Received: 30 December 2024 | Revised: 24 June 2025 | Accepted: 25 June 2025

DOI: 10.1111/epi.18547

Epilepsia®

RESEARCH ARTICLE

Evaluating crowdsourcing for ICU EEG annotation: A comparison with expert performance

Wan-Yee Kong[1,2] | **Fábio A. Nascimento**[3] | **Aaron Struck**[4] | **Erik Duhaime**[5] | **Srishti Kapur**[5] | **Edilberto Amorim**[6] | **Gregory Kapinos**[7] | **Andres Rodriguez**[8] | **Brendan Thomas**[9] | **Masoom Desai**[10] | **Jong Woo Lee**[2,11] | **M. Brandon Westover**[1,2] | **Jin Jing**[1,2]

[1]Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA

[2]Harvard Medical School, Boston, Massachusetts, USA

[3]Washington University School of Medicine, St. Louis, Missouri, USA

[4]University of Wisconsin Hospital and Clinics, Madison, Wisconsin, USA

[5]Centaur Labs, Boston, Massachusetts, USA

[6]University of California San Francisco, San Francisco, California, USA

[7]Icahn School of Medicine at Mount Sinai, New York City, New York, USA

[8]Emory University, Atlanta, Georgia, USA

[9]Massachusetts General Hospital, Boston, Massachusetts, USA

[10]University of new Mexico, Albuquerque, New Mexico, USA

[11]Brigham and Women's Hospital, Boston, Massachusetts, USA

Correspondence
Wan-Yee Kong, 330 Brookline Ave, Boston, MA 02215, USA.
Email: wkong@bidmc.harvard.edu

Funding information
National Science Foundation, Grant/Award Number: 2014431; National Institutes of Health, Grant/Award Number: RF1AG064312, RF1NS120947, R01AG073410 and R01HL161253

Abstract

Objective: Detection of seizures and rhythmic or periodic patterns (SRPPs) on electroencephalography (EEG) is crucial for the diagnosis and management of patients with neurological critical illness. Although automated detection methods are advancing, they require large, high-quality, expert-annotated datasets for training. However, expert annotation is limited by the availability of trained neurophysiologists. Crowdsourcing, or soliciting contributions from a large group of people, may present a potential solution. This study evaluates the feasibility of crowdsourcing annotations of short epochs of EEG recordings by comparing the performance of experts and non-experts in identifying six SRPPs.

Methods: We conducted an EEG scoring contest using a mobile app, involving expert and non-expert participants. Non-experts in our studies include physicians-MD, medical students-MS, nurse-practitioner-NP/physician-assistant-PA/pharmacists, NP-students/PA-students/pharmacy-students/other-healthcare-students, and others. Performance was assessed using pairwise agreement and Fleiss' kappa between experts, and accuracy comparisons between experts and the crowd using individual and weighted majority votes.

Results: A total of 1542 participants (8 experts and 1534 non-experts) answered 478 834 questions across six SRPPs: seizures, generalized and lateralized periodic discharges (GPDs and LPDs), and generalized and lateralized rhythmic delta activity (GRDA LRDA), and "Other." Using individual, non-weighted votes, the crowd's performance was inferior to experts for overall and across six SRPP identification. Using weighted majority votes, the crowd was non-inferior to experts for overall SRPP identification with accuracy of .70, 95% confidence interval [CI]: .69–.70 compared to expert's accuracy of .68, 95% CI: .68–.70.

M. Brandon Westover and Jin Jing are co-senior authors.

The crowd performed comparably or better than experts in identifying most SRPPs, except for LPDs and "Other". No individual expert outperformed the crowd on overall metrics.

Significance: This proof-of-concept highlights the promise of crowd reviewers for obtaining expert-level annotations of SRPPs, which could potentially accelerate the development of large, diverse datasets for training automated detection algorithms. Challenges, such as varying calibration/test splits across crowd participants in the study and the absence of gold standard labels in the real-world settings, remain to be addressed.

KEYWORDS

annotation, crowdsourcing, EEG, machine learning, seizures rhythmic and periodic patterns

1 | INTRODUCTION

In this study, we evaluate and compare the performance of experts and non-experts in identifying six seizure rhythmic and periodic patterns (SRPPs) on electroencephalography (EEG) recordings. We hypothesized that although individual experts will outperform individual non-experts, the collective performance of non-experts using weighted voting would be comparable to that of experts. This comparison aims to assess the feasibility of crowdsourcing as a scalable method for generating high-quality SRPP annotations for training EEG classification algorithms.

Detection of SRPPs is an important component of medical care for patients with neurological critical illness or epilepsy.[1,2] Semiautomated and automated SRPP detection using machine learning (ML) methods have gained traction in recent decades.[3–7] However, large, high-quality, expert-annotated EEG datasets are required to train these algorithms. Current publicly available EEG datasets are mainly annotated for seizure detection/prediction tasks.[8] Only one dataset to date has been curated for SRPPs.[9] To expand on this work and train more generalizable algorithms, larger and more diverse SRPP datasets are needed.

Manual annotation of EEG is labor intensive and expensive, further limited by the availability of trained experts. To facilitate scalable SRPP annotation with the goal of curating large, generalizable datasets, crowdsourcing is an attractive solution.

The idea of crowd-sourcing is not novel. In the 1700s, Marquis de Condorcet in his work "Essay on the Application of Analysis to the Probability of Majority" outlines that majority voting is likely to beat individual votes, assuming that each voter is more likely to vote correctly than incorrectly.[10] This forms the basis of ensemble classifiers in ML, combining the predictions of individual classifiers using majority voting. The value of crowdsourcing has also been established in medical image annotation tasks.[11–15] There are similar studies evaluating crowdsourcing to generate EEG data annotations. Warby et al.[16] investigated the performance of 18 researchers and 695 non-experts compared to 47 experts (polysomnogram technologists) for sleep spindle detection. They found that group consensus from researchers and non-experts outperformed individual experts.

However, it remains unknown whether crowdsourcing can produce annotations of sufficient quality to rival those of experts, particularly for complex EEG patterns like SRPPs. This knowledge gap is critical, as the success of crowdsourcing in this domain could significantly accelerate the development of automated SRPP detection algorithms by providing larger, more diverse training datasets.

Key points

1. The availability of trained neurophysiologists limits expert annotation of seizure rhythmic and periodic patterns (SRPPs) on electroencephalography (EEG).
2. Automated detection methods require large, high-quality, expert-annotated datasets for training.
3. We compare the performance of experts and non-experts in identifying six SRPPs via EEG scoring contest using a mobile app.
4. Using weighted majority scoring, the crowd performed comparably or better than experts in identifying most SRPPs.
5. No individual expert outperformed the crowd on overall metrics.

2 | METHODS

2.1 | Study design and participants

2.1.1 | SRPP scoring contest

This study was conducted at Massachusetts General Hospital and Beth Israel Deaconess Medical Center under institutional review board (IRB) protocols that provided waiver of consent for use of the EEG data for research (Protocol no.: MGH 2013P001024, BIDMC #2024P000804). We conducted an EEG scoring contest using a mobile iOS app (DiagnosUs, https://www.DiagnosUs.com) in collaboration with Centaur Labs (Boston, MA, www.centaurlabs.com). The contest was launched publicly at the Critical Care EEG Monitoring Research Consortium (CCEMRC) annual meeting on June 6, 2021, and subsequently made available to a broader audience of medical professionals and students worldwide. The contest was structured as a short-term competition lasting 1 month, during which participants were presented with EEG images and asked to identify specific SRPPs. Each response was scored for accuracy against a predetermined gold standard, and a real-time leaderboard ranked participants based on performance. Participants were given instant feedback on their responses during the contest.

All participants were required to complete a short pre-contest training using a set of training slides. No post-training assessment or test questions were conducted during the study, as the contest was launched with the goal of capturing a broad range of background and knowledge levels. Throughout the contest, participants assessed six SRPPs: seizures, generalized periodic discharges (GPDs), lateralized periodic discharges (LPDs), generalized rhythmic delta activity (GRDA) lateralized rhythmic delta activity (LRDA), and "Other." The "Other" class encompasses any EEG pattern that does not fall into one of the five primary SRPP categories (LPD, GPD, LRDA, GRDA, or seizure). This class includes normal EEG recordings, muscle artifacts, sleep architecture, and epileptiform discharges that do not meet the specific criteria for the five defined patterns. Each question displayed a 10 s EEG epoch in a bipolar montage, accompanied by a 10 min spectrogram, providing a standardized format for pattern identification (Figure 1). Participants were not allowed to skip questions. To encourage participation, the contest offered small monetary prizes to top performers. Beyond the competitive aspect, the contest served as a novel educational tool, allowing participants to improve their EEG analysis skills through practice and immediate feedback. This dual purpose of competition and education aimed to maximize both the quantity and quality of data collected.

