

Recognizing EEG responses to active TMS vs. sham stimulations in different TMS-EEG datasets: A machine learning approach

Ahmadreza Keihani^a, Francesco L. Donati^{a,b}, Simone Russo^{c,d}, Sara Parmigiani^{e,f}, Michela Solbiati^d, Adenauer G. Casali^g, Matteo Fecchio^h, Omeed Chaichian^a, John Rothwellⁱ, Marcello Massimini^{d,j}, Lorenzo Rocchi^{i,k}, Mario Rosanova^d, Fabio Ferrarelli^{a,*}

^a Department of Psychiatry, University of Pittsburgh, 3501 Forbes Avenue, Pittsburgh, PA 15213, USA

^b Department of Health Sciences, University of Milan, 20142 Milan, Italy

^c Wallace H Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332, USA

^d Department of Biomedical and Clinical Sciences, University of Milan, Milan, Italy

^e Department of Psychiatry & Behavioral Sciences, Stanford University School of Medicine, Stanford, CA, 94305, USA

^f Wu Tsai Neurosciences Institute, Stanford University, Stanford, CA, USA

^g Institute of Science and Technology, Federal University of São Paulo, São José dos Campos 12231-280, Brazil

^h Center for Neurotechnology and Neurorecovery, Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA

ⁱ Department of Clinical and Movement Neurosciences, UCL Queen Square Institute of Neurology, University College London, London WC1N 3BG, UK

^j IRCCS Fondazione Don Carlo Gnocchi Onlus, Milan, Italy

^k Department of Medical Sciences and Public Health, University of Cagliari, Cittadella Universitaria di Monserrato, 09042 Cagliari, Italy

ARTICLE INFO

Keywords:

TMS-EEG
Machine learning
Active TMS
Sham stimulations

ABSTRACT

Transcranial Magnetic Stimulation with simultaneous Electroencephalogram (TMS-EEG) allows for the assessment of neurophysiological properties of cortical neurons. However, TMS-evoked EEG potentials (TEPs) can be affected by components unrelated to TMS direct neuronal activation. Accurate, automatic tools are therefore needed to establish the quality of TEPs. We defined innovative comparisons, including effects of both baseline and post-TMS responses, while employing a sequence-to-sequence machine learning model to objectively ascertain active TMS vs. sham stimulation responses.

Two independent TMS-EEG datasets including TMS and several sham stimulation conditions were obtained from the left motor area of 33 healthy individuals (total: 27,590 trials across datasets). A Bi-directional Long Short-Term Memory (BiLSTM) machine learning network was used to label each time point of the EEG signals as pertaining to TMS or sham conditions.

For TMS conditions, post-stimulus vs. baseline/pre-stimulus EEG comparisons yielded moderate (60 %-75 %) single-trial accuracy and high-accuracy (>75 %) for 20 trials across datasets. For sham conditions, post- vs. baseline/pre-stimulus EEG comparisons yielded lower accuracy rates than for TMS conditions, except for unmasked auditory stimulation. Baseline/pre-stimulus TMS vs. baseline/pre-stimulus sham EEG comparisons showed chance-level accuracy. Conversely, post-stimulus TMS vs. post-stimulus sham EEG comparisons had moderate (single trial) to high (20 trial) accuracy, except for TMS with and without the click noise masking.

Consistently across datasets, TEPs after active TMS are discernible from various sham stimulations after few trials and at the single-subject level using a BiLSTM ML model. This approach offers objective criteria to support TEP authenticity, which may help address ongoing discussions about TEP characteristics in TMS-EEG studies.

1. Introduction

The combination of TMS with simultaneous EEG (TMS-EEG), which can non-invasively directly target cortical neurons, allows for the

investigation of the neurophysiological properties of different cortical areas (Bortoletto et al., 2015; Casali et al., 2013; Massimini et al., 2012). EEG responses to TMS pulses, commonly described as transcranial evoked potentials (TEPs), consist of early components determined by the

* Corresponding author at: Associate Professor of Psychiatry, 3811 O'Hara St. Pittsburgh PA 15213.

E-mail address: ferrarelli@upmc.edu (F. Ferrarelli).

<https://doi.org/10.1016/j.neuroimage.2026.121802>

Received 26 August 2025; Received in revised form 23 December 2025; Accepted 11 February 2026

Available online 11 February 2026

1053-8119/© 2026 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

firing of local excitatory cortical pyramidal neurons and inhibitory interneurons (Kaskie and Ferrarelli, 2018; Manganotti et al., 2015), and later components reflecting the interplay between cortical and subcortical regions (Hernandez-Pavon et al., 2023; Momi et al., 2023; Russo et al., 2024). However, TEPs can also be affected by non-transcranial confounding factors. First, TMS produces a "click" upon discharge, which generates an auditory-evoked potential (AEP) through air and bone conduction (Hernandez-Pavon et al., 2023). Moreover, the TMS-induced electric field can activate craniofacial muscles and free cutaneous nerve endings in the scalp, thus producing somatosensory evoked potentials (SEPs) (Mancuso et al., 2023). These confounds may hinder an accurate interpretation of TEPs. Although tools to address these confounds were recently introduced (Russo et al., 2022), experimental countermeasures such as playing a TMS click- masking noise (Ross et al., 2022), are not always utilized in TMS-EEG studies and have been applied with various degrees of effectiveness (Belardinelli et al., 2019; Biabani et al., 2024; Conde et al., 2019).

Several TMS-EEG studies aimed at differentiating TEPs due to TMS direct cortical activation from sensory responses have yielded contradictory findings. Gordon et al. (Gordon et al., 2018) reported clear distinctions between TEPs and EEG responses to auditory/somatosensory stimulations, while Conde et al. (Conde et al., 2019) found no significant differences between sham and TMS, concluding that TEPs primarily arise from auditory and somatosensory activation. Conversely, two recent studies found that TMS, coupled with appropriate masking of sensory input, resulted in lateralized responses at the stimulation site lasting ~300 ms. Sham stimulations, on the other hand, yielded sensory evoked responses represented by late (100-200 ms) components mostly located at central scalp regions (Fecchio et al., 2025; Rocchi et al., 2021). Based on those results (Fecchio et al., 2025; Rocchi et al., 2021), it was concluded that TMS, when accounting for confounding sources, produces responses primarily from direct neuronal activation. Yet, there is a lack of data-driven approaches that can systematically identify EEG responses to TMS and sham stimulations (Belardinelli et al., 2019; Gordon et al., 2023a).

Machine learning (ML) algorithms have gained increasing attention in EEG research based on their ability to differentiate conditions (Gemein et al., 2020; Hosseini et al., 2020; Keihani et al., 2022a; Keihani et al., 2022b). These algorithms excel in uncovering hidden relationships within complex datasets, providing insights otherwise precluded. ML algorithms are also well-suited for handling large, complex datasets, such as TMS-EEG data, in which the number of inputs surpasses the number of subjects and requires differentiation of temporal and spatial information. In those instances, traditional statistical approaches often yield suboptimal results (Rocchi et al., 2021). Hence, utilization of ML represents an intriguing yet underexplored approach in the TMS-EEG field.

A recent study used ML-based classification and showed that it could disentangle TEPs from the TMS click related AEPs more accurately than conventional statistical analyses (Cristofari et al., 2023). However, ML performance was assessed: 1) on EEG responses across all trials, rather than on responses from single/few trials; 2) on a single TMS-EEG dataset; 3) only on auditory stimulation as a confound; 4) without accounting for baseline differences between TMS and sham sessions; and 5) at the group, but not at the single subject level, thus leaving several unanswered questions concerning TMS vs. sham responses (Cristofari et al., 2023; Hernandez-Pavon et al., 2023). First, can we identify TEPs with at least moderate accuracy at the single trial level using ML? And can high accuracy be achieved after averaging a small number of trials? Second, can we leverage ML to identify TEPs with high accuracy in different datasets? Third, can ML identify TEPs more accurately than both somatosensory and auditory evoked potentials when each TMS and sham condition is compared to its baseline? Fourth, can ML achieve high accuracy in distinguishing post-TMS vs. post-sham stimulation responses, but only chance-level accuracy when classifying the corresponding baseline responses? Finally, can high accuracy for TEPs vs.

several sham responses be achieved at the single-subject level?

Here, we began addressing those questions by comparing TMS and several sham stimulation sessions in different TMS-EEG datasets using bidirectional sequence-to-sequence long-short term memory (BiLSTM) ML (Graves et al., 2005; Mwata-Velu et al., 2021; Salehinejad et al., 2017; Schuster and Paliwal, 1997; Yu et al., 2019; Zhang and Fu, 2020). The BiLSTM examined TMS-EEG data bidirectionally (both forward and backward in time), used sequence-to-sequence (labeled each time point as TMS or sham), analyzed single trial and 5-20 averaged trials, compared TMS and sham sessions at baseline (pre-TMS vs. pre-sham) and after stimulation (post-TMS vs. post-sham), and examined both group and single subject performance.

2. Material and methods

2.1. Datasets

The BiLSTM ML model was applied to two datasets collected on thirty-three healthy subjects (HS) by two different research groups (Fecchio et al., 2025; Rocchi et al., 2021). Dataset 1 (N=14 HS, 12256 trials) included: a) The condition to optimize cortical activation by TMS (effective TMS, eTMS), i.e., TEPs recorded with the TMS coil touching the scalp and the TMS click masked by a custom-made noise (TMS1) (Russo et al., 2022); b) Realistic sham, recorded by tilting the TMS coil 90 degrees and simultaneous electrical stimulation to produce a TMS-like scalp sensation along with TMS- masking noise (S1); c) AEP not masked, obtained with the coil at 90 degrees and no masking noise (AEP1); d) AEP masked, as in c, but with a noise masking sound (AEPm1); and e) High intensity electrical stimulation, obtained with an electrical stimulation producing discomfort in all subjects (HES1). Dataset 2 (N=19 HS, 15334 trials) included: a) TEP, recorded with the TMS coil on the scalp while playing a masking noise (TEP, TMS2); b) Sham, with a sham coil on the scalp and noise masking (S2); c) AEP not masked, with the TMS coil on a 5 cm pasteboard cylinder placed on the scalp, but without noise masking (AEP2); d) AEP masked, as in c but with a noise masking sound (AEPm2); e) Low-intensity ES, with an electrical stimulation of the scalp of intensity similar to that generated by 90 % resting motor threshold (RMT) TMS (LES2); and f) TEP not masked, with the coil on the EEG cap, but no noise masking (TMSnm2). Both datasets utilized a neuronavigation system for the targeted cortical area (i.e., the primary motor cortex, M1) (Fecchio et al., 2025; Rocchi et al., 2021). However, for Dataset 1 a visualization of TEPs (rt-TEP (Casarotto et al., 2022)), which allows optimizing TMS parameters (coil position, stimulation intensity) based on EEG responses, was employed (Fecchio et al., 2025), while for Dataset 2 a fixed TMS intensity (i.e., 90 % RMT) with no rt-TEP was chosen (Rocchi et al., 2021). Furthermore, Dataset 1 TMS-EEG data collected more trials (217.64 ± 30.70 , average for TMS1) than Dataset 2 (106 ± 4.56 , average for TMS2). All procedures were conducted in accordance with the Declaration of Helsinki. Dataset 1 was acquired with the approval of the Ethics Committee Milano Area A, and Dataset 2 was acquired under the approval of the Human Subjects Review Board of University College London. Written informed consent was obtained from all participants prior to the experimental session.

2.2. Preprocessing of TMS-EEG data

2.2.1. Dataset 1

EEG signals were recorded in direct current (DC) mode using a TMS-compatible amplifier (BrainAmp DC, Brain Products GmbH) with a 62-electrode montage. Reference and ground electrodes were placed on the forehead. Signals were sampled at 5 kHz, and impedances were kept below 5 k Ω . EEG data analyses were performed in MATLAB R2016b using custom scripts adapted from EEGLAB (Delorme and Makeig, 2004). Channels and epochs showing artifactual activity were excluded by visual inspection. EEG artifacts due to the high-energy TMS pulse were removed by replacing signal around the TMS pulse (-2 to 8 ms)

with the preceding ten-second time interval (-12 to -2 ms) and then applying a 5th-order moving-average filter (between 6 and 10 ms). Realistic sham and HES1 conditions were high-pass filtered at 0.5 Hz, while for all other sessions, a 1 Hz high-pass filter (3rd-order Butterworth) was utilized. EEG data were divided into epochs spanning -1000 to 1000 ms around the TMS pulses and re-referenced to the average of all

channels. Independent Component Analysis (ICA) was performed and remaining artifacts from eye blinks, saccades, muscle activity, and electrode discharge were removed manually. Signals were then low-pass filtered at 45 Hz (3rd order Butterworth), down sampled to 1 kHz, and segmented from -600 to 600 ms around the stimulus to remove edge effects. Additional details are available in (Fecchio et al., 2025).

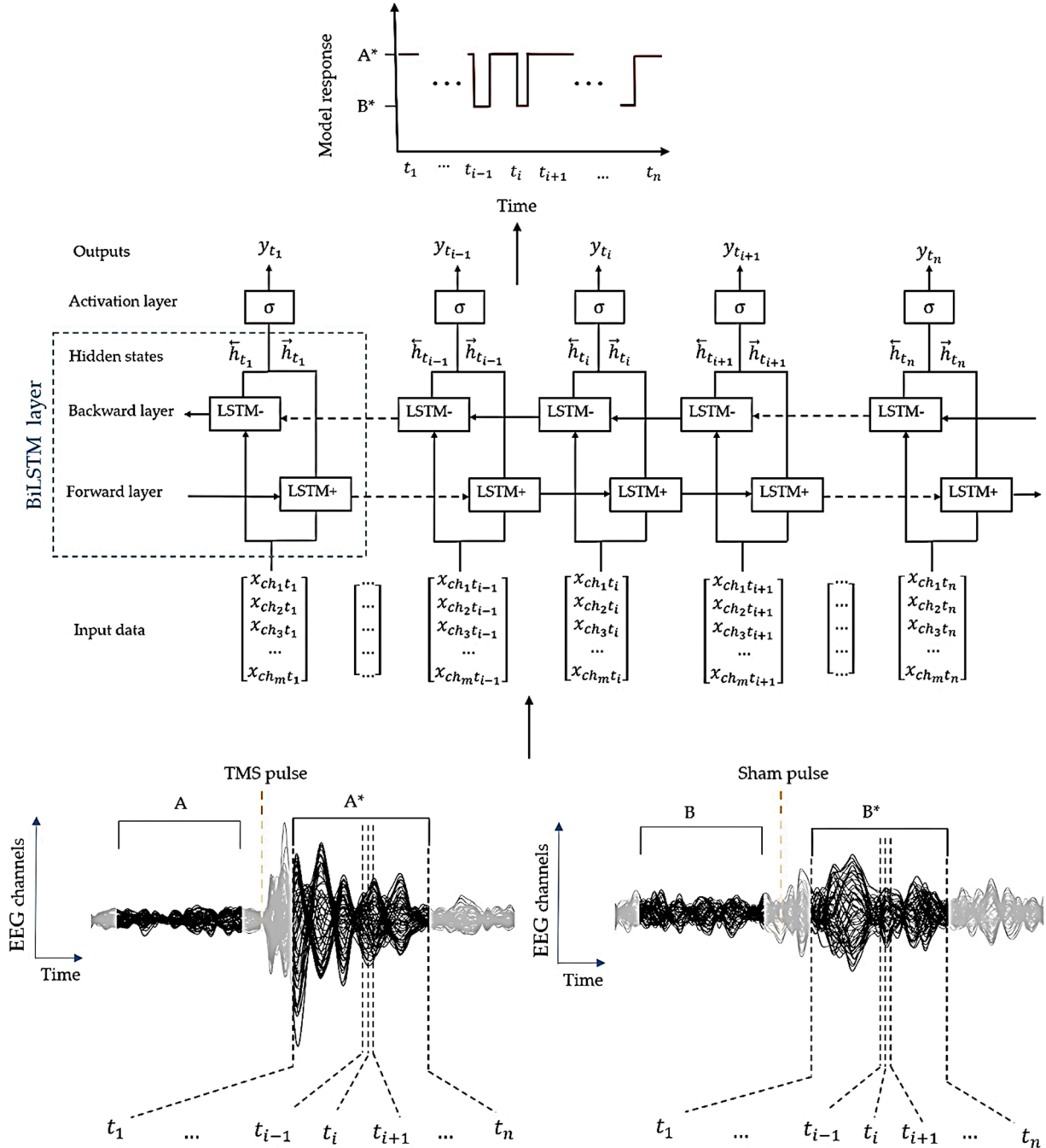


Fig. 1. The bidirectional sequence-to-sequence long-short term memory (LSTM) machine-learning approach. At each time point t , the input multichannel time-series data are processed by a bidirectional LSTM layer. The resulting hidden states from the forward and backward LSTM layers are concatenated and passed through dropout and fully connected layers, followed by a softmax operation (collectively denoted as the 'activation layer' for simplicity in the figure) to produce the output class probabilities. An example of our machine-learning approach applied to TMS-EEG sessions is shown. Bottom panel) TMS-EEG traces from one representative subject from Dataset 1. The model employs user-defined time windows (i.e., 300 ms long) relative to the TEP or sham pulse as input. In this example: A* (left traces, post TMS) vs B* (right traces, post sham). Of note, the model can also be employed to compare time windows within the same condition (e.g.: A vs A* for pre. vs. post TMS). Middle panel) general block diagram of model; the model uses the entire spatio-temporal information (all channels [channel number (Ch_i)] at each time point [sample time (t_i)] as input data, while forward and backward directions are shown as black arrows and h represent the hidden states of the model. The forward and backward layers of the model are displayed as LSTM+ and LSTM-, respectively. Top panel) model output; in this example, the output of the model consists of a binary sequence, which defines whether the response label at each time point is TMS (A*) or sham (B*).

2.2.2. Dataset 2

EEG signals were recorded in DC mode, using an Actichamp amplifier (Brain Products, GmbH) with a 63-electrode montage. Signals were referenced to Oz, and Fpz was used as the ground electrode. Signals were sampled at 5 kHz, with impedances <5 k Ω . EEG data analysis was performed in MATLAB R2016b, employing EEGLAB 14.1.1 (Delorme and Makeig, 2004), TMS-EEG signal analyser (TESA) toolbox (Rogasch et al., 2017), and Fieldtrip (Oostenveld et al., 2011). EEG signals were divided into epochs spanning -1300 to 1300 ms around the TMS pulse. Noisy channels and epochs were removed after visual inspection. TMS pulse-related artifacts were removed by excluding the EEG activity from -5 to 10 ms, followed by cubic interpolation. Then, an initial round of ICA decomposition was run to identify large scalp muscle and discharge artifacts, which were removed after visual inspection. Band-pass (1–100 Hz) and band-stop (48–52 Hz) fourth-order Butterworth filters were then applied. To reduce possible edge artefacts, EEG signals were subsequently split in epochs spanning -1000 to 1000 ms and another round of ICA was performed to remove residual artifacts from eyeblinks, saccades, and muscle activity. Lastly, EEG signals were re-referenced to the average activity across all channels. Further details can be found in (Rocchi et al., 2021). Because the two datasets were obtained from independent, previously published TMS-EEG studies (Fecchio et al., 2025; Rocchi et al., 2021), they differed in acquisition settings, stimulation protocols, sham implementations, and artifact profiles. Therefore, we retained the dataset-specific preprocessing pipelines used in the original publications, rather than enforcing a single harmonized pipeline. This choice was made to preserve comparability with each dataset's original report.

