4-7 Hz), high θ (7-13 Hz), beta (β, 13-30 Hz), low gamma (γ, 30-55 Hz), and high γ (65-100 Hz). While line noise was low, given the recording configuration, we excluded 55-65 Hz to ensure this classifier did not have line noise-related problems if used in other settings. A ratio of delta power to low theta power was also calculated for both FFT magnitude and PSD, paralleling a primary feature used for manual scoring (often 'theta delta ratio'). After taking the six statistical features, the Fourier magnitude and the PSD for each frequency bin are computed and normalized to the broadband Fourier magnitude or PSD for the 2-55 Hz broadband. The delta/theta ratio for FFT magnitude and PSD were also computed. In total, 20 features were gathered for each of the four channels.

Finally, 20 one-second root-mean-squared (RMS) amplitude values are calculated from EMG for each epoch and concatenated with the other channel features for 100 features per epoch.

Three groups of features were selected for comparison to validate our features and provide some direct comparison to manual scoring where applicable. The first consisted solely of the four channels' delta/theta ratios in both FFT magnitude and PSD and the full vector of RMS EMG amplitude for the epoch. This closely reproduces the features most important to a sleep-wake expert scorer when scoring manually. This first feature we refer to as "Delta/Theta and RMS" (**DT/RMS**). The second feature set included only the four channels' broadband-normalized (2-55 Hz) Fourier magnitudes in the selected bins, the delta/theta ratio, and the RMS EMG. No epoch-level statistical features nor Power Spectral Density were used for Feature Set 2, referred to as **FFT/RMS**. The final evaluated feature set contained all statistical features, normalized FFT and normalized PSD magnitudes, delta/theta ratios, and RMS EMG components of the full 100-feature vector and is referred to in subsequent text and figures as the **Full** feature set. See Table 1 for a visual summary of tested feature sets and frequency bins of interest.

| For epoch signal E | $f_1 \dots f_n$ | 2 - 4 Hz | 4 - 7 Hz | 7 - 13 Hz | 13 - 30 Hz | 30 - 55 Hz | 65 - 100 Hz | | EMG |
|---|---|---|---|---|---|---|---|---|---|
| Feature Set, # | Statistical | δ | Low-θ | High-θ | β | Low-γ | High-γ | Ratio | |
| # 1<br>"Delta/Theta & RMS"<br>DT/RMS<br>2 features/channel + 20 RMS<br>V = 28 features | None | None | | | | | | $\frac{\lvert\overline{\mathrm{FFT}(\delta)}\rvert}{\lvert\overline{\mathrm{FFT}(\mathrm{Low}-\theta)}\rvert}, \frac{\lvert\overline{\mathrm{PSD}(\delta)}\rvert}{\lvert\overline{\mathrm{PSD}(\mathrm{Low}-\theta)}\rvert}$ | 1 sec RMS Binning |
| # 2<br>"FFT & RMS"<br>FFT/RMS<br>7 features/channel + 20 RMS<br>V = 48 features | None | $\frac{\lvert\overline{\mathrm{FFT}(f_1 \dots f_n)}\rvert}{\lvert\overline{\mathrm{FFT}(2-55\ \mathrm{Hz})}\rvert}$ | $\frac{\lvert\overline{\mathrm{FFT}(f_1 \dots f_n)}\rvert}{\lvert\overline{\mathrm{FFT}(2-55\ \mathrm{Hz})}\rvert}$ | $\frac{\lvert\overline{\mathrm{FFT}(f_1 \dots f_n)}\rvert}{\lvert\overline{\mathrm{FFT}(2-55\ \mathrm{Hz})}\rvert}$ | $\frac{\lvert\overline{\mathrm{FFT}(f_1 \dots f_n)}\rvert}{\lvert\overline{\mathrm{FFT}(2-55\ \mathrm{Hz})}\rvert}$ | $\frac{\lvert\overline{\mathrm{FFT}(f_1 \dots f_n)}\rvert}{\lvert\overline{\mathrm{FFT}(2-55\ \mathrm{Hz})}\rvert}$ | $\frac{\lvert\overline{\mathrm{FFT}(f_1 \dots f_n)}\rvert}{\lvert\overline{\mathrm{FFT}(2-55\ \mathrm{Hz})}\rvert}$ | $\frac{\lvert\overline{\mathrm{FFT}(\delta)}\rvert}{\lvert\overline{\mathrm{FFT}(\mathrm{Low}-\theta)}\rvert}$ | 1 sec RMS Binning |
| # 3<br>"Full feature set"<br>20 features/channel + 20 RMS<br>V = 100 features | $\bar{x}, \tilde{x}, \sigma, \nu$, skewness, kurtosis | $\frac{\lvert\overline{\mathrm{FFT}(f_1 \dots f_n)}\rvert}{\lvert\overline{\mathrm{FFT}(2-55\ \mathrm{Hz})}\rvert}$ & $\frac{\lvert\overline{\mathrm{PSD}(f_1 \dots f_n)}\rvert}{\lvert\overline{\mathrm{PSD}(2-55\ \mathrm{Hz})}\rvert}$ | $\frac{\lvert\overline{\mathrm{FFT}(f_1 \dots f_n)}\rvert}{\lvert\overline{\mathrm{FFT}(2-55\ \mathrm{Hz})}\rvert}$ & $\frac{\lvert\overline{\mathrm{PSD}(f_1 \dots f_n)}\rvert}{\lvert\overline{\mathrm{PSD}(2-55\ \mathrm{Hz})}\rvert}$ | $\frac{\lvert\overline{\mathrm{FFT}(f_1 \dots f_n)}\rvert}{\lvert\overline{\mathrm{FFT}(2-55\ \mathrm{Hz})}\rvert}$ & $\frac{\lvert\overline{\mathrm{PSD}(f_1 \dots f_n)}\rvert}{\lvert\overline{\mathrm{PSD}(2-55\ \mathrm{Hz})}\rvert}$ | $\frac{\lvert\overline{\mathrm{FFT}(f_1 \dots f_n)}\rvert}{\lvert\overline{\mathrm{FFT}(2-55\ \mathrm{Hz})}\rvert}$ & $\frac{\lvert\overline{\mathrm{PSD}(f_1 \dots f_n)}\rvert}{\lvert\overline{\mathrm{PSD}(2-55\ \mathrm{Hz})}\rvert}$ | $\frac{\lvert\overline{\mathrm{FFT}(f_1 \dots f_n)}\rvert}{\lvert\overline{\mathrm{FFT}(2-55\ \mathrm{Hz})}\rvert}$ & $\frac{\lvert\overline{\mathrm{PSD}(f_1 \dots f_n)}\rvert}{\lvert\overline{\mathrm{PSD}(2-55\ \mathrm{Hz})}\rvert}$ | $\frac{\lvert\overline{\mathrm{FFT}(f_1 \dots f_n)}\rvert}{\lvert\overline{\mathrm{FFT}(2-55\ \mathrm{Hz})}\rvert}$ & $\frac{\lvert\overline{\mathrm{PSD}(f_1 \dots f_n)}\rvert}{\lvert\overline{\mathrm{PSD}(2-55\ \mathrm{Hz})}\rvert}$ | $\frac{\lvert\overline{\mathrm{FFT}(\delta)}\rvert}{\lvert\overline{\mathrm{FFT}(\mathrm{Low}-\theta)}\rvert}, \frac{\lvert\overline{\mathrm{PSD}(\delta)}\rvert}{\lvert\overline{\mathrm{PSD}(\mathrm{Low}-\theta)}\rvert}$ | 1 sec RMS Binning |

**Table 1. Feature Set Selection.** Testing the feature space dependence of the various classification regimes was accomplished by selecting three groups of features to compare. The first (**DT/RMS)** consisted solely of the four channels' delta/theta ratios in both FFT magnitude and PSD, as well as the full vector of RMS EMG amplitude. The second feature set (**FFT/RMS)** included the 2-55 Hz broadband-normalized Fourier magnitudes in the selected delta, low-theta, high-theta, beta, low-gamma, and high-gamma frequency bins, the FFT delta/theta ratio, and the RMS EMG. The final evaluated feature set contained all statistical features, broadband-normalized FFT and PSD magnitudes, delta/theta ratios for FFT and PSD, and RMS EMG components for a 100-feature

vector, and is referred to as the **Full** feature space.

