Confidence in the Qualified Crowd: A Platform for Sourcing EEG Annotations

J. Freitas[1], A. Nguyen[2] and W. Bosl[2,3,4]

1. Department of Computer Science, Loyola Marymount University, Los Angeles, California, USA
2. University of San Francisco, San Francisco, California, USA
3. Boston Children's Hospital, Boston, Massachusetts, USA
4. Harvard Medical School, Boston, Massachusetts, USA

jordan.freitas@lmu.edu, andrew.nguyen@usfca.edu, william.bosl@childrens.harvard.edu

Abstract— The use of new signal processing and machine learning algorithms to find digital biomarkers in electroencephalogram (EEG) signals requires annotations for EEG signals at scale and in an on-going manner. In this paper, we provide an overview of the computational infrastructure to support this with an emphasis on how we efficiently process EEG signal data from multiple sources and manage all of the corresponding information. If each annotator is better than 50% accurate, increasing the number of annotators for annotation tasks will result in higher and higher reliability of the annotation. Our platform enables new approaches to quantifying the accuracy of annotations based on the annotators' accuracy, expertise, and relevant experience.

I. Introduction and Motivation

The motivation for crowdsourcing and annotation research in our case comes from a traditional use: that of reviewing continuous EEG readings in an ICU setting (cEEG-ICU). Continuous EEG monitoring is a relatively new and growing practice used mostly in larger medical centers. The challenge is that busy neurologists and even neurodiagnostic technologists do not have time to adequately review the data as often as medically necessary. A neurologist typically reviews the EEGs every 12 hours, and a technologist may do a brief review every few hours. This is not often enough to catch emergency seizures that occur without any clinical signs ("electrographic" seizures).

As such there is a rather urgent need for algorithms to screen the data continuously and, ideally, send an alert when something appears to require more serious attention. In this case, the algorithms need to continuously monitor the EEG data streams to detect the signal features that a neurologist would be searching for: spikes, slowing, rhythmic activity. These are known indicators of seizure activity.

In order to begin to develop and train algorithms for this task, annotated data must be available to researchers. The same problem arises again. Experts do not have time to annotate data. Furthermore, if this data will be used for research, the annotations must be reliable. How reliable? How reliable is a resident doing the labeling? How reliable is a trained neurophysiologist with 10 years experience? None of these questions have been adequately addressed. This is the primary driving force behind the need for annotation.

In addition, the emergence of a new generation of EEG devices that are easier to use *and* lower in cost will absolutely bring EEG into community and primary clinics in low-income regions. In many places, epilepsy is not treated, not because low cost anti-epileptic drugs (AED) are not available (they are), but because a qualified neurologist is not available to review and diagnose epilepsy [1, 2]. AED's can have powerful side effects, thus should not be given unless warranted. The ability to use algorithms to screen for epilepsy in the low-income country (LIC) settings would be very beneficial. Having annotated EEG data to train screening algorithms will be necessary. The goal in this case is to fill in for the lack of highly trained professionals for epilepsy screening for epidemiological studies, policy planning, and eventually for therapy or medications.

All of these clinical research scenarios require "experts" to review and manually annotate a significant amount of EEG data. Two significant challenges arise. First, given the number of annotations required for a large research dataset, such as the Temple University Hospital EEG Corpus [3], hundreds of experts would be needed to annotate all of the data, each reviewing hundreds or thousands of hours of EEGs. This becomes extremely cost prohibitive and begs the question, "how does one define expert?" This question requires a method to determine the accuracy of annotators, which will vary depending on the annotation tasks, as well as overall accuracy when many raters of varying accuracy are involved. Studies have consistently shown that the interrater reliability of experts is less than one might think is necessary for a gold standard [4]. Typically, two or three experts may confer about a diagnostic interpretation in order to improve reliability. This is the correct approach, however it can be quantified for annotations. In Condorcet's original 1785 paper, he argued that a jury of less accurate voters might arrive at a more reliable decision than a single expert [5]. It is possible to determine the accuracy of a majority vote of annotators if the accuracy of each annotator is known.