2.1.2 | Participants

The contest was open to the public through the DiagnosUs app. For analysis purposes, we grouped participants based

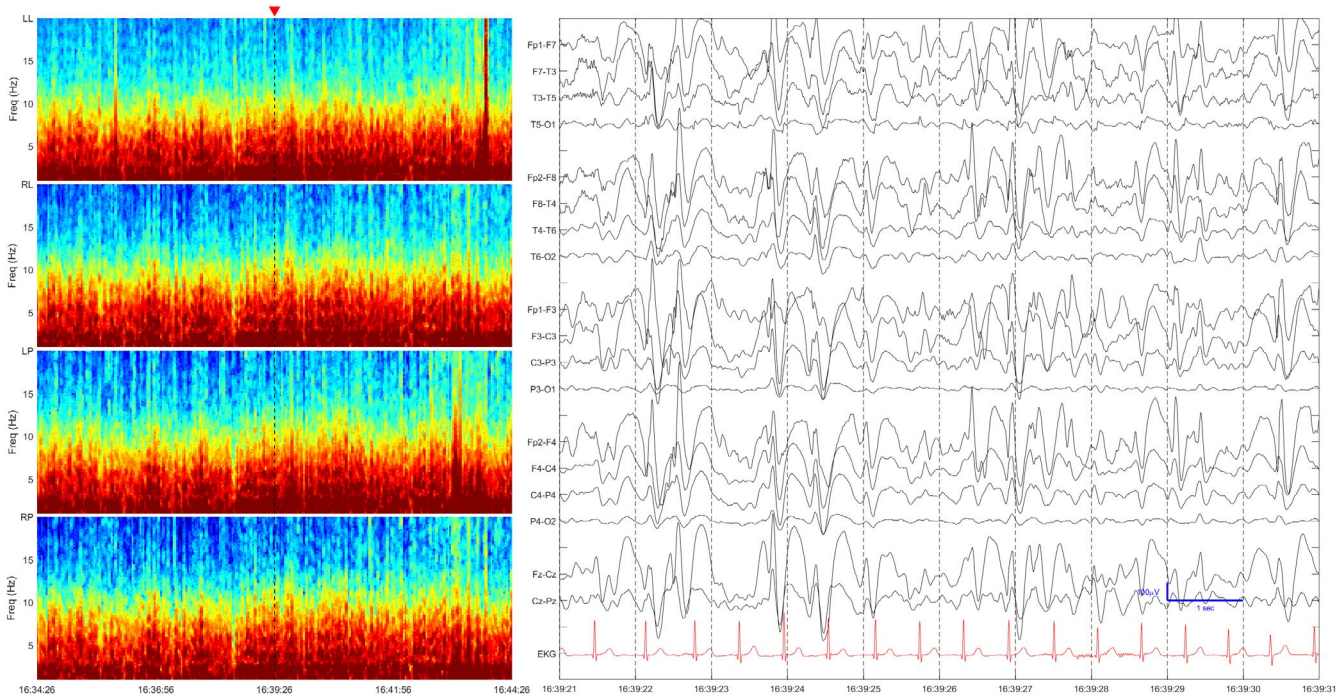

FIGURE 1 Standardized question format. A 10 s electroencephalography (EEG) epoch in a bipolar montage, accompanied by a 10 min spectrogram.

on their experience level. Experts in the study is defined operationally as having completed at least 1 year of specialty training (board certified in epilepsy or clinical neurophysiology) as that is the minimal amount of training required to complete subspecialty fellowship training before going into practice as an independent clinical neurophysiologist. All eight experts who participated in the study had verifiable training and credentials. Level of experience (time in practice) for the experts in our study is between 1 and 15 years (Supplemental S3). Non-expert/crowd participants were divided into five groups: physicians-MD, medical students-MS, nurse practitioner-NP/physician assistant-PA/pharmacists, NP students/PA students/pharmacy students/other healthcare students, and others/experience level not specified.

2.1.3 | Gold standard

In our previous work, 30 expert raters from 18 centers scored 50 697 EEG segments from 2711 patients, for the six SRPPs. Specifics on EEG sources and candidate interictal epileptiform discharges (IED) collection have been described previously.[9] Not all segments were scored by all 30 experts. Only segments with 10 or more expert votes were included in the gold standard for the present study. For each segment, the correct class label was defined to be the one receiving the most expert votes. The certainty of a label for samples with ≥10 votes was published previously: 35.2% were top class received >80% votes, 13.1% where the difference in the top 2 classes were <10%, and 51.7% were in-between.[9] Ties were infrequent; however, when they occurred, they were resolved arbitrarily using the following priority order: Seizure > LPD > GPD > LRDA > GRDA > Other.

2.2 | Interrater agreement between experts in the study and gold standard experts

The overall level of agreement between pairs of raters was calculated based on all the cases where both raters answered a common question. Only pairs of raters who answered at least five questions in common were included in the analysis. The pattern-specific pairwise agreement was computed using the same approach.

Fleiss' kappa (κ) was obtained by comparing the observed agreement (mean item-level agreement score) with the expected agreement (overall distribution of ratings), adjusting for chance agreement. For subgroup analyses, both Fleiss' kappa and Gwet's AC1 (AC1) were computed for each subgroup. In subgroup analyses, the distribution of ratings is often skewed, since each subgroup is defined by a dominant category, which inflates the expected chance agreement and results in a lower Fleiss' kappa. Gwet's AC1 is less sensitive to these prevalence effects and provides a more accurate estimate of inter-rater agreement.[17,18]

2.3 | Calibration and test dataset

We first group the data by user, questions, and experience level, aggregating the unique responses from the user's answer and gold standard (correct answer) into sets. For each user, we compute two greedy set covers—one based on the user's answer and one based on the gold standard—to determine the minimal set of questions that together cover all required patterns in both settings; the union of these two covers forms the calibration set. Only users who achieve full coverage in both user's answer and gold standard and who have at least one additional unused question remaining were qualified. Finally, we split the data into two sets: a calibration dataset for each qualified user, which contains the user-question pairs reserved for calibration; and a test dataset, which includes all the remaining unused questions combinations for those same users.

We did not split the gold standard randomly into two parts—one for computing weights and one for assessing performance—as each user answered a different set of questions, and such an approach would dramatically reduce the available user-question set. Instead, this approach selects a unique set of calibration questions for each user, which were completely excluded from the test dataset for that user, ensuring that the calibration and test sets are mutually exclusive. The calibration dataset will be used only to compute accuracy, whereas the test dataset will serve for performance and sensitivity analyses.

2.4 | Performance of experts vs crowd using individual, non-weighted votes

We fitted two mixed-effects models using Restricted Maximum Likelihood (REML) estimation. In the overall (across all SRPPs) model (Model 1), the outcome was accuracy—the proportion of correct answers for each question by each group (Expert vs Crowd). Fixed effects included the group indicator and the average number of questions answered to adjust for the learning effect, and a random intercept was included for each problem to account for between-problem variability. In the by-pattern model (Model 2), we included the SRPP as a fixed effect, along with an interaction between group and SRPP.

We extracted the group coefficient and its standard error from the model's covariance matrix. We then computed a one-sided z-statistic with a non-inferiority margin of .05 (with a null hypothesis that Expert has at least 5% higher accuracy than Crowd). In this analysis, a non-inferiority margin (δ) of 5% was selected as a conservative threshold—this choice was made in the absence of an established non-inferiority margin in the literature. We computed a one-sided p-value (pattern-specific p-values were adjusted with Bonferroni correction to account for multiple comparisons) with $p < .025$ considered statistically significant.

2.5 | Performance of experts vs crowd using weighted majority votes

To compare the performance of experts and the crowd, the majority vote for each group was compared against the gold standard answer.[19] These votes were weighted using the individual expert's and the crowd's accuracy, defined below.