***2.8. Machine Learning Packages -*** Machine learning was implemented using the Python libraries TensorFlow (Google, version 2.9.2 (Abadi et al., 2016)) and Keras (version 2.9.0 (Chollet, 2018)), with multiple tests of feature sets and hyperparameters performed to determine the optimal model for sleep-wake and seizure scoring. The class distribution (Data Distribution, section 2.9) contained several imbalanced classes, which were accounted for using a modified version of sklearn's compute_class_weight (version 1.0.2 (Pedregosa et al., 2011)) function. All file conversion and feature extraction was performed on a desktop PC (Intel i7-6950X at 3.00 GHz, 128 GB RAM, Nvidia GTX960 2GB). In contrast, all training of machine learning algorithms was performed on a Google Colab Pro+ instance (Intel Xeon 12-core CPU at 2.20GHz, ~88 GB RAM, Nvidia A100 40GB). The training was repeated on the local machine for timing purposes only.

***2.9. Dataset Composition -*** The dataset, including unscored data, contains 3,770 files (1,885 days) from experimental KA mice and 650 files (325 days) from either pre-treatment baseline recordings or saline-injected controls. Of those, a total of 900 files (450 days) from experimental KA mice and 279 files (139.5 days) from either pre-treatment baseline recordings or saline-injected controls had been manually scored (~ 27% of the dataset). Data containing either uninterpretable recording errors (i.e., values that are "not a number" [NaNs] after computation) or text-encoded markers containing sleep scores that were not part of the target scoring states (such as markers for scoring on which experts disagreed) were not used to ensure error-free computation and data sharing.

***2.10. Data Distribution -*** Approximately two-thirds of the scored recordings were used for training, with a final train/test/validation split of 61%/12%/27%. This data split was effective for training with manually-scored data while reliably producing an accurate classification. Sleep state distribution (wake, NREM, REM) was consistent in IAKA animals, whether from the VGAT-Cre line [variance of wake: 3.57%, NREM: 2.52%, REM: 0.116%, seizure and post-ictal: <.0001%] or wild-type background [variance of wake: 4.71%, NREM: 3.41%, REM: 0.125%, seizure and post-ictal: <.0001%]. Saline-treated animals had state distributions consistent with existing large-scale datasets in the literature (Yamabe et al., 2019), regardless of background, as well as low variance per file [VGAT-Cre variance of wake: 2.62%, NREM: 1.83%, REM: 0.093%; wild-type variance of wake: 4.38%, NREM: 3.071%, REM: 0.119%]. Data used

for training (n=45) contained animals of all types, with ~80% being VGAT-Cre KA animals (n=36), ~15.55% being wild-type KA animals (n=7), and ~4.5% being VGAT-Cre saline control animals (n=2). The test dataset contained 75% VGAT-Cre saline control animals (n=3) and 25% wild-type saline control animals (n=1) to properly emphasize our less-represented cases on the loss function optimization. The validation dataset contained only animals from our most recent three cohorts (n=11), of which ~45.45% were VGAT-Cre KA (n=5), ~27.27% were wild-type KA (n=3), and ~27.27% were wild-type saline controls (n=3). In building this additional validation dataset from these recent cohorts and evaluating the classifier's performance on it, in addition to the test dataset, we can ensure that the classifier covers all experimental conditions, VGAT-Cre, and wild-type backgrounds.

## *3. Results*

***3.1. Support Vector Machine (SVM) Performance -*** SVMs are similar to a common approach to human scoring where discriminations are made principally on theta-delta ratio and EMG activity. We hypothesize this would not perform well for epileptic mice, despite reasonable classification for controls. The SVM, trained with an iteration count of 1000, was trained and validated against all three feature sets.

This testing found that the **DT/RMS** feature set, using a linear discriminator, produced inferior classification precision (or positive predictive value), very poor recall (or sensitivity), and a low F1 score (the harmonic mean of precision and recall; a metric that takes into account true positives, false positives, and false negatives to assess overall accuracy) for REM as compared to both NREM and wake in the mixed-treatment validation dataset. In addition, seizure state precision, recall, and F1 scores were near-zero [Precision: 0.0417, Recall: 0.00926, F1: 0.0152], and all post-ictal classification measures were 0. The **FFT/RMS** feature set improved classification for the three sleep states substantially in the validation dataset. However, classification was still undesirable for REM, seizure, and post-ictal states. Finally, the **Full** feature set, when tested against the validation dataset, performed quite well for wake [F1: 0.940] and NREM [F1: .902], but still achieved below human accuracies for REM [F1: 0.786], seizure [F1: 0.856], and post-ictal [F1 0.417] classes. This confirms that the SVM, while useful, can potentially be improved upon by other models. Reporting of precision, recall, and F1 score results from SVM trials with all three feature sets are included in Supplemental Table 1. The rankings of SVM performance F1 score results are summarized in Table 2 and the accompanying legend, alongside other models described below.

***3.2. Model Architectures -*** We next sought to evaluate other classifier architectures. The model architecture space for testing machine learning implementations included a variable-size Dense layer with by 5-node Dense layer classifier, as this is a widely-used and relatively straightforward architecture. Given the need to classify time-series data where epochs are not independent of prior and future epochs, we next tried networks containing either LSTM or BiLSTM recurrent neural network layers. BiLSTM layers can use forward- and backward-looking recurrent prediction to evaluate inputs, whereas LSTM layers are unidirectional. We hypothesized that a classifier that does not treat epochs as independent would perform better and may implicitly incorporate Rechtschaffen and Kales criteria. The LSTM architecture variants tested were a variable-size long-short-term-memory (LSTM) layer with a 5-node Dense layer classifier, a variable-size BiLSTM with a Dense layer classifier, and a stacked variant of the BiLSTM with a Dense layer classifier, with each of four BiLSTM layers halving in size as they progress.

These models were tested against one another using a grid search technique, where training is executed using each possible combination of various manually specified parameters to tune the model's inputs and features. The first feature tested in our grid search of models was the three feature vectors. The second feature tested in this grid search was the layer size of the initial layer, at variants of 50, 100, and 200 nodes. For testing purposes, the same layer size input was used in LSTM as in BiLSTM, resulting in the BiLSTM layer of these architectures consisting of [2 x layer size] input nodes where layer size is denoted in figures and tables. Our third feature evaluated was input sequence length, where the epoch of interest was padded with several preceding and following epochs. The assessed variants of input sequence length were 1, 3, 5, and 7. The windowing protocol for creating input arrays of varying sequence lengths is presented graphically in Figure 3 and described in the accompanying legend.

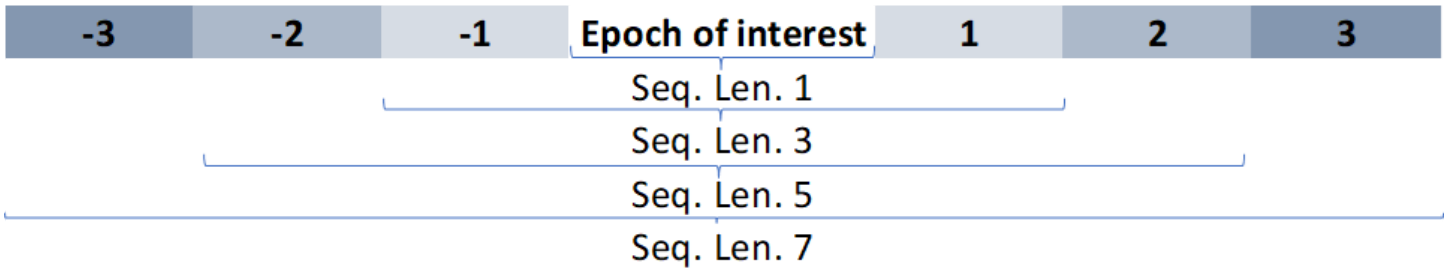


**Figure 3. Windowing Protocol.** The windowing protocol for inputs is visually represented above, where Sequence Length 1 in subsequent tables refers to the input of solely the epoch to be scored. Each additional sequence of a given length is centered on a given epoch, with an equal number of preceding and following epoch feature vectors padding it.

