Toward unified and comprehensive automated electroencephalogram interpretation: a multicentre development and validation of an electroencephalogram foundation model

Chenxi Sun, Ioannis Karakis, Aline Herlopian, Marcus C Ng, Gamaleldin Osman, Zubeda Sheikh, Olga Taraschenko, Ji Yeoun Yoo, Brian L Appavu, Lakshman Arcot Jayagopal, Peter W Kaplan, Jong Woo Lee, Olga Selioutski, Hiba A Haider, Jonathan J Halford, Daniel B Hoch, Alice Lam, Fabio A Nascimento, Jay Pathmanathan, Sarah Schmitt, Pia De Stefano, Uwrs Fisch, Jeroen Gijs, Humberto Castro-Lima, Wan-Yee Kong, Mackenzie C Cervenka, Monica B Dhakar, Safoora Fatima, Nicolas Gaspard, Emily J Gilmore, Susan T Herman, Manisha G Holmes, Emily L Johnson, Carlos F Muniz, Eric S Rosenthal, Andres A Rodriguez Ruiz, Rani A Sarkis, Mouhsin M Shafi, Christa B Swisher, Mohammad Tabaeizadeh, Selim R Benbadis, Fonda Chan, Catherine J Chu, Marjan Dolatshahi, Adam S Greenblatt, Roohi Katyal, Chinasa Nwankwo, Edilberto Amorim, William O Tatum, Dan Weber, Tobias Loddenkemper, Jurriaan M Peters, Umakanth Katwa, Kiran Maski, Robert Joseph Thomas, Shenda Hong, Doyle Yuan, Sydney S Cash, Andrew J Cole, Daniel M Goldenholz, Charlotte Stow, Jennifer A Kim, Sahar F Zafar*, Aaron F Struck*, M Brandon Westover*, Jin Jing**

Summary

Background Electroencephalogram (EEG) interpretation is essential for neurological diagnosis, but expert interpretation is limited globally, and existing AI methods address narrow tasks. This study aimed to develop and externally validate a broadly applicable foundation model capable of expert-level performance across diverse EEG tasks and clinical settings.

Methods In this multicentre study we developed a multidomain omnibus for reading and generalising over thorough EEG interpretation (MORGOTH), a foundation model that supports broad clinical interpretation across all major settings. We developed MORGOTH using EEGs from 18 677 patients across Massachusetts General Hospital, Brigham and Women's Hospital, Beth Israel Deaconess Medical Center, and Boston Children's Hospital, collected between Jan 1, 2003, and Feb 1, 2025, and validated it internally on 13 334 patients and externally on 1573 patients from 48 institutions, spanning diverse clinical settings and ages (0 years to >90 years). Test datasets annotated by six to 30 experts enabled inter-rater reliability (IRR) analysis by comparing model–expert and expert–expert agreement. MORGOTH was compared against both human experts and state-of-the-art models using area under the curve (AUC) and the percentage of experts' operating points under the curve (EUC) for receiver operating characteristic (ROC) and precision-recall curves, as well as IRR and statistical calibration.

Findings MORGOTH achieved expert-level performance with AUC-ROC scores of 0·86 to 0·98 across 17 EEG findings. MORGOTH outperformed at least 90% of experts on three of seven multi-expert-annotated datasets and exceeded at least 20% of experts on each of the 17 tasks. Event-level performance was especially strong for seizure and ictal–interictal–injury continuum detection (EUC=96·6%) and spike detection (EUC=100%). IRR analysis showed that MORGOTH matched or exceeded expert consensus. External validation confirmed consistent performance with modest declines from internal to external test sets (event-level AUC –1·21%, EUC –3·33%; EEG-level AUC –2·12%, and EUC –9·52%). MORGOTH performance also remained robust across age, sex, and moderate channel loss, with lower age sensitivity (20·90% *vs* 30·90%) and fewer sex-related differences (33·33% *vs* 44·00%) than SPaRCNet.

Interpretation MORGOTH advances automated EEG interpretation with expert-level performance across clinical settings, offering improved diagnostic accuracy in low-resource environments and greater efficiency in high-volume centres.

Funding US National Institutes of Health.

Introduction

Electroencephalography is a cornerstone diagnostic tool for evaluating patients with neurological disorders.[1] When interpreted by skilled clinicians, electroencephalograms (EEGs) provide crucial information that guides diagnosis and therapeutic decision making.[2] Despite the prevalence of conditions requiring EEG interpretation—with epilepsy alone affecting more than 70 million people

Lancet Digit Health 2026

Published Online https://doi.org/10.1016/j.landig.2026.101039

*Co-senior authors

Beth Israel Deaconess Medical Center, Boston, MA, USA (C Sun PhD, W-Y Kong MD PhD, M M Shafi MD, Prof R J Thomas MD, Prof S S Cash MD, D M Goldenholz MD PhD, Prof M B Westover MD PhD, J Jing PhD); **Department of Neurology and Neurological Sciences, Stanford University School of Medicine, Palo Alto, CA, USA** (C Sun, Prof M B Westover); **Emory University School of Medicine, Atlanta, GA, USA** (I Karakis MD PhD, A A Rodriguez Ruiz MD); **Yale School of Medicine, New Haven, CT, USA** (A Herlopian MD, Prof E J Gilmore MD, J A Kim MD PhD); **University of Manitoba, Winnipeg, MB, Canada** (M C Ng MD FRCPC); **Mayo Clinic-Rochester, Rochester, MN, USA** (Gamaleldin Osman MD); **Virginia Commonwealth University, Richmond, VA, USA** (Z Sheikh MD); **University of Nebraska Medical Center, Omaha, NE, USA** (O Taraschenko MD PhD, L Arcot Jayagopal MD); **Icahn School of Medicine, New York, NY, USA** (Prof J Y Yoo MD); **University of Arizona College of Medicine, Phoenix, AZ, USA** (B L Appavu MD); **Johns Hopkins School of Medicine,**

Baltimore, MD, USA (Prof P W Kaplan MD FRCP, Prof M C Cervenka MD, E L Johnson MD); **Brigham and Women's Hospital, Boston, MA, USA** (J W Lee MD PhD, U Fisch MD PhD, R A Sarkis MD); **Stony Brook University, Stony Brook, NY, USA** (Prof O Selioutski DO); **University of Chicago, Chicago, IL, USA** (H A Haider MD); **Ralph H Johnson Veterans Affairs Medical Center, Charleston, SC, USA** (J J Halford MD); **Massachusetts General Hospital, Boston, MA, USA** (D B Hoch MD PhD, A Lam MD PhD, E S Rosenthal MD, C J Chu MD, A J Cole MD, S F Zafar MD); **Washington University in St Louis, St Louis, MO, USA** (F A Nascimento MD, A S Greenblatt MD); **Beacon Biosignals/Boston VA Healthcare, Boston, MA, USA** (J Pathmanathan MD PhD); **University Hospital of Geneva, Geneva, Switzerland** (P De Stefano MD); **University Hospital Basel, Basel, Switzerland** (U Fisch); **Department of Neurosciences, Experimental Neurology, Laboratory for Epilepsy Research, KU Leuven, Leuven Brain Institute, Leuven, Belgium** (J Gijs MD); **Bahiana School of Medicine and Public Health, Salvador, Brazil** (H Castro-Lima MD); **Warren Alpert School of Medicine of Brown University, Providence, RI, USA** (M B Dhakar MD); **University of Wisconsin-Madison, Madison, WI, USA** (S Fatima MD, A F Struck MD); **Belgium University Libre de Bruxelles, Hopital Universitaire de Bruxelles - Hopital Erasme, Brussels, Belgium** (N Gaspard MD PhD); **Barrow Neurological Institute, Phoenix, AZ, USA** (Prof S T Herman MD); **Westchester Medical Center, New York Medical College, Valhalla, NY, USA** (M G Holmes MD); **SUNY Upstate Medical University, Syracuse, NY, USA** (C F Muniz MD); **Atrium Health, Charlotte, NC, USA** (C B Swisher MD); **University of South Florida, Tampa, FL, USA** (Prof S R Benbadis MD); **Neurology Consultants of Dallas, Dallas, TX, USA** (F Chan MD); **William Osler Health System, Brampton, ON,**

Research in context

Evidence before this study

We searched PubMed and arXiv using the following terms: ("EEG" OR "electroencephalogram") AND ("artificial intelligence" OR "deep learning" OR "automated interpretation" OR "foundation model"). We reviewed studies published between Jan 1, 2015, and July 1, 2025, without language restrictions. We included studies that developed or validated machine learning, deep learning, or foundation models for clinical electroencephalogram (EEG) interpretation in any setting and excluded those focused solely on non-clinical applications such as brain–computer interfaces or emotion recognition. Existing models, such as SCORE-AI, SPaRCNet, TEEGNet, SpikeNet, U-Sleep, and others, have demonstrated strong performance within specific clinical domains. However, to our knowledge, no previous study has introduced a unified model capable of expert-level interpretation across all major clinical contexts while supporting a broad spectrum of clinically relevant tasks. Moreover, few previous efforts have conducted external validation against multi-expert annotations, particularly in comparison to inter-rater reliability, a crucial benchmark for evaluating clinical utility.

Added value of this study

To our knowledge, this is the first study to present a foundation model for EEGs that enables expert-level interpretation across all major clinical settings (including routine outpatient clinics, epilepsy monitoring units, and critical care environments) while supporting both event-level and EEG-level classification. The model achieved state-of-the-art performance across 17 clinically relevant EEG tasks. Its performance was benchmarked against multi-expert annotations, demonstrating agreement comparable with or exceeding expert consensus, and consistently outperforming most human experts. Validation was conducted on large and diverse datasets from 52 institutions in eight countries, encompassing patients of all ages and EEG recordings from routine, critical care, and sleep study settings. This combination of broad task coverage, expert-level performance, and strong generalisability represents a substantial advancement over existing task-specific models.

Implications of all the available evidence

This study shows that a single foundation model can deliver robust, expert-level EEG interpretation across the full range of clinical contexts. This approach might improve access to diagnostic-quality EEG interpretation in under-resourced settings, reduce interpretation workload in high-volume centres, and enable continuous monitoring in critical care environments. Our results suggest that future clinical artificial intelligence systems might benefit from unified foundation architectures, validated against expert consensus across diverse populations and tasks, to ensure safe and scalable deployment in real-world practice.

worldwide[3]—expertise in clinical EEG interpretation remains scarce.[4] Most EEGs are interpreted by physicians without specialised fellowship training,[5] and most hospitals cannot provide real-time EEG monitoring in hospital or inpatient settings. This expertise gap contributes to preventable harm, with EEG misinterpretation being the leading cause of epilepsy overdiagnosis and misdiagnosis.[6]

Artificial intelligence (AI) has the potential to address these unmet clinical needs by providing accurate EEG interpretation where expertise is unavailable, reducing expert workload, and improving diagnostic consistency. However, existing AI approaches for EEG interpretation have addressed only a limited number of EEG interpretation tasks, typically focusing on a single finding or clinical application. For example, SCORE-AI[7] focuses exclusively on routine outpatient EEGs, SPaRCNet[8] targets only critical care settings, SpikeNet[9] addresses only epileptiform discharge detection, and U-Sleep[10] is used for sleep staging. No previous work has simultaneously addressed all clinical EEG settings and the full spectrum of clinically relevant EEG patterns. Additionally, few studies have externally validated these models, particularly regarding algorithm performance assessment relative to expert inter-rater reliability (IRR).[11]

Foundation models adaptable to diverse tasks have gained attention, as shown by the success of large language models. Inspired by the success of large language models, EEG foundation models such as LaBraM,[12] EEGFormer,[13] and BrainBERT[14] have emerged since 2023, leveraging transformer architectures and self-supervised learning on large-scale EEG data. However, these models have primarily focused on non-clinical domains such as brain–computer interfaces and emotion recognition, with limited relevance to clinical EEG interpretation.