For the weighted analysis, analogous models were estimated. In the overall weighted model (Model 3), the outcome was a binary weighted accuracy—computed by taking a weighted majority vote for each problem. Fixed effects again included group and the average number of questions answered, with a random intercept for problem capturing between-problem variability. The by-pattern weighted model (Model 4) incorporated SRPP as a fixed effect, along with an interaction between group and pattern, to investigate whether the weighted accuracy differences between experts and the crowd differ across SRPP.

We then computed a one-sided z-statistic with a non-inferiority margin of .05. We computed a one-sided p-value (pattern-specific p-values were adjusted with Bonferroni correction) with $p < .025$ considered significant. The weighted majority vote was calculated as follows:

$$\text{Weighted Majority Vote} = \arg\max_{c \in C} \sum_{i=1}^{N} w_i \times I\left(c_{i,j} = c\right),$$

where

$$\omega_i = \frac{\sum_{p=1}^{P} \text{Accuracy}_{i,p}}{P}$$

Where C is the set of all possible classes (LPD, LRDA, GRDA, GPD, Seizures, other), N is the total number of participants, w_i is the weight for the i-th participant, defined as their average accuracy across all patterns. Arg_max{$c \in C$} is the value of the class C with the maximum weighted votes, and P is the total number of patterns (in this case, $p = 6$). Accuracy$_{i,p}$ is the accuracy of the i-th participant for the p-th pattern. $I\left(c_{i,j} = c\right)$ is an indicator function that equals 1 if a participant i's vote for problem j matches class c, and 0 otherwise.

2.6 | Accuracy

Accuracy for each participant was calculated for each pattern. Responses were grouped by participant and pattern, and the values for True Positives (TPs), True Negatives (TNs), False Positives (FPs), and False Negatives (FNs) were aggregated accordingly. Accuracy is defined as the sum of TPs and TNs divided by the total of TPs, TNs, FPs, and FNs: $Accuracy = (TP + TN)/(TP + TN + FP + FN)$. For each participant, the average accuracy across all patterns was obtained by summing the accuracy for the six patterns, then dividing that total by 6. The average accuracy was computed using the calibration dataset and used as a weight for the calculation of weighted majority scoring on the test dataset. The distribution of the crowd's accuracy for each SRPP, along with the average that was used as a weight, is shown in Supplemental S11.

2.7 | Micro- and macro-averaged F1 scores

For the non-weighted approach, we compared each user's response directly to the corresponding gold standard label for each EEG segment and calculated the F1 scores. The micro F1 score is computed by aggregating the overall counts of TPs, FPs, and FNs across all classes, whereas the macro F1 score is obtained by averaging the F1 scores calculated independently for each class. For the weighted approach, we first derived a single prediction for each EEG segment using a majority weighted voting method that selects the answer with the highest cumulative score. We then computed the F1 scores using the same definitions as in the non-weighted approach.

2.8 | Performance of experts vs crowd subgroups

We fit a mixed-effects model using REML, with experts and crowd subgroups (distinguishing between Neurology and Non-Neurology) as fixed effects. The model also includes the average number of questions answered per user as an additional fixed effect and included a random

intercept for each problem to control for between-problem variability. For each crowd subgroup, we performed an analysis of covariance (ANCOVA) to estimate the adjusted difference between expert and crowd performance. Two-sided *p*-values were adjusted for multiple testing using Bonferroni correction, with *p*-value <.05 considered statistically significant.

2.9 | Performance of individual expert vs crowd

The performance of each expert was also compared against the crowd's majority votes computed above, using independent *t* tests. The *p*-value is computed by calculating the Z-score from the difference in accuracy between each expert and crowd, with $p < .05$ considered statistically significant. For this assessment, any expert who responded to fewer than five questions for any given SRPP was excluded from the comparison.

2.10 | Sensitivity analyses

To evaluate whether crowd performance is influenced by the number of user responses per question or the crowd's experience level, we calculated the weighted majority accuracy of the crowd after sequentially removing different groups from each experience level, one at a time.

To evaluate the number of crowd consensus labels needed to match the performance of experts, we use a bootstrap resampling technique to assess how the accuracy of the crowd changes across different sample sizes. For a sample size of 5–50, repeated sampling with replacement was performed to generate multiple datasets, each containing a randomly selected subset of users. For each of these resampled datasets, weighted majority accuracy was computed. This was repeated across 1000 bootstrap iterations. The mean accuracy of the bootstrapped samples and the 95% confidence intervals (CIs) were derived from the empirical distribution of bootstrap estimates, which were compared against expert accuracy and treated as a fixed benchmark across sample sizes.

To assess whether hard-filtering improves crowd performance compared to the weighted majority approach, we filtered the crowd participants that excluded the bottom 20% of crowd participants based on calibration set accuracy, effectively removing the poorly performing participants. We then compared the average performance of this filtered crowd on the test dataset using non-majority weighted scoring.

3 | RESULTS

3.1 | Participants

A total of 1542 participants were eligible for calibration and test dataset split. For the calibration dataset, the range of question count answered were 6–11 (Table 1). For the test dataset, the number of participants stratified by participation level (number of questions answered/number of users) are as follows: expert (8433/8), MD/DO (56 025/77), medical student-MS (249 827/538), nurse practitioner-NP/physician assistant-PA/pharmacist (12 960/61), NP student/PA student/pharmacy student/other healthcare student (56 034/378), others/experience level not specified (95 555/480) (Table 1). The calibration dataset had 17 619 answered questions, whereas the test dataset had 478 834. Consequently, the calibration set represents 4% of the overall dataset.

Survey responses on participants' place of residence revealed that many live in the Philippines and the United States, with representation spanning all continents. However, most participants did not provide their location (Supplemental S1).

The most common pattern (count/percentage) in the questions set was 'Other' (102 693/21.4), followed by seizure (89 868/18.8), LPD (77 718/16.2), GPD (73 058/15.3), GRDA (67 989/14.2), and LRDA (67 508/14.1). Most participants answered between 1 and 1000 questions with accuracy ranges between .67 and .92; participants answering between 1000 and 2000 had accuracy ranges between .67 and .86, and those who answered >2000 questions had accuracies in the range .72–.87. The correlation between accuracy and number of questions answered per participants is .75. (Supplemental S2).

3.2 | Inter-rater agreement between experts

The inter-rater agreement (pairwise agreement-PA/κ/AC1) for 8 experts in the study and 30 gold standard experts (Supplemental S3) were as follows: Overall (PA-.57/.53; κ-.54/.43); GPD (PA-.60/.46; AC1-.7/.56); GRDA (PA-.48/.39; AC1-.53/.44); LPD (PA-.41/.54; AC1-.51/.57); LRDA (PA-.44/.31; AC1-.54/.34); Other (PA-.57/.54; AC1-.61/.49); and Seizure (PA-.5/.4; AC1-.6/.5). Interrater agreement among the gold standard experts was lower than that of the experts in the study—both overall and for most SRPP categories—across various metrics, with the exception of the LPD category. However, the distribution of the agreement rate for gold standard experts was more widely dispersed (Supplemental S4 and S5). Among the gold standard

TABLE 1 Number of questions answered and number of users in each group (stratified by experience level) for calibration and test datasets.

Calibration dataset		
Group	**Experience level**	**Range of questions answered per user/ number of users**
Expert	Expert	8–11/8
MD	MD	7–11/73
	DO	8–10/4
Medical student	Medical student	6–11/538
NP/PA/pharmacist	NP	7–11/45
	PA	8–11/7
	Pharmacist	8–11/9
Other students	NP student	7–11/38
	PA student	9–11/5
	Pharmacy student	8–11/14
	Other healthcare student	7–11/321
Others	Others	6–11/480
Test dataset		
Group	**Experience level**	**Number of questions answered/ number of users**
Expert	Expert	8433/8
MD	MD (all)	55 829/73
	MD (neurology)	786/3
	DO	196/4
Medical student	Medical student (all)	249 827/538
	Medical student (neurology)	38 959 /28
NP/PA/pharmacist	NP (all)	10 185/45
	NP (neurology)	801/2
	PA	1909/7
	Pharmacist	866/9
Other students	NP student	2003/38
	PA student	725/5
	Pharmacy student (all)	3939/14
	Pharmacy student (neurology)	1529/1
	Other healthcare student (all)	49 367/321
	Other healthcare student (neurology)	1772/21
Others	Others (all)	95 555/480
	Others (neurology)	2918/30

Note: Among the crowd subgroups that identified neurology as their subspecialty, this information was also displayed.