The phrase "wisdom of the crowd" has become common

in recent culture, but it is not a new idea. The Marquis de Condorcet originally derived the mathematical theory in the late 1700s in an effort to determine precisely how likely a consensus vote from a jury of N would arrive at the true decision. The theorem has become known as Condorcet's Jury theorem. The basic idea behind this theorem is that the majority opinion is always more likely to be correct than any individual in the group of voters, assuming that each individual is more often correct than incorrect [5]. This work forms the foundation for modern crowd-sourcing methods for decision making and annotating data. Modern ensemble-based machine learning methods are also based on this theorem. Ensemble methods are essentially a majority vote of many individual classifiers and will mathematically perform better than any single classifier as long as all are greater than 50% accurate [6].

The following sections offer an explanation of what we consider essential supporting infrastructure for the sourcing of EEG annotations from a qualified crowd. We first propose the design requirements derived from the background described in the next section. The methods section then discusses the data we use to inform and validate our platform architecture, along with the design and implementation of our annotation platform (illustrated in Figure 1.

II. Related Work

The term crowdsourcing, with respect to labeling data, can imply scenarios ranging from the solicitation of online masses to complete simple labeling or validation tasks at enormous scales, e.g., Amazon's Mechanical Turk, to the recruitment of tens of highly-trained experts within a professional network of known colleagues to interpret complex datasets, such as medical images or signal data. We use the term *qualified crowd* to capture two meanings. First, the annotators have been trained to annotate. They are qualified. Second, the annotation itself can be qualified based on the experience and expertise of the annotator. In other words, the suggested annotation can and should be weighted if the annotator has experience and expertise relevant to the nature of that annotation task.

We share an interest with Warby et al. [7] in evaluating annotation methods. Their work involves a comparison of expert, non-expert, and automated annotations collected for the study. Our current focus however is on the computational infrastructure needed for collecting various kinds of EEG annotations in an on-going manner with the intention to scale over time (see Section III), and we do not differentiate between annotators a priori.

Interrater or interreader agreement is another important concept for confidence measures of annotations. It is particularly important when no true "gold standard" annotation exists. Interrater agreement can serve as a proxy for individual rater accuracy. Key studies of this (for example, [8–11]) have studied neurology residents, neurophysiology fellows, or board-certified neurophysiologists. We see an opportunity to engage EEG expertise in a broader community of neurodiagnostic professionals. Annotation accuracy will only increase as more participation yields more suggested annotations.

III. Design Requirements

Collecting and analyzing EEG signal annotations involves segments of EEG recordings being displayed to EEG technologists, residents in neurology, fellows in neurophysiology–or those in training–who submit one or more annotations for each segment of the data they are shown. These contributions of annotations from several professionals over time build up a repository of EEG data and corresponding annotations which may or may not be correct. Additional information about an EEG professional such as specialty, experience, and previous annotation accuracy within the platform can be used to weigh the annotations in determining which is correct.

In order to study annotation accuracy, the initial EEG recordings shown to users have corresponding annotations (see Section IV-A). Cleaning the data for use in the labeling platform involves formatting and filtering the data by types of labels or other corresponding file metadata. The volume and complexity of the raw EEG recordings make complete data processing a heavy task for personal computers. In order to study the categories of labels in a comparable way, it is important to reproduce the data cleaning steps exactly even though there may be several months between analysis of different types of labels. The data set is also publicly available and other researchers may also be interested in studying labels, it is important to document the data provenance in a readable way. Although data cleaning and filtering is an important step, the primary goal of the platform is to collect annotations from users and identify correctness. Clinical researchers can then use the annotations with measurable confidence.

The labeling platform has potential to ultimately enable new collaboration and data sharing practices in which researchers may contribute data or data processing components and workflows, in exchange for the annotations, additional data, or new data processing components and workflows. Although leveraging the platform for data sharing is a future consideration, we are mindful of user roles, access control, and data privacy concerns.