To address these limitations, we aimed to develop a multidomain omnibus for reading and generalising over thorough EEG interpretation (MORGOTH), a unified and generalisable foundation model for comprehensive clinical EEG interpretation across routine outpatient clinics, epilepsy monitoring units, critical care environments, and sleep laboratories. MORGOTH performs both event-level detection and EEG-level interpretation within a single architecture and supports a broad range of clinically relevant EEG findings, including seizures, epileptiform discharges, rhythmic and periodic patterns along the ictal–interictal–injury continuum (IIIC), pathological slowing, and sleep staging. We also aimed to benchmark the performance of MORGOTH against expert interpretation and IRR.

Methods

Study design and data sources

We developed the MORGOTH model to support the clinical interpretation of EEGs across all major settings. MORGOTH was developed and validated using EEG datasets from diverse sources (figure 1, appendix p 2;

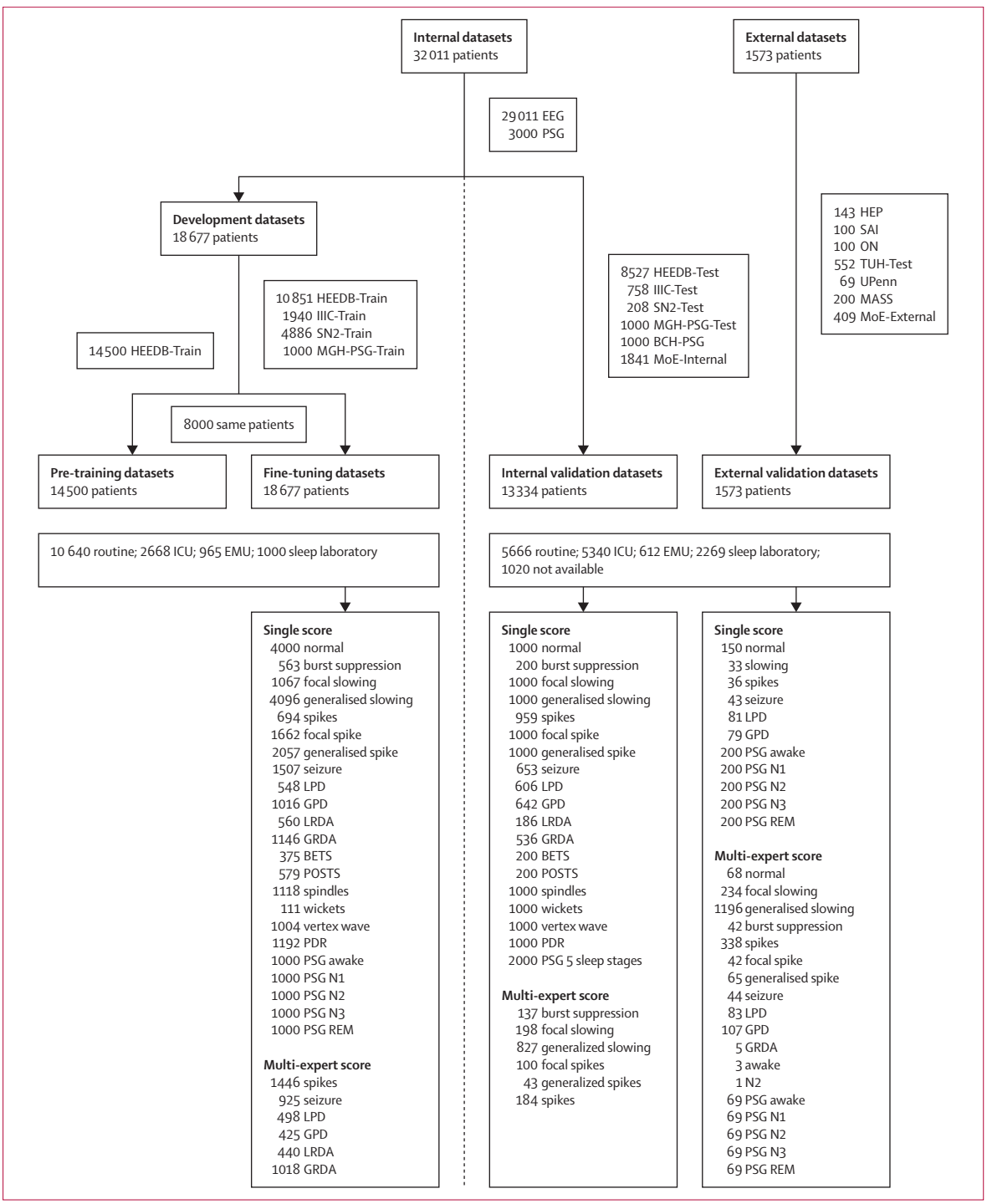


Figure 1: **Dataset partitioning consort diagram**
The internal data was used to pretrain and fine tune the model, and the internal holdout datasets and external datasets were used to validate it. BCH-PSG=Boston Children's Hospital polysomnography dataset. BETS=benign epileptiform transients of sleep. EEG=electroencephalogram. EMU=epilepsy monitoring unit. GPD=generalised periodic discharges. GRDA=generalised rhythmic delta activity. HEEDB=Harvard Electroencephalography Database. HEP=Human Epilepsy Project. ICU=intensive care unit. IIIC=ictal–interictal–injury continuum. LPD=lateralised periodic discharges. LRDA=lateralised rhythmic delta activity. MASS=Montreal Archive of Sleep Studies. MGH-PSG=Massachusetts General Hospital polysomnography dataset. MoE=Map of Everything. N1=non-rapid eye movement sleep stage 1. N2=non-rapid eye movement sleep stage 2. N3=non-rapid eye movement sleep stage 3. ON=Occasion Noise. PDR=posterior dominant rhythm. POSTS=positive occipital sharp transients of sleep. PSG=polysomnography. REM=rapid eye movement sleep. SAI=SCORE-AI. SN2=SpikeNet 2. TUH=Temple University Hospital. UPenn=University of Pennsylvania.

Canada (M Dolatshahi MD); **Louisiana State University Health Shreveport, Shreveport, LA, USA** (R Katyal MD); **Akron Children's Hospital, Akron, OH, USA** (C Nwankwo MD); **University of California San Francisco, San Francisco, CA, USA** (E Amorim MD); **Mayo Clinic Florida, Jacksonville, FL, USA** (Prof W O Tatum DO); **St Louis University, St Louis, MO, USA** (D Weber DO); **Boston Children's Hospital, Boston, MA, USA** (Prof T Loddenkemper MD, J M Peters MD, U Katwa MD, K Maski MD); **National Institute of Health Data Science, Peking University, Beijing, China** (S Hong PhD); **University of Texas Southwestern, Dallas, TX, USA** (M Tabaeizadeh MD); **Department of Neurology, University of Texas Southwestern Medical Center, Dallas, TX, USA** (D Yuan MD); **Peter O'Donnell Jr Brain Institute, University of Texas Southwestern Medical Center, Dallas, TX, USA** (D Yuan); **University of Crete School of Medicine, Heraklion, Greece** (I Karakis); **Medical University of South Carolina, Charleston, SC, USA** (S Schmitt MD); **Division of Neurology, University Hospitals Leuven, Leuven, Belgium** (J Gijs); **Trustee TeleEEG, London, UK** (C Stow BA)

Correspondence to:
Dr Chenxi Sun, Department of Neurology and Neurological Sciences, Stanford University School of Medicine, Palo Alto CA 94304, USA
cxsun@stanford.edu

	Number of hospitals	Number of experts	Number of experts that labelled each sample	Number of patients	Number of EEGs	Female	Age, years	Ethnicity*	Event labels†	EEG labels‡	Routine outpatient clinics§	Intensive care units§	Epilepsy monitoring units§	Sleep laboratories¶	Findings‖
Pre-training dataset															
HEEDB-Pretrain**	4	4	1 (1–3)	14 500	14 500	50%	49 (0–100)	63·51%; 7·92%; 0·14%; 2·88%; 0·04%; 0·91%; 5·65%	Yes	Yes	7793	2942	765	0	All 17 findings
Fine-tuning datasets															
HEEDB-Train**	4	1	1 (1–3)	10 851	10 851	49%	49 (0–100)	63·60%; 7·90%; 0·15%; 2·90%; 0·05%; 0·90%; 5·70%	Yes	Yes	6695	3064	591	0	All 17 findings
MGH-PSG-Train††	1	7	1 (1–1)	1000	1000	50%	60 (2–94)	67·80%; 8·60%; 0·05%; 2·10%; 0·01%; 0·60%; 4·80%	Yes	Yes	0	0	0	1000	5 sleep stages
IIIC-Train	1	124	3 (1–30)	1940	1940	49%	54 (0–96)	69·20%; 8·50%; 0·08%; 2·10%; 0·01%; 0·55%; 4·40%	Yes	Yes	367	1295	243	0	Seizure and 4 IIICs
SN2-Train	2	24	8 (1–23)	4886	4886	49%	53 (0–99)	72·75%; 9·20%; 0·05%; 1·80%; 0·01%; 0·35%; 3·65%	Yes	Yes	3578	899	131	0	1-second spike
Internal validation datasets															
HEEDB-Test**	4	1	1 (1–3)	8527	8527	51%	59 (0–100)	63·56%; 7·90%; 0·15%; 2·90%; 0·10%; 0·90%; 5·69%	Yes	Yes	4464	3404	278	0	All 17 findings
IIIC-Test	2	30	3 (1–30)	758	758	48%	60 (0–95)	68·95%; 8·60%; 0·09%; 2·15%; 0·02%; 0·48%; 4·55%	Yes	Yes	135	565	45	0	Seizure and 4 IIICs
SN2-Test	2	24	8 (8–23)	208	208	48%	37 (0–96)	70·40%; 8·30%; 0·06%; 1·95%; 0·01%; 0·42%; 4·15%	Yes	Yes	127	64	17	0	1-second spike
MGH-PSG-Test††	1	7	1 (1–1)	1000	1000	50%	59 (5–90)	69·50%; 8·40%; 0·08%; 2·05%; 0·01%; 0·52%; 4·25%	Yes	Yes	0	0	0	1000	5 sleep stages
BCH-PSG††	1	5	1 (1–1)	1000	1000	50%	7 (0–35)	73·20%; 9·10%; 0·05%; 1·60%; 0·01%; 0·35%; 3·85%	Yes	Yes	0	0	0	1000	5 sleep stages
MoE-Internal‡‡	3	21	13 (10–19)	1841	2125	53%	54 (0–95)	82·30%; 8·10%; 0·03%; 0·90%; 0·01%; 0·15%; 2·50%	Yes	Yes	597	898	272	0	Slowing, burst suppression, spike location, seizure and 4 IIICs, awake, N1, and N2

(Table continues on next page)

table). This study was conducted with ethical approval from the appropriate institutional review boards and was approved by the Beth Israel Deaconess Medical Center Institutional Review Board (protocols #2022P000481 and #2022P000417), with a waiver of informed consent for retrospective analysis. All data were collected and analysed in accordance with relevant ethical guidelines and regulations.