Abbreviation: DO, doctor of osteopathic medicine; MD, doctor of medicine; NP, nurse practitioner; PA, physician assistant.

experts, the mean agreement with the consensus vote was 68.2% (95% CI: 65.5%–70.8%).

3.3 | Performance of individual experts vs individual members of the crowd

We compare the average performance of all individual experts with the average performance of all individual members of the crowd. Overall, the average crowd performance was not non-inferior to the average experts in identifying SRPP, with accuracy difference: −.22, 95% CI: −.23 to −.21 and one-sided non-inferiority *p*-value: 1. Crowd's vs Experts' performance for SRPPs were as follows (accuracy difference, 95% CI): LPD (−.26, −.28 to −.24); GPD (−.22, −.24 to −.2); LRDA (−.22, −.25 to −.19); GRDA (−.2, −.23 to −.18); Seizure (−.11, −.13 to −.09); Other (−.32, −.34 to −.3). The adjusted one-sided non-inferior *p*-value for all SRPPs was 1. The 95% CI lies entirely below 0 for overall and all SRPPs, indicating that the experts were superior to the crowd (Figures 2 and 3).

3.4 | Performance of experts vs crowd using weighted majority votes

Overall SRPP classification performance of the crowd vs the experts was measured using weighted majority voting within each group. The crowd was non-inferior to the experts, with an accuracy difference of −.001 (95% CI −.015 to .013, one-sided non-inferior *p*-value <.001). (Figures 2 and 3).

For SRPP subgroup identification, the crowd was not non-inferior (accuracy difference/adjusted one-sided non-inferior *p*-value) to the experts for LPDs (−.04, 1) and Other (−.17, 1). The crowd was non-inferior to experts for GPDs (.07, <.001), GRDA (.14, <.001), LRDA (.1, <.001), and Seizure (.06, <.001) classification. The 95% CI for GPDs, GRDA, LRDA, and Seizure were entirely above 0, which indicates that the crowd's performance was statistically superior to that of experts (Figures 2 and 3).

3.5 | Effect of number of questions answered on performance

Derived from the model's coefficient, for each additional question answered by a rater, the overall (non-weighted and weighted) and by-pattern (non-weighted) accuracy increased by .001% (*p*-value <.001); the by-pattern (weighted) accuracy has no significant association with number of questions answered.

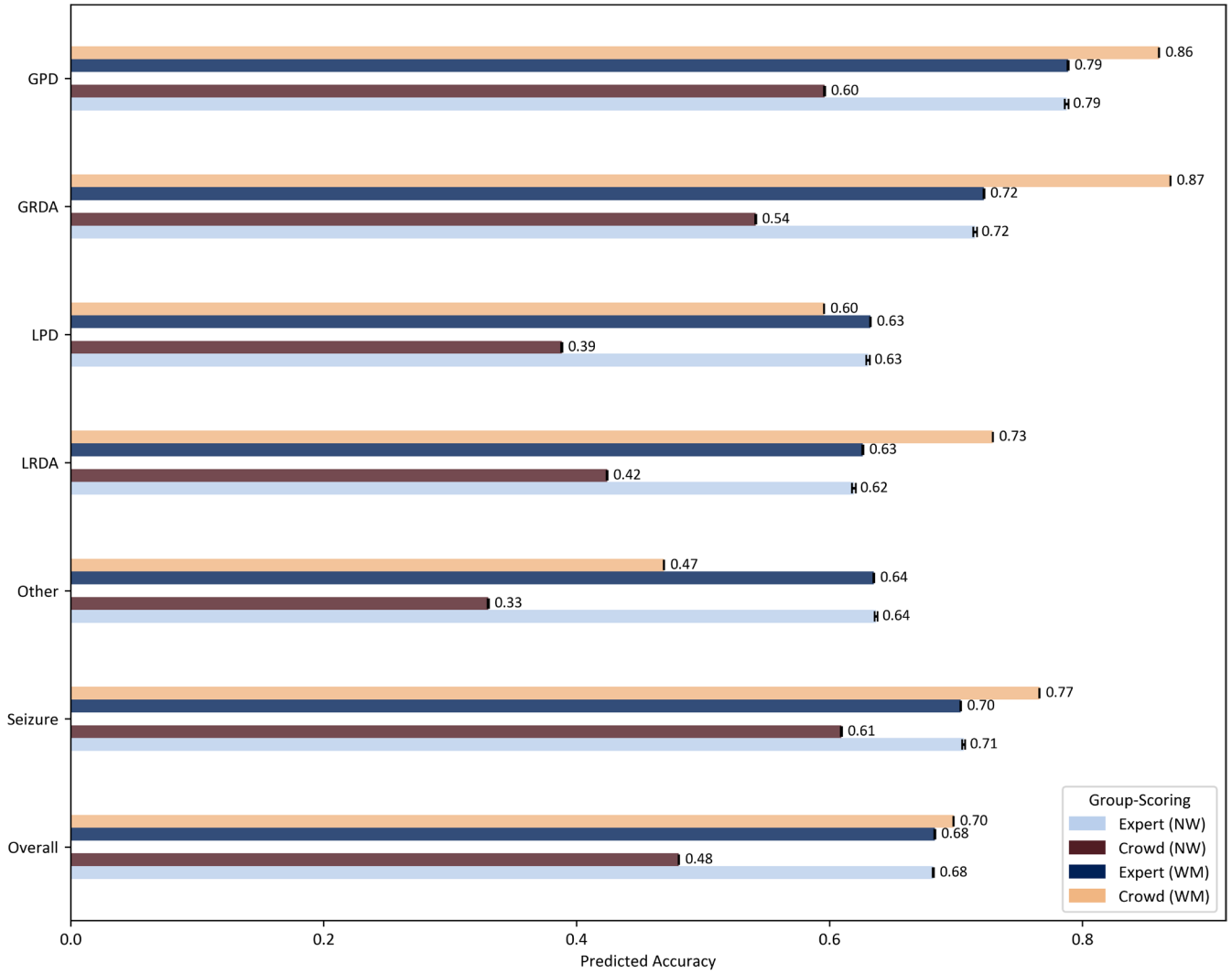

FIGURE 2 Performance of crowd vs experts using individual, non-weighted votes and weighted majority vote. The predicted accuracy (NW, non-weighted; WM, weighted majority accuracy) derived from the mixed-effects models (represented by bar graph) and the corresponding 95% confidence intervals (represented by error bars) were calculated for expert and crowd for overall and each SRPP (x-axis). GPD, generalized periodic discharge; GRDA, generalized rhythmic delta activity; LPD, lateralized periodic discharge; LRDA, lateralized rhythmic delta activity; NW, non-weighted, average proportion of correct answers; WM, weighted majority accuracy.

3.6 | Comparison of the crowd vs individual experts

Crowd majority accuracy for overall SRPP identification was higher than three experts with an accuracy difference range of .06–.11 ($p < .05$); no significant difference was found for the other five experts. None of the experts performed better than the crowd when scored individually (Figure 4). Crowd majority accuracy was either higher than or not significantly different from any individual expert accuracy for GRDA, LRDA, and seizure identification.

Only one expert performed better than the crowd for GPD identification, with an accuracy difference of .14 ($p < .05$). Three experts performed better than the crowd for LPD identification, with an accuracy difference ranged .11 and .4 ($p < .05$). Six of eight experts performed better than the crowd for the category Other, with an accuracy difference range of .07–.38 ($p < .05$) (Supplemental S6).

3.7 | Micro- and macro-averaged F1 score for overall results

Using unweighted metrics, Expert's vs Crowd's F1 scores for overall performance were as follows: micro-averaged F1 score (.68/.47) and macro-averaged F1 score (.68/.47).

Using weighted majority votes, Expert's vs Crowd's F1 score for overall performance were as follows: micro-averaged F1 score (.69/.70) and macro-averaged F1 score (.68/.69).