2.3. Analytical method, machine-learning approach

A BiLSTM ML network, which retains information for longer periods of time and allows for capturing long-term dependencies, was employed (Graves et al., 2005; Hochreiter and Schmidhuber, 1997; Salehinejad et al., 2017; Schuster and Paliwal, 1997; Yu et al., 2019). BiLSTM is well suited for TMS-EEG data because it leverages both forward and backward information during training and inference, which can be utilized to distinguish temporal patterns (Fig. 1). We used the Statistics and Machine Learning Toolbox in MATLAB 2022a along with a BiLSTM architecture. The network input consisted of samples from all EEG channels within a 300-ms window, extracted for both pre-stimulus (-500 to -200 ms) and post-stimulus (20 to 320 ms) periods. The BiLSTM layer included 5 hidden state units and was set to output a sequence. To mitigate overfitting, a dropout layer with a 0.4 probability rate was added after the BiLSTM layer. A fully connected layer then mapped the BiLSTM output to two classes (e.g., TEP vs. sham), followed by a softmax layer to generate the final classification output. The model used data from all EEG channels for the defined time windows. Forward and backward activations were used to calculate the output at each time point (i.e., y_t), which was labeled as sham or TEP as a binary sequence with the same length as the input sequence from each trial.

The network was trained using the Adam optimizer (Kingma and Ba, 2014) with a maximum of 60 epochs and a learning rate of 0.001. Model validation was performed during training using the 'ValidationData' option, which provided an early-stopping mechanism based on validation performance. The training progress was executed in a parallel computing environment for efficient processing. To assess the model's performance, we employed both leave-one-out (LOO) subject-independent validation, which involves removing one subject from the training data and using such data for testing iteratively, and a subject-dependent five-fold cross-validation (CV) approach. For CV, trials from each dataset were divided into five-folds and in every fold 80 % of the data (i.e., all trials of all subjects for sham and TEP) were used as a training set while the remaining 20 % of the data were used as a test set.

We first used LOO on the conditions shared across the two datasets (i.

e., TEP [TMS1, TMS2], sham [S1, S2], AEP not masked [AEP1, AEP2], and AEP masked [AEPm1, AEPm2]) to assess the consistency of ML performance. For these conditions, we performed within-condition (i.e., post-stimulus vs. baseline/pre-stimulus) and between-condition (i.e., baseline/pre-stimulus vs. baseline/pre-stimulus and post-stimulus vs. post-stimulus) comparisons for both datasets. Within-condition comparisons established how well the ML algorithm differentiated post stimulus responses from their baselines, whereas between-condition comparisons examined whether BiLSTM could discern TMS1 and TMS2 from their respective sham conditions when looking at post stimulus responses (i.e., post-TMS vs. post-sham comparisons), but not when comparing their baselines (i.e., pre-TMS vs. pre-sham comparisons). We then applied the same approach to the conditions unique for Dataset 1 (HES1), and Dataset 2 (Low intensity ES [LES2] and TEP not masked [TMSnm2]). Accuracy (overall proportion of correct predictions), which was categorized as low (<60 %), moderate (60 %-75 %), and high (>75 %) (Li et al., 2019), was the primary outcome measure for single trial and for 5-20 trial analyses. When averaging, we accounted for the number of trials (i.e., less trials available for training and testing by averaging) and adjusted performance accordingly (see supplementary figures 2-19). In addition to accuracy, we computed sensitivity, specificity, area under the curve (AUC), and F1 score at both group level and single subject level. Due to differences in experimental setup and sham conditions across datasets (e.g., rt-TEP-guided dosing vs. fixed intensity, trial counts per condition, and distinct sham implementations), the BiLSTM model was applied to each dataset separately. Rather than training a model on one dataset and testing it on another, we evaluated the ML model's ability to distinguish TEPs from sham responses within each dataset, thereby minimizing confounds due to between-dataset heterogeneity.

3. Results

3.1. Within-condition (post-stimulus vs. pre-stimulus for TMS and sham) comparisons

3.1.1. TMS1 and TMS2 had single trial moderate accuracy and high accuracy with 20 trials

In Dataset 1, the post vs. pre TMS comparison (TMS1) yielded moderate accuracy for single trial and high accuracy for 20 trials (Fig. 2, top left panel, blue solid line, and Table 1), Acc range [mean Acc of single trial-mean Acc of 20 trials averaged across subjects]=[71.23 %-100 %]. Similarly, in Dataset 2 the post vs. pre TMS comparison (TMS2) showed moderate to high accuracy from single trial to 20 trials (Fig. 2, top right panel, blue solid line, and Table 1, Acc range=[61.32 %-77.62 %]). However, accuracy levels for TMS2 were lower than for TMS1 (Fig. 2, Table 1). This finding can be explained by a higher signal to noise ratio (SNR) for TMS1 relative to TMS2 [TMS1 SNR=4.24 $\pm$ 0.56, TMS2 SNR=2.96 $\pm$ 1.56, tstat=2.35, pval=0.025], likely due to using the rt-TEP to optimize TMS doses for TMS1 vs. fixed TMS intensities for TMS2. We also performed surrogate data analysis (i.e., artificial TEPs created through random circular shift of the data) for TMS1 and TMS2, which yielded chance level accuracy for both datasets (Supp. Figure 1), further suggesting that accuracy differences between TMS1 and TMS2 were related to TEPs. Combined, these findings indicate that TEPs, especially when collected with high SNR, can be detected even after a handful of TMS-EEG trials.

3.1.2. Sham conditions shared across datasets showed variable accuracy levels but, except for AEP2, had lower single and averaged trial accuracy relative to TMS conditions

For Dataset 1, realistic sham (S1) (Fig. 2, top left panel, black solid line, and Table 1, Acc range=[67.76 %-84.54 %]) and AEP not-masked (AEP1) (Fig. 2, middle left panel, green solid line, and Table 1, Acc range=[61.32 %-77.62 %]) had moderate to high accuracy from single trial to 20 trials, whereas AEP masked (AEPm1) showed low accuracy

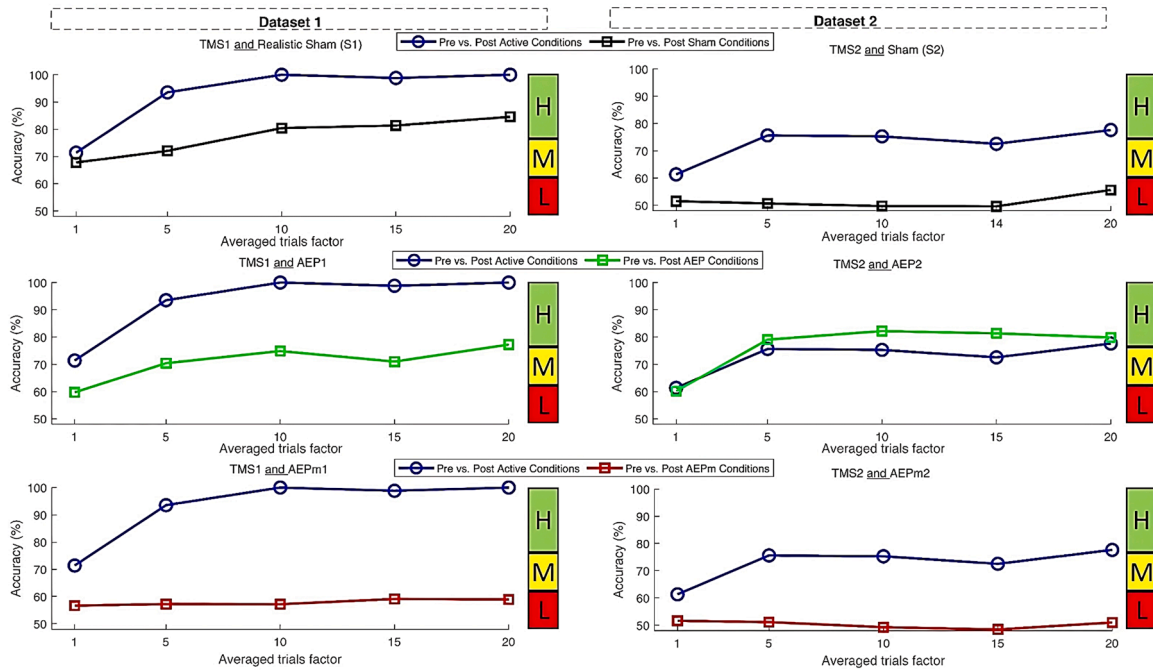


Fig. 2. TMS1 and TMS2 had moderate (single trial) to high (average of 20 trials) accuracy for within-condition, post- vs. pre- comparisons that was higher than the accuracy of the shared sham conditions, except AEP2, across datasets. Accuracy rates (% , y axes) after averaging an increasing number of trials (N, x axes) for post- vs. pre-stimulus comparisons of TEP (blue traces) and sham conditions (black, green, and red traces) shared across datasets. Color bar chart shows Acc ranges: Low (L < 60 %), Medium (60 % ≤ M ≤ 75 %), and High (H > 75 %).

Table 1

Main findings from within-condition comparisons of TEP and sham stimulations. L: low accuracy (<60 %), M: moderate accuracy (60–75 %), H: high accuracy (>75 %) for single trial Acc vs. average of 20 trials Acc across subjects.

Comparisons	Model performance	
	Post-stimulus vs. Pre-stimulus	
	Single trial (Acc [%])	Average of 20 trials (Acc [%])
Dataset 1		
TEP (TMS1)	M(71.23)	H(100)
Realistic sham (S1)	M(67.76)	H(84.54)
AEP not masked (AEP1)	L(59.70)	H(77.24)
AEP masked (AEPm1)	L(56.63)	L(57.75)
High intensity ES (HES1)	M(68.34)	H(90.55)
Dataset 2		
TEP (TMS2)	M(61.32)	H(77.62)
Sham (S2)	L(51.48)	L(55.61)
AEP not masked (AEP2)	M(60.15)	H(79.82)
AEP masked (AEPm2)	L(51.60)	L(50.94)
ES (LES2)	L(50.81)	L(59.32)
TEP not masked (TMSnm2)	M(69.73)	H(93.28)

and negligible differences between single and averaged trials (Fig. 2, lower left panel, red solid line, and Table 1, Acc range=[56.63 %–57.75 %]). Furthermore, S1, AEP1 and AEPm1 all had lower accuracy for both single and averaged trials relative to TMS1 (Table 1).

For Dataset 2, post vs. pre AEP2 showed moderate single trial accuracy that increased to high accuracy with averaging (Fig. 2, middle right panel, green solid line, and Table 1, Acc range=[60.15 %–79.82 %]), whereas low, close-to-chance accuracy levels were obtained for AEPm2 (Acc range=[50.94 %–51.50 %], Fig. 2, bottom right panel, red solid line and Table 1), and S2 in both single trial and 20 trials (Fig. 2, top right panel, black solid line, and Table 1, Acc range=[51.48 %–55.61 %]). All those sham conditions, except AEP2, had lower accuracy for both single trial and 5–20 trial analyses relative to TMS2 (Table 1).

3.1.3. Unique sham conditions showed moderate to high accuracy for HES1 and TMSnm2 and chance-level accuracy for LES2 vs. their baselines

Regarding sham conditions unique to each dataset, in Dataset 1 HES1 showed moderate single trial accuracy that increased with averaging (accuracy range=[68.34 %–90.55 %], Supp. Figures 2, and Table 1), while in Dataset 2 post- vs. pre- comparison for LES2 showed chance-level single trial accuracy that slightly increased with averaging (accuracy range=[50.81 %–59.32 %], Supp. Figures 3, and Table 1). Both sham conditions had lower single and averaged trial accuracy compared to TMS. Dataset 2 also included TMS not masked (TMSnm2), which yielded moderate single trial accuracy that increased with averaging (accuracy range=[69.73 %–93.28 %], Supp. Figure 4 and Table 1). As expected, given that it included both TEP and AEP components, this sham condition reached higher accuracy than TMS2.

3.1.4. Single subject and other performance parameters confirmed group accuracy findings

For all the within-condition comparisons, single-subject analyses confirmed what was observed in group analyses and showed that group effects were not driven by just a few individuals but were present in most subjects (Fig. 3, single trial in blue circles and 20 trials in green asterisk). Nonetheless, some variability was observed in individual accuracy for TEP and sham conditions across datasets.

Across both datasets, sensitivity, specificity, AUC, and F1-score closely paralleled accuracy results (Supp. Figures 2–10). Furthermore, CV yielded performance estimates consistent with LOO (see Supp. Figures 11–19), supporting the robustness of the model across validation approaches.

3.2. Between-condition (pre-TMS vs. pre-sham and post-TMS vs. post-sham) comparisons

3.2.1. Baseline/pre-stimulus comparisons of TMS vs. sham conditions shared across datasets resulted in low/chance level accuracy

In Dataset 1, baseline/pre-stimulus comparisons yielded low accuracy rates for single trial with marginal increases after 20 trials for TMS1

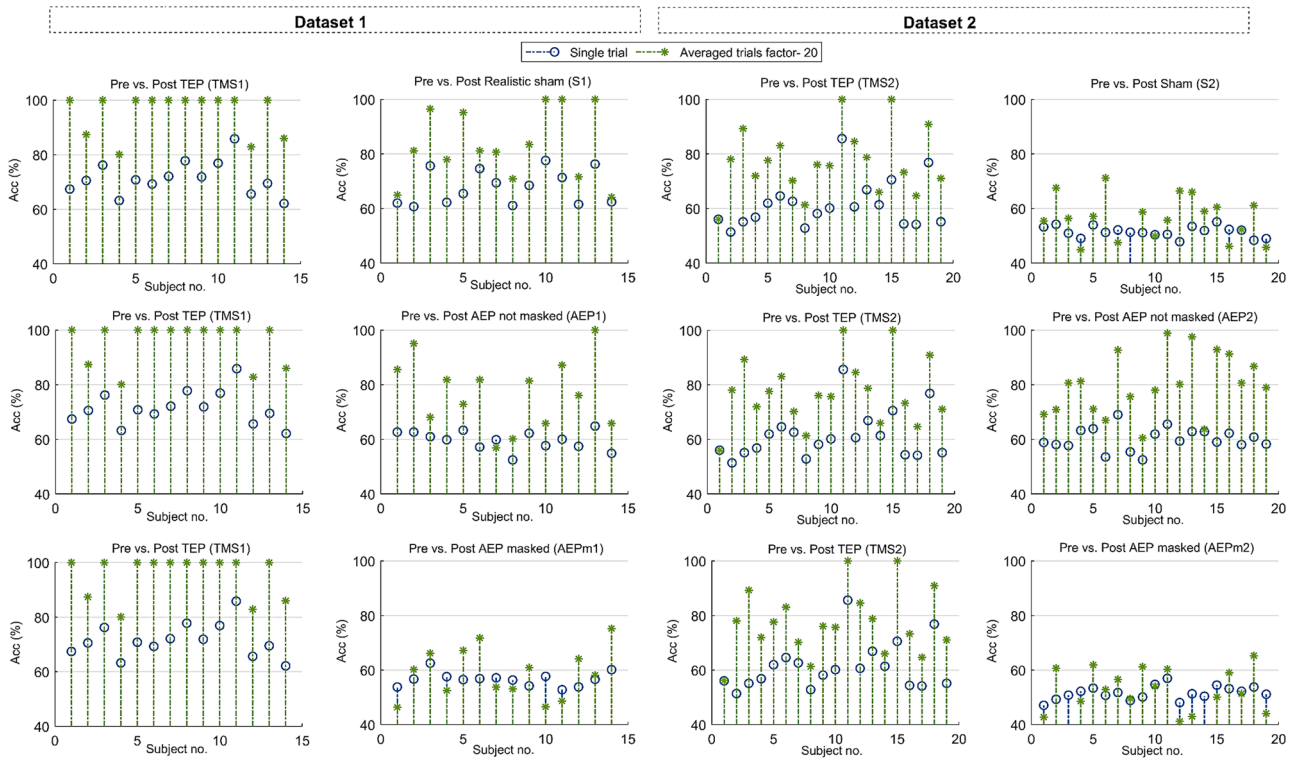


Fig. 3. Single subject confirmed group-level findings for within-condition, post- vs. pre- comparisons of TMS and shared sham conditions across datasets. Individual accuracy rates for single trial (blue circles) and 20 trials (green asterisks) of TMS (first and third column) and shared sham conditions (second and fourth column) are shown for both datasets.

vs. S1 (Fig. 4, top left panel, black dotted line, and Table 2, Acc range=[54.87 %-55.49 %]), TMS1 vs. AEP1 (Acc range=[55.62 %-59.69 %],

Fig. 4, middle right panel, green dotted line, and Table 2), and TMS1 vs. AEPm1 (Acc range=[57.41 %-62.79 %], Fig. 4, bottom left panel, red

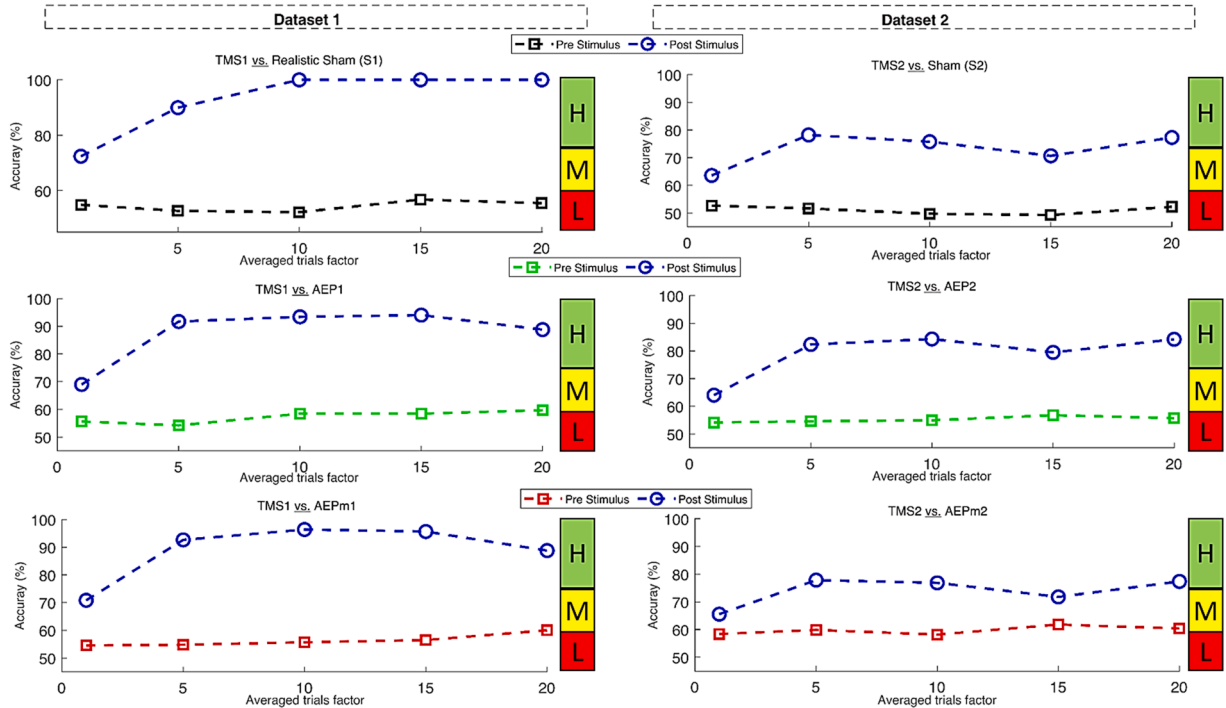


Fig. 4. Baseline/pre-stimulus comparisons between TMS and shared sham conditions yielded low (single trial) to low/moderate (20 trials) accuracy, whereas post-stimulus comparisons showed moderate (single trial) to high (20 trials) accuracy across datasets. Accuracy rates (%, y axes) after averaging an increasing number of trials (x axes) for baseline/pre TEP vs sham comparisons (black, green and red dotted traces) and for post TMS vs. sham comparisons (blue dotted traces) for both datasets. Color bar chart shows Acc ranges: Low (L < 60 %), Medium (60 % ≤ M ≤ 75 %), High (H > 75 %).