With these considerations in mind, we offer the following design requirements for the extract, transform, load (ETL) of EEG files and metadata management compo-

nents of a crowdsourcing approach to EEG annotations. We elaborate on these requirements and how we satisfy them in Section IV. There are considerations for the web interface (part (e) of Figure 1), captured in additional related work [12, 13]. Besides collecting annotations, a successful implementation of the web interface will also provide value to annotators.

Support time-series and corresponding data The time-series EEG recordings are well-suited for a column-oriented database, while database requirements are significantly different for the corresponding information related to the patient, context, and reader.

Scalability We are aware of several hundreds of terabytes of EEG files in need of annotations belonging to a single large epilepsy center, with similar numbers at other large medical centers. Partnerships are being discussed with national professional organizations of neurodiagnostic professionals who have expressed an openness towards inviting members to provide annotations as a part of training and practice exercises. This scale of participation is necessary to build confidence in the quantification of trends in annotation accuracy. The amount of data and desired extent of participation necessitates parallelizing the ETL and annotation tasks.

Standardization and flexibility In order to achieve impactful scale, it is also necessary to be able to incorporate data from different sources. Our ETL pipelines in Cloud Dataflow involve customized operations for uploading different types of raw data files such that any EEG data and corresponding clinical information can be loaded into our schemas.

Provenance The need for data and analysis provenance plays out on two main fronts. First, as annotations standards are updated over time, ideally the previously annotated data can still be included in clinical research. Second, our mechanisms for data filtering must be consistent so results are comparable. Provenance enables researchers to determine whether results were derived from consistent processes or not.

Consistency Input data files of the same type must be processed with the same steps. Inconsistencies in data cleaning and filtering jeopardize reproducibility of proceeding analysis results.

Sharing and collaboration So far, we simply distinguish those providing and those receiving EEG annotations. Our platform could be contained to tightly maintain data ownership. However ultimately there are several stakeholders across several institutions including technologists, physicians, clinical researchers, informatics researchers, and patients or study participants (see Section VI about future work).

Privacy Corresponding clinical information such as age and sex are highly relevant to the reading and proposal of EEG annotations, and can be sufficiently de-identified when presented to annotators. We host raw data files in Google Cloud Storage, which is encrypted at rest and HIPAA compliant. The privacy concerns of annotators are an equally important consideration. We need to keep track of detailed performance indicators related to the annotators in order to qualify the annotations submitted through the platform–i.e., to weigh some more heavily than others depending on the nature of the tasks and the expertise of the annotators. Using the data related to annotators requires informed consent, use agreements, and fine-grained control over the data management.

Efficiency The expertise needed to annotate EEG data is in high demand for clinical care. As we describe in Section I, collecting annotations for existing data is also a bottleneck for neurodiagnostic research. In any context where resources are constrained, efficiency is both practically and morally imperative.

Preliminary feedback from neurodiagnostic professionals raised the question of whether our platform would facilitate automating the work of technologists all together. In the shorter term it may be possible to automate the most simple annotations, in which case technologists would be able to focus on tasks related to more complex, subtle signal patterns.

IV. Methods

IV-A. TUH EEG Corpus

The initial use case is focused around the Temple University Hospital EEG Corpus, and specifically the Events Corpus subset (TUEV) [3]. This subset contains six annotation types: spike and slow wave, generalized periodic epileptiform discharge, periodic lateralized epileptiform discharge, eye movement, artifact, and background. This subset provides a set of data that have been manually annotated by three neurologists from Temple University School of Medicine.

IV-B. Signal Data

While the initial focuses on the annotation of EEG data that largely comes in European Data Format (EDF), data often come in other formats (e.g., Matlab). Additionally, this project aims to be a general biomedical signal annotation platform which expands the number of formats even more. Most of these file formats are designed to facilitate the storage and transmission of waveform data so they are designed to be compact. In order to read a segment of the waveform, the entire file must be read into memory and then a seek performed on the appropriate subset of data.