Model development framework

MORGOTH automatically processes 19-channel EEG recordings, with the international 10–20 system at 200 Hz (data at other rates are resampled). When one or more channels are unavailable, the missing channels are imputed using the mean of the available channels. Signals are bandpass filtered (0·5–70 Hz), notch filtered at 50 Hz and 60 Hz, clipped at plus or minus 500 µV, and normalised to [–1, 1] using a common average montage (appendix p 15). Although the electrocardiogram channel is excluded, the model can reject electrocardiogram artifacts (appendix p 34).

This EEG foundation model comprises a tokeniser, an EEG transformer, and specialised task heads (appendix p 13). The tokeniser converts continuous EEG signals into 8192 discrete tokens via contrastive learning[15] and vector quantisation. The EEG transformer includes 12 encoder blocks with multihead attention and temporal position encoding and spatial position encoding.[16] Event-level tasks use fully connected heads for pattern detection, and EEG-level tasks use convolutional and attention-based heads for whole-recording classification (appendix p 12).

MORGOTH performs seven event-level tasks: three binary classifications (normal *vs* abnormal, burst suppression, and spike detection), two three-class tasks (focal *vs* generalised *vs* none for slowing and spike localisation), one five-class sleep staging task (American Academy of Sleep Medicine defined[17]), and one six-class seizure and IIIC task (seizure, lateralised periodic discharges [LPDs], generalised periodic discharges [GPDs], lateralised rhythmic delta activity [LRDA], generalised rhythmic delta activity, and other; appendix p 9). EEG-level tasks include 17 binary classifications, one per EEG finding. Labels used for training and evaluation are assigned to 10-s segments (event-level), 1-s segments (event-level spikes), 10-min segments (EEG-level internal datasets), or full recordings (EEG-level external datasets). The model supports variable-length EEG inputs at inference.

Model training

We proposed a two-stage training approach: first, pre-training the tokeniser and transformer on EEGs from 14 500 Harvard Electroencephalography Database (HEEDB)[18] patients using self-supervised learning; then, fine-tuning task-specific heads on expert-labelled datasets. Expert annotations were provided by board-certified

(Continued from previous page)

	Number of hospitals	Number of experts	Number of experts that labelled each sample	Number of patients	Number of EEGs	Female	Age, years	Ethnicity*	Event labels†	EEG labels‡	Routine outpatient clinics§	Intensive care units§	Epilepsy monitoring units§	Sleep laboratories¶	Findings‖
External validation datasets															
MoE-External‡‡	4	21	13 (10–19)	409	636	30%	63 (16–93)	NA	Yes	Yes	0	409	0	0	Slowing, burst suppression, spike location, seizure and 4 IIICs, awake, N1, and N2
HEP§§	29	13	3 (3–3)	143	143	NA	NA	NA	Yes	Yes	143	0	0	0	1-second spike
MASS††	3	1	1 (1–1)	200	200	51%	40 (18–76)	NA	Yes	Yes	0	0	0	200	5 sleep stages
TUH-Test¶¶	1	1	1 (1–1)	552	552	51%	52 (9–90+)	NA	Yes	Yes	Unavailable	Unavailable	Unavailable	0	4 tasks
UPenn††	3	6	6 (6–6)	69	69	100%	51 (39–63)	NA	Yes	Yes	0	0	0	69	5 sleep stages
SA‖‖	3	14	11 (11–11)	100	100	61%	26 (0–95)	NA	No	Yes	100	0	0	0	Normal, slowing, and spike location
ON	5	15	15 (15–15)	100	100	53%	26 (0–99)	NA	No	Yes	100	0	0	0	Slowing and spike location

Data are n, median (minimum to maximum), or proportion, unless stated otherwise. BCH-PSG=Boston Children's Hospital polysomnography dataset. EEG=electroencephalogram. HEEDB=Harvard Electroencephalography Database. HEP=Human Epilepsy Project. IIIC=ictal–interictal–injury continuum. MASS=Montreal Archive of Sleep Studies. MGH-PSG=Massachusetts General Hospital polysomnography dataset. MoE=Map of Everything. N1=non-rapid eye movement sleep stage 1. N2=non-rapid eye movement sleep stage 2. NA=not available. ON=Occasion Noise. SAI=SCORE-AI. SN2=SpikeNet 2. TUH=Temple University Hospital. UPenn=University of Pennsylvania. *White; Black; American Indian or Alaska Native; Asian; Native Hawaiian or Pacific Islander; Multiracial; Other. Following US Census Bureau classifications. †Indicates whether annotations were available at the event level. ‡Indicates whether annotations were available for the entire EEG recording. §The number of EEG recordings from these setting. ¶Number of polysomnography recordings. ‖The types of labels included in the dataset. **Based on HEEDB database, we extracted 14 500 10-min EEGs from 14 500 patients to pretrain the model, 17 677 patients' EEGs (10 851 for HEEDB-Train, 1940 for IIIC-Train, and 4886 for SN2-Train) to fine tune the model, and 9493 patients' EEGs (8527 for HEEDB-Train, 758 for IIIC-Train, and 208 for SN2-Train) to validate the model. HEEDB-Pretrain and HEEDB-Train have 8000 patients in common, but both are strictly separated from HEEDB-Test, with no patient overlap. In HEEDB-Train and HEEDB-Test, 1 [1–3] in the number of experts that labelled each sample column means that 10 082 samples were labelled by 1–2 clinical experts using reports or annotations, or both, with 9296 of them receiving a confirmatory label from a third expert. ††For sleep-related datasets, we used overlapping EEG channels where available: 19 EEG channels for HEEDB data with sleep labels, six channels for the MGH-PSG and BCH-PSG datasets, 16 channels for the MASS dataset, and two channels for the UPenn dataset. ‡‡MoE combines both internal and external validation data. We separately report counts for MoE-Internal and MoE-External. The MoE-Internal dataset includes data from MGH, BWH, and Beth Israel Deaconess Medical Center, all of which are part of HEEDB. MoE-External includes data from International Cardiac Arrest Research Consortium Electroencephalography Database sites. §§For HEP, each event has three labels from 13 experts, with the three experts varying. ¶¶The TUH is split into train and evaluation, with testing on the evaluation subset, termed TUH-Test. The exact clinical settings are not available, the settings are a mix of EMU, ICU, and routine. ‖‖For SAI, each event has 11 labels from 14 experts, with the 11 experts varying.

Table: **Dataset information statistics**

neurologists, epileptologists, and sleep medicine specialists.

For multi-expert labelled datasets, we used soft labels (expert vote distributions) to approximate class probabilities and capture consensus. To address class imbalance and sample difficulty, we applied focal loss[19] with curriculum learning.[20] Training was optimised using AdamW with learning rate scheduling and early stopping. Additional training details in the appendix (p 12).

See Online for appendix

For more on **AdamW** see https://docs.pytorch.org/docs/main/generated/torch.optim.AdamW.html

For more on the **Human Epilepsy Project** see https://www.humanepilepsyproject.org

For more on **TELEEEG** see https://www.teleeeg.org/

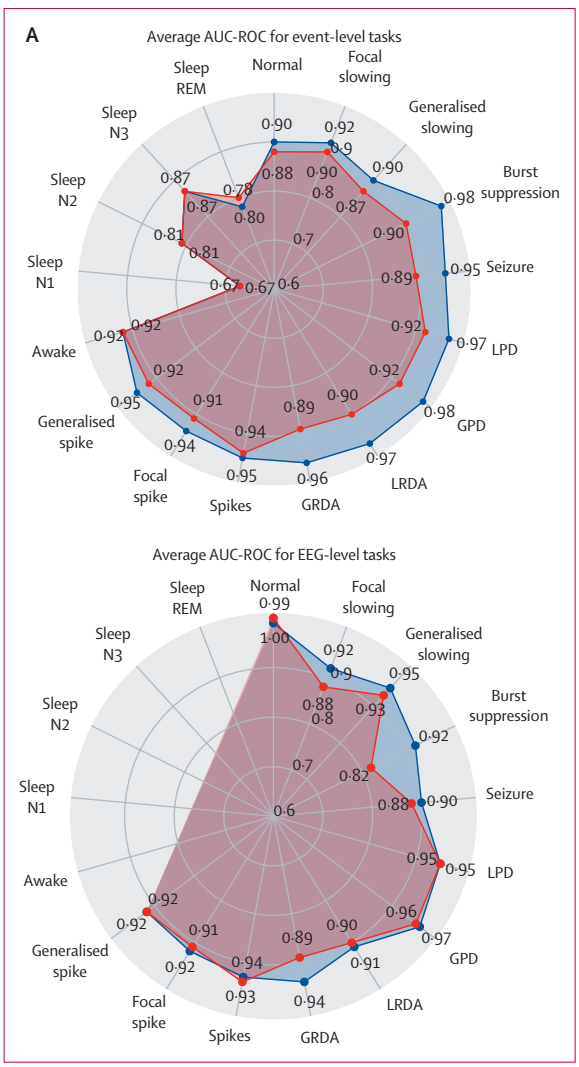
Validation strategies

We validated MORGOTH using a comprehensive four-tier approach. First, internal validation using holdout data from HEEDB-Test, IIIC-Test, SpikeNet 2-Test (SN2-Test), Massachusetts General Hospital polysomnography test (MGH-PSG-Test), and Boston Children's Hospital polysomnography (BCH-PSG) datasets, with no patient overlaps with training data (appendix p 7). Second, external validation using six independent datasets from 48 institutions, including the Human Epilepsy Project (HEP), SCORE-AI (SAI),[7] Occasion Noise (ON), Temple University EEG Corpus (TUH),[21] University of Pennsylvania (UPenn),[22] Montreal Archive of Sleep Studies (MASS),[23] and TELEEEG datasets (appendix p 8). Third, IRR analysis, using specialised multi-expert annotated datasets, including the Map of Everything (MoE) dataset (21 experts), IIIC-Test (30 experts), SN2-Test (24 experts), HEP (14 experts), SAI (14 experts), ON (15 experts), and UPenn (six experts; appendix p 9). Fourth, comparison against state-of-the-art (SOTA) models, including SpikeNet 2[9] for spike detection; SPaRCNet[8] and the Kaggle competition winner[24] for seizure and IIIC classification; U-Sleep[10] for sleep staging; a burst suppression detector; SCORE-AI[8] for EEG-level tasks; and EEG foundation models (LaBraM,[12] EEGFormer,[13] and BrainBERT[14]) on TUH benchmarks (appendix p 19).

Statistical analysis

Model performance was evaluated using receiver operating characteristic (ROC) and precision-recall (PR) curves. To compare with experts, we computed experts' operating points under the curve (EUC),[8] which was the fraction of expert operating points (ie, each expert's sensitivity and specificity or precision and recall pair) that lie below the model's ROC or PR curve, indicating the proportion of experts the model outperforms across clinically relevant thresholds. We calculated 95% CIs using 10 000 bootstrap iterations and considered differences significant when their intervals did not overlap.

To assess reliability, we report IRR based on Cohen's κ, comparing expert–expert and expert–model agreement patterns.[12] We evaluated statistical calibration by fitting parametric models to derive a calibration index ranging from –1 (maximal undercalling) to 1 (maximal overcalling). We applied Platt scaling[25] and isotonic regression[26] to calibrate model outputs to better reflect true probabilities (appendix p 20).