Table 2

Main findings from between-condition comparisons of TEP and different sham stimulations. L: low Acc (<60%), M: moderate Acc (60-75 %), and H: high Acc (>75 %) for single trial and 20 trials Acc across subjects.

Comparisons	Model performance			
	Pre TMS vs. Pre sham		Post TMS vs. Post sham	
	Single trial (Acc [%])	Average of 20 trials (Acc [%])	Single trial (Acc [%])	Average of 20 trials (Acc [%])
Dataset 1				
TEP (TMS1) vs. Realistic sham (S1)	L(54.87)	L(55.49)	M(72.41)	H(100)
TEP (TMS1) vs. AEP not masked (AEP1)	L(55.62)	L(59.69)	M(68.96)	H(88.78)
TEP (TMS1) vs. AEP masked (AEPm1)	L(57.41)	M(62.79)	M(70.82)	H(88.77)
TEP (TMS1) vs. High intensity Sham (HES1)	L(50.81)	L(59.32)	M(74.25)	H(100)
Dataset 2				
TEP (TMS2) vs. Sham (S2)	L(52.64)	L(52.25)	M(63.57)	H(77.27)
TEP (TMS2) vs. AEP not masked (AEP2)	L(54.15)	L(55.72)	M(64.03)	H(84.20)
TEP (TMS2) vs. AEP masked (AEPm2)	L(58.37)	M(60.36)	M(65.54)	H(77.37)
TEP (TMS2) vs. ES (LES2)	M(70.20)	M(73.56)	H(77.79)	H(93.20)
TEP (TMS2) vs. TEP not masked (TMSnm2)	L(50.20)	L(47.10)	L(57.36)	M(66.87)

dotted line, and Table 2). Dataset 2 showed similar results (Fig. 4, right panels, and Table 2).

3.2.2. Post-stimulus comparisons of TMS vs. sham conditions shared across datasets yielded moderate (single trial) to high (20 trials) accuracy

In Dataset 1, moderate to high accuracy levels were obtained for post vs. post comparisons, including TMS1 vs. S1 (Fig. 4, top left panel, blue dotted line, and Table 2, Acc range=[72.41 %-100 %]), TMS1 vs. AEP1 (Acc range=[68.96 %-88.78 %], Fig. 4, middle left panel, blue dotted line, and Table 2), and TMS1 vs. AEPm1 (Acc range=[70.82 %-88.77 %], Fig. 4, left bottom panel, blue dotted line, and Table 2). Similar findings were obtained in Dataset 2 (Fig. 4, right panels, blue dotted lines, and Table 2).

3.2.3. TMS vs. unique sham conditions confirmed between-condition trends except for post TMS2 vs. post TMSnm2

Low/moderate accuracy rates for the pre-TMS vs. pre-sham comparisons along with moderate to high accuracy for post-TMS vs. post-sham comparisons were also observed for the sham conditions unique to each dataset (Supp. Figures 2-3 and Table 2), except for post TMS2 vs. post TMSnm2, which included TEP with and without noise masking and yielded low to moderate accuracy levels (Supp. Figure 4, Table 2).

3.2.4. Single subject and other performance parameters confirmed group accuracy findings

Single-subject analyses replicated the group-level pattern for accuracy (Fig. 5). Furthermore, sensitivity, specificity, AUC, and F1-score provided results comparable to those obtained with accuracy for both datasets (Supp. Figures 2-10).

4. Discussion

We applied a sequence-to-sequence BiLSTM ML algorithm to identify

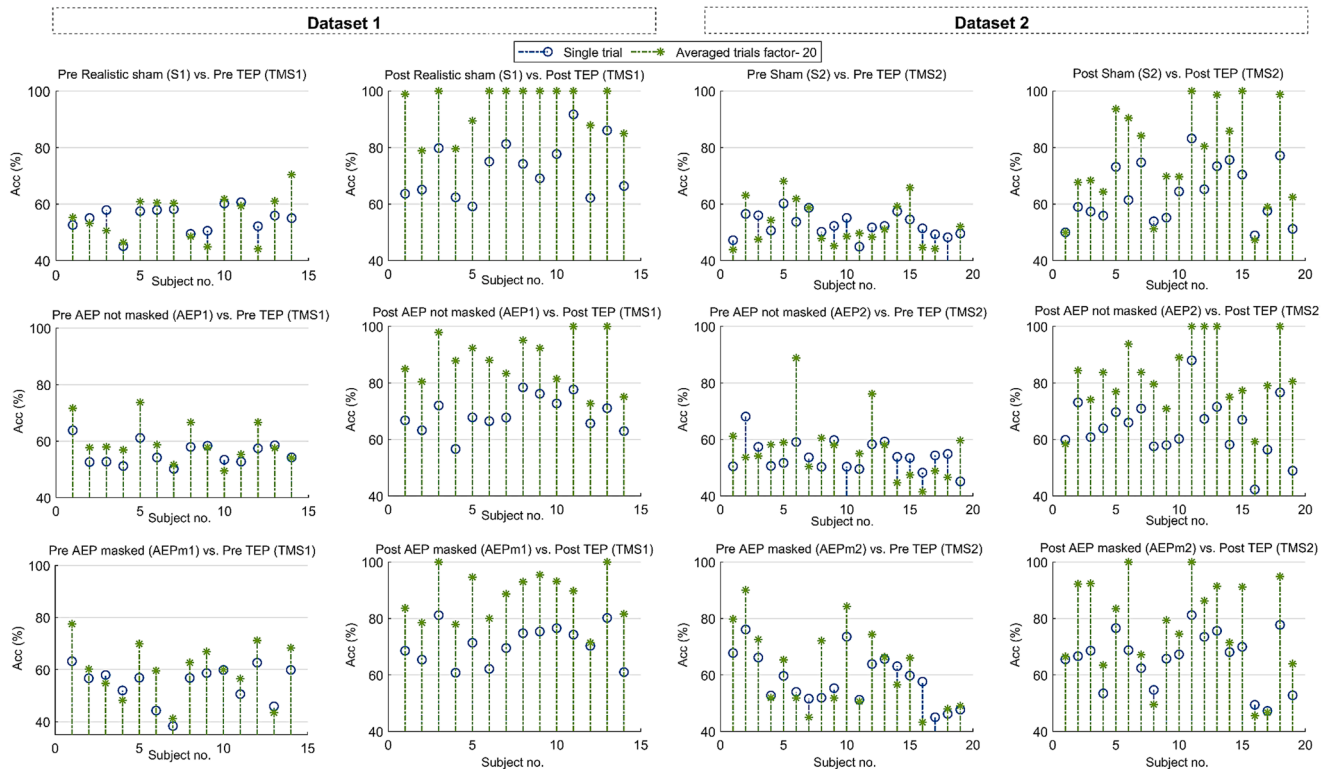


Fig. 5. Single subject confirmed group-level findings for pre- vs. pre- and post- vs. post-comparisons between TMS and shared sham conditions across datasets. Individual accuracy rates for single trial (blue circles) and the average of 20 trials (green asterisks) of pre-TMS vs. pre-sham (first and third column) and post-TMS vs. post-sham (second and fourth column) comparisons are shown for both datasets.

EEG responses to TMS and several sham stimulation conditions using two different TMS-EEG datasets. Our ML approach yielded: 1) at least moderate accuracy for post- vs. pre-TMS comparisons at the single trial level; 2) high accuracy for the post- vs. pre-TMS comparisons for 20 trials, which was higher than the correspondent post- vs. pre-sham comparisons; 3) low accuracy for comparisons between pre-TMS and pre-sham conditions, for both single trial and 20 trials; 4) moderate accuracy for post-TMS vs. post-sham comparisons at single trial and high accuracy for 20 trials. These findings were consistent across datasets as well as for single subject analyses, thus indicating that TEPs can be discerned from sham responses after just a few trials and at the single subject level.

The BiLSTM model achieved moderate accuracy in differentiating pre- from post-stimulus responses for TMS1 and TMS2 at the single trial level. As with conventional event-related potentials, TEPs typically require averaging across multiple trials to emerge reliably. Thus, while the present single trial findings suggest that our ML algorithm can extract informative spatio-temporal TEP features from individual TMS pulses, these moderate single trial accuracies (60-75 %) warrant cautious interpretation. Additional work is needed to characterize false positive and false-negative patterns, assess decision confidence, and define practical thresholds for experimental use at the single trial level. Importantly, the performance of the BiLSTM model increased markedly with averaging (i.e., from one to twenty trials), yielding high accuracy (>75 %) for both datasets, consistent with standard TMS-EEG acquisition practices. Whether achieving higher accuracy levels (e.g., >80 %) requires increasing the number of trials (e.g., $N > 40$) may critically depend on the overall quality of the TMS-EEG setup. Optimized TMS-EEG procedures, such as those employing the rt-TEP graphical user interface, individual TMS dosing (Casarotto et al., 2022), and appropriate auditory masking (Russo et al., 2022), substantially improve SNR and can yield accuracies above 80 % even after only a handful of trials (see Fig. 2).

Our ML algorithm reached high accuracy levels in differentiating pre- vs. post-stimulus for both TMS1 and TMS2. However, while TMS1 yielded 100 % accuracy, TMS2 reached ~80 % accuracy. The higher performance of the former is likely related to the between-dataset differences in experimental paradigms. TMS1 data were collected using a real-time TEP monitoring tool (rt-TEP), which allows obtaining TEPs with higher SNR (TEPs peak-to-peak amplitude $\geq 6 \mu V$ (Casarotto et al., 2022)), by optimizing TMS parameters (e.g., coil position, stimulation intensity) during TMS-EEG sessions while minimizing stimulation-related artifacts (e.g., scalp muscle activation) and also checking for effectiveness of noise-masking (Casarotto et al., 2022; Fecchio et al., 2025). In contrast, TMS2 data were acquired 1) without real-time monitoring and 2) with a TMS intensity set to 90 % RMT (Rocchi et al., 2021). Furthermore, consistent with the assumption that higher SNR in TMS1 vs. TMS2 was due to the TEPs, rather than to overall differences in TMS-EEG data quality acquisition between those sessions, we also performed surrogate data analysis (i.e., artificial TEPs created through random circular shift of the data) for TMS1 and TMS2, which yielded chance level accuracy for both datasets (Supp. Figure 1). These findings indicate that classifier performance was driven by post-stimulus signal structure rather than preprocessing artifacts or trivial temporal cues.

S1 and S2 showed lower accuracy levels compared to TMS1 and TMS2 respectively (Fig. 2). However, while S2 had low accuracy levels both for single trial and after averaging 20 trials, S1 yielded moderate to high accuracy levels. A possible explanation for this discrepancy is that S1, but not S2, delivered an electrical stimulation to superficial scalp muscles, and a significant activation of these muscles evoked a saliency-related multimodal response contaminating S1 (Mancuso et al., 2023). Although some scalp muscle activation after TMS is unavoidable, this finding suggests that, when collecting TMS-EEG data, it is important to minimize the activation of such muscles.

Concerning the other sham conditions shared across datasets, the

AEP and TEP not masked conditions yielded high accuracy in discerning post- vs pre-stimulus responses, while low accuracy was reported for AEP masked. Recent studies showed that AEP is a major source of contamination in TMS-EEG responses (Farzan et al., 2016; Pasiadis et al., 2023; Poorganji et al., 2021; Russo et al., 2022; Ter Braack et al., 2015), and there is an ongoing discussion about the best approach to mitigate its effects on TEPs (Belardinelli et al., 2019; Conde et al., 2019; Hernandez-Pavon et al., 2023; Pasiadis et al., 2023; Poorganji et al., 2021; Russo et al., 2022; Ter Braack et al., 2015). Here, AEP not masked generated a response that was detected with moderate to high accuracy; however, when applying a noise masking sound, the accuracy of BiLSTM dropped to low/chance level. This is an important confirmation that EEG responses due to the TMS click can be suppressed when noise masking is properly used (Belardinelli et al., 2019; Fecchio et al., 2025; Hernandez-Pavon et al., 2023).

Notably, the BiLSTM could discern with high accuracy TMS from sham conditions involving auditory and/or somatosensory stimulation in both TMS-EEG datasets. Differentiating TEPs due to direct cortical activation from other brain-evoked responses is a topic of great interest (Biabani et al., 2024; Gordon et al., 2021; Gordon et al., 2023a; Rocchi et al., 2021). Conde et al. found no significant differences in TEPs between sham and TMS sessions, leading the authors to propose that TEPs derive mostly from auditory and somatosensory responses, rather than from TMS direct neuronal cortical activation. Conversely, Gordon et al. found no evidence for interaction between TMS-EEG responses and sensory inputs in early TEPs (i.e., ≤ 100 ms) (Gordon et al., 2021; Gordon et al., 2023a). Two recent TMS-EEG studies, which also controlled for TMS-related click and somatosensory scalp activation (Rocchi et al., 2021) could reliably distinguish TEPs from sham responses up to 300 ms after TMS (Fecchio et al., 2025). Furthermore, a recent ML study showed that AEP-contaminated responses could be differentiated from TEPs (Cristofari et al., 2023). Here, we demonstrated that TEPs could be discerned from several sham conditions with high accuracy. A notable exception was TMSnm2, which involved TMS delivered without noise masking. The fact that TMS responses without noise masking were less discernable from (i.e., more similar to) real TMS with noise masking relative to other sham conditions is an important finding, further corroborating the notion that TMS does indeed evoke TEPs via direct cortical activation, rather than through the auditory pathway.

Along with high accuracy, group analyses showed high sensitivity (i.e., the ability to correctly identify a TMS condition), specificity (i.e., the ability to detect any sham condition), AUC (i.e., the degree of separability between TMS and sham conditions), and F1 score (i.e., the harmonic mean of precision, or % of TMS instances correctly identified, and recall, or % of TMS instances correctly classified). Together, the convergence of high accuracy, sensitivity, specificity, AUC, and F1 score indicates that the BiLSTM reliably distinguished active TMS from sham conditions. Nevertheless, the absence of an independent external test set represents an important limitation, and future studies employing nested cross-validation and multisite or external cohorts will be critical for establishing broader generalizability.

Single-subject findings largely mirrored group-level effects. Most individuals showed similar trends when comparing TMS1 and TMS2 vs. their baselines as well as vs. sham conditions. Our approach also revealed some across-subject variability in the responses, which likely reflects inter-individual differences in SNR and/or possible contamination of TEPs by sensory components. Given that this effect was detectable after just a handful of trials, our method offers a unique opportunity to characterize the quality of every TMS-EEG recording by evaluating the accuracy in detecting TEPs.

Collectively, these findings suggest that objective criteria for TMS-EEG studies can be established to ensure that TEPs are not substantially confounded by AEP or other artifacts. Our results further underscore the importance of effective noise masking and online monitoring of TMS efficacy and confounder contributions, facilitated by our ML-based framework, to assist TMS-EEG operators. This capability

supports the development of objective, data-driven tools for assessing cortical reactivity and neuromodulation effects. By providing a robust framework for verifying the authenticity of TEPs, the proposed approach holds potential translational value in clinical contexts where accurate characterization of cortical responsiveness is essential.

4.1. Limitations/Future directions

Future work is needed to determine whether these findings are reproducible across repeated recording sessions, different hardware configurations, and alternative preprocessing pipelines.

To begin with, multi-session and multi-center studies with larger TMS-EEG datasets will be critical for establishing robustness and generalizability. Future studies should also employ standardized pipelines to analyze TMS-EEG data to maximize comparability and replicability across datasets. While methodological consistency, including comparable stimulation parameters, recording setups, and data analysis steps, would facilitate replication and cross-site comparisons, broader validation across diverse experimental environments will ultimately confirm the reliability of the proposed approach. It is also important to point out that our ML model was not implemented or tested online. Thus, feasibility of integrating BiLSTM-based TEP verification into real-time tools, such as rt-TEP (Casarotto et al., 2022), will need to be assessed in future studies.

The present findings were obtained exclusively from stimulation of the primary motor cortex, which remains the most extensively characterized region in TMS-EEG research. As a result, it is currently unclear whether comparable levels of TEP vs. sham discriminability would be observed when stimulating other cortical areas, such as the dorsolateral prefrontal cortex, which exhibits different network dynamics and distinct sensory contamination profiles. Given the relevance of these regions for understanding the neurobiology of neuropsychiatric disorders (Donati et al., 2023; Kaskie and Ferrarelli, 2018), future studies are needed to determine whether the current results generalize to additional cortical targets and clinical populations. Of note, the present study included only healthy participants; thus, it remains to be determined whether the BiLSTM approach maintains comparable sensitivity, specificity, and overall robustness in clinical or neuropsychiatric populations, in which cortical excitability, signal-to-noise ratios, and artifact profiles may differ substantially. Future work extending this approach to patient populations and to stimulation of non-motor cortical regions (e.g., the dorsolateral prefrontal cortex) will be essential for establishing the broader clinical relevance and translational utility of ML-based TEP verification.