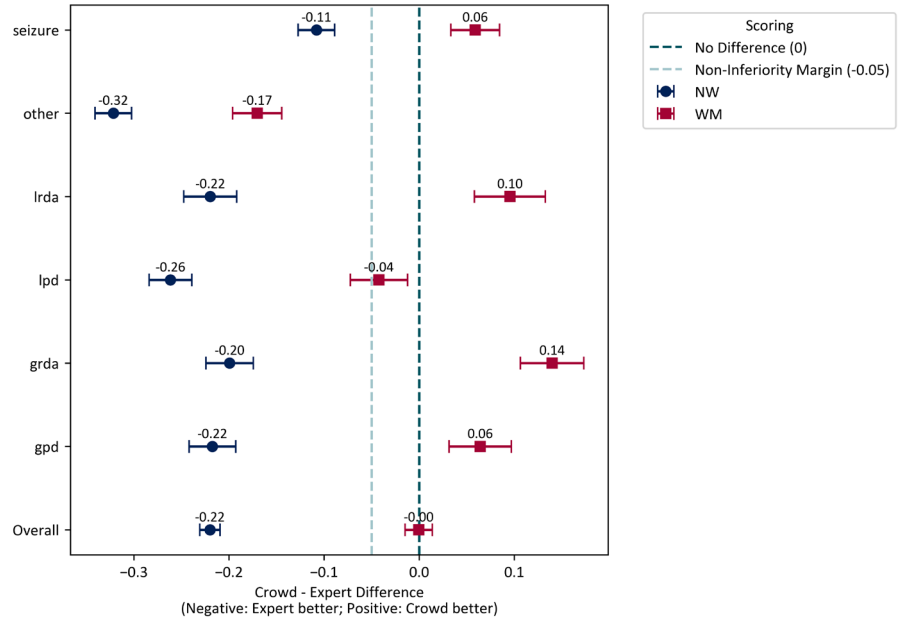

FIGURE 3 Combined forest plot: Point estimates of accuracy difference between crowd and experts (NW, WM). This forest plot displays the adjusted difference in accuracy between crowd and expert (Crowd—Expert) across various SRPPs (y-axis), for both non-weighted (NW) and weighted majority accuracy (WM). Each dark blue circle (NW) or red square (WM) represents the point estimate from the mixed-effects model, with horizontal error bars indicating the 95% confidence interval. A value below zero (to the left) indicates that experts perform better, whereas a value above zero (to the right) indicates that the crowd performs better. The green dashed vertical line at 0 represents "no difference," and the light blue dashed vertical line at .05 is the non-inferiority margin.

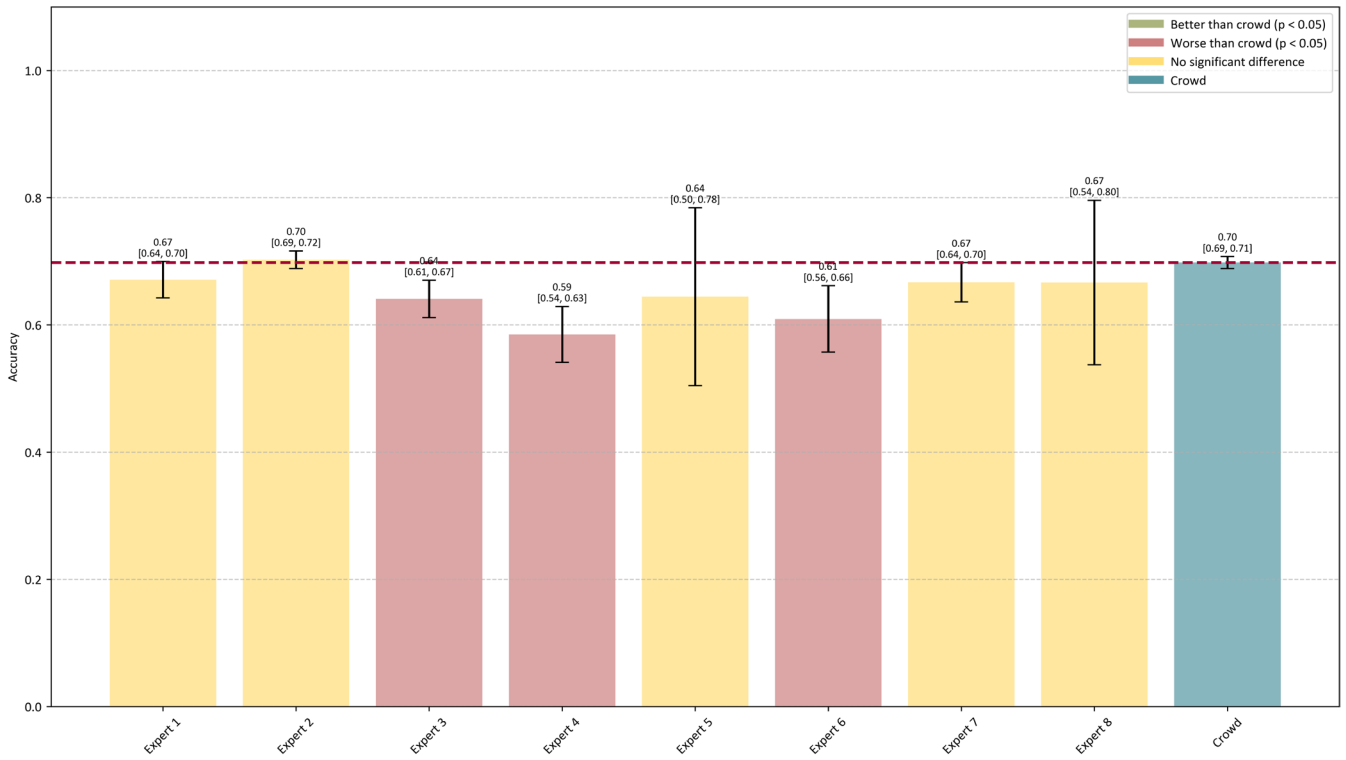

FIGURE 4 Performance of individual expert vs crowd as a group. Each expert's accuracy (represented by either a red/yellow bar graph), the crowd's weighted majority votes (represented by a blue bar graph), and the corresponding 95% confidence intervals (represented by error bars) were shown. Red dotted line represents the crowd's weighted majority accuracy.

3.8 | Performance of experts vs crowd subgroups

Experts generally outperformed the crowd ($p < .05$) for most subgroups—with accuracy differences ranging from .16 to .36—except for pharmacy student (Neurology). Although pharmacy student (Neurology) displayed a positive difference, this finding was not statistically significant (adjusted p-value = .14). A neurology subspecialty improved accuracy for some non-expert groups—MDs and pharmacy students—while offering either no advantage or reduced performance for PAs, "Other" groups, NPs, and medical student subgroups (Supplemental S10).

3.9 | Sensitivity analyses using the leave-one-group-out approach

The crowd's weighted majority accuracy for overall and each SRPP identification was not affected by the removal of any participants subgroups of the crowd (Supplemental S7 and S8).

An assessment of crowd performance across various sample sizes revealed improvement in accuracy as the number of user participants per question increased. The crowd's accuracy steadily approached that of expert consensus as the sample size grew, with crowd performance equaling expert majority accuracy when the sample size reached 20 participants. (Figure 5).

By applying a hard filter that excludes the bottom 20% of crowd raters based on calibration set accuracy—their average performance on the test dataset (using non-majority weighted scoring) gained improvement ranging from .02 to .05 (Overall-.03; Seizure/Other-.02; LRDA/LPD-.03; GRDA/GPD-.05) (Supplemental S9).