(Figure 2 continues on next page)

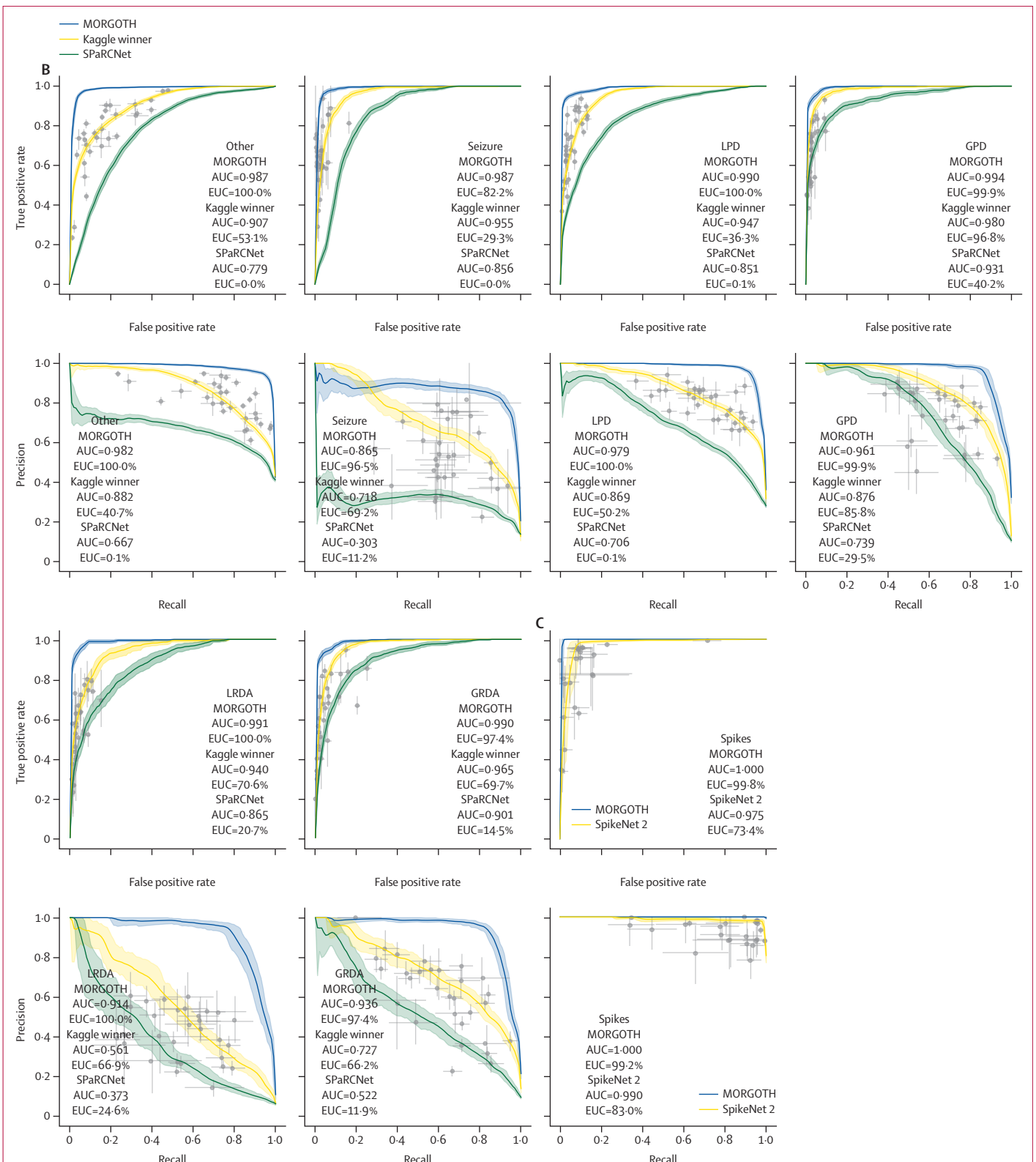
Role of the funding source

The funder had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

Figure 2A presents the overall results on both event-level and EEG-level tasks, showing that MORGOTH matches or exceeds expert and SOTA model performance across all test datasets and tasks. Among tasks with baseline model comparisons, MORGOTH achieved improved AUC values for five out of seven tasks (spike detection, seizure and IIIC classification, burst suppression detection, slowing classification, and abnormal detection) and in EUC for three out of seven tasks (spike detection, seizure and IIIC classification, and burst suppression detection; appendix p 50).

In internal validation (HEEDB-Test, IIIC-Test, SN2-Test, MoE, MGH-PSG-Test, and BCH-Test), MORGOTH



(Figure 2 continues on next page)

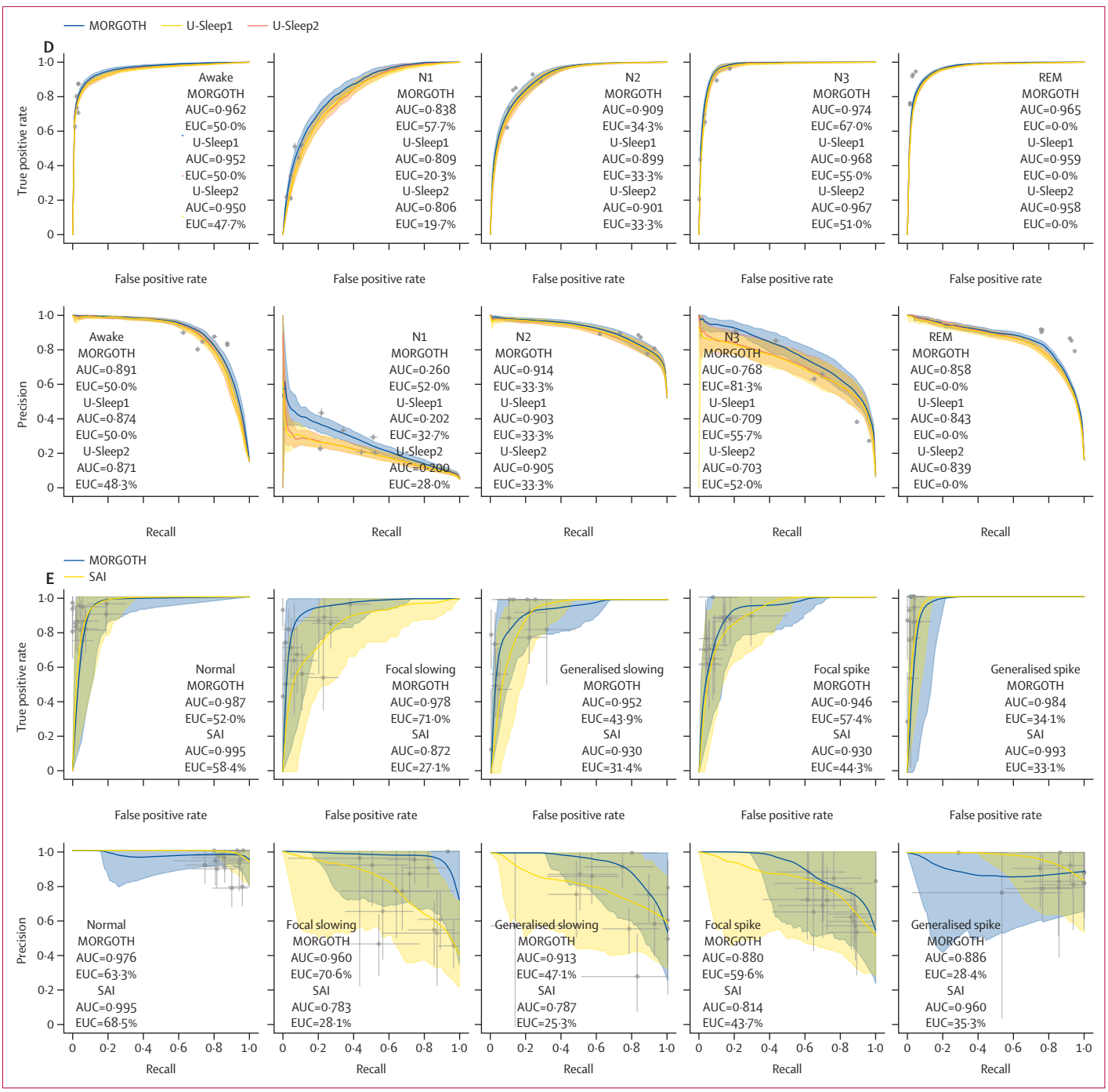
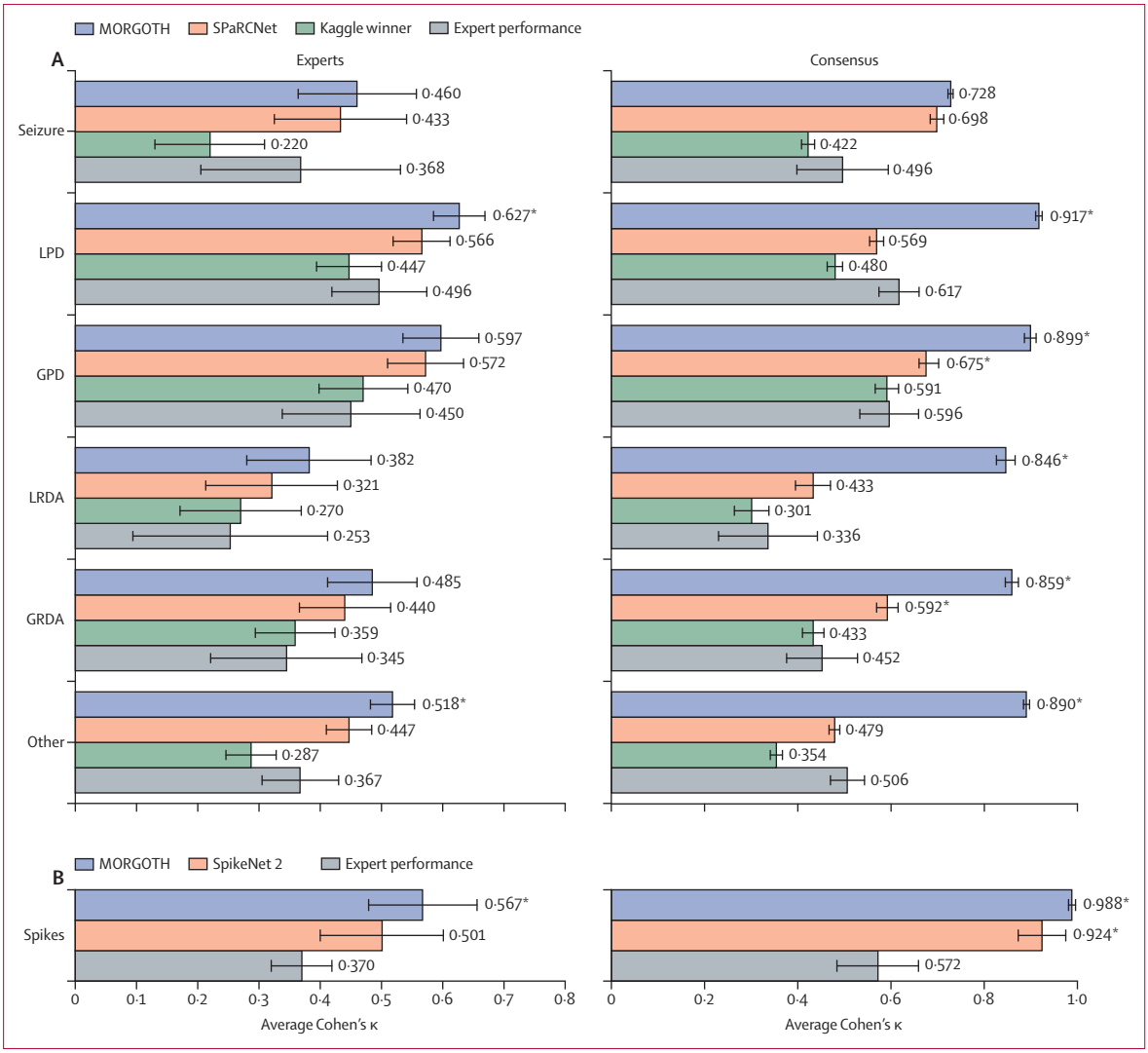