In the present study we utilized a sequence-to-sequence ML approach based on LSTM architectures (Graves et al., 2005; Hochreiter and Schmidhuber, 1997; Kudo et al., 1999; Salehinejad et al., 2017; Schuster and Paliwal, 1997; Yu et al., 2019). While this approach demonstrated robust performance on our datasets, future work, particularly on larger TMS-EEG datasets, should systematically compare benchmark BiLSTM with other ML methods, including attention-based networks and large language models (Fan et al., 2023; Kotei and Thirunavukarasu, 2023; Liu et al., 2023; Luong et al., 2015; Vaswani et al., 2017; Wen et al., 2022; Yun et al., 2019)). Such comparisons will help determine whether higher classification accuracy can be achieved and whether self-supervised or unsupervised ML strategies further improve differentiation between TEPs and sham responses. Nonetheless, the present findings demonstrate that LSTM-based models can achieve moderate to high accuracy between active TMS and multiple sham conditions.

Finally, auditory and somatosensory evoked potentials are known to contribute to TEPs (Du et al., 2017; Gordon et al., 2023b; Hernandez-Pavon et al., 2022; Mancuso et al., 2023). In addition to employing auditory noise masking and other experimental/data analysis approaches to mitigate these confounds (Hernandez-Pavon et al., 2022), future ML studies should quantify the extent to which each contamination source is present in the TEPs. Achieving this goal will require

large, specifically annotated TMS-EEG datasets enabling multilabel classification or the development of advanced unsupervised approaches capable of disentangling overlapping sensory components. As an initial step, the present findings demonstrate that TEPs elicited after active TMS can be reliably discerned from multiple sham stimulations, even after a handful of trials and at the single-subject level, using a BiLSTM ML framework.

Data and code availability

The data and code that support the findings of this study are available upon request.

Funding Source Declarations, Author Agreements/Declarations, or Permission Notes

This work was supported by the National Institute of Mental Health (NIMH) under grant R01MH125816 awarded to FF. MF was additionally supported by the Mass General Neuroscience Transformative Scholar Award in Brain Health and the Tiny Blue Dot Foundation.

All authors have reviewed and approved the final version of the manuscript.

CRediT authorship contribution statement

Ahmadreza Keihani: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Francesco L. Donati:** Writing – review & editing, Visualization, Investigation, Data curation, Conceptualization. **Simone Russo:** Writing – review & editing, Visualization, Validation, Methodology, Data curation, Conceptualization. **Sara Parmigiani:** Writing – review & editing, Visualization, Validation, Resources, Methodology, Investigation, Data curation, Conceptualization. **Michela Solbiati:** Writing – review & editing, Visualization, Validation, Investigation, Data curation. **Adenauer G. Casali:** Writing – review & editing, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Matteo Fecchio:** Writing – review & editing, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Omeed Chaichian:** Writing – review & editing, Visualization, Investigation, Data curation. **John Rothwell:** Writing – review & editing, Validation, Resources, Investigation, Data curation, Conceptualization. **Marcello Massimini:** Writing – review & editing, Visualization, Validation, Supervision, Project administration, Investigation, Conceptualization. **Lorenzo Rocchi:** Writing – review & editing, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Mario Rosanova:** Writing – review & editing, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Data curation, Conceptualization. **Fabio Ferrarelli:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

S.R. is the Chief Medical Officer of Manava Plus. Marcello Massimini is co-founder and shareholder of Intrinsic Powers, a spin-off of the University of Milan. Mario Rosanova is the advisor of Intrinsic Powers. These affiliations in no way affect the content of this article.

Acknowledgments

This work was supported by R01MH125816 NIMH grant awarded to

FF. MF was supported by the Mass General Neuroscience Transformative Scholar Award in Brain Health and the Tiny Blue Dot Foundation.

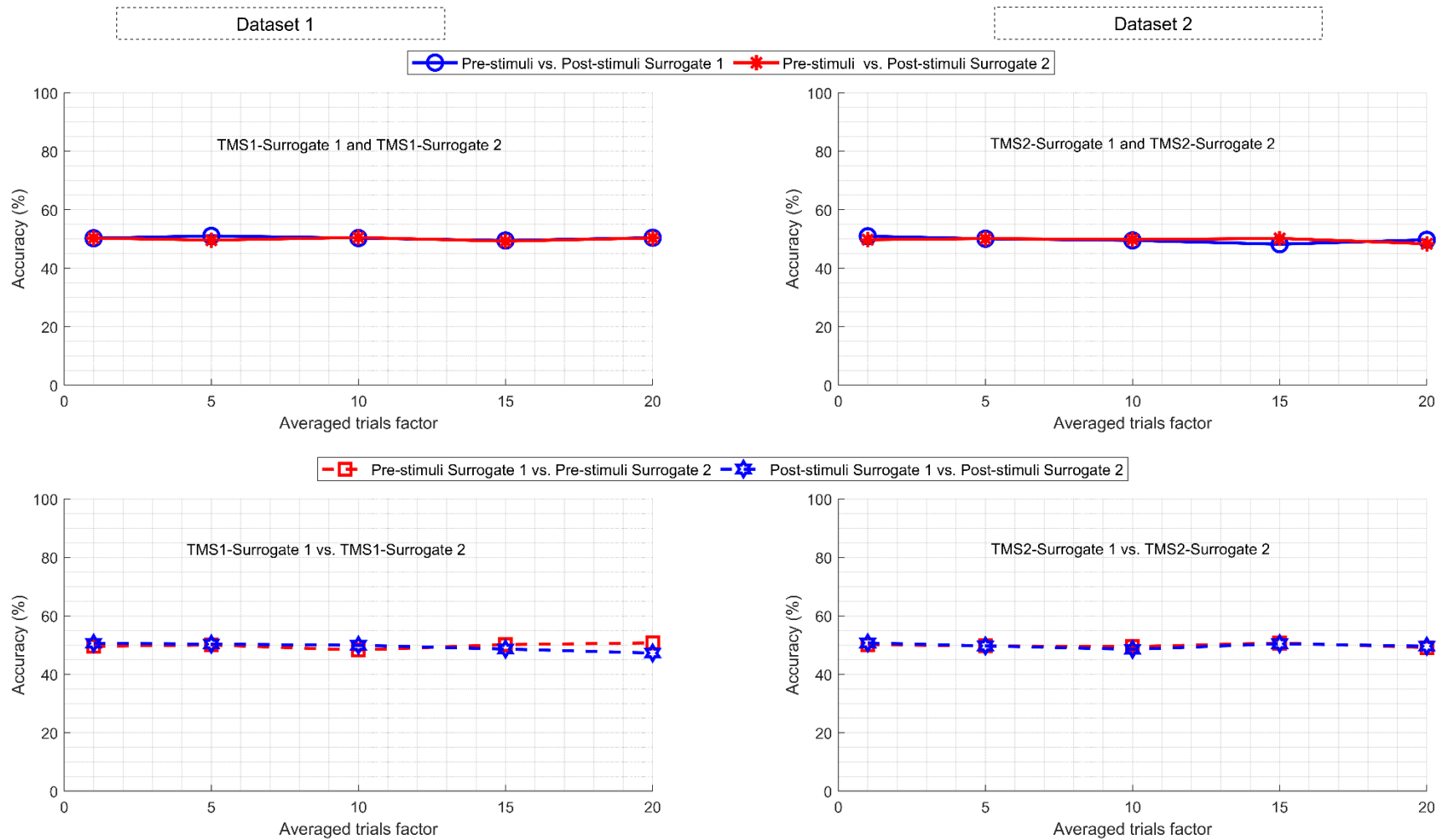
Supplementary materials

Supplementary material associated with this article can be found, in the online version, at [doi:10.1016/j.neuroimage.2026.121802](https://doi.org/10.1016/j.neuroimage.2026.121802).

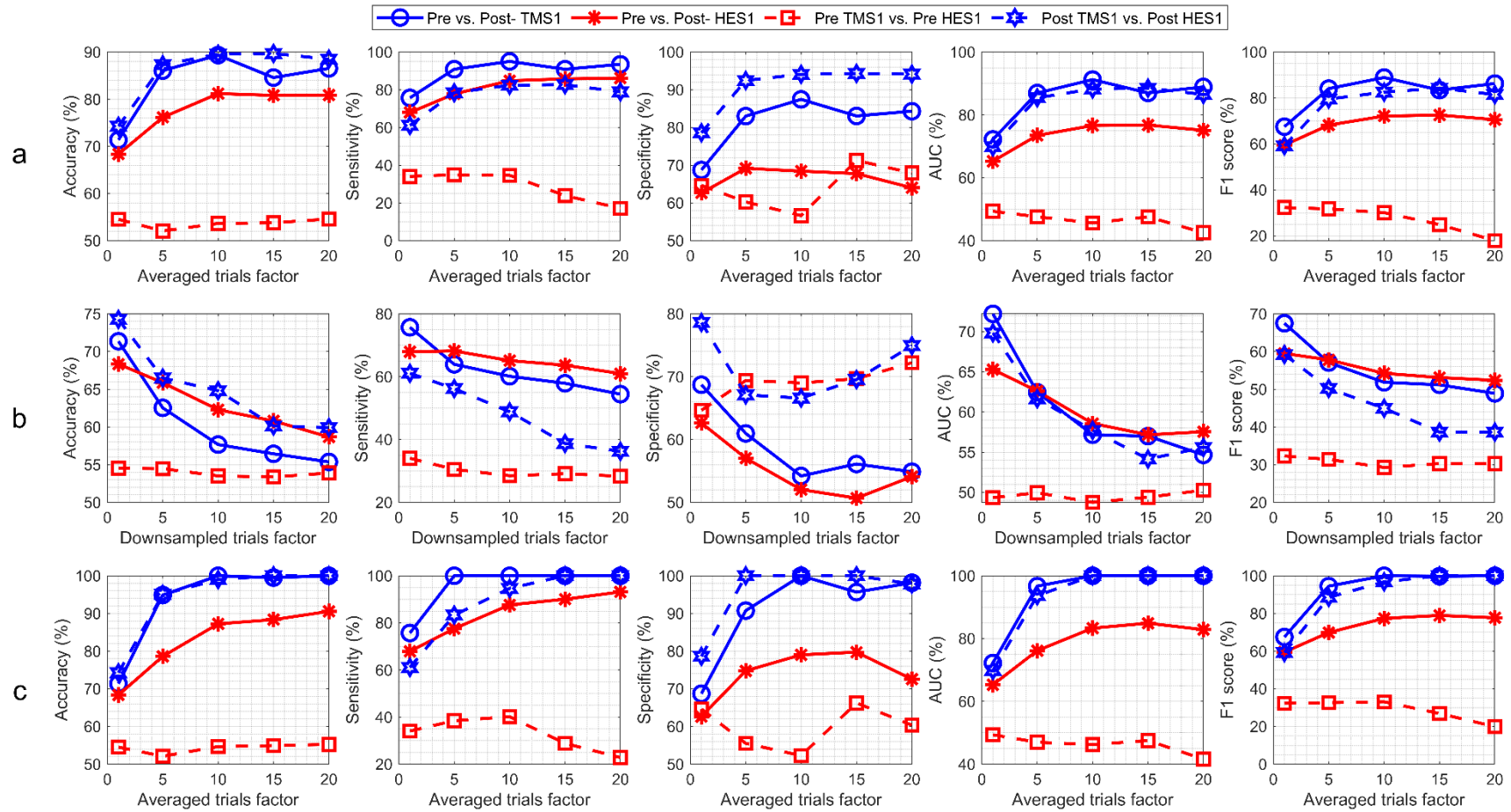
References

- Belardinelli, P., Biabani, M., Blumberger, D.M., Bortoletto, M., Casarotto, S., David, O., Desideri, D., Etkin, A., Ferrarelli, F., Fitzgerald, P.B., 2019. Reproducibility in TMS-EEG studies: A call for data sharing, standard procedures and effective experimental control. *Brain Stimul.: Basic Transl. Clin. Res. Neuromodulation* 12, 787–790. <https://doi.org/10.1016/j.brs.2019.01.010>.
- Biabani, M., Fornito, A., Goldsworthy, M., Thompson, S., Graetz, L., Semmler, J.G., Opie, G.M., Bellgrove, M.A., Rogasch, N.C., 2024. Characterising the contribution of auditory and somatosensory inputs to TMS-evoked potentials following stimulation of prefrontal, premotor, and parietal cortex. *Imaging Neurosci.* 2, 1–23. https://doi.org/10.1162/imag_a.00349.
- Bortoletto, M., Veniero, D., Thut, G., Miniussi, C., 2015. The contribution of TMS-EEG coregistration in the exploration of the human cortical connectome. *Neurosci. Biobehav. Rev.* 49, 114–124. <https://doi.org/10.1016/j.neubiorev.2014.12.014>.
- Casali, A.G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K.R., Casarotto, S., Bruno, M.-A., Laureys, S., Tononi, G., 2013. A theoretically based index of consciousness independent of sensory processing and behavior. *Sci. Transl. Med.* 5, <https://doi.org/10.1126/scitranslmed.3006294>, 198ra105–198ra105.
- Casarotto, S., Fecchio, M., Rosanova, M., Varone, G., D'Ambrosio, S., Sarasso, S., Pigorini, A., Russo, S., Comanducci, A., Ilmoniemi, R.J., 2022. The rt-TEP tool: real-time visualization of TMS-Evoked Potentials to maximize cortical activation and minimize artifacts. *J. Neurosci. Methods* 370, 109486. <https://doi.org/10.1016/j.jneumeth.2022.109486>.
- Conde, V., Tomasevic, L., Akopian, I., Stanek, K., Saturnino, G.B., Thielscher, A., Bergmann, T.O., Siebner, H.R., 2019. The non-transcranial TMS-evoked potential is an inherent source of ambiguity in TMS-EEG studies. *Neuroimage* 185, 300–312. <https://doi.org/10.1016/j.neuroimage.2018.10.052>.
- Cristofari, A., De Santis, M., Lucidi, S., Rothwell, J., Casula, E.P., Rocchi, L., 2023. Machine Learning-Based Classification to Disentangle EEG Responses to TMS and Auditory Input. *Brain Sci.* 13, 866. <https://doi.org/10.3390/brainsci13060866>.
- Delorme, A., Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* 134, 9–21. <https://doi.org/10.1016/j.jneumeth.2003.10.009>.
- Donati, F.L., Mayeli, A., Sharma, K., Janssen, S.A., Lagoy, A.D., Casali, A.G., Ferrarelli, F., 2023. Natural Oscillatory Frequency Slowing in the Premotor Cortex of Early-Course Schizophrenia Patients: A TMS-EEG Study. *Brain Sci.* 13, 534. <https://doi.org/10.3390/brainsci13040534>.
- Du, X., Choa, F.-S., Summerfelt, A., Rowland, L.M., Chiappelli, J., Kochunov, P., Hong, L. E., 2017. N100 as a generic cortical electrophysiological marker based on decomposition of TMS-evoked potentials across five anatomic locations. *Exp. Brain Res.* 235, 69–81. <https://doi.org/10.1007/s00221-016-4773-7>.
- Fan, L., Li, L., Ma, Z., Lee, S., Yu, H., Hemphill, L., 2023. A bibliometric review of large language models research from 2017 to 2023. *arXiv preprint arXiv:2304.02020*. <https://doi.org/10.48550/arXiv.2304.02020>.
- Farzan, F., Vernet, M., Shafi, M.M., Rotenberg, A., Daskalakis, Z.J., Pascual-Leone, A., 2016. Characterizing and modulating brain circuitry through transcranial magnetic stimulation combined with electroencephalography. *Front. Neural Circuits* 10, 73. <https://doi.org/10.3389/fncir.2016.00073>.
- Fecchio, M., Russo, S., Couto, B.A.N., Mikulan, E., Pigorini, A., Furegoni, G., Hassan, G., D'Ambrosio, S., Solbiati, M., Viganò, A., Parmigiani, S., Sarasso, S., Casarotto, S., Massimini, M., Casali, A.G., Rosanova, M., 2025. The specific spatiotemporal evolution of TMS-evoked potentials reflects the engagement of cortical circuits. *bioRxiv*. <https://doi.org/10.1101/2025.06.25.661535>, 2025.06.25.661535.
- Gemein, L.A., Schirmer, R.T., Chrabaszcz, P., Wilson, D., Boedeker, J., Schulze-Bonhage, A., Hutter, F., Ball, T., 2020. Machine-learning-based diagnostics of EEG pathology. *Neuroimage* 220, 117021. <https://doi.org/10.1016/j.neuroimage.2020.117021>.
- Gordon, P.C., Desideri, D., Belardinelli, P., Zrenner, C., Ziemann, U., 2018. Comparison of cortical EEG responses to realistic sham versus real TMS of human motor cortex. *Brain Stimul.* 11, 1322–1330. <https://doi.org/10.1016/j.brs.2018.08.003>.
- Gordon, P.C., Jovellar, D.B., Song, Y., Zrenner, C., Belardinelli, P., Siebner, H.R., Ziemann, U., 2021. Recording brain responses to TMS of primary motor cortex by EEG—utility of an optimized sham procedure. *Neuroimage* 245, 118708. <https://doi.org/10.1016/j.neuroimage.2021.118708>.
- Gordon, P.C., Song, Y., Jovellar, B., Belardinelli, P., Ziemann, U., 2023a. No evidence for interaction between TMS-EEG responses and sensory inputs. *Brain Stimul.: Basic Transl. Clin. Res. Neuromodulation* 16, 25–27. <https://doi.org/10.1016/j.brs.2022.12.010>.
- Gordon, P.C., Song, Y.F., Jovellar, D.B., Rostami, M., Belardinelli, P., Ziemann, U., 2023b. Untangling TMS-EEG responses caused by TMS versus sensory input using optimized sham control and GABAergic challenge. *J. Physiol.* 601, 1981–1998. <https://doi.org/10.1113/jp283986>.
- Graves, A., Fernández, S., Schmidhuber, J., 2005. Bidirectional LSTM networks for improved phoneme classification and recognition. *International conference on artificial neural networks*. Springer, pp. 799–804. https://doi.org/10.1007/11550907_163.
- Hernandez-Pavon, J.C., Kugiumtzis, D., Zrenner, C., Kimiskidis, V.K., Metsomaa, J., 2022. Removing artifacts from TMS-evoked EEG: A methods review and a unifying theoretical framework. *J. Neurosci. Methods* 376, 109591. <https://doi.org/10.1016/j.jneumeth.2022.109591>.
- Hernandez-Pavon, J.C., Veniero, D., Bergmann, T.O., Belardinelli, P., Bortoletto, M., Casarotto, S., Casula, E.P., Farzan, F., Fecchio, M., Julkunen, P., 2023. TMS combined with EEG: Recommendations and open issues for data collection and analysis. *Brain Stimul.* <https://doi.org/10.1016/j.brs.2023.02.009>.
- Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. *Neural Comput.* 9, 1735–1780.
- Hosseini, M.-P., Hosseini, A., Ahi, K., 2020. A review on machine learning for EEG signal processing in bioengineering. *IEEE Rev. Biomed. Eng.* 14, 204–218. <https://doi.org/10.1109/RBME.2020.2969915>.
- Kaskie, R.E., Ferrarelli, F., 2018. Investigating the neurobiology of schizophrenia and other major psychiatric disorders with Transcranial Magnetic Stimulation. *Schizophr. Res.* 192, 30–38. <https://doi.org/10.1016/j.schres.2017.04.045>.
- Keihani, A., Mohammadi, A.M., Marzbani, H., Nafissi, S., Haidari, M.R., Jafari, A.H., 2022a. Sparse representation of brain signals offers effective computation of corticomuscular coupling value to predict the task-related and non-task sEMG channels: A joint hEEG-sEMG study. *PLoS One* 17, e0270757. <https://doi.org/10.1371/journal.pone.0270757>.
- Keihani, A., Sajadi, S.S., Hasani, M., Ferrarelli, F., 2022b. Bayesian optimization of machine learning classification of resting-state EEG microstates in schizophrenia: a proof-of-concept preliminary study based on secondary analysis. *Brain Sci.* 12, 1497. <https://doi.org/10.3390/brainsci12111497>.
- Kingma, D.P., Ba, J., 2014. Adam: Method Stoch. Optim. *arXiv preprint arXiv:1412.6980*.
- Kotei, E., Thirunavukarasu, R., 2023. A Systematic Review of Transformer-Based Pre-Trained Language Models through Self-Supervised Learning. *Information* 14, 187. <https://doi.org/10.48550/arXiv.1412.6980>.
- Kudo, M., Toyama, J., Shimbo, M., 1999. Multidimensional curve classification using passing-through regions. *Pattern Recognit. Lett.* 20, 1103–1111. [https://doi.org/10.1016/S0167-8655\(99\)00077-X](https://doi.org/10.1016/S0167-8655(99)00077-X).
- Li, J., Gao, M., D'Agostino, R., 2019. Evaluating classification accuracy for modern learning approaches. *Stat. Med.* 38, 2477–2503. <https://doi.org/10.1002/sim.8103>.
- Liu, X., McDuff, D., Kovacs, G., Galatzer-Levy, I., Sunshine, J., Zhan, J., Poh, M.-Z., Liao, S., Di Achille, P., Patel, S., 2023. Large Lang. Models are Few-Shot Health Learn. *arXiv preprint arXiv:2305.15525*.
- Luong, M.-T., Pham, H., Manning, C.D., 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*. <https://doi.org/10.48550/arXiv.2305.15525>.
- Mancuso, M., Cruciani, A., Sveva, V., Casula, E., Brown, K., Rothwell, J., Di Lazzaro, V., Koch, G., Rocchi, L., 2023. Somatosensory input in the context of transcranial magnetic stimulation coupled with electroencephalography: An evidence-based overview. *Neurosci. Biobehav. Rev.* 155, 105434. <https://doi.org/10.1016/j.neubiorev.2023.105434>.
- Manganotti, P., Acler, M., Masiero, S., Del Felice, A., 2015. TMS-evoked N100 responses as a prognostic factor in acute stroke. *Funct. Neurol.* 30, 125. <https://doi.org/10.11138/fneur/2015.30.2.125>.
- Massimini, M., Ferrarelli, F., Sarasso, S., Tononi, G., 2012. Cortical mechanisms of loss of consciousness: insight from TMS/EEG studies. *Arch. Ital. Biol.* 150, 44–55. <https://doi.org/10.4449/aib.v150i2.1361>.
- Momi, D., Wang, Z., Griffiths, J.D., 2023. TMS-evoked responses are driven by recurrent large-scale network dynamics. *Elife* 12. <https://doi.org/10.7554/eLife.83232>.
- Mwata-Velu, T.Y., Avina-Cervantes, J.G., Cruz-Duarte, J.M., Rostro-Gonzalez, H., Ruiz-Pinales, J., 2021. Imaginary finger movements decoding using empirical mode decomposition and a stacked BiLSTM architecture. *Mathematics* 9, 3297. <https://doi.org/10.3390/math9243297>.
- Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.-M., 2011. FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.* 2011, 156869. <https://doi.org/10.1155/2011/156869>.
- Pastiadis, K., Vlachos, I., Chatzikiriakou, E., Roth, Y., Zibman, S., Zangen, A., Kugiumtzis, D., Kimiskidis, V.K., 2023. Auditory Fine-Tuned Suppressor of TMS-Clicks (TMS-Click AFTS): A Novel, Perceptually Driven/Tuned Approach for the Reduction in AEP Artifacts in TMS-EEG Studies. *Appl. Sci.* 13, 1047. <https://doi.org/10.3390/app13021047>.
- Poorganji, M., Zomorodi, R., Hawco, C., Hill, A.T., Hadas, I., Rajji, T.K., Chen, R., Voineskos, D., Daskalakis, A.A., Blumberger, D.M., 2021. Differentiating transcranial magnetic stimulation cortical and auditory responses via single pulse and paired pulse protocols: A TMS-EEG study. *Clin. Neurophysiol.* 132, 1850–1858. <https://doi.org/10.1016/j.clinph.2021.05.009>.
- Rocchi, L., Di Santo, A., Brown, K., Ibáñez, J., Casula, E., Rawji, V., Di Lazzaro, V., Koch, G., Rothwell, J., 2021. Disentangling EEG responses to TMS due to cortical and peripheral activations. *Brain Stimul.* 14, 4–18. <https://doi.org/10.1016/j.brs.2020.10.011>.
- Rogasch, N.C., Sullivan, C., Thomson, R.H., Rose, N.S., Bailey, N.W., Fitzgerald, P.B., Farzan, F., Hernandez-Pavon, J.C., 2017. Analysing concurrent transcranial magnetic stimulation and electroencephalographic data: A review and introduction to the open-source TESA software. *Neuroimage* 147, 934–951. <https://doi.org/10.1016/j.neuroimage.2016.10.031>.
- Ross, J.M., Sarkar, M., Keller, C.J., 2022. Experimental suppression of transcranial magnetic stimulation-electroencephalography sensory potentials. *Hum. Brain Mapp.* 43, 5141–5153. <https://doi.org/10.1002/hbm.25990>.
- Russo, S., Claar, L., Marks, L., Krishnan, G., Furegoni, G., Zauli, F.M., Hassan, G., Solbiati, M., d'Orto, P., Mikulan, E., 2024. Thalamic feedback shapes brain responses