4 | DISCUSSION

Our results demonstrate that crowdsourcing is a feasible and scalable method for annotating SRPPs, achieving performance levels comparable to those of expert annotations for most patterns. Although individual expert rater outperformed individual crowd rater, we found that weighted

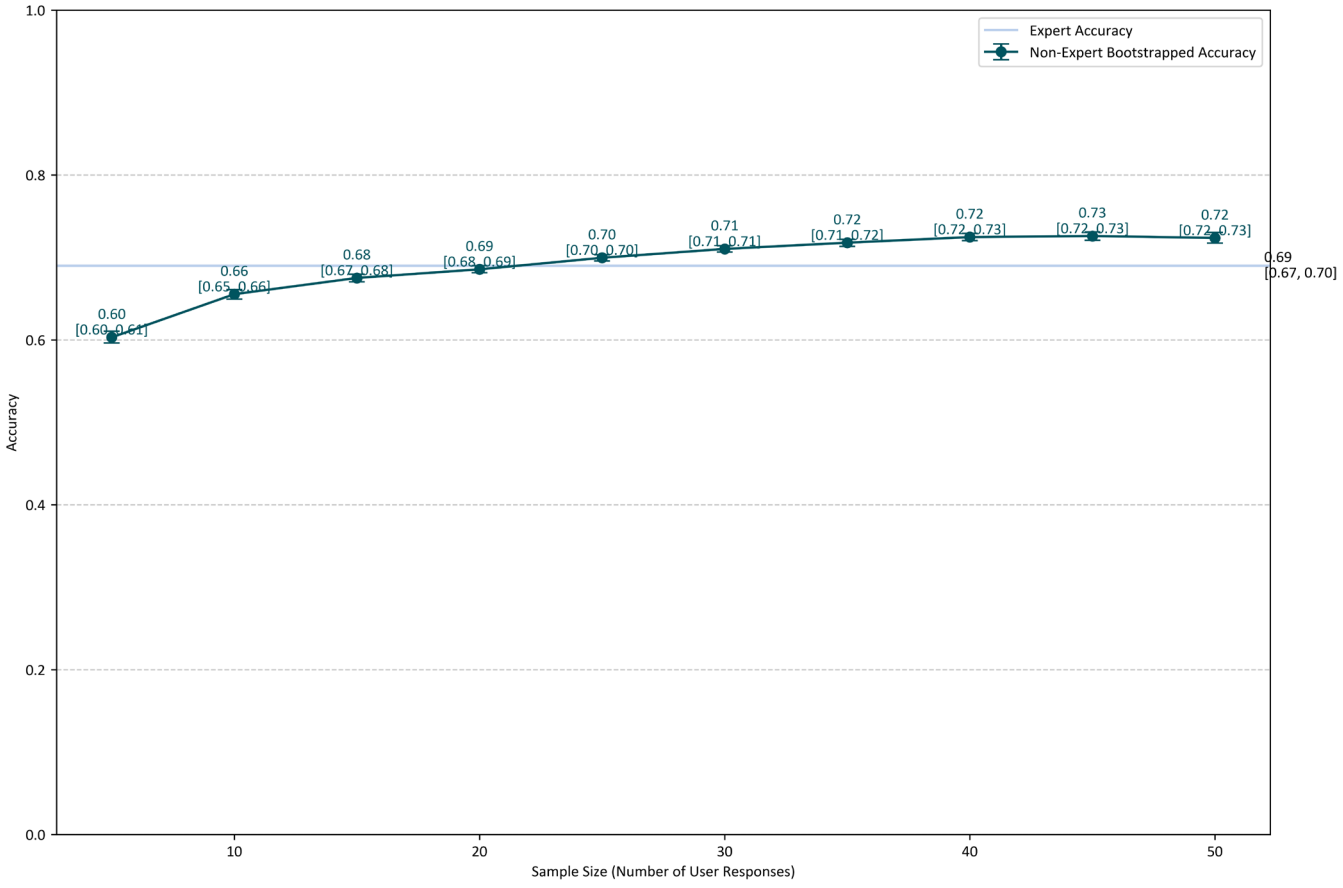

FIGURE 5 Crowd's performance across various sample sizes. Bootstrap resampling technique to assess how the accuracy of crowd changes across different sample sizes (x-axis). For each of these resampled datasets, weighted majority accuracy (represented by green dots), and the corresponding 95% confidence intervals (represented by error bars) were shown. Expert weighted majority accuracy, treated as a fixed benchmark across multiple sample sizes, is represented by a blue line.

majority voting based on the crowd matched or exceeded the accuracy of experts in identifying several SRPP types, including generalized periodic discharges and rhythmic delta activity. The crowd slightly underperformed with lateralized periodic discharges, indicating a need for refined strategies to enhance crowd performance in these cases. Overall, our findings suggest that crowdsourcing can serve as an effective tool for generating large, high-quality datasets for developing robust machine-learning algorithms for automated EEG analysis, potentially accelerating advancements in the field.

We found that, in most cases, the crowd performed as well as or better than individual experts in identifying SRPP (except LPD and Other). Moreover, the crowd's performance was comparable to or surpassed the weighted majority of expert votes. Potential challenges in the crowd's identification of LPD may be attributed to the limitation in montage selection. However, even experts find LPD difficult to identify, as these had the lowest pairwise agreement among evaluated SRPPs. Crowd performance on "Other" may reflect unfamiliarity with what it entails, hence selecting one of the five predefined options rather than accurately classifying the pattern as "Other." Although prior studies have compared the "wisdom of the crowd" against individual experts,[11,16] we believe that this comparison alone is insufficient. There is no clear definition for a true gold standard for EEG labeling, although visual inspection by human domain experts is traditionally used. Studies have shown that even expert consensus can be challenging to establish, as demonstrated by only moderate inter-rater reliability for pattern identification between experts.[9] Therefore, the gold standard for EEG labeling should ideally be established through expert consensus rather than by relying on individual expert judgments.[20,21] This consensus, rather than individual expert performance, should serve as the baseline for evaluating crowd performance. Our study's findings are particularly encouraging because the crowd not only outperformed individual experts but also matched or exceeded this more stringent standard of expert consensus.

In our study, the crowd consisted of individuals from diverse backgrounds, including health care professionals and those without medical expertise. The crowd had lower individual average accuracy compared to experts. We sought to answer several pivotal questions: What strategies can be employed to aggregate crowd wisdom? Is there a critical threshold in the number of crowd votes needed to achieve accuracy comparable to expert performance? How does the diversity within the crowd impact the quality of their collective decisions? Do learning effects occur, and do they influence the way the contest is conducted? To address the first question, we reviewed the literature on methods for aggregating crowd performance to identify the most effective strategies for harnessing the collective intelligence of diverse groups. Methods range from simple majority voting[15] to more sophisticated Bayesian models accounting for individual error patterns.[22] We employed a weighting model, with more weight given to participants with higher accuracy.[23] This approach was superior to unweighted averaging, as the crowd's performance was inferior to experts without weighting, but further exploration of other strategies might further enhance crowd performance.[11,19,24] Because our voting task involved multiclass classification, we elected to use an "average accuracy across different classes" approach as a weighting criterion to select better performers within the crowd. Therefore, even if a participant happens to be correct by chance for a particular pattern, their contribution to the weighted vote would remain limited unless their overall averaged accuracy across all patterns is consistently high. By applying a hard filter that excludes the bottom 20% of crowd raters based on calibration set accuracy, effectively removing the poorly performing raters, their average performance on the test dataset, as measured by non-majority weighted scoring, improved slightly but still remained below the crowd's weighted average performance. Although increasing the hard filtering threshold might improve the crowd's measured accuracy further, it ultimately defeats the purpose of crowdsourcing—since it ends up eliminating most participants and essentially selecting only those who perform at an expert level. Unlike previous studies that filtered out subgroups of users based on higher accuracy thresholds,[22] we chose not to exclude any participants. By including all users, we aimed to capture the full spectrum of crowd wisdom, ensuring that the aggregation process remained inclusive and reflective of the diverse perspectives within the crowd.

Crowdsourcing offers a potentially less costly and more time-efficient alternative to the expert annotation of EEG data; however, it is essential to analyze the optimal number of crowd participants required for voting. This analysis is crucial to ensure that the process is reliable. Warby et al. found that aggregating scores of 3 researchers or 10 non-experts provide comparable performance to that of an average individual expert in the binary task of identifying sleep spindles.[16] In our study, we observed a clear trend of accuracy improvement as the number of crowd participants grew. Crowd performance was comparable to experts after aggregating 20 or more participants, and surpasses that of experts with 30 participants. Accuracy continues to improve with the inclusion of more participants. Nevertheless, a threshold of 20 appears to be a practical benchmark for designing a crowdsourcing platform, with the potential for further accuracy gains with increased participation. The threshold can serve as a potential reference to construct a second-opinion system for

EEG pattern identification, where a given reader drops an EEG example on a crowdsourcing platform to engage at least 20 participants for a collective opinion that matches expert-level accuracy.

A diverse crowd is thought to be a strength in crowdsourcing due to problem-solving diversity.[25] To examine this, we used a leave-one-group-out approach to examine the effect of excluding crowd subgroups on the overall crowd's accuracy and found no difference in the crowd's performance. This suggested that although crowd subgroups were heterogeneous, they did not differ significantly in their contributions. In addition, our analysis comparing experts vs crowd subgroups revealed a heterogeneous impact of a neurology background on performance across different crowd subgroups. Although we currently lack additional characteristics to classify participants into distinct performance groups, identifying both high and low performance represents an intriguing direction for future research.