The final grid search consisted of a combinatorial search of 3 feature vectors, four machine learning models, three-layer sizes, and four input sequence lengths. With the addition of the three feature vectors tested for the SVM evaluation, the total number of models evaluated for this problem was 147. To reduce computational time, the initial training for the first round of evaluation was performed out to 20 epochs before classification matrices and reports were saved. Examples of layer architectures, with all assessed variables represented in the architecture diagrams, are presented in Figure 4 and the accompanying legend.

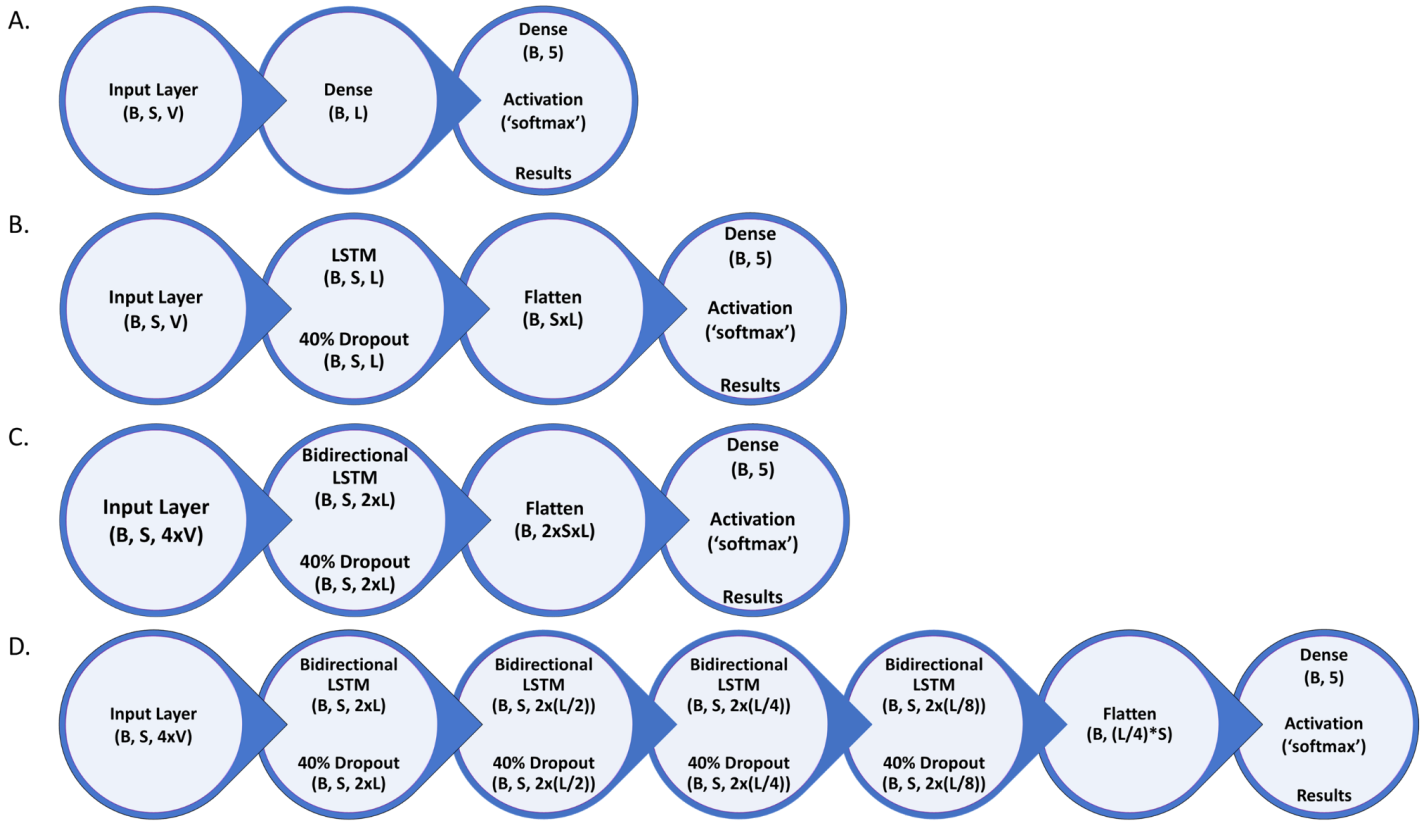



**Figure 4. Layer Architecture Search Space.** Graphical representation of the layer architectures tested and the parameters applied. Variables tested in the grid search are represented here with single letters. B represents the batch size of 2160, the length in 20-second epochs of one full 12-hour electrographic recording. S represents the variable sequence length of epochs in the input vector, as described in Figure 3. V represents the variable length of the input vector itself, as described in Table 1**.** L represents the variable layer sizes of 50, 100, or 200.

***3.3. Classification Metrics*** - The distribution of classification metrics varied substantially by class. Focusing on F1 Score as a well-rounded metric described previously, we evaluated the F1 scores for all classifiers across our classes to determine the best-performing classifiers tested in our grid search. F1 means and standard deviations of class performance across all models were high for wake and NREM but lower than desired for REM, seizure, and post-ictal states, showing a relative lack of performance on the classification of these states for many of our tested model variants (Overall mean and SD of F1 of wake [.928 ± .040], NREM [.872 ± .074], REM [.631 ± .236], Seizure [.244 ± .307], and post-ictal [.286 ± .227]). See Figure 5 for the plotted distribution of class-specific precision, recall, and F1 scores across all tested classifiers.