Given the need to support multiple file formats and to read arbitrary subsets of the waveform, we opted to

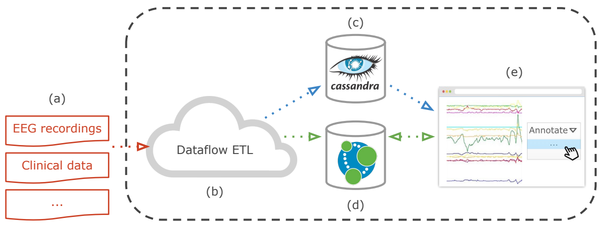

Figure 1. The annotation platform architecture. The Dataflow ETL (b) can be configured to accept the input data files (a) in any format, and load them into databases separately optimized for signal data in NoSQL (c) and everything else in a graph database (d)–described in subsections IV-B and IV-C respectively. Segments of the EEG data along with only the necessary corresponding clinical data (e.g., age) are loaded into our web application (e) for annotators to review. The annotations are then added to the graph database (d).

store the data in a column store (Apache Cassandra, (b) in Figure 1). The data were stored as key/value pairs where the value is the measurement of the signal. The row identifier is the timestamp of the value. Each column represents the signal from a particular sensor on a particular patient or research subject.

This approach allows us to store waveform / signal data efficiently while also enabling fast reads of subsets of signal data. There are two primary use cases from the perspective of reading signal data. First, the web application needs to read small segments of waveform data in order to produce the annotation tasks. Second, we need to bulk load signal data into the Google Dataflow environment for signal processing and feature extraction. Both use cases (high volume reading of small segments, and low volume reading of entire signals) are both well supported by Cassandra.

The annotation data (described in the next section) and the signal data are linked via UUIDs and timestamps (see Figure 2).

Patient ID	Signal ID	Patient ID	Signal ID	Patient ID	Signal ID
Timestamp, Value		Timestamp, Value		Timestamp, Value	
Timestamp, Value		Timestamp, Value		Timestamp, Value	
Timestamp, Value		Timestamp, Value		Timestamp, Value	
Timestamp, Value		Timestamp, Value		Timestamp, Value	
Timestamp, Value		Timestamp, Value		Timestamp, Value	
Timestamp, Value		Timestamp, Value		Timestamp, Value	
Timestamp, Value		Timestamp, Value		Timestamp, Value	
Timestamp, Value		Timestamp, Value		Timestamp, Value	

Figure 2. Schema for EEG signal data

IV-C. Annotation Data

The core of the system is the storage and management of the annotation data, sometimes referred to as metadata. This includes all data related to annotations including annotator demographics, signal metadata (e.g., sensor location, sampling rate), and the actual annotations.

While most of the data could be stored in any type of database–relational or NoSQL, we use a property graph database to manage the annotations. Specifically, the Neo4j graph database provides ACID semantics and enables the ability to capture semantic relationships within a fully transactional database that supports the annotation web application. Any database with ACID semantics would provide the necessary functionality (e.g, ensuring that all data related to a particular annotation are atomically recorded to the database) to ensure accurate collection of annotations and metadata. However, non-graph databases would increase the complexity of the data schema when trying to manage the associated terminologies.

EEG annotations are complex and the underlying terminologies are constantly evolving [14]. Consequently, the system was designed to accommodate current terminologies while also accounting for future changes. Property graph databases provide two key features that help address this challenge: relationships and properties within the relationship.

At its core, a graph database consists of nodes and relationships that connect the nodes. Within Neo4j, both nodes and relationships have labels. A node label is akin to the name of a table. For example, there are nodes corresponding to individual annotators, each with the Annotator label. Each node with the Annotator label is comparable to a row within an Annotator table of a relational DB.

Within the graph database, properties are essentially key/value pairs that can be attached to either nodes or relationships. Properties of nodes are essentially columns of a table and have similar semantics. Properties of relationships are essentially the columns within a join table. For example, if there are two tables, A and B, that are connected by a join table, the properties of nodes would be columns in tables A and B and the properties of the relationship are the columns of the join table between A and B.

However, as a NoSQL database, there are fewer restrictions on the underlying schema. A single node can have multiple labels, and it may also have properties that are not common to any other nodes. Properties can be added or removed on a per-node basis without needing to change a schema or anything else.