***Figure 2*: Performance on multi-expert scored validation datasets**

(A) The overall performance of MORGOTH compared to state-of-the-art models or human experts at the event level (upper) and EEG level (lower). EEG-level sleep staging performances are not reported here due to the absence of available baselines for comparison. The curve plots report AUC-ROC, EUC-ROC, AUC-PR, and EUC-PR for IIIC classification on internal IIIC-Test dataset (seizure *vs* LPD *vs* GPD *vs* LRDA *vs* GRDA versus other [seizure and IIIC free]; B), spike detection on internal SN2-Test dataset (C), sleep staging on external UPenn dataset (D), and five EEG-level tasks on external SAI dataset (E). Dashed boxes highlight the results from EEG-level tasks; the others correspond to event-level tasks. MORGOTH is shown in blue, with baseline models in yellow and green: SPaRCNet and the Kaggle winner (B), SpikeNet 2 (C), U-Sleep (D), and SAI (E). Expert performance is represented by grey dots. Shaded areas indicate 95% CIs from 10 000 bootstrap iterations, and expert results are marked with crosses. AUC= area under the curve. EEG=electroencephalogram. EUC=experts' operating points under the curve. GPD=generalised periodic discharges. GRDA=generalised rhythmic delta activity. IIIC=ictal–interictal–injury continuum. LPD=lateralised periodic discharges. LRDA=lateralised rhythmic delta activity. MORGOTH=multidomain omnibus for reading and generalising over thorough EEG interpretation. N1=non-rapid eye movement sleep stage 1. N2=non-rapid eye movement sleep stage 2. N3=non-rapid eye movement sleep stage 3. PR=precision–recall. REM=rapid eye movement sleep. ROC=receiver operating characteristic. SAI=SCORE-AI. SN2=SpikeNet 2. UPenn=University of Pennsylvania.



(Figure 3 continues on next page)

achieved an average AUC-ROC of 0·921 (95% CI 0·780–0·981) across seven event-level and 17 EEG-level tasks, outperforming an average of 81·0% (46·9–100; 17 of 21) of experts on three multi-expert-annotated datasets. In external validation (HEP, ON, SAI, TUH, UPenn, and MASS), MORGOTH achieved an average AUC-ROC of 0·908 (0·705–0·980) across seven event-level and eight EEG-level tasks, outperforming an average of 50·8% (14·3–98·0; 18 of 35) of experts on four multi-expert-annotated datasets. Subgroup analyses across age and sex showed that MORGOTH exhibited greater robustness than baseline models across demographic groups. For example, on the IIIC-Test dataset, MORGOTH showed lower average age sensitivity than SPaRCNet (20·90% *vs* 30·90%) and fewer sex-related differences (33·33% *vs* 44·00% for SPaRCNet and 48·00% for the Kaggle winner; appendix p 71).

Overall, we used 12 datasets from 34 602 patients across 52 hospitals in eight countries (appendix p 9).

For development (figure 1), we used 18 677 EEGs from 18 677 patients (51% male and 49% female; median age 56 years [range 0 to >90]) across four hospitals (Massachusetts General Hospital, Brigham and Women's Hospital, Beth Israel Deaconess Medical Center, and Boston Children's Hospital). Of these, 14 500 10-min EEGs were used for self-supervised pretraining; 102 442 expert-labelled 10-s segments for fine-tuning the seven event-level task heads; and 18 677 10-min EEGs for training the 17 EEG-level task heads.

For validation, we compared MORGOTH's performance against 11 different SOTA models. For internal validation, we used 13 618 EEGs from 13 334 patients (51% male and 49% female; median age 59 years [range 0 to >90]), annotated by one to 30 experts across the same four hospitals described earlier. Ethnicity was reported as 63·51% White, 7·92% Black, 0·14% American Indian or Alaska Native, 2·88% Asian, 0·04% Native Hawaiian or Pacific Islander,

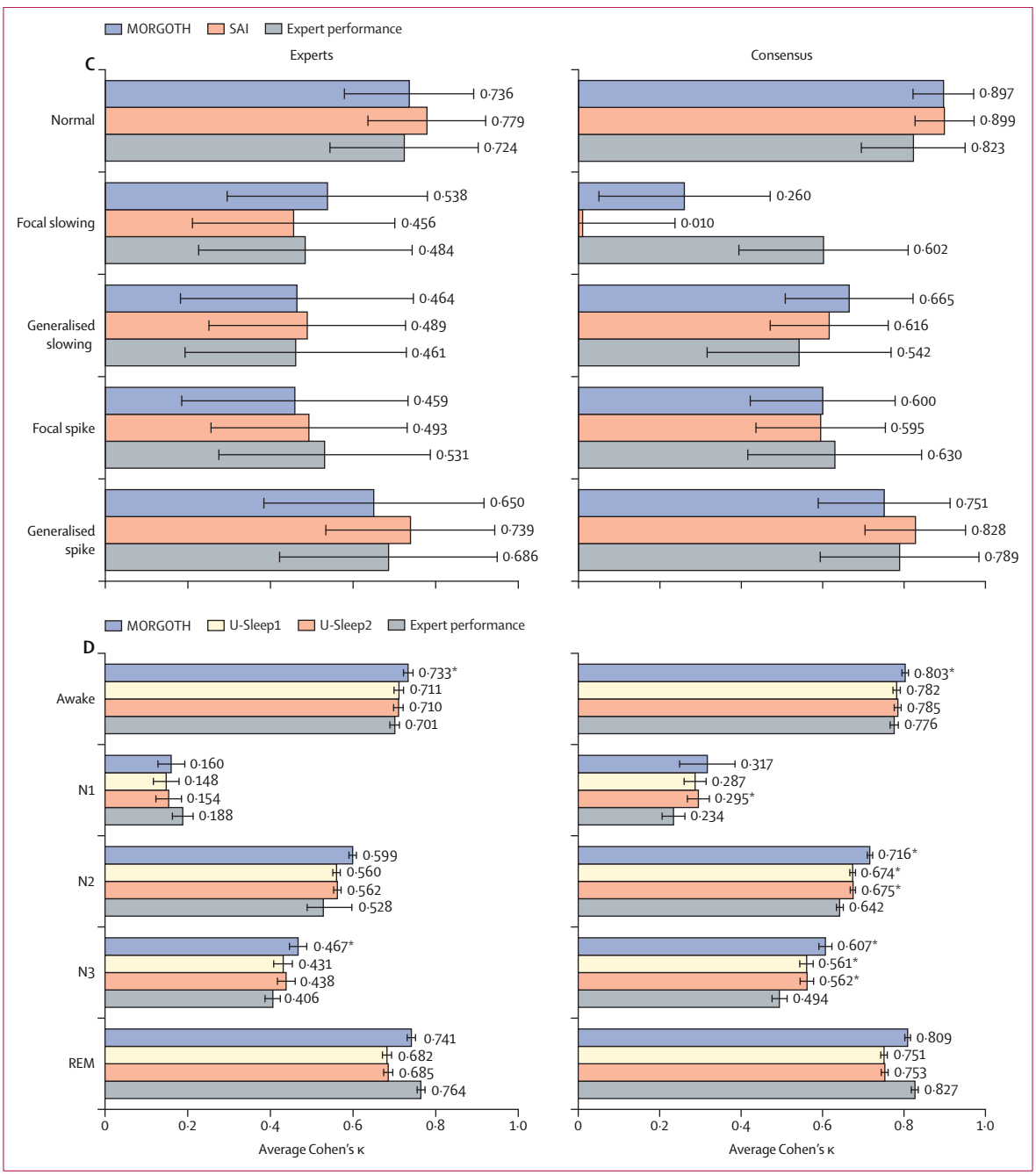


Figure 3: **Reliability analyses on multi-expert scored validation datasets**

Agreement with experts (left) and agreement with consensus (right) on Internal IIIC-Test dataset (A), Internal SN2-Test dataset (B), External SAI dataset (C), and External UPenn dataset (D). MORGOTH is represented by blue, with baseline models by yellow, orange, and green. U-Sleep requires one EEG channel and one electro-oculography channel. Because the UPenn polysomnography dataset contains two EEG channels (C3 and C4) and one electro-oculography channel, U-Sleep was evaluated separately using each EEG channel. The results from the C3 EEG channel are labelled U-Sleep1 and the C4 channel are labelled U-Sleep2 (appendix p 20). Expert performance is represented by grey bars. Error bars indicate 95% CIs from 10 000 bootstrap iterations. Agreement with experts (pairwise inter-rater reliability) analysis is limited to experts who labelled at least 1000 samples in common per class for event-level tasks, and at least 30 patients in common per class for EEG level classes: IIIC-Test dataset has 30 experts with 772 expert pairs, SN2-Test dataset has eight experts with 56 expert pairs, SAI dataset has 14 experts with 164 expert pairs, and UPenn dataset has six experts with 30 expert pairs. EEG=electroencephalogram. GPD=generalised periodic discharges. GRDA=generalised rhythmic delta activity. IIIC=ictal–interictal–injury continuum. LPD=lateralised periodic discharges. LRDA=lateralised rhythmic delta activity. MORGOTH=multidomain omnibus for reading and generalising over thorough EEG interpretation. N1=non-rapid eye movement sleep stage 1. N2=non-rapid eye movement sleep stage 2. N3=non-rapid eye movement sleep stage 3. REM=rapid eye movement sleep. SAI=SCORE-AI. SN2=SpikeNet 2. UPenn=University of Pennsylvania. *The lower limit of the model's 95% CI exceeds the upper limit of the expert-performance 95% CI.

0·91% Multiracial, and 5·65% Other. For external validation, we used seven datasets totalling 1800 EEGs from 1573 patients (51% male and 49% female; median age 45 years [range 0 to >90]) across 48 institutions. Within each individual subset, only one EEG per patient was used for evaluation to ensure the independence of samples within that subset.