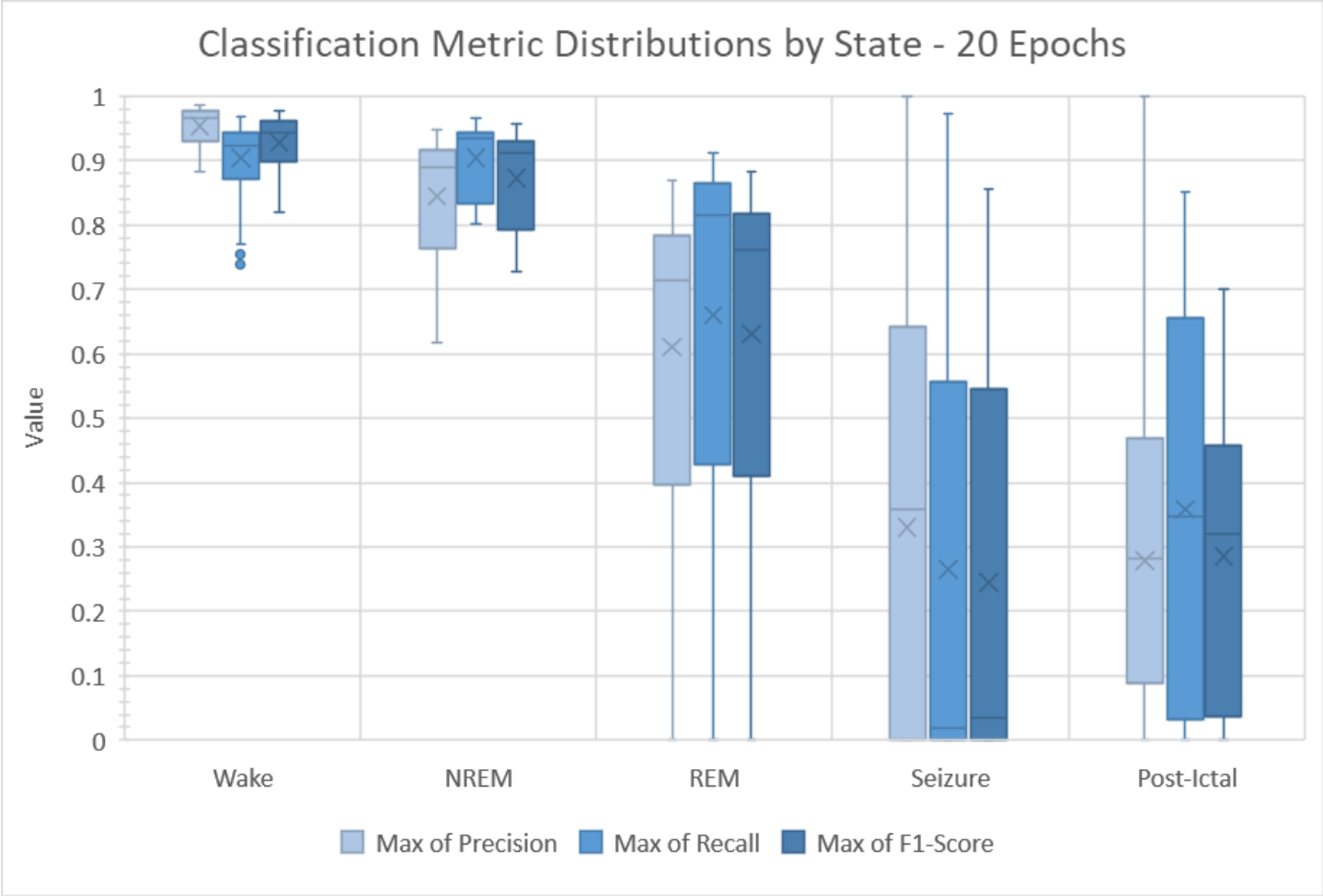



**Figure 5 - Classification Metric Distributions by State at 20 Epochs.** Box-and-whisker plotting of all classification metrics for all models evaluated at 20 epochs of training showed tight ranges of classification of wake and NREM across most classifiers but large ranges of classification F1 scores for REM, seizure, and post-ictal states, with some failing to classify entirely. Precision reaches one

on seizure and post-ictal states in some models due to a lack of false positives on some non-seizure-containing test sets (detail can be found in Supplemental Table 1) but is not representative of the true maximum precision of validation datasets.

Our top two most well-rounded classifiers across all three classification metrics, and all classes, were trained with a layer size of 200 on the **Full** feature space at a sequence length of 7. The first of the top classifiers with these parameters, the BiLSTM, achieved higher F1 scores than the SVM on wake [0.977], NREM [0.956], REM [0.883], and post-ictal [0.674] states, though there was a slight dip when compared on seizure F1 score [0.782]. The runner-up, the LSTM, performed similarly [F1 scores: wake [0.975], NREM [0.954], REM [0.878], seizure [0.785], post-ictal [0.690]].

As mentioned in the SVM subsection, the full results of the above grid search are summarized and ranked in Table 2 and the accompanying legend. Complete breakdowns of all metrics for all models tested can be found in Supplemental Table 1.

| Max of F1-Score | | | Class | | | | | |
|---|---|---|---|---|---|---|---|---|
| Variables | Layer Size | Classifier | Wake | NREM | REM | Seizure | Post Ictal | Weighted Average |
| Full | 200 | **Single-Layer BiLSTM, Sequence Length 7** | 0.9767 | 0.9562 | 0.8827 | 0.7819 | 0.6741 | 0.9634 |
| | | **Single-Layer LSTM, Sequence Length 7** | 0.9749 | 0.9535 | 0.8777 | 0.7950 | 0.6901 | 0.9612 |
| | | Single-Layer BiLSTM, Sequence Length 5 | 0.9743 | 0.9535 | 0.8796 | 0.7722 | 0.6553 | 0.9609 |
| | 100 | Single-Layer BiLSTM, Sequence Length 7 | 0.9738 | 0.9521 | 0.8707 | 0.7561 | 0.7011 | 0.9596 |
| | | Single-Layer BiLSTM, Sequence Length 5 | 0.9710 | 0.9492 | 0.8677 | 0.8205 | 0.6453 | 0.9567 |
| | | Single-Layer LSTM, Sequence Length 7 | 0.9713 | 0.9481 | 0.8618 | 0.6008 | 0.6281 | 0.9561 |
| | 50 | Single-Layer BiLSTM, Sequence Length 7 | 0.9700 | 0.9463 | 0.8516 | 0.5556 | 0.5501 | 0.9540 |
| | | Single-Layer BiLSTM, Sequence Length 5 | 0.9673 | 0.9425 | 0.8490 | 0.6271 | 0.5603 | 0.9510 |
| | | Quad-Layer BiLSTM, Sequence Length 7 | 0.9672 | 0.9380 | 0.8166 | 0.2054 | 0.5324 | 0.9472 |
| | | SVM | 0.9403 | 0.9029 | 0.7861 | 0.8560 | 0.4178 | 0.9172 |
| FFT/RMS | 200 | Single-Layer BiLSTM, Sequence Length 7 | 0.9681 | 0.9390 | 0.8216 | 0.4497 | 0.4567 | 0.9484 |
| | | Single-Layer LSTM, Sequence Length 7 | 0.9656 | 0.9351 | 0.8167 | 0.2920 | 0.4622 | 0.9452 |
| | | Single-Layer BiLSTM, Sequence Length 5 | 0.9644 | 0.9339 | 0.8178 | 0.4231 | 0.4378 | 0.9442 |
| | 100 | Single-Layer BiLSTM, Sequence Length 7 | 0.9654 | 0.9336 | 0.8087 | 0.2537 | 0.4387 | 0.9441 |
| | | Single-Layer BiLSTM, Sequence Length 5 | 0.9617 | 0.9294 | 0.8066 | 0.2576 | 0.4133 | 0.9403 |
| | | Single-Layer LSTM, Sequence Length 7 | 0.9625 | 0.9292 | 0.7988 | 0.0342 | 0.4496 | 0.9402 |
| | 50 | Single-Layer BiLSTM, Sequence Length 7 | 0.9622 | 0.9280 | 0.7924 | 0.0820 | 0.3805 | 0.9391 |
| | | Single-Layer BiLSTM, Sequence Length 5 | 0.9587 | 0.9240 | 0.7861 | 0.0364 | 0.3566 | 0.9353 |
| | | Single-Layer LSTM, Sequence Length 7 | 0.9594 | 0.9229 | 0.7818 | 0.0000 | 0.3888 | 0.9351 |
| | | SVM | 0.9281 | 0.8826 | 0.6751 | 0.4943 | 0.3226 | 0.8961 |
| DT/RMS | 50 | Single-Layer BiLSTM, Sequence Length 7 | 0.9065 | 0.8051 | 0.4368 | 0.0000 | 0.1399 | 0.8411 |
| | | Single-Layer LSTM, Sequence Length 7 | 0.9046 | 0.8000 | 0.4219 | 0.0000 | 0.0617 | 0.8372 |
| | | Quad-Layer BiLSTM, Sequence Length 7 | 0.9051 | 0.7954 | 0.4292 | 0.0000 | 0.0349 | 0.8362 |
| | 200 | Single-Layer BiLSTM, Sequence Length 7 | 0.9058 | 0.8058 | 0.4437 | 0.0000 | 0.1569 | 0.8414 |
| | | Single-Layer LSTM, Sequence Length 7 | 0.9054 | 0.8055 | 0.4424 | 0.0000 | 0.1452 | 0.8409 |
| | | Quad-Layer BiLSTM, Sequence Length 7 | 0.9055 | 0.8016 | 0.4427 | 0.0175 | 0.1953 | 0.8396 |
| | 100 | Single-Layer BiLSTM, Sequence Length 7 | 0.9055 | 0.8045 | 0.4329 | 0.0000 | 0.1405 | 0.8401 |
| | | Single-Layer LSTM, Sequence Length 7 | 0.9046 | 0.8041 | 0.4335 | 0.0000 | 0.1634 | 0.8395 |
| | | Quad-Layer BiLSTM, Sequence Length 7 | 0.9044 | 0.8009 | 0.4362 | 0.0000 | 0.1361 | 0.8383 |
| | | SVM | 0.8279 | 0.7351 | 0.2181 | 0.0152 | 0.0000 | 0.7573 |

| Color Key | 0 | 0.25 | 0.5 | 0.75 | 1 |
|---|---|---|---|---|---|

**Table 2. Parameter Space Search: F1 Score Results.** The F1-score ranking of the Validation dataset results from a 20-epoch gd search of models across the full parameter space found excellent candidates. The three best-performing models were the 200-node layer size Single-Layer BiLSTM with Sequence Length 7, the 200-node layer size Single-Layer LSTM with the same sequence length, and the 200-node layer size Single-Layer BiLSTM with Sequence Length 5. Though the latter performed slightly more effectively on seizure scoring than the former models in F1 score, the difference in effective scoring, as reported by confusion matrices, was negligible.