- evoked by cortical stimulation in mice and humans. *bioRxiv*. <https://doi.org/10.1101/2024.01.31.578243>.
- Russo, S., Sarasso, S., Puglisi, G.E., Dal Palù, D., Pigorini, A., Casarotto, S., D'Ambrosio, S., Astolfi, A., Massimini, M., Rosanova, M., 2022. TAAC-TMS Adaptable Auditory Control: A universal tool to mask TMS clicks. *J. Neurosci. Methods* 370, 109491. <https://doi.org/10.1016/j.jneumeth.2022.109491>.
- Salehinejad, H., Sankar, S., Barfett, J., Colak, E., Valaee, S., 2017. Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078*. doi: <https://doi.org/10.48550/arXiv.1801.01078>.
- Schuster, M., Paliwal, K.K., 1997. Bidirectional recurrent neural networks. *IEEE trans. Signal Process.* 45, 2673–2681. <https://doi.org/10.1109/78.650093>.
- Ter Braack, E.M., de Vos, C.C., van Putten, M.J., 2015. Masking the auditory evoked potential in TMS-EEG: a comparison of various methods. *Brain Topogr.* 28, 520–528. <https://doi.org/10.1007/s10548-013-0312-z>.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems* 30. doi: <https://doi.org/10.48550/arXiv.1706.03762>.
- Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., Sun, L., 2022. Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*. doi: <https://doi.org/10.48550/arXiv.2202.07125>.
- Yu, Y., Si, X., Hu, C., Zhang, J., 2019. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput.* 31, 1235–1270. https://doi.org/10.1162/neco_a_01199.
- Yun, C., Bhojanapalli, S., Rawat, A.S., Reddi, S.J., Kumar, S., 2019. Are transformers universal approximators of sequence-to-sequence functions? *arXiv preprint arXiv:1912.10077*. doi: <https://doi.org/10.48550/arXiv.1912.10077>.
- Zhang, Y., Fu, Z., 2020. The study of EEG Recognition of Depression on Bi-LSTM based on ERP P300. *E3S Web of Conferences*. EDP Sci., 02007 <https://doi.org/10.1051/e3sconf/202018502007>.

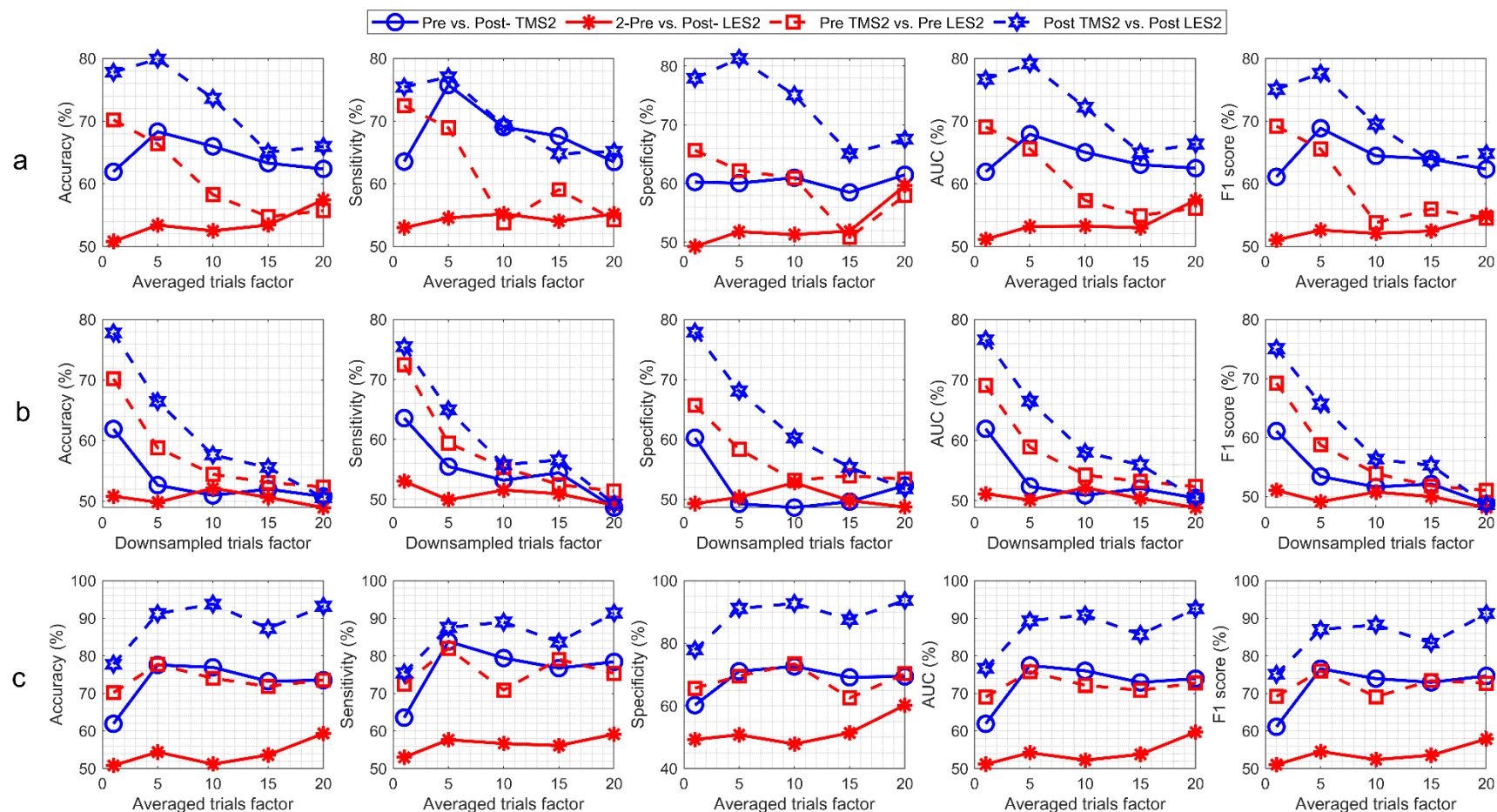


Supplementary Figure 1: Surrogate analysis results. TMS1 and TMS2 surrogate analyses showed chance-level accuracy with both single trial and average of 20 trials analyses.



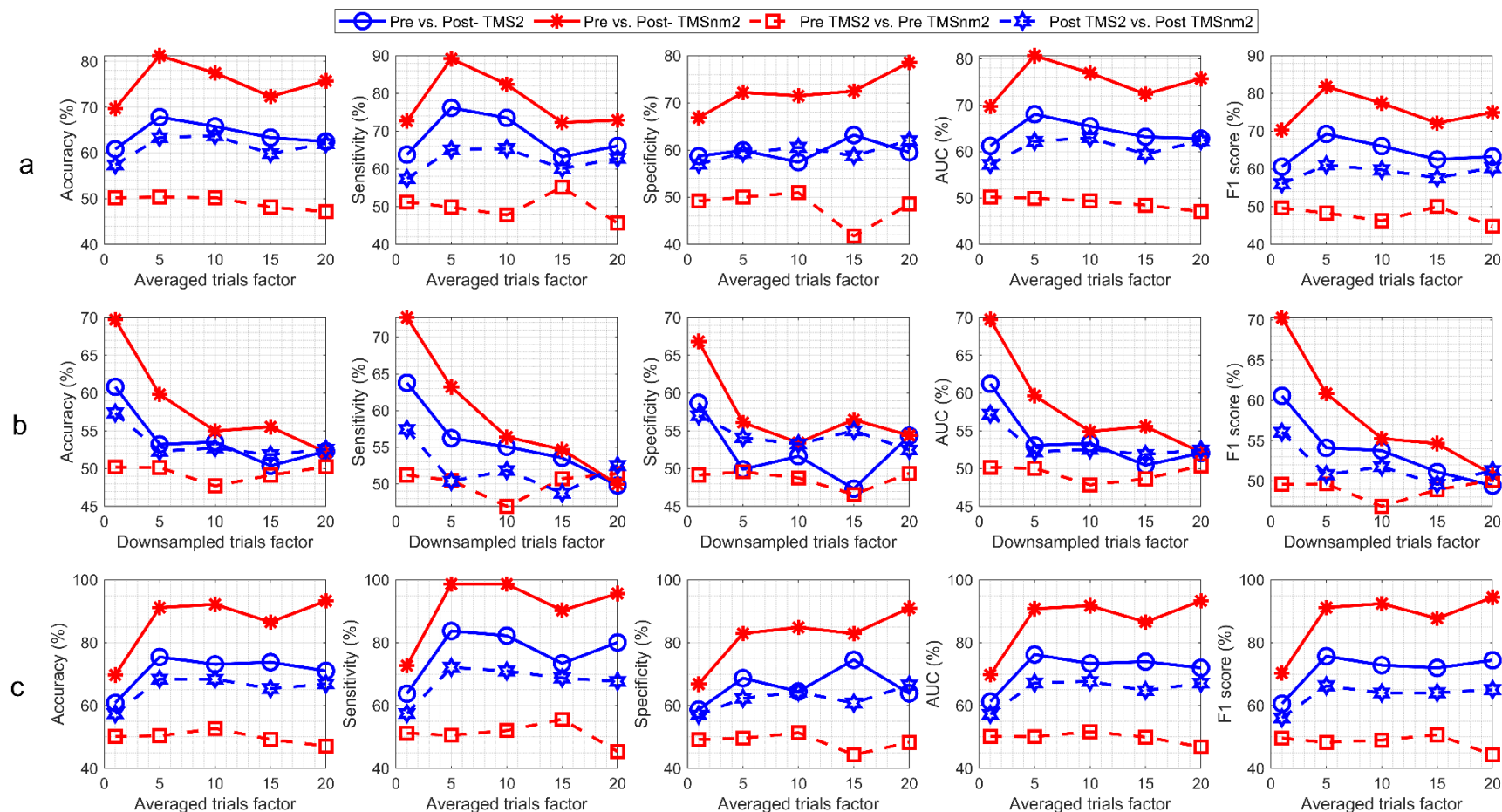
Supplementary Figure 2: Model LOO results for TMS1vs. HES1 for Dataset 1. From left to right, accuracy, sensitivity, specificity, AUC, and F1-score are shown. Panel (a) displays the group validation results for single trial and averaged trials (i.e., $n=5$, etc.) in between- and within-session comparisons. Panel (b) displays the same results after accounting for each

down-sampled trials factor (i.e., $n=5$, etc.) in between- and within-session comparisons. Panel (c) shows the adjusted values, after accounting for the reduction in the total number of trials due to the averaged trials factor (i.e., $n=5$, etc.).



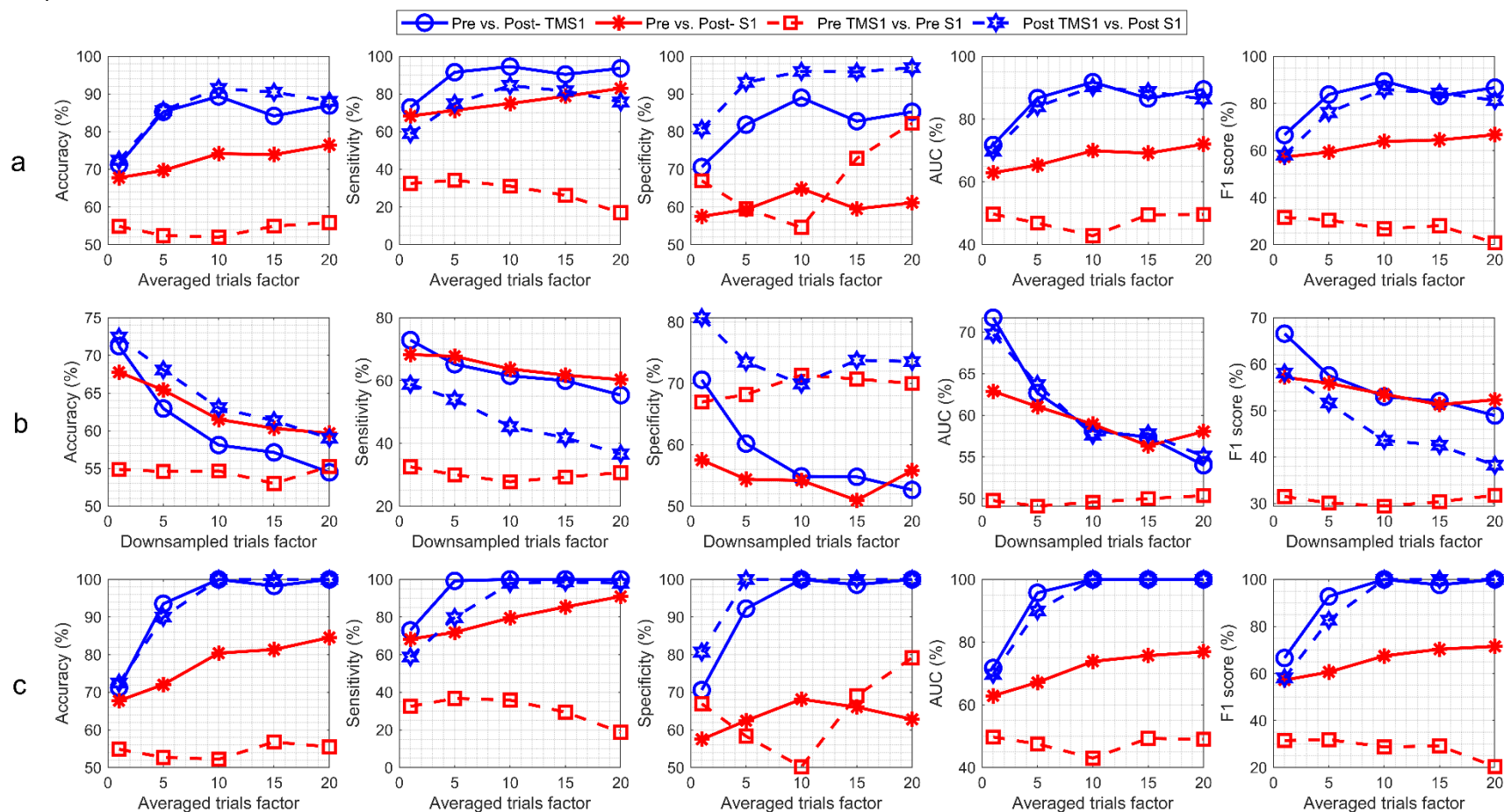
Supplementary Figure 3: Model LOO results for single subjects for TMS2 vs. LES2 comparison for Dataset 2; the top panel shows the accuracy rates for single trial and averaged trials (i.e., $n=5$, etc.) in between- and within-session comparisons. The middle panel displays the same results after accounting for each down-sampled trials factor (i.e., $n=5$,

etc.), while panel (c) shows the adjusted values after accounting for the reduction in the total number of trials due to the averaged trials factor (i.e., $n=5$, etc.).