To explore learning effect, we used mixed-effects models to compare expert and crowd performance overall and by SRPP, incorporating the number of questions answered as a proxy for learning. Across the models, an increased number of questions answered was associated with a statistically significant yet very small improvement in accuracy—except in the by-pattern weighted model, where the relationship is not significant. This suggests that, although there is a slight improvement in performance with practice, the learning effect is minimal. We propose that a calibration process—where raters complete a limited set of calibration questions with feedback—should be sufficient to qualify raters.

Another key consideration was the benchmark used to compare crowd performance—namely, the performance of the experts in our study and the quality of the gold standard label. The overall inter-rater agreement among experts in the study was moderate, and slightly higher than the gold standard experts. The gold standard involved 30 experts compared to only 8 in our study, this discrepancy may have partly contributed to the lower interrater agreement observed. The performance of experts in our study against the gold standard was 68% in consensus and individually. Of interest, this was similar to the gold standard metric, where the gold standard expert average agreement with group consensus was 68%. The average agreement with group consensus is a comparable measure of individual expert accuracy relative to the consensus and reflects the inherent variability between different experts' assessments. Previous studies have described interrater reliability in EEG interpretation[26–29] and although not perfect, the expert performance in our study supports the validity of these findings. Given that even experts can exhibit interrater agreement issues, the requirement of 10 or more responses is intended to ensure that the gold standard dataset is robust.

Our study was facilitated by a crowdsourcing gamified platform, enabling us to launch massive EEG annotation tasks to a large diverse group consistently and uniformly.[30] In addition, a unique advantage of a gamified platform used in this study is the potential for performance-based incentivization.[15] There are other crowdsourcing platforms for medical data annotations.[31–33] Through Amazon's Mechanical Turk crowdsourcing platform, the ImageNet database was created, containing 3.2 million annotated images. This database serves as a training resource and benchmark for many computer vision machine learning models and revolutionizes the field.[34] Collecting a similar magnitude of medical data annotations would not be feasible if the task were limited to expert annotation alone. EEG signal annotation is a nuanced process, complicated by factors such as varying viewing options (e.g., montage selection and signal gain adjustments) and the use of evolving terminologies. These challenges must be considered carefully when designing a crowdsourcing platform to ensure that the crowd can perform effectively. The goal is to strike a balance by providing sufficient guidance to enhance crowd performance without overwhelming them with overly technical or complex tasks. A crowdsourcing infrastructure for SRPP annotation has been proposed[35] but to date, there is little empirical evidence demonstrating the feasibility of this approach. Our study contributes to the literature by providing concrete data that support the feasibility of using a crowdsourcing platform for SRPP annotation.

Our study has several important limitations. First, the provision of instant feedback during the contest likely improved participant performance over time through learning effects.[36,37] Although we split the data into calibration and test sets, participants in the test set continued to be exposed to the correct answers, which may have further reinforced these learning effects. Although our mixed-effects model attempted to control for this factor, we cannot fully assess crowd performance without the influence of that feedback. Second, we did not investigate how question difficulty influenced crowd performance; excluding challenging questions (those with divided expert opinions) could help but might undermine crowdsourcing's purpose of handling varying difficulty levels. Third, our study design limited participants to six forced choices among common SRPP patterns, not the full spectrum of possibilities, and additional pattern classes could affect crowd performance in unknown ways. Future studies should investigate crowd performance on additional or more granular EEG pattern categories to better understand and potentially improve classification in the 'Other' category. For example, accurate identification of non-SRPP patterns and normal EEG recordings is essential. Without reliable identification of these patterns, SRPP subtype classification alone offers limited clinical benefits. Fourth, the data present a limited

EEG sample, which is not reflective of actual clinical settings, especially in the critical care unit, where extended monitoring offers the opportunity to better elucidate patterns as they develop over time. Fifth, individual crowd reviewers received different calibration and test splits, which diverge from how a fully implemented system would function without gold-standard annotations. Demonstrating the feasibility of crowdsourcing under these real-world constraints, such as ensuring consistent data partitioning and validating performance without gold standard labels for the entire dataset, should be addressed in future work to translate this approach beyond the research setting. Sixth, one might question whether these "gold standard" labels truly represent a gold standard, since it is based solely on EEG consensus, despite poor agreement even among gold standard experts. Demanding very high consensus among experts could, in fact, restrict the range of EEG findings included in a contest like this. Although crowd-based and expert-based approaches might yield similar consensus results, simply increasing the number of raters will not eliminate inherent bias in how the gold standard is defined. In a related publication, we examined interrater reliability among 30 experts and found that expert disagreement is driven primarily by differences in decision thresholds rather than a complete lack of shared understanding.[9] That study revealed substantial variability in expert ratings—which contributes to diagnostic inconsistency—but also demonstrated strong underlying agreement when receiver-operating characteristic and precision-recall curves were evaluated. This indicates that the consensus (our gold standard) reflects a common interpretation of EEG patterns, even if decision thresholds differ. Although this approach does not fully remove the inherent bias in the consensus, it provides a rationale for using the gold standard as a benchmark. Ideally, an external clinical outcome metric—such as a risk gradient for future seizures—would be necessary to validate the clinical significance of the EEG findings. However, because such an external standard has not yet been established in the literature, it is beyond the scope of our work to incorporate this level of validation. Finally, our study's generalizability may be limited by the use of EEG data from a single institution.

5 | CONCLUSION

On average, experts outperformed the crowd in overall SRPP detection. However, using weighted majority votes, the crowd matched or exceeded experts accuracy in identifying most SRPPs, except for LPDs and "Other." Our study is a proof-of-concept investigation that demonstrated crowdsourcing as a feasible and scalable approach to obtaining expert-level annotations of certain SRPPs. We highlighted the promise of crowd reviewers while acknowledging that challenges, such as varying calibration/test splits across crowd participants in the study and the absence of gold standard labels in the real-world settings, remain to be addressed.

AUTHOR CONTRIBUTIONS

All authors contributed to the study conception and design. W.Y.K. conducted the data analysis and drafted the manuscript. J.J. and M.B.W. contributed to project supervision, data interpretation, and manuscript revision, and provided critical feedback. All authors read and approved the manuscript.

ACKNOWLEGDEMENT

We grafefully ackowledge the contributions of all participants in the SRPP scoring contest.

CONFLICT OF INTEREST STATEMENT

This work was supported by grants from the National Institutes of Health (NIH; RF1AG064312, RF1NS120947, R01AG073410, R01HL161253, R01NS126282, R01AG073598, R01NS131347, R01NS130119) and the National Science Foundation (NSF; 2014431). Dr. Westover is a co-founder, scientific advisor, consultant to, and has personal equity interest in Beacon Biosignals. He also receives royalties for authoring *Pocket Neurology* from Wolters Kluwer and *Atlas of Intensive Care Quantitative EEG* by Demos Medical. Erik Duhaime and Srishti Kapur are employees of and have personal equity interest in Centaur Labs. There are no conflicts of interest for the other authors. We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Jong Woo Lee https://orcid.org/0000-0001-5283-7476
M. Brandon Westover https://orcid.org/0000-0003-4803-312X

REFERENCES

1. Hirsch LJ, Fong MWK, Leitinger M, LaRoche SM, Beniczky S, Abend NS, et al. American clinical neurophysiology Society's standardized critical care EEG terminology: 2021 version. J Clin Neurophysiol. 2021;38(1):1–29. https://doi.org/10.1097/WNP.0000000000000806