Given the success of the **Full** feature space at a 7-sequence window length for the two primary LSTM variants, all three of the LSTM variants for this set of parameters were trained for an additional 40 epochs, or until early stopping criteria of a lack of .001 change in loss function were not met in 5 training epochs. The F1 score metrics achieved by these models were used for the final evaluation. F1 scores for all training, testing, and validation sets across all states for these three models are ranked and presented in Table 3 and the accompanying legend**.**

Our grid search of the four-variable parameter space found that the 7-sequence-length input vector using the **Full** feature space, input to a classifier with an architecture consisting of a 200-node layer size, single BiLSTM layer with a 40% Dropout, and a 5-node Dense classifier (Figure 6**)** was the most effective for classification of sleep as well as seizure and post-ictal state.

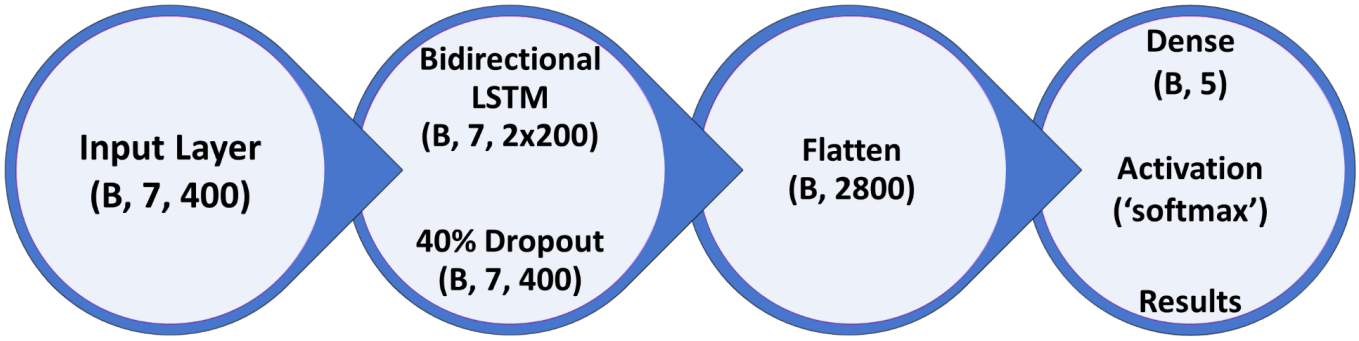



**Figure 6. Final Model Architecture.** The classification section of the model consists of an input layer for epoch sequences, followed by a BiLSTM layer of 200 units in each direction. Generalization of the classifier is improved by using an activity regularization function with an L1 of .0001 on the BiLSTM layer, then a 40% Dropout layer. The output of this section is then flattened, and classification is performed by a 5-unit Dense layer with softmax activation.

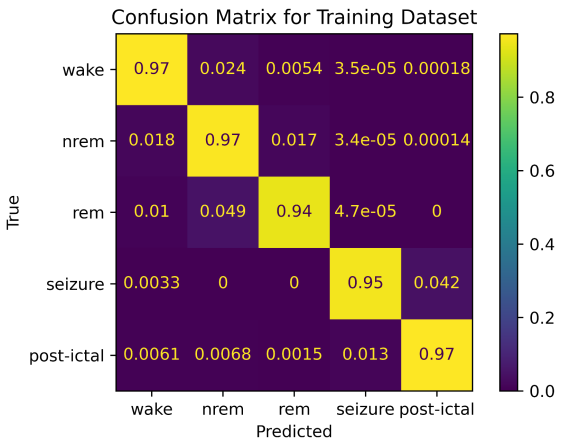
F1 scores for this final classifier for our validation dataset were [wake: 0.978, NREM: 0.960, REM: .890, seizure: 0.858, and post-ictal: 0.747]. The weighted average F1 score for this model, which weights the F1 score by the occurrence of each class and produces an average, is 96.58%. This data is presented graphically alongside the other two fully-trained LSTM variants in Table 3.

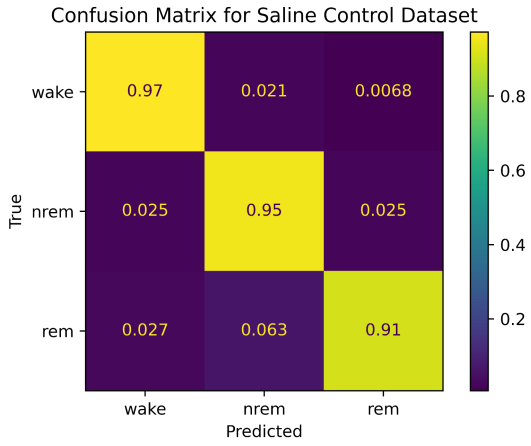
| Max of F1-Score | | | | Class | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Variables** | **Layer Size** | **Classifier** | **Dataset** | **Wake** | **NREM** | **REM** | **Seizure** | **Post Ictal** | **Accuracy** | **Macro Average** | **Weighted Average** |
| **Full** | **200** | **Single-Layer BiLSTM, Sequence Length 7** | Test | 0.9724 | 0.9582 | 0.8638 | | | 0.9577 | 0.9315 | 0.9581 |
| | | | Training | 0.9773 | 0.9635 | 0.9026 | 0.8716 | 0.9010 | 0.9662 | 0.9232 | 0.9665 |
| | | | Validation | 0.9782 | 0.9592 | 0.8896 | 0.8584 | 0.7469 | 0.9657 | 0.8864 | 0.9658 |
| | | Quad-Layer BiLSTM, Sequence Length 7 | Test | 0.9718 | 0.9558 | 0.8554 | | | 0.9556 | 0.9277 | 0.9562 |
| | | | Training | 0.9770 | 0.9626 | 0.8994 | 0.7877 | 0.9033 | 0.9654 | 0.9060 | 0.9657 |
| | | | Validation | 0.9781 | 0.9580 | 0.8853 | 0.6853 | 0.7505 | 0.9648 | 0.8514 | 0.9651 |
| | | Single-Layer LSTM, Sequence Length 7 | Test | 0.9725 | 0.9576 | 0.8598 | | | 0.9572 | 0.9299 | 0.9576 |
| | | | Training | 0.9758 | 0.9612 | 0.8970 | 0.8140 | 0.8937 | 0.9641 | 0.9084 | 0.9644 |
| | | | Validation | 0.9780 | 0.9585 | 0.8905 | 0.7661 | 0.7295 | 0.9653 | 0.8645 | 0.9655 |

| Color Key | 0 | 0.25 | 0.5 | 0.75 | 1 |
|---|---|---|---|---|---|

**Table 3. Top Candidate Evaluation: F1 Score Results.** The F1 scores reported here are the fully-trained top-ranked models from Table 2, now trained for either 60 epochs or until early stopping criteria were met. The results of this further training emphasized the conclusions from Table 2 that the 200-node layer size Single-Layer BiLSTM using the **Full** feature vector for input performed the best on our classification problem for all classes on which it was trained.

Confusion matrices produced for all train/test/validation sets and all classes for the fully-trained BiLSTM are presented in Figure 7**,** showing the true classification of wake, NREM, REM, and seizure above 91%, with variation in the post-ictal state accounted for its more qualitatively defined nature described earlier in this manuscript.



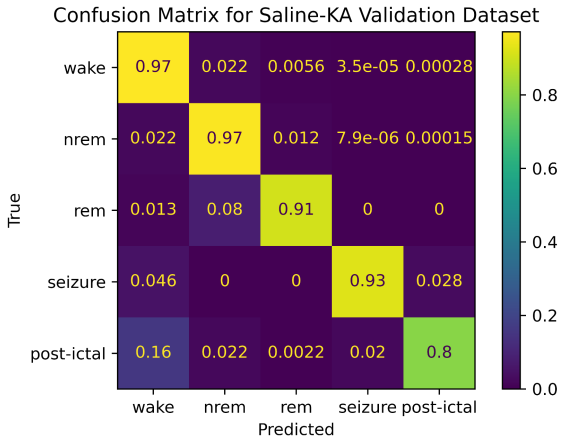



**Figure 7. Performance after 60 Training Epochs.** With the 2x200-node initial layer size, the training and validation datasets were scored accurately compared to expert scoring (low 90%), mirroring its performance on the training and control datasets without overfitting.