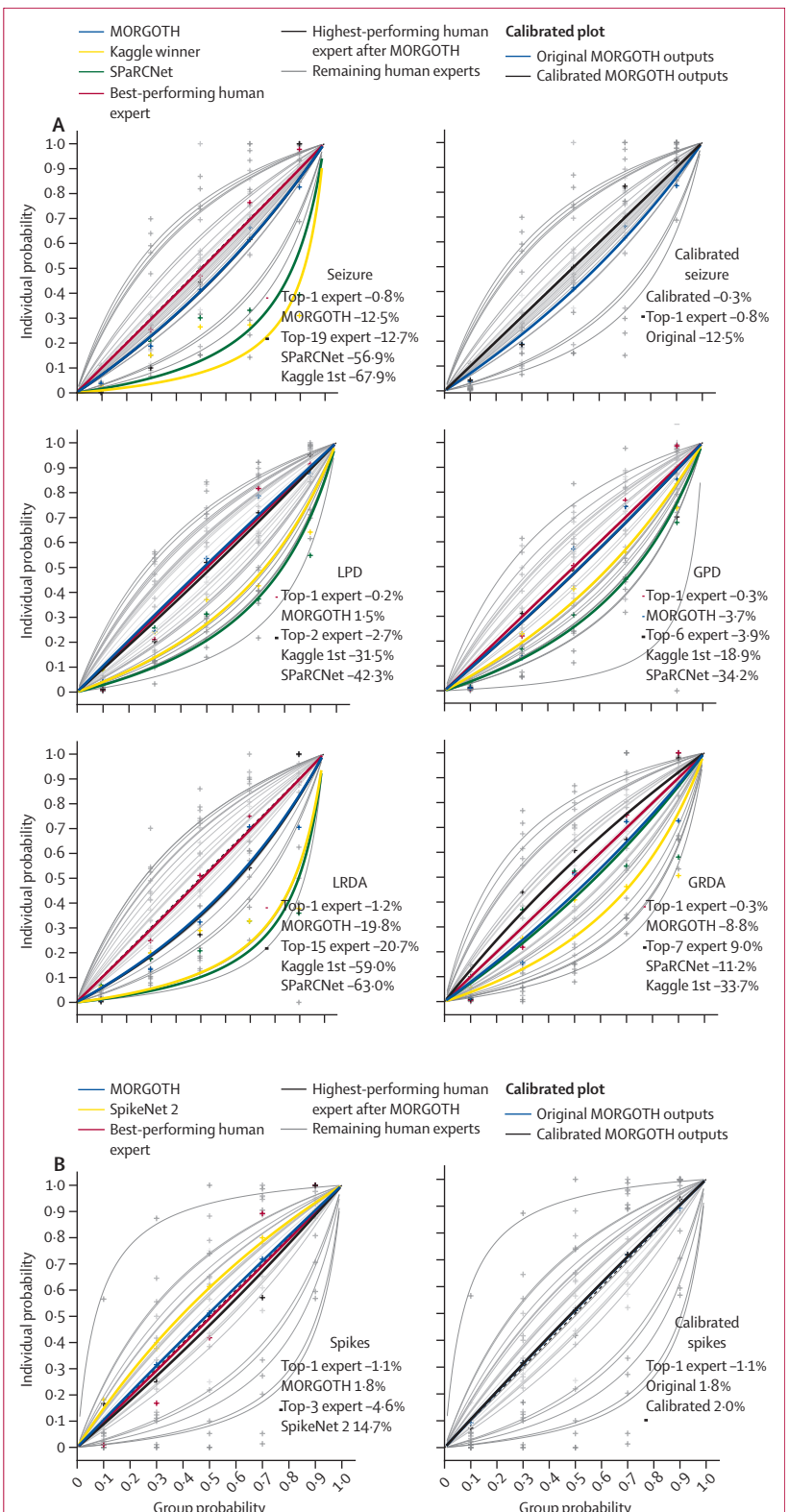
On internal test datasets with single-expert labels (appendix p 24), MORGOTH achieved high accuracy with AUC-ROC values of 0·86–0·98 for event-level tasks and 0·89–0·98 for EEG-level tasks across all detection categories. The average difference between training and test accuracy was 2·75%, never exceeding 5%, indicating minimal overfitting.

On external single-scored datasets (appendix p 26), MORGOTH achieved AUC-ROC values of 0·883 (95% CI 0·881–0·885) on the Temple University Seizure Corpus, 0·729 (0·695–0·763) on the Temple University EEG Events Corpus, 0·747 (0·715–0·779) on the Temple University Slowing Corpus, and 0·902 (0·900–0·904) on the Temple University Abnormal EEG Corpus. MORGOTH also outperformed the baseline model U-Sleep by an average AUC-ROC of 0·030 (0·022–0·038) on the MASS sleep staging task.

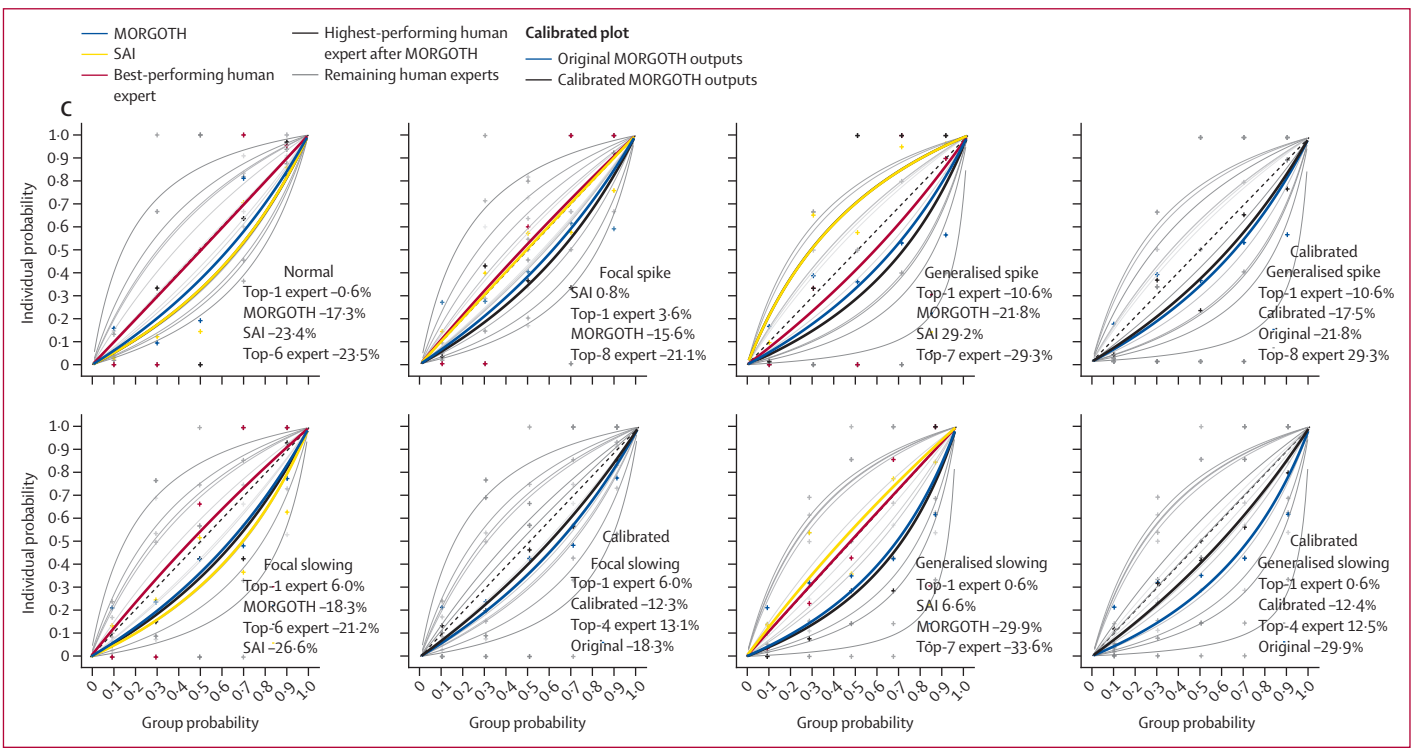
On the MoE dataset (appendix p 27), MORGOTH outperformed 81·0% (95% CI 46·9–100; 17 of 21) of experts across 13 of 14 tasks. On the IIIC-Test set (figure 2B), MORGOTH outperformed 96·6% (82·1–100; 29 of 30) of experts, surpassing all in LPD, GPD, and LRDA and all 24 experts in spike detection on the SN2-Test set (figure 2C).

On the MoE dataset (appendix p 27), MORGOTH achieved an average AUC-ROC of 0·945 (95% CI 0·925–0·962) and AUC-PR of 0·803 (95% CI 0·731–0·860) across 14 categories, outperforming the automated burst suppression detector by 0·084 and 0·428 (statistically significant), respectively. On the IIIC-Test set (figure 2B), MORGOTH reached an AUC-ROC of 0·990 (0·988–0·991) and AUC-PR of 0·940 (0·926–0·953), exceeding SPaRCNet by 0·041 and 0·168, respectively, and the Kaggle winner by 0·126 and 0·387, respectively. On the SN2-Test set (figure 2C), MORGOTH outperformed SpikeNet 2 with perfect AUC-ROC and AUC-PR of 1·00.

On the SAI dataset (figure 2E), MORGOTH outperformed half of the experts, with an average EUC across tasks of 50·62% (95% CI 24·16–77·14) for ROC and 53·50% (26·36–77·58) for PR curves. On the ON dataset with 15 experts (appendix p 28), EUC-ROC was 43·8% (14·3–78·6) and EUC-PR was 39·0% (8·9–75·0). For sleep staging on the UPenn dataset with six experts (figure 2D), EUC-ROC was 41·8% (36·7–46·7) and EUC-PR was 42·6% (40·0–46·7). Despite a low AUC-PR (0·262 [0·248–0·278]) for N1=non-rapid eye movement sleep stage 1 (N1), MORGOTH still outperformed 52·7% of experts; for rapid eye movement (REM) sleep (AUC-PR 0·857 [0·849–0·864]), it did not surpass any. From internal test to external test sets, the event-level AUC dropped by 1·21% (0·00–3·94) and EUC



(Figure 4 continues on next page)



(Figure 4 continues on next page)

by 3·33% (0·00–33·33); at the EEG level, the AUC dropped by 2·12% (0·00–5·83) and EUC by 9·52% (0·00–75·00).

On the SAI dataset (figure 2E), MORGOTH matched SCORE-AI overall, slightly outperforming it on slowing and spike tasks, and achieving similar performance on normal classification. MORGOTH outperformed 14 (50·0%) of 28 experts on average, compared with five (35·7%) of 14 for SCORE-AI. On the HEP dataset (appendix p 28), MORGOTH showed near-perfect event-level performance (AUC-ROC and AUC-PR: 0·999), surpassing SpikeNet 2 and leading at the EEG level as well. On the UPenn dataset (figure 2D), MORGOTH outperformed U-Sleep in both AUC-ROC (0·930 [95% CI 0·927–0·933] *vs* 0·917 [0·914–0·921]) and AUC-PR (0·739 [0·730–0·748] *vs* 0·706 [0·698–0·715]), despite U-Sleep using an additional electro-oculography channel.

Across multi-expert annotated datasets, MORGOTH consistently demonstrated expert-level reliability. On the IIIC-Test set (figure 3A), it outperformed experts on all IRR metrics, with a substantial lead for LPDs. On the MoE dataset (appendix p 30), MORGOTH showed substantial reliability for burst suppression and awake, fair for slowing and seizures, and surpassed expert–expert κ scores in all but N1. On SN2-Test (figure 3B), MORGOTH achieved moderate agreement with individual experts and near-perfect agreement with consensus. On the SAI dataset (figure 3C), it matched expert reliability without substantial advantage. On the ON (appendix p 31) and UPenn (figure 3D) datasets, MORGOTH showed IRR comparable with that of experts. Across the seven multi-expert-annotated datasets, MORGOTH outperformed at least 90% of experts on three datasets (IIIC-Test, SN2-Test, and MoE), outperformed 43 of 65 individual experts (66·2% [95% CI 14·3–100·0]) overall, and exceeded at least 20% of experts on every one of the 17 evaluated tasks, further demonstrating expert-level reliability across diverse EEG interpretation tasks. The calibration curves for MORGOTH and experts across multi-expert annotated datasets are in figure 4 and the appendix (p 29). MORGOTH achieved near-perfect calibration, with calibration indices between –20% and 20% in ten of 16 event-level tasks and four of five EEG-level tasks (0·1–19·8%), outperforming 40·0–96·7% of experts and all baseline models. However, MORGOTH undercalled generalised spikes, generalised slowing, awake, N1, N3, and REM at the EEG level. Post-calibration improvements were observed for generalised spikes (+4·3%) and slowing (+17·5%) using isotonic regression, and for awake (+25·0%), N3 (+30·4%), and REM (+16·6%) using Platt scaling. N1 showed no improvement.

MORGOTH's performance on representative EEGs across tasks, including spike detection and seizure or IIIC classification are in the appendix (pp 36–49). The figures highlight MORGOTH's ability to detect epileptiform discharges, seizures, and rhythmic or periodic patterns while ignoring artifacts and benign variants.