2. Gallotto S, Seeck M. EEG biomarker candidates for the identification of epilepsy. Clin Neurophysiol Pract. 2022;8:32–41. https://doi.org/10.1016/j.cnp.2022.11.004
3. Machine learning for detection of interictal epileptiform discharges. Clin Neurophysiol. 2021;132(7):1433–43. https://doi.org/10.1016/j.clinph.2021.02.403
4. Zhang L, Wang X, Jiang J, Xiao N, Guo J, Zhuang K, et al. Automatic interictal epileptiform discharge (IED) detection based on convolutional neural network (CNN). Front Mol Biosci. 2023;10:1146606. https://doi.org/10.3389/fmolb.2023.1146606
5. Heers M, Böttcher S, Kalina A, Katletz S, Altenmüller DM, Baroumand AG, et al. Detection of interictal epileptiform discharges in an extended scalp EEG array and high-density EEG—A prospective multicenter study. Epilepsia. 2022;63(7):1619–29. https://doi.org/10.1111/epi.17246
6. Jing J, Sun H, Kim JA, Herlopian A, Karakis I, Ng M, et al. Development of expert-level automated detection of epileptiform discharges during electroencephalogram interpretation. JAMA Neurol. 2020;77(1):103–8. https://doi.org/10.1001/jamaneurol.2019.3485
7. Thomas J, Comoretto L, Jin J, Dauwels J, Cash SS, Westover MB. EEG CLassification via convolutional neural network-based interictal epileptiform event detection. in 2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC), 2018, pp. 3148–3151. https://doi.org/10.1109/EMBC.2018.8512930
8. Wong S, Simmons A, Rivera-Villicana J, Barnett S, Sivathamboo S, Perucca P, et al. EEG datasets for seizure detection and prediction—a review. Epilepsia Open. 2023;8(2):252–67. https://doi.org/10.1002/epi4.12704
9. Jing J, Ge W, Struck AF, Fernandes MB, Hong S, An S, et al. Interrater reliability of expert electroencephalographers identifying seizures and rhythmic and periodic patterns in EEGs. Neurology. 2023;100(17):e1737–e1749. https://doi.org/10.1212/WNL.0000000000201670
10. Boland PJ. Majority systems and the Condorcet jury theorem. J R Stat Soc Ser D-Stat. 1989;38(3):181–9. https://doi.org/10.2307/2348873
11. Duhaime EP, Jin M, Moulton T, Weber J, Kurtansky NR, Halpern A, et al. Nonexpert crowds outperform expert individuals in diagnostic accuracy on a skin lesion diagnosis task. in 2023 IEEE 20th international symposium on biomedical imaging (ISBI), Apr. 2023, pp. 1–5. https://doi.org/10.1109/ISBI53787.2023.10230646
12. Kentley J, Weber J, Liopyris K, Braun RP, Marghoob AA, Quigley EA, et al. Agreement between experts and an untrained crowd for identifying Dermoscopic features using a gamified app: reader feasibility study. JMIR Med Inform. 2023;11(1):e38412. https://doi.org/10.2196/38412
13. McNeil AJ, Parks K, Liu X, Jiang B, Coco J, McCool K, et al. Crowdsourcing skin demarcations of chronic graft-versus-host disease in patient photographs: training versus performance study. JMIR Dermatol. 2023;6:e48589. https://doi.org/10.2196/48589
14. Duggan NM, Jin M, Duran Mendicuti MA, Hallisey S, Bernier D, Selame LA, et al. Gamified crowdsourcing as a novel approach to lung ultrasound data set labeling: prospective analysis. J Med Internet Res. 2024;26(1):e51397. https://doi.org/10.2196/51397
15. Skinner G, Chen T, Jentis G, Liu Y, Mc Culloh C, Harzman A, et al. Real-time near infrared artificial intelligence using scalable non-expert crowdsourcing in colorectal surgery. NPJ Digit Med. 2024;7(1):99. https://doi.org/10.1038/s41746-024-01095-8
16. Lacourse K, Yetton B, Mednick S, Warby SC. Massive online data annotation, crowdsourcing to generate high quality sleep spindle annotations from EEG data. Sci Data. 2020;7(1):190. https://doi.org/10.1038/s41597-020-0533-4
17. t-Test, Chi-Square, ANOVA, Regression, Correlation. [cited 2025 Apr 04]. Available from: https://datatab.net/tutorial/fleiss-kappa
18. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC Med Res Methodol. 2013;13(1):61. https://doi.org/10.1186/1471-2288-13-61
19. Collins RN, Mandel DR, Budescu DV. "Performance-weighted aggregation: ferreting out wisdom within the crowd," In: Seifert M, editor. International Series in Operations Research & Management Science, in Judgment in Predictive Analytics, Cham: Springer International Publishing; 2023. p. 185–214. https://doi.org/10.1007/978-3-031-30085-1_7
20. Halford JJ, Arain A, Kalamangalam GP, LaRoche SM, Leonardo B, Basha M, et al. Characteristics of EEG interpreters associated with higher interrater agreement. J Clin Neurophysiol. 2017;34(2):168–73. https://doi.org/10.1097/WNP.0000000000000344
21. Bagheri E, Dauwels J, Dean BC, Waters CG, Westover MB, Halford JJ. Interictal epileptiform discharge characteristics underlying expert interrater agreement. Clin Neurophysiol. 2017;128(10):1994–2005. https://doi.org/10.1016/j.clinph.2017.06.252
22. Hasan E, Duhaime E, Trueblood JS. Boosting wisdom of the crowd for medical image annotation using training performance and task features. Cogn Res Princ Implic. 2024;9(1):31. https://doi.org/10.1186/s41235-024-00558-6
23. Bai S, Wang D, Muller T, Cheng P, Chen J. Stability of weighted majority voting under estimated weights. 2024 arXiv: arXiv:2207.06118. https://doi.org/10.48550/arXiv.2207.06118
24. Budescu DV, Chen E. Identifying expertise to extract the wisdom of crowds. Manag Sci. 2015;61(2):267–80.
25. Hong L, Page SE. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. Proc Natl Acad Sci. 2004;101(46):16385–89. https://doi.org/10.1073/pnas.0403723101
26. Gerber PA, Chapman KE, Chung SS, Drees C, Maganti RK, Ng YT, et al. Interobserver agreement in the interpretation of EEG patterns in critically ill adults. J Clin Neurophysiol. 2008;25(5):241–9. https://doi.org/10.1097/WNP.0b013e318182ed67
27. Ronner HE, Ponten SC, Stam CJ, Uitdehaag BMJ. Inter-observer variability of the EEG diagnosis of seizures in comatose patients. Seizure. 2009;18(4):257–63. https://doi.org/10.1016/j.seizure.2008.10.010
28. Mani R, Arif H, Hirsch LJ, Gerard EE, LaRoche SM. Interrater reliability of ICU EEG research terminology. J Clin Neurophysiol. 2012;29(3):203–12. https://doi.org/10.1097/WNP.0b013e3182570f83
29. Gaspard N, Hirsch LJ, LaRoche SM, Hahn CD, Westover MB, Critical Care EEG Monitoring Research Consortium. Interrater

agreement for critical care EEG terminology. Epilepsia. 2014;55(9):1366–73. https://doi.org/10.1111/epi.12653
30. Centaur Labs—Medical Data Labeling Solutions. [cited 2024 Aug 19]. Available from: https://www.centaurlabs.com/
31. Amazon Mechanical Turk. [cited 2024 Aug 19]. Available from: https://www.mturk.com/
32. Improves AI with Data—Powering AI Innovation, Appen. [cited 2024 Aug 19]. Available from: https://www.appen.com/
33. Generative AI and Computer Vision Data Annotation, Sama. [cited 2024 Aug 19]. Available from: https://www.sama.com/
34. ImageNet. [cited 2024 Aug 19]. Available from: https://www.image-net.org/index.php
35. Freitas J, Nguyen A, Bosl W. Confidence in the qualified crowd: a platform for sourcing EEG annotations. 2020 IEEE signal processing in medicine and biology symposium (SPMB). Philadelphia, PA, USA: IEEE; 2020. p. 1–6. https://doi.org/10.1109/SPMB50085.2020.9353617
36. Barfuss JD, Nascimento FA, Duhaime E, Kapur S, Karakis I, Ng M, et al. On-demand EEG education through competition—a novel, app-based approach to learning to identify interictal epileptiform discharges. Clin Neurophysiol Pract. 2023;8:177–86. https://doi.org/10.1016/j.cnp.2023.08.003
37. Nascimento FA, Jing J, Traner C, Kong WY, Olandoski M, Kapur S, et al. A randomized controlled educational pilot trial of interictal epileptiform discharge identification for neurology residents. Epileptic Disord. 2024;26(4):444–59. https://doi.org/10.1002/epd2.20229

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Kong W-Y, Nascimento FA, Struck A, Duhaime E, Kapur S, Amorim E, et al. Evaluating crowdsourcing for ICU EEG annotation: A comparison with expert performance. Epilepsia. 2025;66:4366–4380. https://doi.org/10.1111/epi.18547