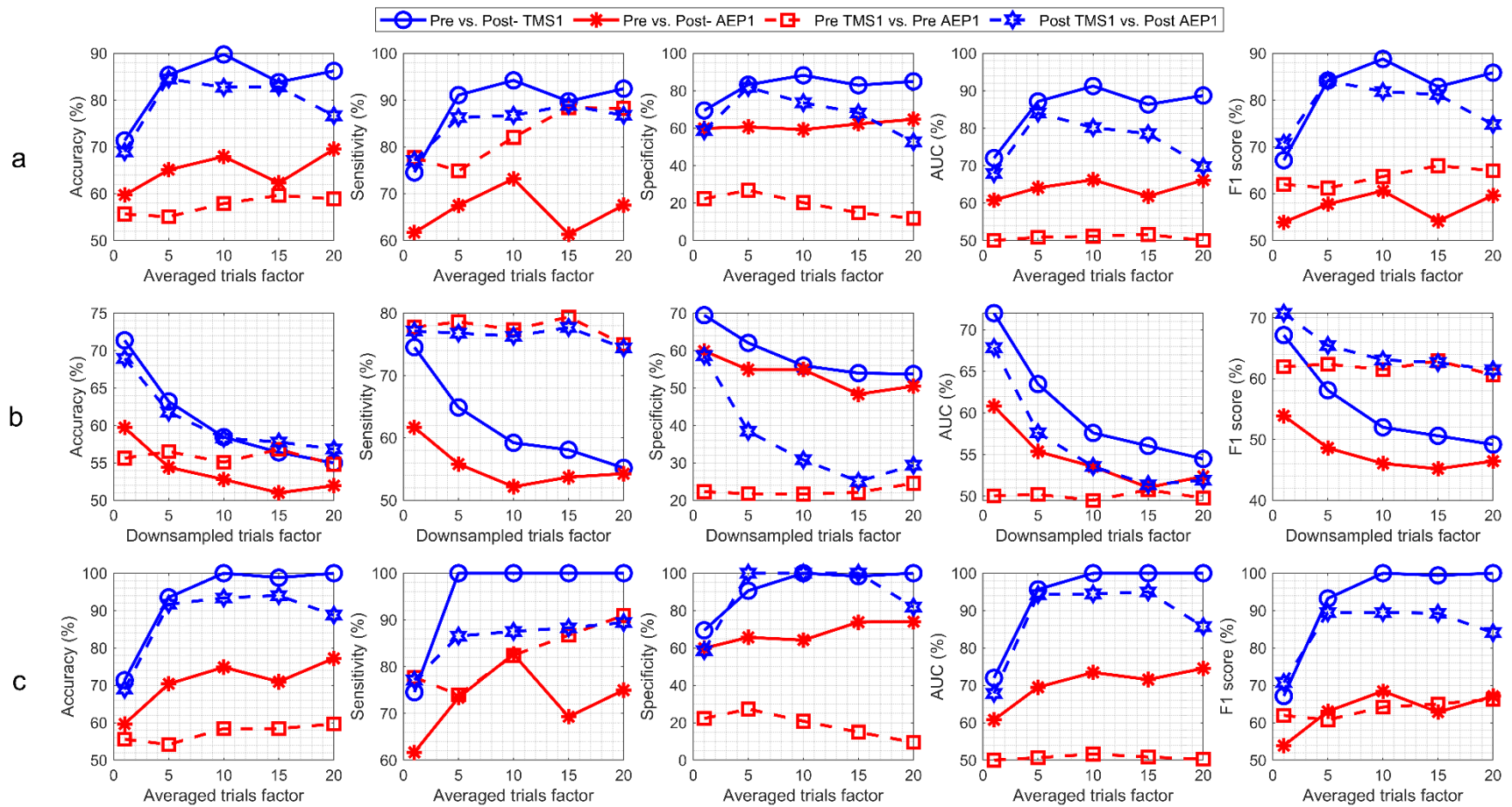


Supplementary Figure 4: Model LOO results for TMS2vs. TMSnm2 for Dataset 2. The top panel shows the accuracy rates for single trial and averaged trials (i.e., $n=5$, etc.) in between- and within-session comparisons. The middle panel displays the same results after accounting for each down-sampled trials factor (i.e., $n=5$, etc.), while panel (c) shows the

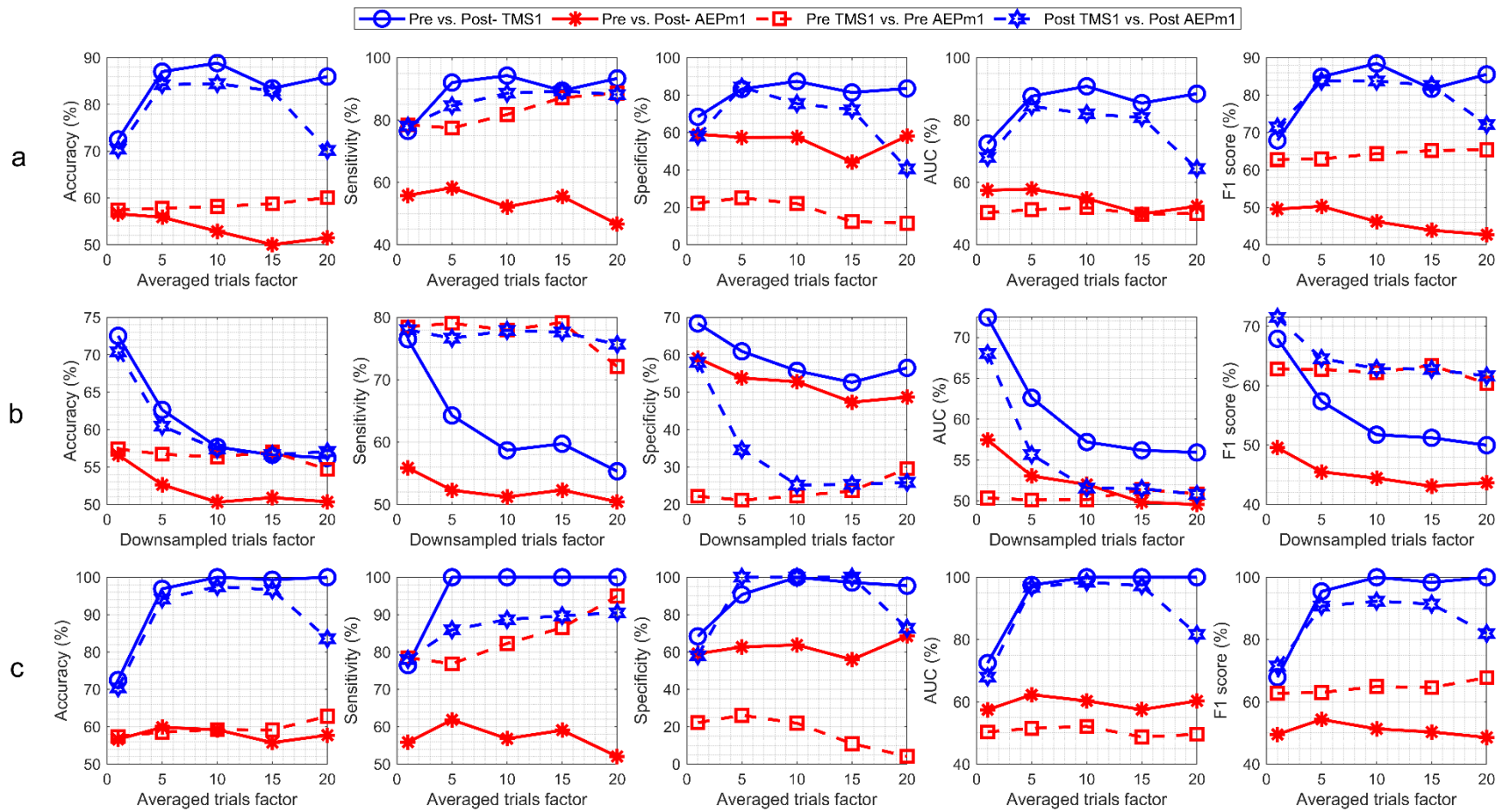
adjusted values after accounting for the reduction in the total number of trials due to the averaged trials factor (i.e., $n=5$, etc.).



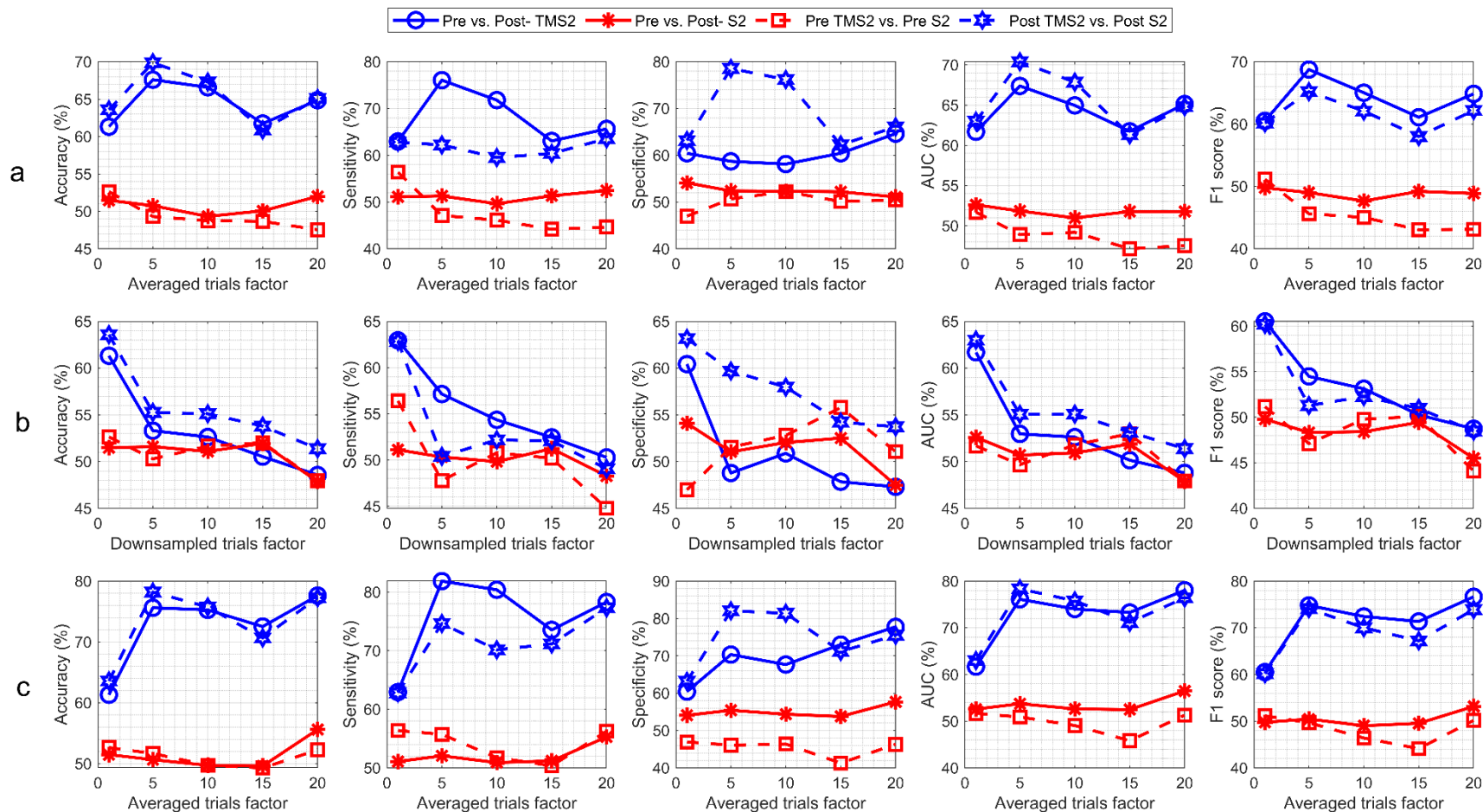
Supplementary Figure 5: Model LOO results for TMS1vs. S1 for Dataset 1.



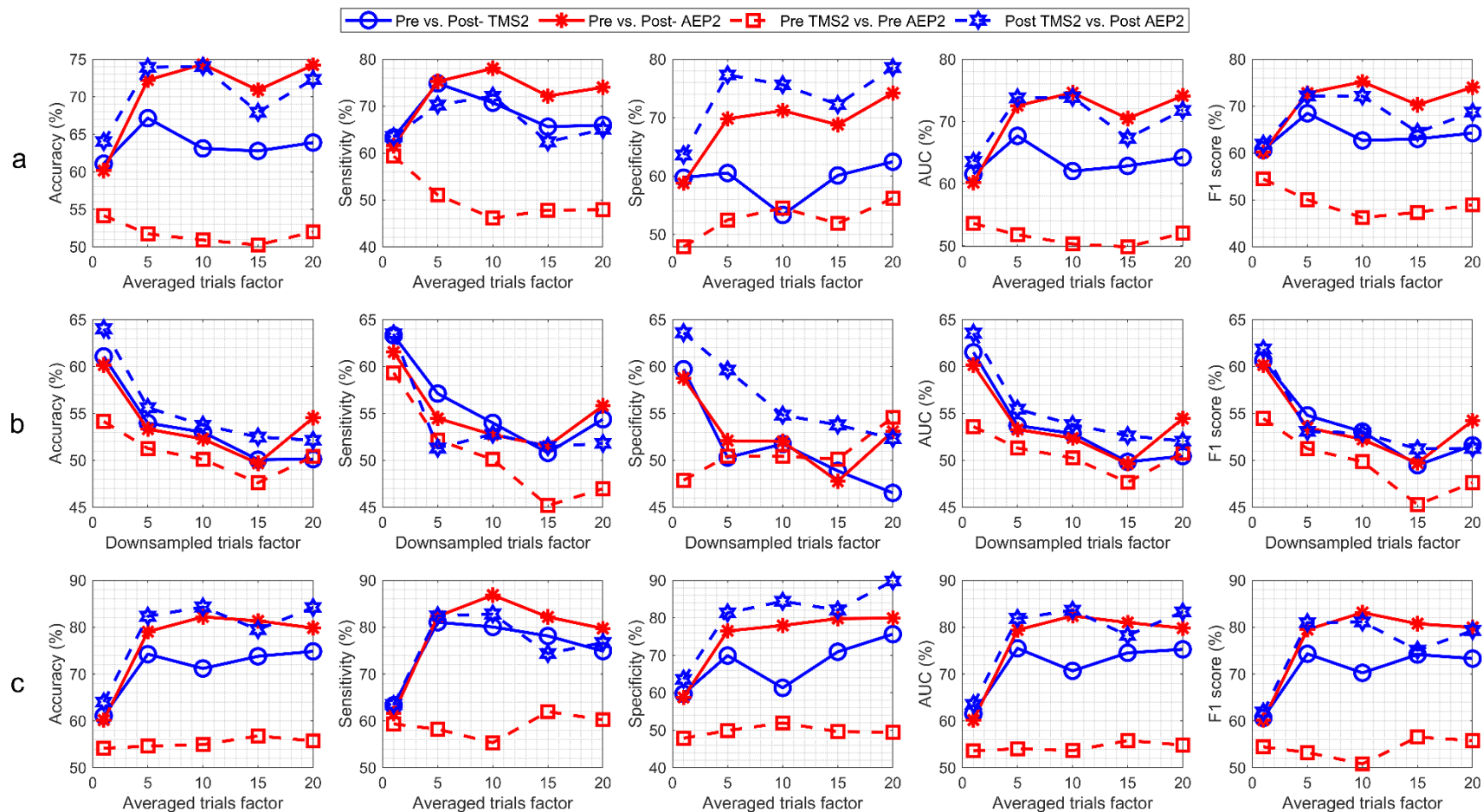
Supplementary Figure 6: Model LOO results for TMS1 vs. AEP1 for Dataset 1.