To better understand the contribution of the individual electrophysiological channels to the scoring accuracy of the chosen architecture, we systematically removed individual channels and then trained new instances of the winning BiLSTM model out to 60 epochs or early stopping, to properly evaluate these instances' performance against the full dataset.

Scoring based on hippocampal channels alone was highly effective. Training this architecture with solely our bilateral hippocampal channels, using all 20 statistical and spectral features present for each channel, produced F1 scores on our validation dataset of [wake: 0.954, NREM: 0.933, REM: 0.832, seizure: 0.790, post-ictal: 0.750]. Even with half of the original classifier's inputs, robust classification is possible using this architecture. Adding EMG statistical and spectral features and RMS EMG to this montage produced similar F1 scores [wake: 0.974, NREM: 0.954, REM: 0.881, seizure: 0.810, post-ictal: 0.757].

Scoring based on the input of the ECog channel alone retained useful classification. F1 scores in the validation dataset of this variant still reached usable levels [wake: 0.961, NREM: 0.932, REM: 0.778, seizure: 0.826, post-ictal: 0.655]. The saline test dataset in this variant achieved sleep classification F1 scores of [wake: 0.908, NREM: 0.888, REM: 0.739], showing that expansion of this classifier to sleep studies in much simpler montages in non-epileptic mice is achievable, broadening the potential reach of this classifier even further.

As expected, scoring on EMG alone poorly discriminated forebrain-related states. This version of the classifier still achieved high F1 scores for wake [0.944] and NREM [0.857] but faltered for REM [0.512] and failed for seizure [0.107] and post-ictal states [0.169]. Thus, EMG could crudely classify sleep and wake states but displayed poor performance with REM, seizure, and the post-ictal state. Confusion matrices for the validation dataset for all of these interpretable machine-learning variants are displayed in Figure 8.

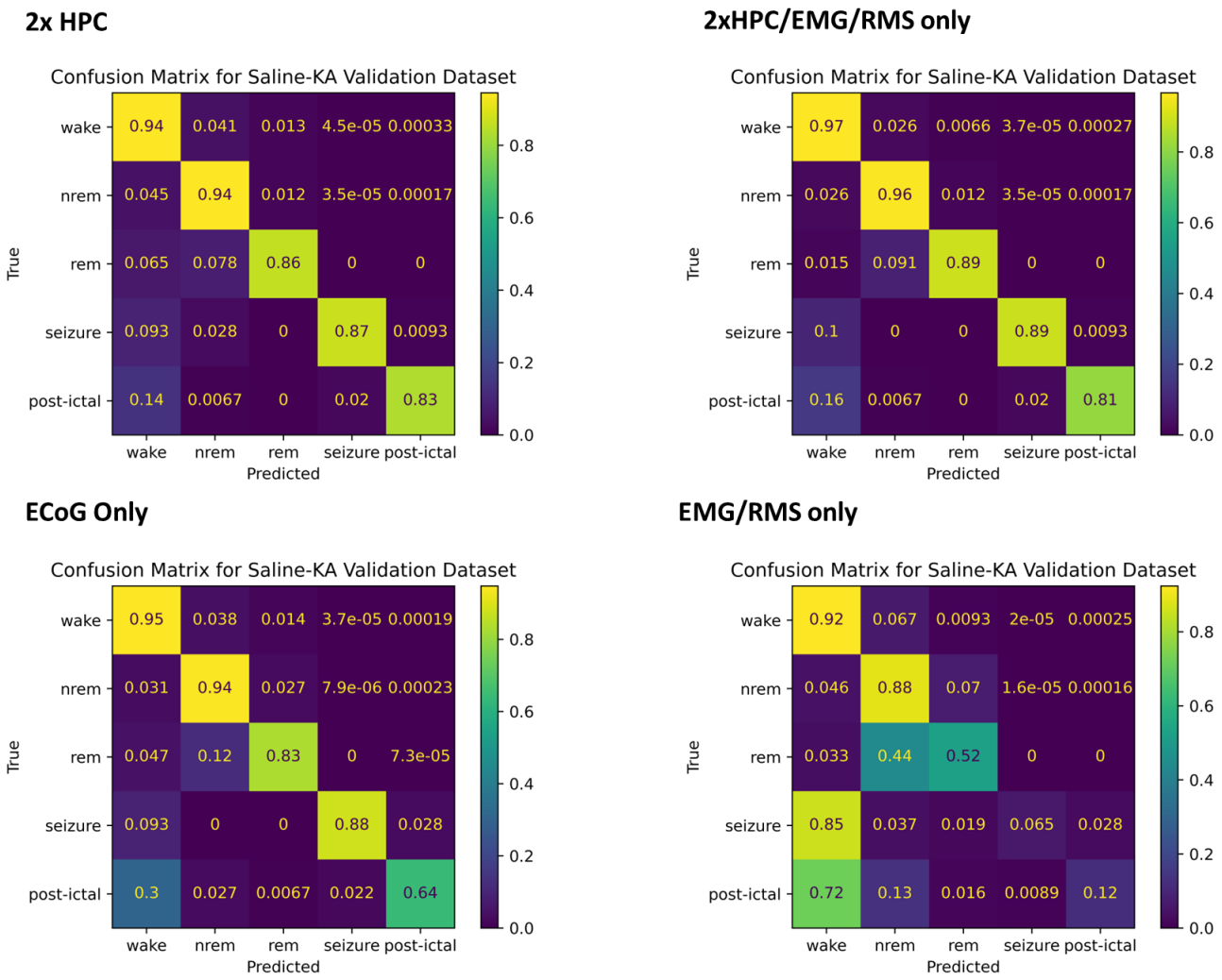
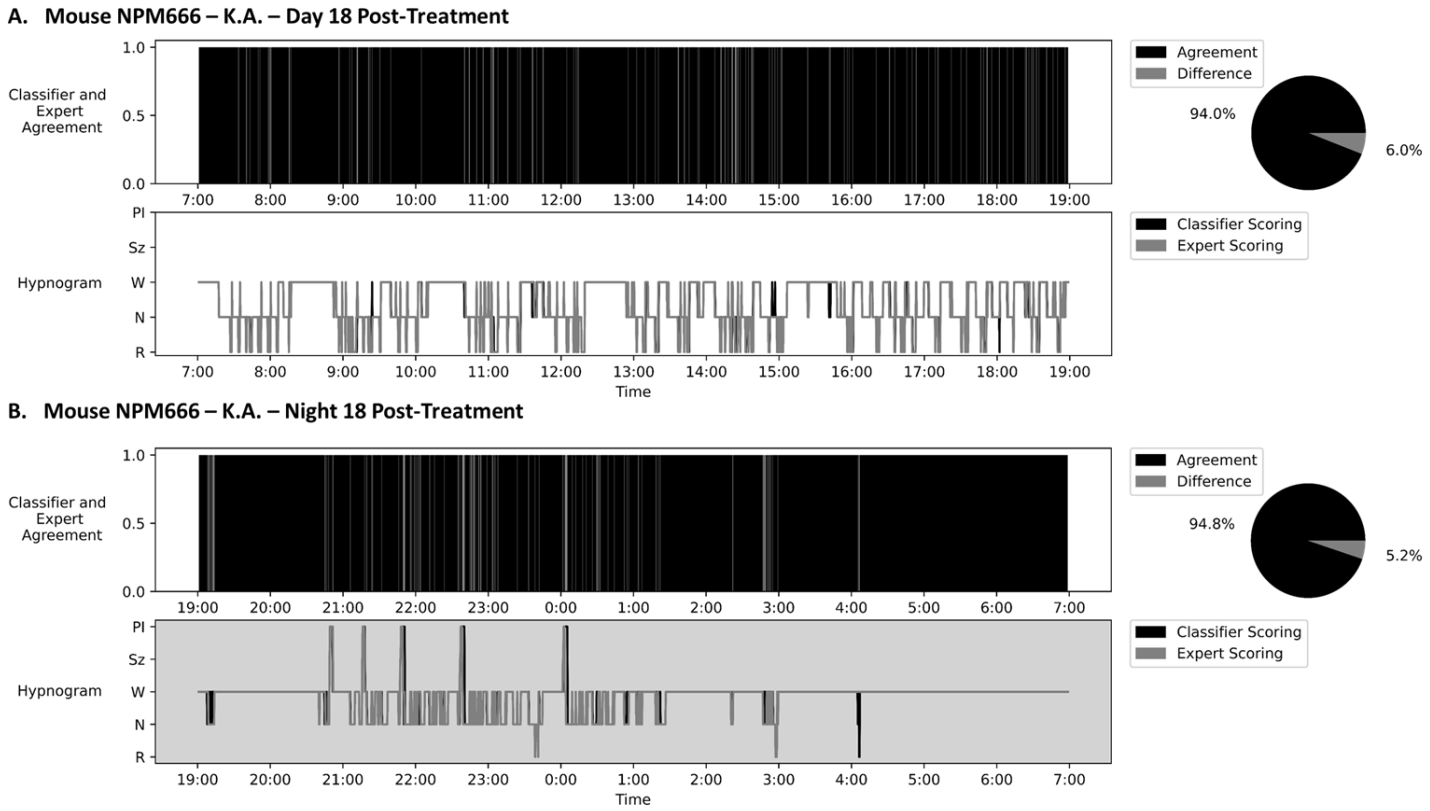