MORGOTH showed a more robust performance across age and sex subgroups than all baselines, with some findings such as spike detection being most stable and slowing

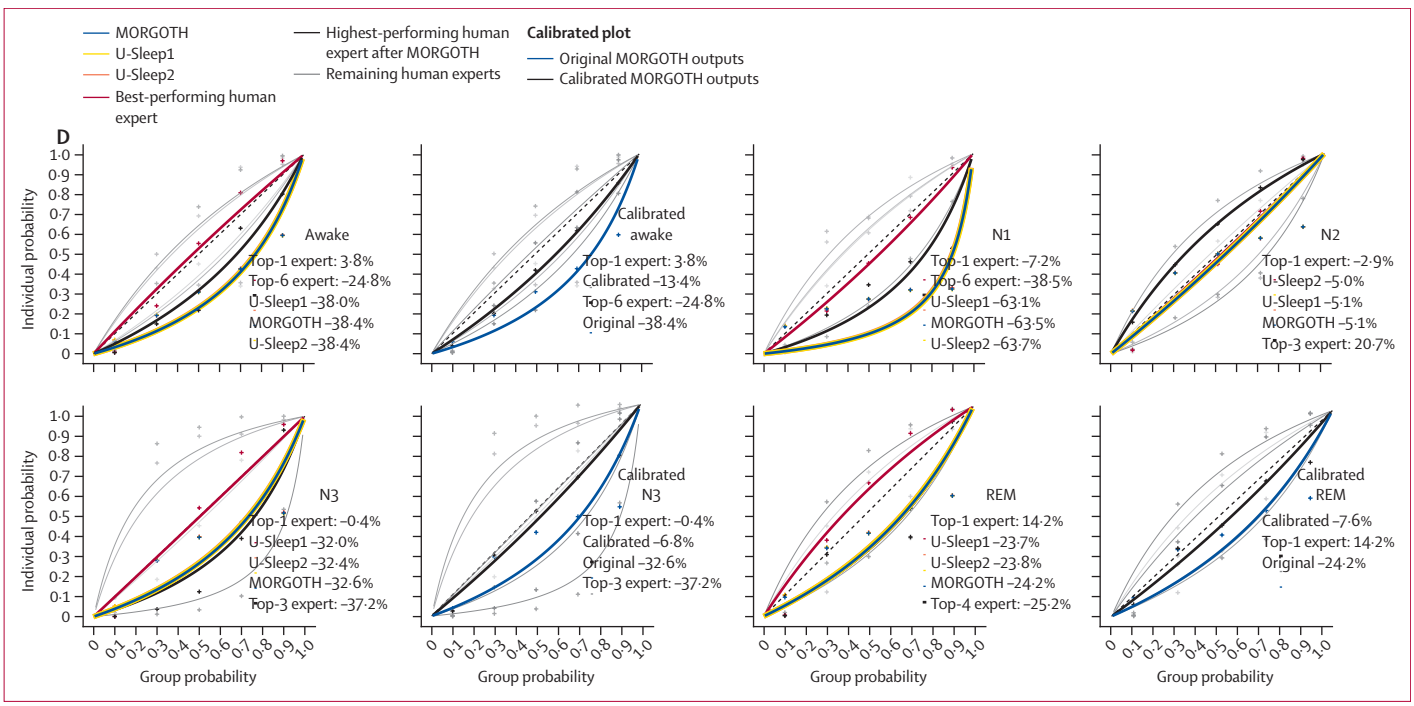


Figure 4: **Calibration curves on multi-expert scored validation datasets**

Type segments were assigned to one of the five probability bins 0–20%, 20–40%, 40–60%, 60–80%, and 80%–100% based on the proportion of experts who classified them into that category. (A) The internal IIIC-Test dataset includes seizure, LPD, GPD, LRDA, and GRDA patterns with calibrations from 30 experts. (B) The internal SN2-Test dataset has spike detection calibration curves with 24 experts. (C) The external SAI dataset covers normal, focal slowing, generalised slowing, focal spike, and generalised spike, with 14 experts' calibration curves involved. (D) The external UPenn dataset includes awake, N1, N2, N3, and REM sleep stages, along with calibration curves from six experts. The dashed diagonal lines indicate the lines of identity (y=x). EEG=electroencephalogram. GPD=generalised periodic discharges. GRDA=generalised rhythmic delta activity. IIIC=ictal–interictal–injury continuum. LPD=lateralised periodic discharges. LRDA=lateralised rhythmic delta activity. MORGOTH=multidomain omnibus for reading and generalising over thorough EEG interpretation. N1=non-rapid eye movement sleep stage 1. N2=non-rapid eye movement sleep stage 2. N3=non-rapid eye movement sleep stage 3. REM=rapid eye movement sleep. SAI=SCORE-AI. SN2=SpikeNet 2. UPenn=University of Pennsylvania.

detection showing greater variability among age groups (appendix p 71).

We also evaluated MORGOTH on the TELEEEG dataset, a global telemedicine platform for remote electroencephalogram interpretation, comprising 8099 EEGs from 23 low-income and middle-income countries. Although performance remained good (AUC-ROC >0·7 in nine [75%] of 12 tasks), some decline was observed, likely due to missing channels and label noise, highlighting the need for further expert review (appendix p 121).

MORGOTH was trained on 19 EEG channels but can handle missing ones by imputing them with the mean of available channels. Local features are more affected by missing channels, and global patterns are more robust (appendix p 126).

Discussion

MORGOTH is the first and most comprehensive AI system to provide fully automated, comprehensive EEG interpretation across all clinical settings. MORGOTH accepts raw EEGs as input and performs automatic channel selection and preprocessing, requiring no additional user intervention, making it user-friendly for clinical implementation. Unlike previous models[8–13] that focus on specific contexts or tasks, MORGOTH achieves expert-level performance on both event-level detection and EEG-level interpretation across a wide range of clinically relevant patterns. Its strong generalisability suggests potential to improve care in low-resource settings and enhance efficiency in specialised centres.[27]

Our methodologically rigorous development process ensured robust performance and generalisability while minimising biases. We developed MORGOTH using diverse EEG data from multiple hospitals and validated it on broad, multi-institutional external datasets across all clinical settings. Expert-labelled test sets with six to 30 annotations per EEG enabled blinded, automated comparison to expert agreement, ensuring robust, generalisable performance with minimal bias.

MORGOTH aids epilepsy diagnosis by detecting epileptiform discharges and distinguishing focal from generalised abnormalities, supporting appropriate treatment decisions. In epilepsy monitoring units, MORGOTH summarises long-term recordings and detects ictal and interictal events for presurgical evaluation.[9] In critical care, MORGOTH identifies seizures and rhythmic and periodic patterns, tracks discharge trends for early ischaemia detection,[28] and monitors pattern evolution for post-cardiac arrest prognosis,[29] matching expert performance. With high specificity (95% for seizures, 92% for rhythmic

patterns, and 99% for spikes), MORGOTH outperforms previous models. The sleep staging capability of MORGOTH further extends its use to sleep labs, enabling broad application across clinical neurophysiology.

A key finding from our validation is the comprehensive assessment of expert agreement across diverse EEG tasks and settings, offering new insight into variability in clinical interpretation. Unlike previous studies limited to specific patterns or segments,[6–9,12–14] our analysis spanned multiple recording types and revealed substantial variation by pattern. Consistent with the Results, MORGOTH matched or exceeded expert consensus, outperforming 66·2% of individual experts across seven multi-expert-annotated datasets. This broad comparison marks a crucial step toward establishing AI reliability for clinical use.

Despite its advances, MORGOTH has limitations. The current model uses only the standard 19 EEG channels and does not incorporate additional electromyography, electrooculography, or other polysomnography signals. Upcoming versions will integrate these modalities to improve MORGOTH's performance on sleep-related tasks. Although trained on a large dataset, the model might reflect biases from tertiary epilepsy centres. MORGOTH does not explicitly detect some normal variants or artifacts (eg, breach rhythm and photoparoxysmal responses), which even experts interpret variably; expert-curated data and hard-example mining might help. The model distinguishes focal versus generalised activity but does not distinguish between finer localisation or seizure subtypes, which limits presurgical utility. MORGOTH also does not use clinical context (eg, treatment response) and currently focuses on pattern recognition rather than outcome prediction such as epilepsy risk or syndrome classification. Although the sleep staging performance is promising, full polysomnography integration and detection of arousals, apnoeas, and limb movements remain future goals.

From an implementation perspective, MORGOTH is being integrated with commercial EEG viewers and automated reporting, although not yet deployed. Aligned with clinical guidelines, MORGOTH is designed to augment, not replace, expert interpretation. MORGOTH 2 reflects this approach by generating preliminary reports linked to raw EEGs, allowing clinicians to review and edit before finalisation. This approach preserves clinical oversight while improving efficiency and consistency. Prospective trials will be key to validating this human–AI workflow for regulatory use.

These capabilities are key goals for MORGOTH 2, as we expand the foundation model toward its full potential in predictive analytics and broader clinical use.

MORGOTH demonstrates expert-level performance in automated EEG interpretation across all clinical settings, simultaneously addressing a broad spectrum of clinically relevant EEG patterns. The AI system could change neurological care by improving access to high-quality interpretation in resource-limited settings, enhancing diagnostic accuracy where specialised training is limited, increasing efficiency in high-volume centres, and enabling continuous monitoring in critical care environments.[30] As the first AI system validated across the entire spectrum of clinical neurophysiology applications, MORGOTH establishes a new standard for comprehensive, expert-level automated EEG interpretation with broad implications for improving neurological care globally.

Contributors

All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication. JAK, SFZ, AFS, and MBW provided administrative, technical, and material support. JJ and MBW oversaw and supervised the study. CSu, JJ, and MBW verified the data and take responsibility for the integrity of the data and accuracy of the analyses. MBW, JJ, CSu, AFS, JAK, and SFZ conceived and designed the study. All authors contributed to data acquisition and interpretation. CSu, JJ, and MBW performed the statistical analyses and verified the analytical results. CSu, JJ, and MBW drafted the manuscript, and all authors critically revised it for important intellectual content.