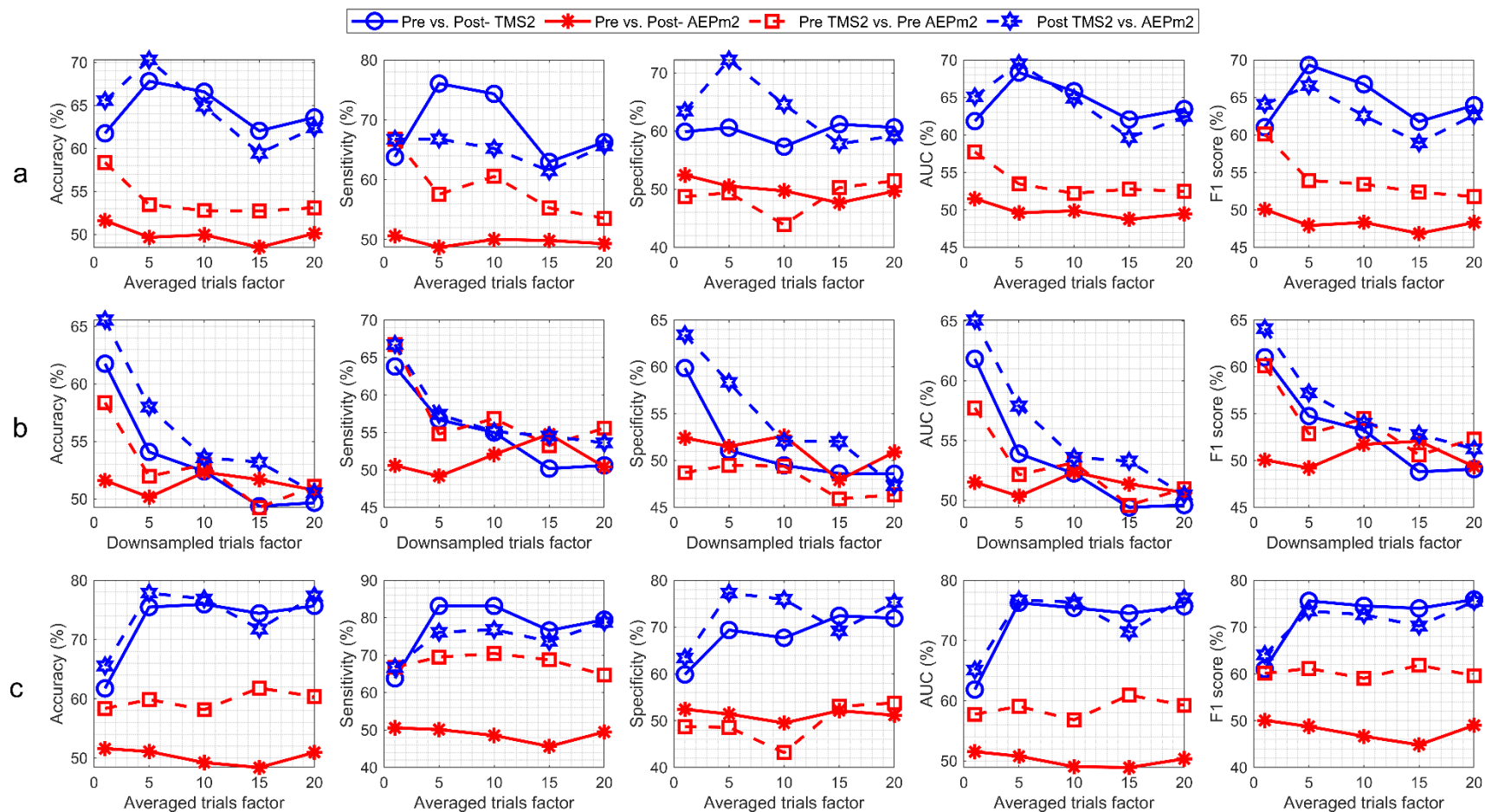
Supplementary Figure 7: Model LOO results for TMS1 vs. AEPm1 for Dataset 1.



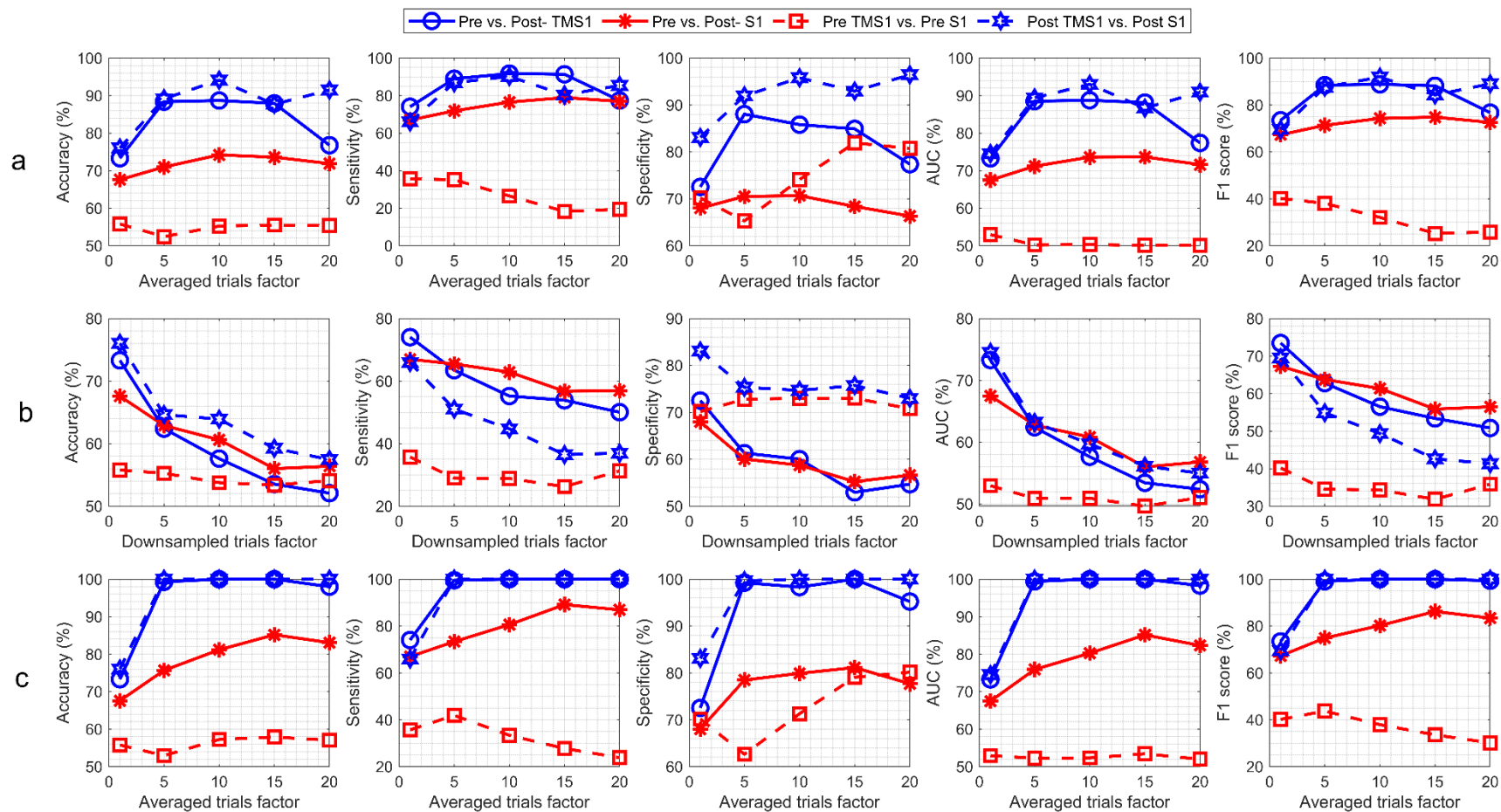
Supplementary Figure 8: Model LOO results for TMS2 vs. S2 for Dataset 2.



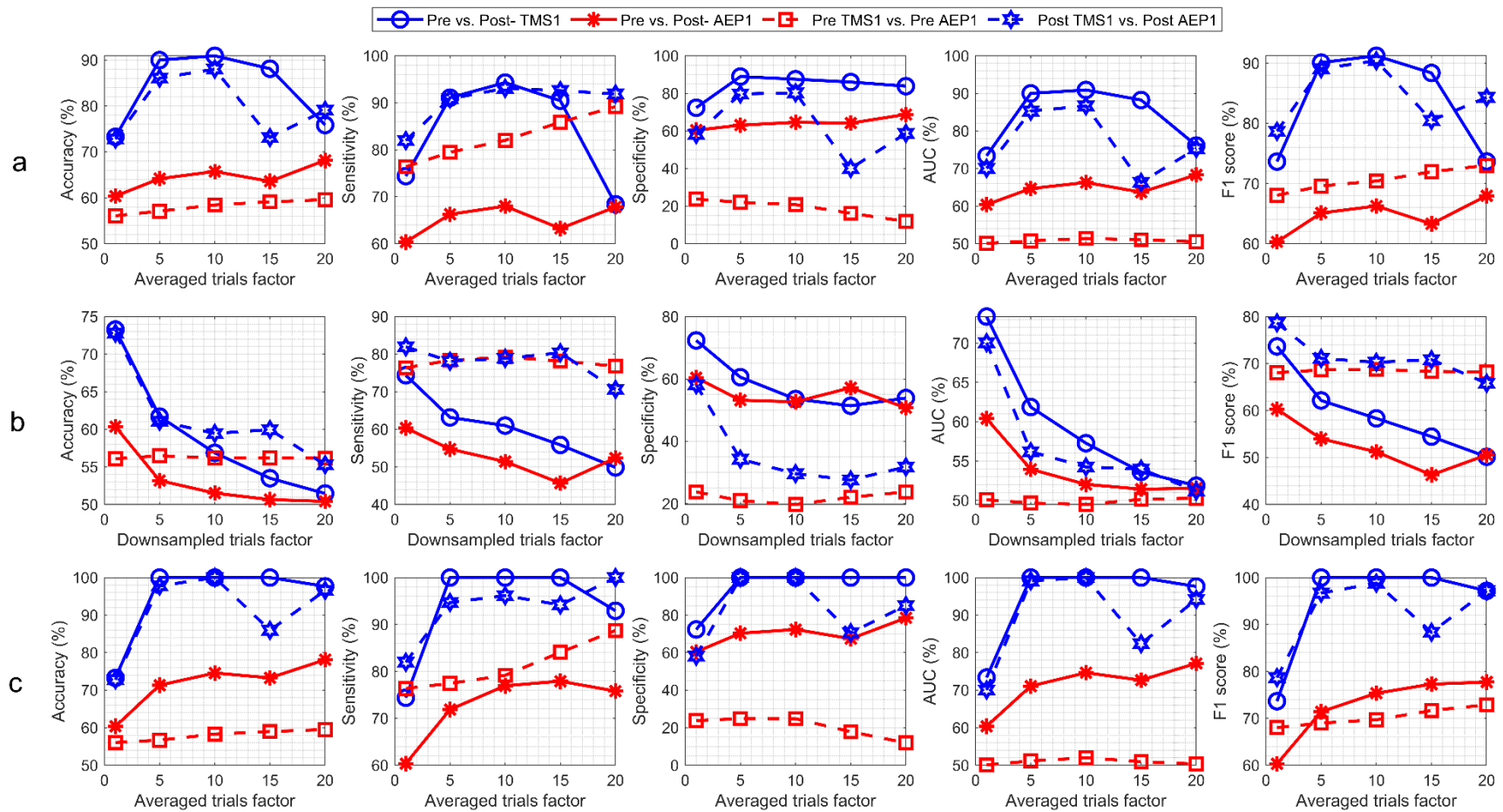
Supplementary Figure 9: Model LOO results for TMS2vs. AEP2 for Dataset 2.



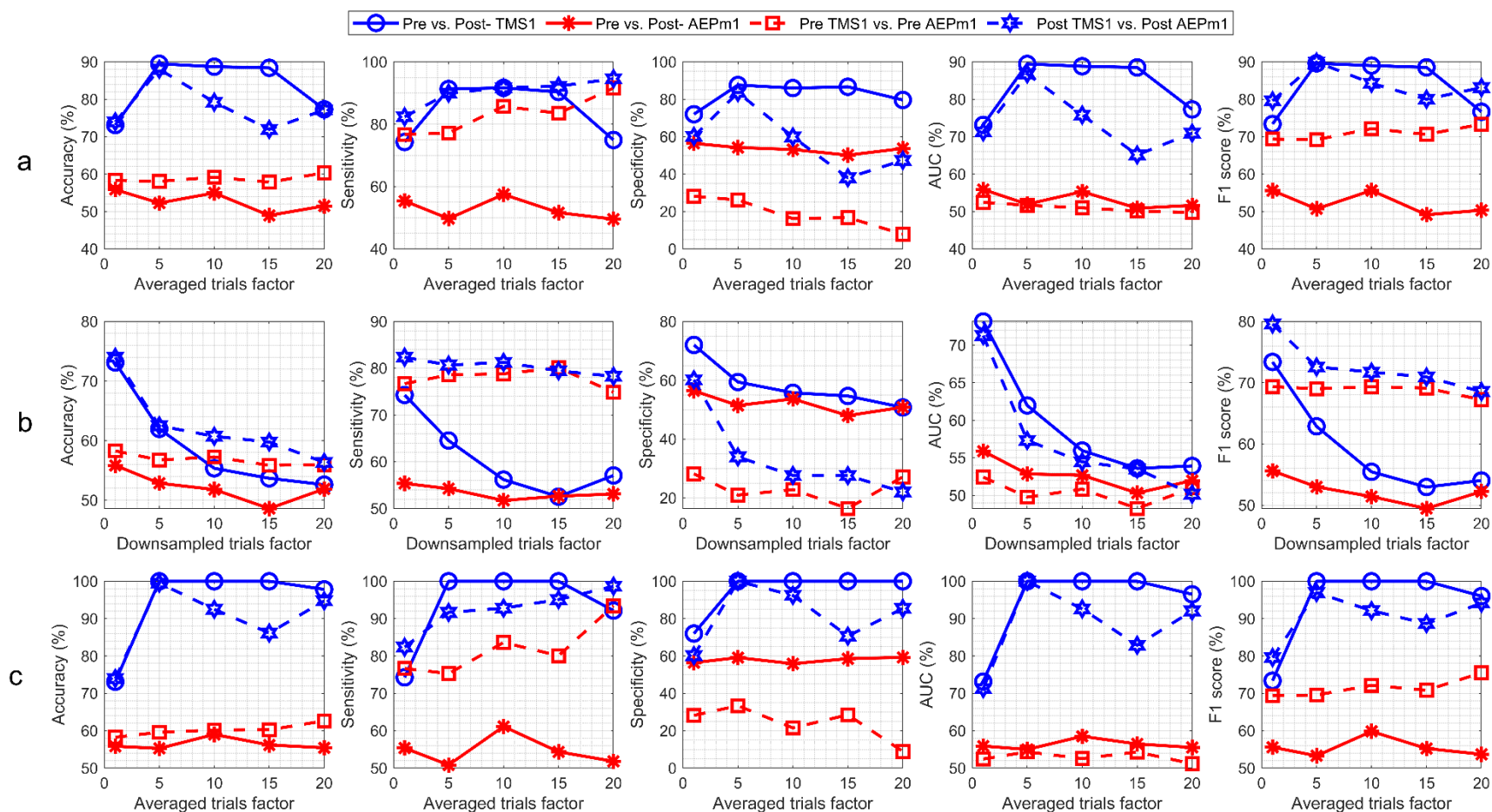
Supplementary Figure 10: Model LOO results for TMS2 vs. AEPm2 for Dataset 2.



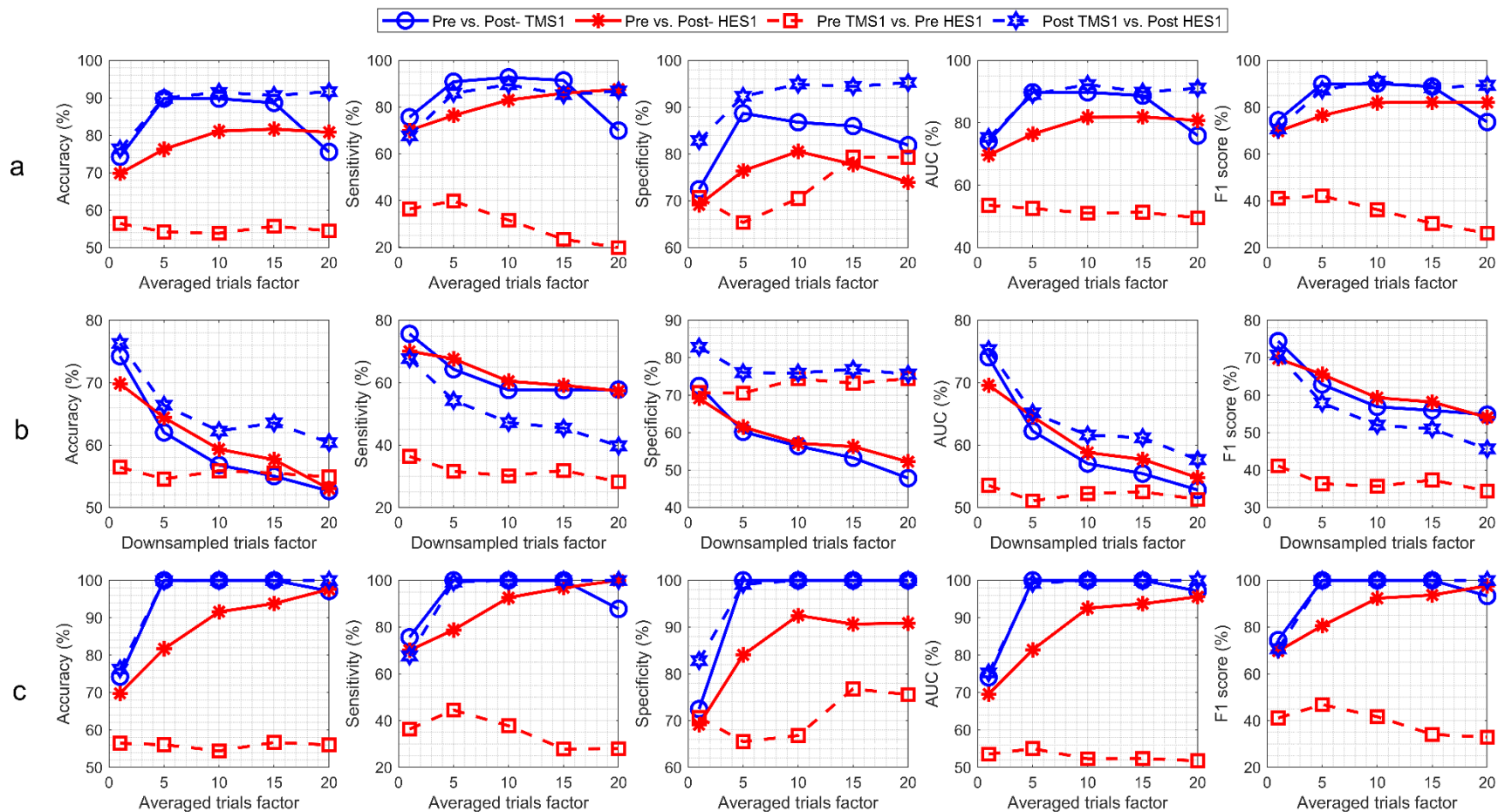
Supplementary Figure 11: Model CV results for TMS1 vs. S1 for Dataset 1.