**Figure 8. Interpretable Machine Learning via Masking.** Masking specific data channels during training, a hands-on method of interpretable machine learning, showed reliable classification of the winning model when used in any configuration that included both hippocampal channels. Classification accuracy did not drop significantly until all but one of the most information-rich channels were removed.

### *3.4 Scoring Results Comparison*

When scores from the validation dataset are compared to those produced by the full classifier and rated for agreement across time epochs, we see agreement mirroring the accuracy scores in the confusion matrices in Figure 9**.** Classification agreement with our ground-truth expert scores for our epileptic mice is >93% when all states are accounted for, with a >97% agreement in scores for recordings from saline-treated mice. Figure 9 contains graphical representations of these agreements and the corresponding hypnograms from expert scores and the classifier to confirm the classifier's accuracy visually.



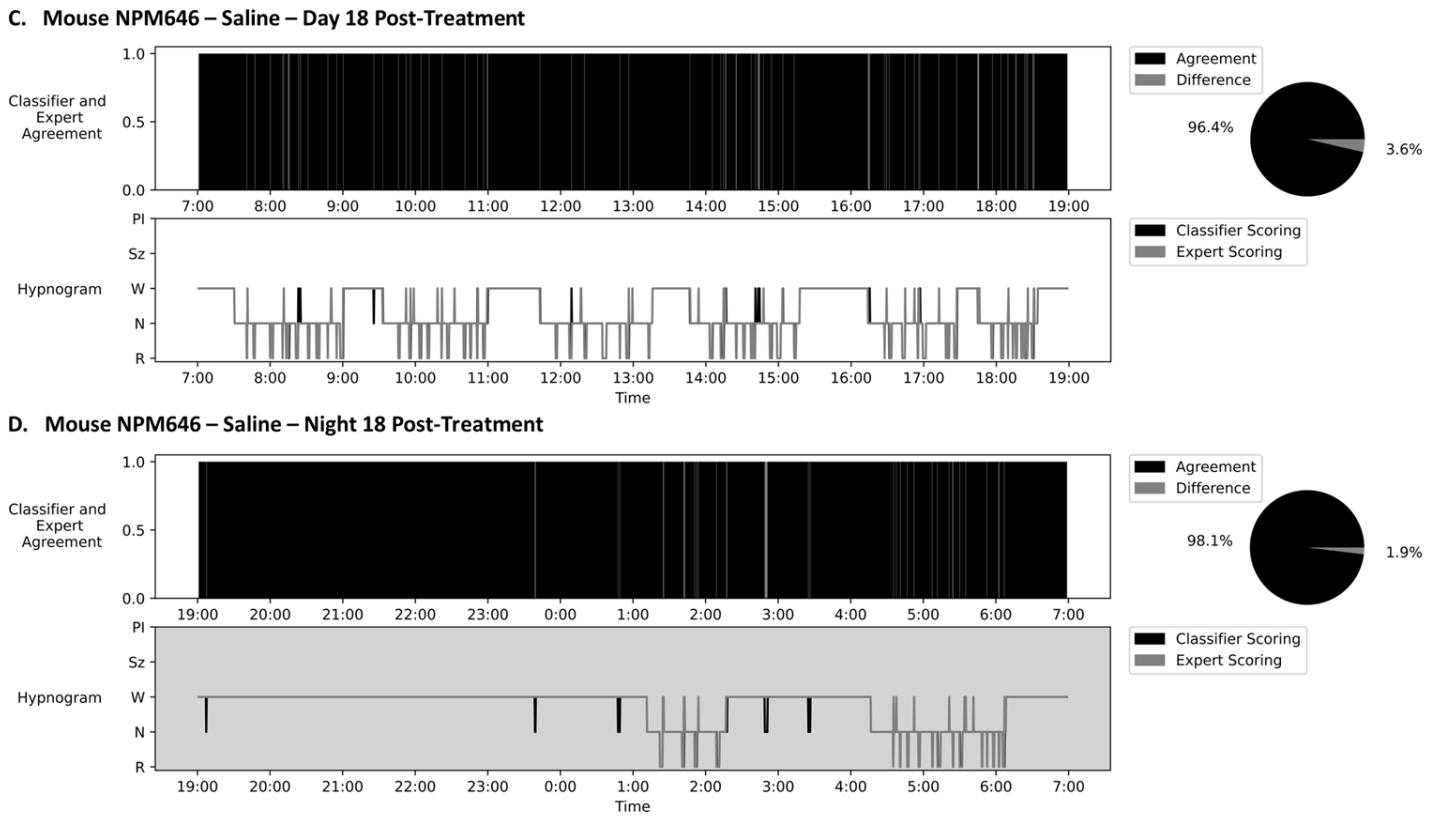



**Figure 9. Scoring Comparisons.** To test the classifier, we chose to visually and computationally compare the agreement between expert scorers and the classifier for two representative and consecutive 12-hour records of sleep-wake. In a KA-treated animal (**A**. day, **B**. night), the classifier performs with ~94% overall accuracy relative to expert scoring, which is comparable to inter-rater reliability on sleep-wake scoring tasks as demonstrated in Kloefkhorn, et al. 2020, where 93% agreement was reported between 3 expert scorers**.** In a saline-treated animal, lacking an abnormal EEG background and with reduced sleep fragmentation, (**C**. day, **D**. night), the scoring agreement exceeds 96% and shows hypnogram progression typical of mouse sleep-wake cycles. Hypnogram legend for states: post-ictal: PI; seizure: Sz; W: wake; N: NREM; R: REM.

Scoring results were also evaluated according to Rechtshaffen and Kales' criteria regarding state transitions. These conditions, as originally described and since adapted by the AASM, hold that sequences of REM must follow NREM and cannot follow wake and that a single NREM- or wake-scored epoch cannot occur alone. The evaluation of our automatic scoring for these two rules found no violations, suggesting that the classifier has implicitly learned these rules. Therefore, we did not create an additional layer to make these corrections as we had previously (Exarchos et al., 2020).

### *4. Discussion*

We present the successful creation of a machine-learning-based classifier for the automated, accurate, and rapid scoring of sleep-wake states, seizures, and post-ictal states in mice with epilepsy. While previously reported classifiers effectively classified sleep in various populations of mice (Exarchos et al., 2020; Grieger et al., 2021; Lampert et al., 2015; Yamabe et al., 2019), none have demonstrated scoring proficiency in any rodent models of epilepsy. To our knowledge, our classifier is the first to achieve this goal. This classifier also performs well with ILAE-recommended data elements for mouse epilepsy studies and implicitly follows AASM scoring criteria. This is a key step toward modeling and investigating the mechanisms of the clinically meaningful interaction between sleep disturbances and epilepsy. Furthermore, this classifier could readily be applied to other settings with or without substantial EEG abnormalities. The classifier performs well with ECog alone, ECog and EMG, or hippocampal recordings alone, making this applicable to various recording configurations and experimental settings.