Declaration of interests

MBW is a co-founder, scientific advisor, consultant to, and has personal equity interest in Beacon Biosignals, and has received grants from the US National Institutes of Health (NIH; RFG064312, RF1NS120947, R01AG073410, R01HL161253, R01NS126282, R01AG073598, R01NS131347, and R01NS130119). JJ receives author royalties from Springer Publishing. AFS received a grant from ceribell and NIH (R01NS111022). IK serves as a consultant for Epitel, Ceribell, Neurotech, UCB, and GSK. TL is listed on patents and patent applications related to the detection, prediction, diagnosis, and treatment of neurological conditions, including epilepsy and seizures (patent numbers: 12150771, 20230397876, 20230386025, 20230141496, 11564617, 10959662, 20210038143, 20200383627, 20190298248, 10278608, 20180206776, and20150216436); his laboratory has received NIH grant funding (grant numbers: U44 NS121562, R01NS111022, R01NS088627, R21NS101381, U01NS090415, U24NS107200, and K23NS076550); has received device donations for research from companies including Epitel and Empatica; has received travel support for scientific presentations from the American Academy of Neurology (AAN), American Clinical Neurophysiology Society (ACNS), and American Education Services; and his laboratory is supported by international academic fellows. AARR serves on the advisory board of SK Life Science and holds ownership in Rodzi Homes and Lourbeth. KM has received a grant from NIH (R61 NS130215-02). AH has received consultancy fees from UCB and Medtronic; honoraria from Natus; royalties from Springer Nature; and has received research funding from the NIH (R01NS132121-01A1; R21HD115285, AWD0012923) and the Swebilius Foundation. MCN receives author royalties from Demos Publishing; research grants from Eisai Canada and Paladin; serves as a consultant for UCB Canada and Eisai Canada (proceeds donated to a hospital charity foundation); and has received grants from CIHR (PJT178217 and DC0190GP). JYY receives author royalties from Elsevier, and has received NIH grants (1UH3NS109557-01A1). JP receives salary from Beacon Biosignals and owns stock in Beacon Biosignals. MCC has served as a consultant and speaker for Nutricia North America/Danone, a speaker for Nestle Health Science, and receives author royalties from Demos/Springer Publishing Company. Nicolas Gaspard serves on the scientific board of Bioserenity; and has received research funding from the Fonds National pour la Recherche Scientifique, Brussels Institute for Research and Innovation (INNOVIRIS), Fonds Erasme pour la Recherche Médicale, and Fonds Jaumotte, all provided to his institution. RJT is co-inventor and patent holder of the ECG/PPG-derived sleep spectrogram, licensed by Beth Israel Deaconess Medical Center (BIDMC) to MyCardio; receives royalties through the BIDMC; has an unlicensed patent on the positive airway pressure gas modulator for treatment of central/complex sleep apnoea; and is a co-inventor of a licensed (by BIDMC to Sleepcare Asia) submitted patent for respiratory self-similarity and inventor for a

licensed (by BIDMC to Sleepcare Asia) enhanced expiratory rebreathing space for treatments for high loop gain sleep apnoea, these two patents have currently no royalties; consults for Guidepoint Global and GLG Councils; and is a co-inventor of licensed auto-continuous positive airway pressure software to DeVilbiss-Drive, without current royalties. DMG is an unpaid advisor for Epilepsy AI and Eysz and has been a paid advisor for Magic Leap; has received speaker fees from AAN, the American Epilepsy Society, ACNS, National Neurotrauma Society, and AI in Epilepsy and Neurology; has previously served as a paid consultant for Neuro Event Labs, Insights Driven Research, LivaNova, Health Advances, Duke University, and Bloom Insights; has received research funding from NIH and BIDMC; and has received a grant from National Institute of Neurological Disorders and Stroke (NINDS; K23NS124656). PWK serves as an expert witness on qEEG, EEG, neurology, and epilepsy; is a member of the IFCN Executive Committee; consults for Natus; and receives author royalties from Demos and Wiley-Blackwell. JWL is cofounder of Soterya (with no financial ties); has performed contract work for Teladoc; and has received an investigator-initiated study grant from SK Biopharmaceuticals. OS is a council member of ACNS and has received research support through a subaward on an NIH grant (R01NS131967). HAH receives author royalties from UpToDate and Springer Publishing; and has received grants from NINDS (5R21NS137117) and NCATS (5UG3TR004501). STH reports grants to her institution from Neuroelectrics, the Epilepsy Foundation, and the NORSE Institute; and has received research support through a subaward on an NIH Small Business Innovation Research grant (NS121559). ELJ serves as an associate editor for *Neurology* and has received research funding from NIH (K23AG063899). JJH has received research funding from the Veterans Affairs Office of Research and Development (I01HX003107-01A2). BLA has received research grant support from the NIH (R01 NS133037) and the Pediatric Epilepsy Research Foundation, with funding provided to his institution. ESR has received research grant support from NIH/NINDS (R01NS117904) and NIH Office of the Director (OT2OD032701), with funding provided to his institution. OT receives salary and research support from the NIH (P20GM130447), the Cognitive Neuroscience and Development of Aging (CONDA) Award, and the US Department of Health and Human Services LB606 Nebraska Stem Cell Grant. EJG has received NIH research grant support (R01NS117904), with funding provided to her institution. MMS has received research funding from the NIH (R01AG060987, R01EB032820). EA has received research funding from the NIH (K23NS119794, 1OT2OD032701, and R01NS128342), the Department of Defense (HT9425-23-1-0242, HT9425-25-1-0170, W81XWH-19-1-0861, and W81XWH-21-C-0075), the American Heart Association Faculty Development Program (AMFDP) 843457, (20CDA35310297 and 24DIVSUP1274116), the Regents of the University of California, Cures Within Reach (2022CAL-Amorim), the Zoll Foundation, and the Hellman Foundation. JAK has received research funding from NINDS (K23NS112596, R01NS117904, and R01NS126282), the Brain Aneurysm Foundation, and the Swebilius Foundation. All other authors declare no competing interests.

Data sharing

All data needed to reproduce the results is available at https://bdsp.io/content/harvard-eeg-db/4.1/. Code is available at https://github.com/bdsp-core/morgoth/. Usage of data and code are restricted to use by the research community.

Acknowledgments

This work was supported by grants from the NIH (RFG064312, RF1NS120947, R01AG073410, R01HL161253, R01NS126282, R01AG073598, R01NS131347, R01NS130119, R61NS130215-02, R01NS132121-01A1, R21HD115285 AWD0012923, P20GM130447, R01AG060987, R01EB032820, K23NS119794, 1OT2OD032701, R01NS128342, R01NS111022, R01NS131967, K23AG063899, and NS121559). This work was also supported by a Swebilius Foundaion Grant; CONDA Award; DHHS LB606 Nebraska Stem Cell Grant, Fonds National pour la Recherche Scientifique; Brussels Institute for Research and Innovation; Fonds Erasme pour la Recherche Médicale; Fonds Jaumotte; Department of Defense (HT9425-23-1-0242, HT9425-25-1-0170, W81XWH-19-1-0861, and W81XWH-21-C-0075); AMFDP 843457 (20CDA35310297 and 24DIVSUP1274116); Regents of the University of California; Cures Within Reach (2022CAL-Amorim); Zoll Foundation; Hellman Foundation; NINDS (K23NS124656, K23NS112596, R01NS117904, and 5R21NS137117); Brain Aneurysm Foundation; and the Veterans Affairs Office of Research and Development (I01HX003107-01A2).

References

1 Thijs RD, Surges R, O'Brien TJ, Sander JW. Epilepsy in adults. *Lancet* 2019; **393:** 689–701.
2 Smith SJM. EEG in the diagnosis, classification, and management of patients with epilepsy. *J Neurol Neurosurg Psychiatry* 2005; **76** (suppl 2)**:** ii2–7.
3 Saxena S, Li S. Defeating epilepsy: a global public health commitment. *Epilepsia Open* 2017; **2:** 153–55.
4 WHO. Global burden of epilepsy and the need for coordinated action at the country level to address its health, social, and public knowledge implications. Resolution WHA68.20. World Health Organization, 2015.
5 Nascimento FA, Gavvala JR. Education research: neurology resident EEG education: a survey of US neurology residency program directors. *Neurology* 2021; **96:** 821–24.
6 Benbadis SR. Errors in EEGs and the misdiagnosis of epilepsy: importance, causes, consequences, and proposed remedies. *Epilepsy Behav* 2007; **11:** 257–62.
7 Tveit J, Aurlien H, Plis S, et al. Automated interpretation of clinical electroencephalograms using artificial intelligence. *JAMA Neurol* 2023; **80:** 805–12.
8 Jing J, Ge W, Hong S, et al. Development of expert-level classification of seizures and rhythmic and periodic patterns during EEG interpretation. *Neurology* 2023; **100:** e1750–62.
9 Li J, Goldenholz DM, Alkofer M, et al. Expert-level detection of epilepsy markers in EEG on short and long timescales. *NEJM AI* 2025; **2**. https://doi.org/10.1056/aioa2401221.
10 Perslev M, Darkner S, Kempfner L, Nikolic M, Jennum PJ, Igel C. U-Sleep: resilient high-frequency sleep staging. *NPJ Digit Med* 2021; **4:** 72.
11 Jing J, Ge W, Struck AF, et al. Interrater reliability of expert electroencephalographers identifying seizures and rhythmic and periodic patterns in EEGs. *Neurology* 2023; **100:** e1737–49.
12 Jiang W, Zhao L, Lu B. Large brain model for learning generic representations with tremendous EEG data in BCI. International Conference on Learning Representations; May 7–11, 2024.
13 Chen Y, Ren K, Song K, et al. EEGFormer: towards transferable and interpretable large-scale EEG foundation model. *arXiv* 2024; published online Jan 11. https://doi.org/10.48550/arXiv.2401.10278 (preprint).
14 Wang C, Subramaniam V, Yaari AU, et al. BrainBERT: Self-supervised representation learning for intracranial recordings. *arXiv* 2023; published online Feb 28. https://doi.org/10.48550/arXiv.2302.14367 (preprint).
15 Sun C, Li H, Li Y, et al. TEST: text prototype aligned embedding to activate LLM's ability for time series. International Conference on Learning Representations; May 7–11, 2024.
16 Sun C, Hong S, Song M, et al. TE-ESN: time encoding echo state network for prediction based on irregularly sampled time series data. Thirtieth International Joint Conference on Artificial Intelligence; Aug 19–27, 2021.
17 Berry R, Quan S, Abreu A. The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications, version 2.6. American Academy of Sleep Medicine, 2020.
18 Sun C, Jing J, Turley N, et al. Harvard Electroencephalography Database: a comprehensive clinical electroencephalographic resource from four Boston hospitals. *Epilepsia* 2025; **66:** 3411–25.
19 Lin T, Goyal P, Girshick R, et al. Focal loss for dense object detection. IEEE International Conference on Computer Vision; Oct 22–29, 2017.
20 Sun C, Li H, Song M, Cai D, Zhang B, Hong S. Continuous diagnosis and prognosis by controlling the update process of deep neural networks. *Patterns (N Y)* 2023; **4:** 100687.

21 Buckwalter G, Chhin S, Rahman S, et al. Recent advances in the TUH EEG corpus: improving the interrater agreement for artifacts and epileptiform events. IEEE Signal Processing in Medicine and Biology Symposium; Dec 4, 2021.

22 West LC, Summers M, Tang S, et al. Evaluation of consensus sleep stage scoring of dysregulated sleep in Parkinson's disease. *Sleep Med* 2023; **107:** 236–42.

23 O'Reilly C, Gosselin N, Carrier J, Nielsen T. Montreal Archive of Sleep Studies: an open-access resource for instrument benchmarking and exploratory research. *J Sleep Res* 2014; **23:** 628–35.

24 Jing J, Lin Z, Yang C, et al. HMS – harmful brain activity classification. Kaggle Competition. 2024. https://kaggle.com/competitions/hms-harmful-brain-activity-classification (accessed May 17, 2025).

25 Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola AJ, Bartlett P, BSchölkopf B, Schuurmans D, eds. Advances in large margin classifiers. MIT Press, 1999: 61–74.

26 Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates. Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. July 23–26, 2002.

27 Li R, Liu L, Lu B. Discrimination of decision confidence levels from EEG signals. 10th International IEEE/EMBS Conference on Neural Engineering. May 4–6, 2021.

28 Kim JA, Rosenthal ES, Biswal S, et al. Epileptiform abnormalities predict delayed cerebral ischemia in subarachnoid hemorrhage. *Clin Neurophysiol* 2017; **128:** 1091–99.

29 Amorim E, Zheng WL, Jing J, et al. Neurophysiology state dynamics underlying acute neurologic recovery after cardiac arrest. *Neurology* 2023; **101:** e940–52.

30 Nascimento FA, Jing J, Strowd R, et al. Competency-based EEG education: a list of "must-know" EEG findings for adult and child neurology residents. *Epileptic Disord* 2022; **24:** 979–82.