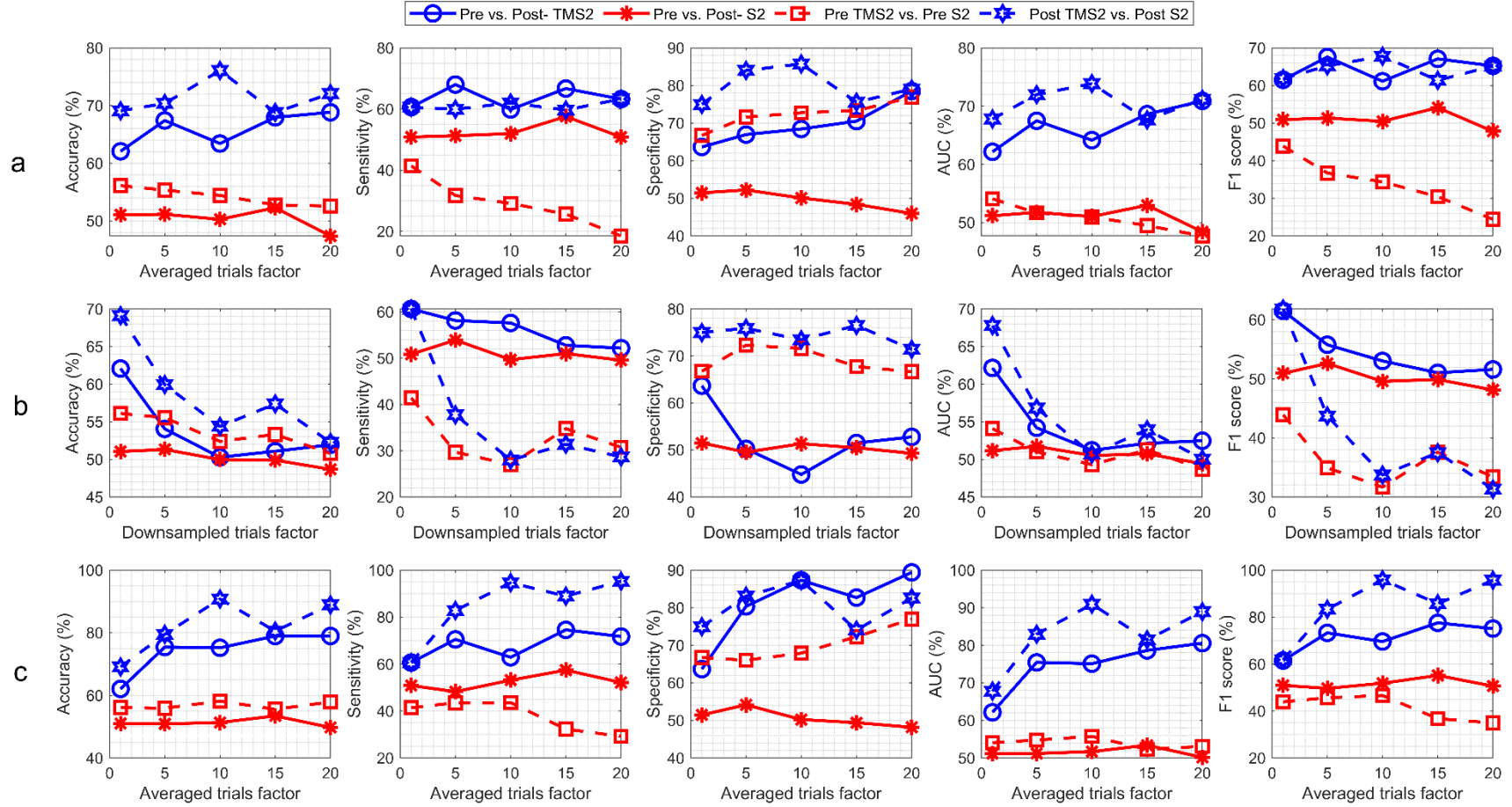
Supplementary Figure 12: Model CV results for TMS1 vs. AEP1 for Dataset 1.



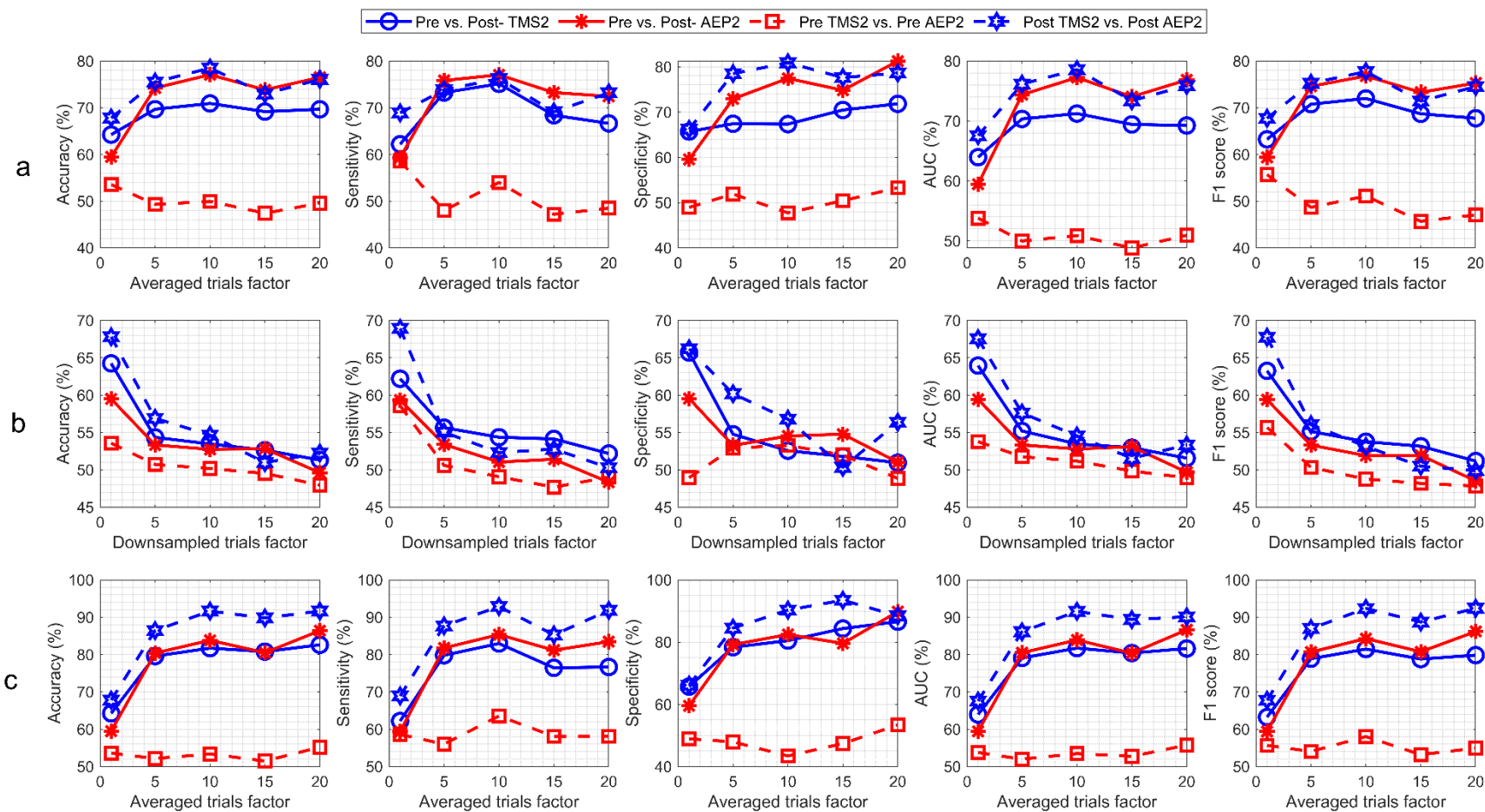
Supplementary Figure 13: Model CV results for TMS1 vs. AEPm1 for Dataset 1.



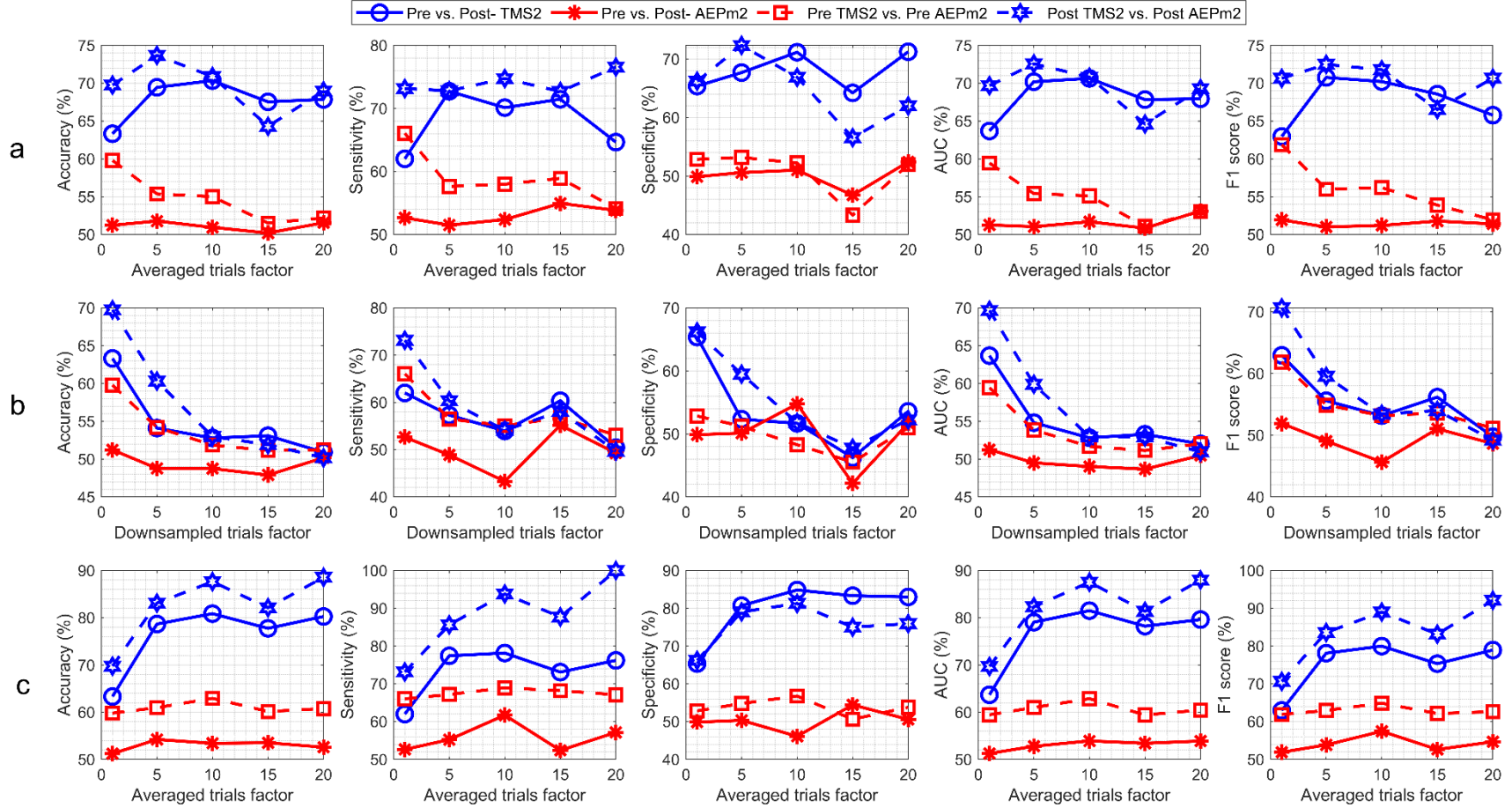
Supplementary Figure 14: Model CV results for TMS1 vs. HES1 for Dataset 1.



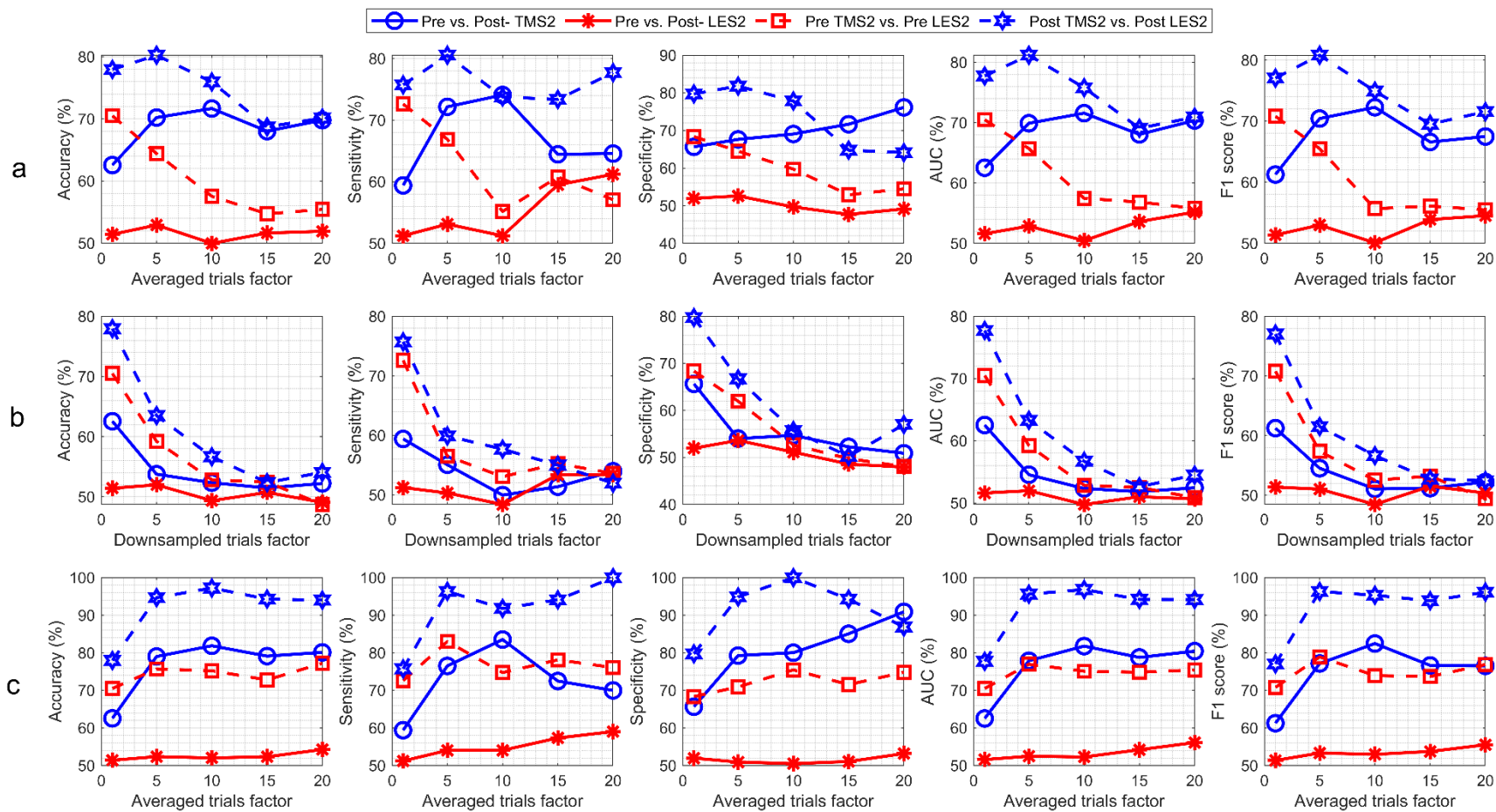
Supplementary Figure 15: Model CV results for TMS2 vs. S2 for Dataset 2.



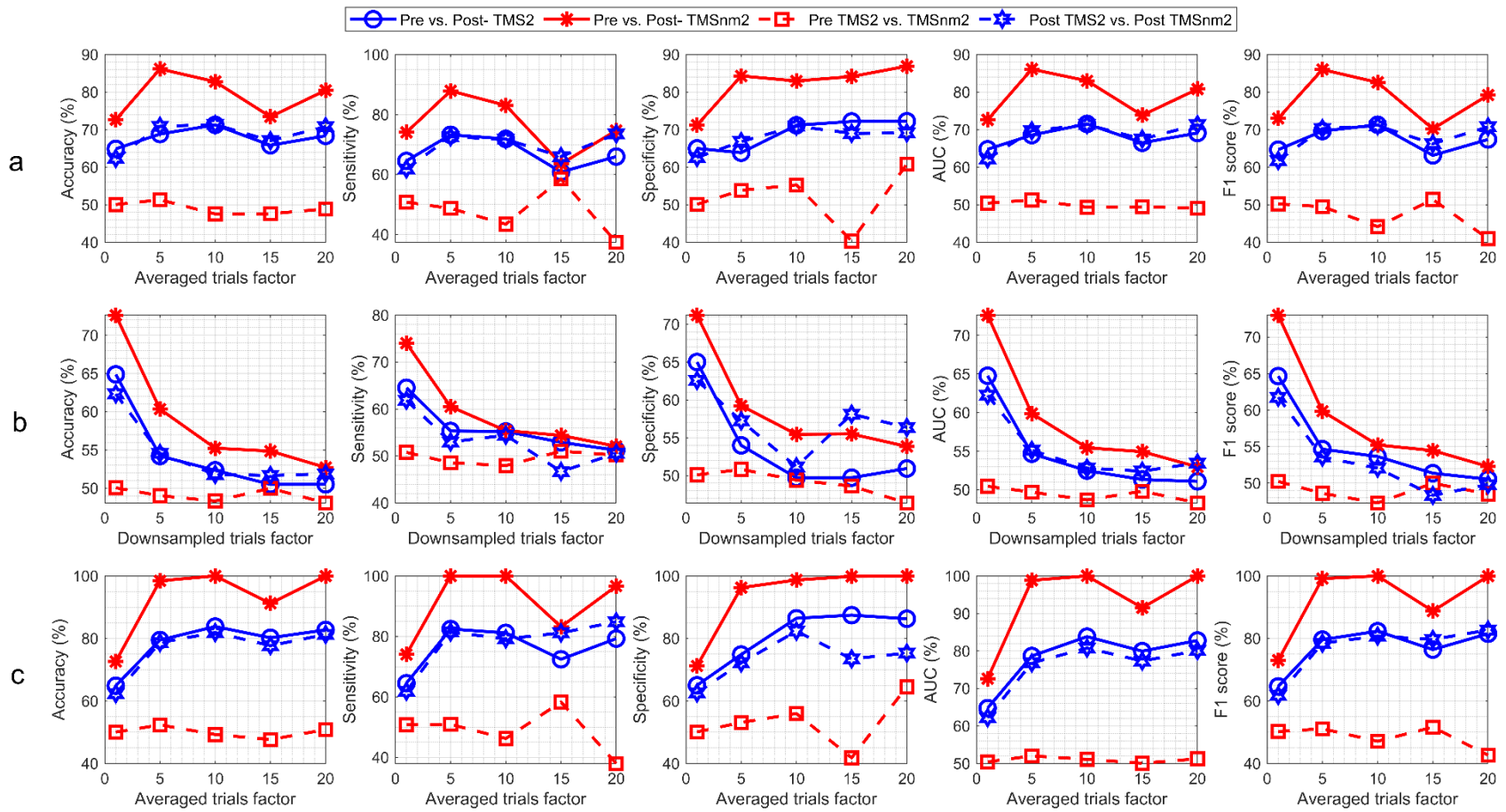
Supplementary Figure 16: Model CV results for TMS2 vs. AEP2 for Dataset 2.



Supplementary Figure 17: Model CV results for TMS2 vs. AEPm2 for Dataset 2.



Supplementary Figure 18: Model CV results for TMS2 vs. LES2 for Dataset 2.



Supplementary Figure 19: Model CV results for TMS2 vs. TMSnm2 for Dataset 2.