Manual data scoring is onerous. With our classifier, scoring time can be reduced to *less than 1 second per 12-hour recording file*. There is no human input or time needed other than visual and statistical assessment of the results. This time saving will make the analysis of larger datasets feasible, as is often required in epilepsy studies due to the large individual variations in epilepsy severity. These time savings are also not traded off for any loss of precision in the sleep scores. The scoring accuracies of the classifier, with a weighted average of > 96%, exceed the 93% agreement in human scores (Kloefkorn et al., 2020) for AASM/Rechtshaffen and Kales sleep scoring, making it a well-rounded classifier for sleep. Although seizures in rodents are variably operationalized, inter-rater agreements are high in our experience, as is the performance of our classifier.

Other sleep classifiers, such as MC-SleepNet, have achieved exceptional levels of accuracy (96.6%) in murine sleep scoring via the use of massive datasets (35.7 million epochs total, ~14.5 million for training) and architectures which take significant time to train (160 hours), even on very high-performance computing clusters. Our model advances classification in several ways: First, comparable accuracy (96.6% on MC-SleepNet, weighted average F1 score 96.58% reported here) with far fewer physiological data points. Second, successful sleep scoring was proven possible in a rodent model of epilepsy with an abnormal electrophysiological background. Third, the classifier additionally identifies the seizure and post-seizure states using the same scoring layers used for sleep, eliminating the need for using two classifiers' input formats and installation dependencies, as well as the arduous synthesis of their output formats. Finally, our classifier

required far less initial data (2.5 million epochs, ~1.8 million for training), thus training in significantly less time (25 minutes on a Google Colab server, ~50 minutes on the PC mentioned in section 2.8).

The network architecture chosen fits well with the design intent. The innovation of BiLSTM is that it adds further information to the classifier by adding prior and future epochs in time series analysis rather than classifying the epoch at hand without sequence information (Graves and Schmidhuber, 2005). For time series classification, this recurrent neural network architecture outperforms the classical LSTM that does not receive input from flanking events and classical Support Vector Machines and Dense Neural Net classifiers. Our architecture also permits the implicit learning of Rechtschaffen and Kales-like behavior without an additional post-processing step. Our comparison to a CNN model is limited by the approach to feature extraction. Running CNN-based, rather than *a priori*, feature extraction classification on this whole dataset by conventional means would be difficult given memory constraints. Still, some incremental improvement in feature selection efficiency could be possible with this approach.

In summary, this classifier can provide a rapid, accurate, robust, multi-featured sleep-wake and seizure scoring platform to those with access to basic computing resources. This tool will benefit the epilepsy research community as we conduct studies to characterize better and investigate the relationship between sleep, epilepsy, and other comorbidities.

**Funding and Acknowledgements**

CURE Epilepsy Award (NPP)

NIH R21NS122011 (NPP)

NIH K08NS105929 (NPP)

We greatly appreciate the input of Matthew Rowan.

**Conflicts of Interests**

No relevant conflicts.

**References**

Abadi M et al. (2016) TensorFlow: a system for large-scale machine learning In: Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation, OSDI'16 , pp265–283. USA: USENIX Association.

Almeida Silva LF, Engel T, Reschke CR, Conroy RM, Langa E, Henshall DC (2016) Distinct behavioral and epileptic phenotype differences in 129/P mice compared to C57BL/6 mice subject to intraamygdala kainic acid-induced status epilepticus. Epilepsy Behav 64:186–194.

Barger Z, Frye CG, Liu D, Dan Y, Bouchard KE (2019) Robust, automated sleep scoring by a compact neural network with distributional shift correction. PLoS One 14:e0224642–e0224642.

Chollet F (2018) Keras: The Python Deep Learning library. Astrophysics Source Code Library.

Costa-Miserachs D, Portell-Cortés I, Torras-Garcia M, Morgado-Bernal I (2003) Automated sleep staging in rat with a standard spreadsheet. J Neurosci Methods 130:93–101.

Crespel A, Baldy-Moulinier M, Coubes P (1998) The relationship between sleep and epilepsy in frontal and temporal lobe epilepsies: practical and physiopathologic considerations. Epilepsia 39:150–157.

Exarchos I, Rogers AA, Aiani LM, Gross RE, Clifford GD, Pedersen NP, Willie JT (2020) Supervised and unsupervised machine learning for automated scoring of sleep-wake and cataplexy in a mouse model of narcolepsy. Sleep 43.

Frigo M (1999) A fast Fourier transform compiler. SIGPLAN Not 34:169–180.

Gomersall H (2016) pyFFTW.

Graves A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional LSTM networks In: Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005. , pp2047–2052 vol. 4.

Grieger N, Schwabedal JTC, Wendel S, Ritze Y, Bialonski S (2021) Automated scoring of pre-REM sleep in mice with deep learning. Sci Rep 11:12245.

Henshall DC, Sinclair J, Simon RP (2000) Spatio-temporal profile of DNA fragmentation and its relationship to patterns of epileptiform activity following focally evoked limbic seizures. Brain Res 858:290–302.

Jang H-J, Cho K-O (2019) Dual deep neural network-based classifiers to detect experimental seizures. Korean J Physiol Pharmacol 23:131–139.

Janz D (1962) The grand mal epilepsies and the sleeping-waking cycle. Epilepsia 3:69–109.

Kloefkorn H, Aiani LM, Lakhani A, Nagesh S, Moss A, Goolsby W, Rehg JM, Pedersen NP, Hochman S (2020)

Noninvasive three-state sleep-wake staging in mice using electric field sensors. J Neurosci Methods 344:108834.

Lampert T, Plano A, Austin J, Platt B (2015) On the identification of sleep stages in mouse electroencephalography time-series. J Neurosci Methods 246:52–64.

Moser D, Anderer P, Gruber G, Parapatics S, Loretz E, Boeck M, Kloesch G, Heller E, Schmidt A, Danker-Hopfe H, Saletu B, Zeitlhofer J, Dorffner G (2009) Sleep classification according to AASM and Rechtschaffen & Kales: effects on sleep scoring parameters. Sleep 32:139–149.

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

Pitkänen A, Buckmaster PS, Galanopoulou AS, Moshé SL (2017) Models of Seizures and Epilepsy, Second Edition. Elsevier Inc.

Rechtschaffen A, Kales A (1968) A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects. United States Government Printing Office, Washington, DC.

Rytkönen K-M, Zitting J, Porkka-Heiskanen T (2011) Automated sleep scoring in rats and mice using the naive Bayes classifier. J Neurosci Methods 202:60–64.

Shouse MN, Sterman MB (1982) Acute sleep deprivation reduces amygdala-kindled seizure thresholds in cats. Exp Neurol 78:716–727.

Tieng QM, Anbazhagan A, Chen M, Reutens DC (2017) Mouse epileptic seizure detection with multiple EEG features and simple thresholding technique. J Neural Eng 14:066006.

Wei L, Boutouil H, R Gerbatin R, Mamad O, Heiland M, Reschke CR, Del Gallo F, F Fabene P, Henshall DC, Lowery M, Morris G, Mooney C (2021) Detection of spontaneous seizures in EEGs in multiple experimental mouse models of epilepsy. J Neural Eng 18.

Wei L, Gerbatin R, Mamad O, Boutouil H, Reschke C, Lowery M, Henshall D, Morris G, Mooney C (2020) XGboost-based Method for Seizure Detection in Mouse Models of Epilepsy In: 2020 IEEE Signal Processing in Medicine and Biology Symposium (SPMB) , pp1–3.

Yamabe M, Horie K, Shiokawa H, Funato H, Yanagisawa M, Kitagawa H (2019) MC-SleepNet: Large-scale Sleep Stage Scoring in Mice by Deep Neural Networks. Sci Rep 9:1–12.

**Supplementary Information**

The full classification metrics spreadsheet and python notebooks can be found on github:

https://github.com/epilepsylab/SWISC