Figure 3 contains the current schema used to manage the annotation data. The key relationship that mitigates issues of terminology evolution is the “IS A” relationship between AnnotationType nodes. For example, what was referred to as a “Periodic Lateralized Epileptiform Discharge (PLED)” was renamed to be a “Lateralized

Periodic Discharge (LPD)" as part of the 2012 ACNS Standardized Critical Care EEG Terminology. Within the database, two nodes are created, one for PLED and another for LPD and they are connected by an "IS A" relationship. When querying for annotations matching a particular term, the Cypher query follows the "IS A" relationship to collect all equivalent terms.

The additional benefit of a property graph database is the ability to add properties to the relationships (and not just the nodes). These properties can include rules or criteria that allow for additional filtering. For example, only those terms marked as equivalent by a particular mapping scheme may be used.

IV-D. Dataflow

Google Cloud Dataflow (Dataflow) is the bridge between raw data files and our schemas described above. Dataflow offers a pipeline paradigm for data processing along with managed parallelism, supporting several of our design requirements.

Raw data files are staged in Google Cloud Storage, and then the data of interest is loaded into Neo4j and Cassandra using Dataflow to execute the data processing pipelines. The pipelines are directed graphs of operators, in which each operator is one step of the data processing. As long as the output data types of each operator match the input data type of the next operator in the pipeline graph, the pipeline will be executed. As such we are able to write custom operators for loading new sources of data with varied formatting conventions. The remaining ETL operators are consistent across pipelines.

While data sharing remains tricky, pipelines or individual operators can be shared among researchers more easily, especially if a supporting platform is easy to use. As a future interest, we plan to explore how our work can be built upon to help facilitate collaboration by leveraging Dataflow. While our infrastructure is set up in Google Cloud, Dataflow is also available as Apache Beam which can run on any cloud or local environment.

Dataflow pipelines are written in Python or Java. Rather than being executed directly, the code describes a pipeline of operators which Dataflow compiles into an optimized program for parallel processing. Pipeline code authors then enjoy the runtime efficiency without worrying about how to achieve that parallelization.

V. Discussion

V-A. Cypher Queries

The use of Neo4j as a graph database provides the ability to easily manage evolving terminologies through the use of simple Cypher queries. Since Cypher queries are very similar to SQL queries, they are relatively easy to interpret by those without training in computer science or programming.

For example, one example discussed previously is the *Periodic Lateralized Epileptiform Discharge* (PLED). This is an old term and was replaced in 2012 with *Lateral Periodic Discharge* (LPD). Our annotation and signal databases might have data encoded with the former while the researcher may be looking for the latter. In this situation, the following Cypher query would retrieve all equivalent terms:

```
MATCH (:SignalType text: "Lateralized
Periodic Discharge")-[r:IS_A
*0..]-(st:SignalType) RETURN st
```

The above query searches the database for LPD (the annotation of interest) and returns all equivalent SignalTypes within the database. Of particular note are use of `*0..` and `-[r:IS_A]-`.

The syntax `*0..` tells Cypher to perform a recursive query, exhaustively traversing the specified relationship. Cypher allows for the query to include/exclude the initial term and to limit the depth of recursion. However, in our use cases, we typically do not want to limit either.

The synatx `-[r:IS_A]-` tells Cypher to follow the `IS_A` relationship while ignoring direction. This essentially treats the relationships as bidirectional.

The above example traversed the graph and extracted all terms regardless of the underlying terminology versions. However, if we wanted to only identify equivalent terms that were changed as part of the 2012 update, we could modify the Cypher query to include an additional restraint (in bold):

```
MATCH (:SignalType text: "Lateralized
Periodic Discharge")-[r:IS_A *0..
{version: "2012"}]-(st:SignalType)
RETURN st
```

The use of a graph database and the Cypher query language makes it simple to manage multiple versions of terminologies.

VI. Future Work

Much of our future work revolves around achieving a larger scale of participation in order to better investigate open questions related to annotation accuracy. We also imagine this platform being mutually beneficial for the technologists-in-training or those want to practice as well as the clinical researchers who will use the annotations. As participation increases, features can be developed to offer feedback to annotators related to how their annotations compare with the qualified crowd.

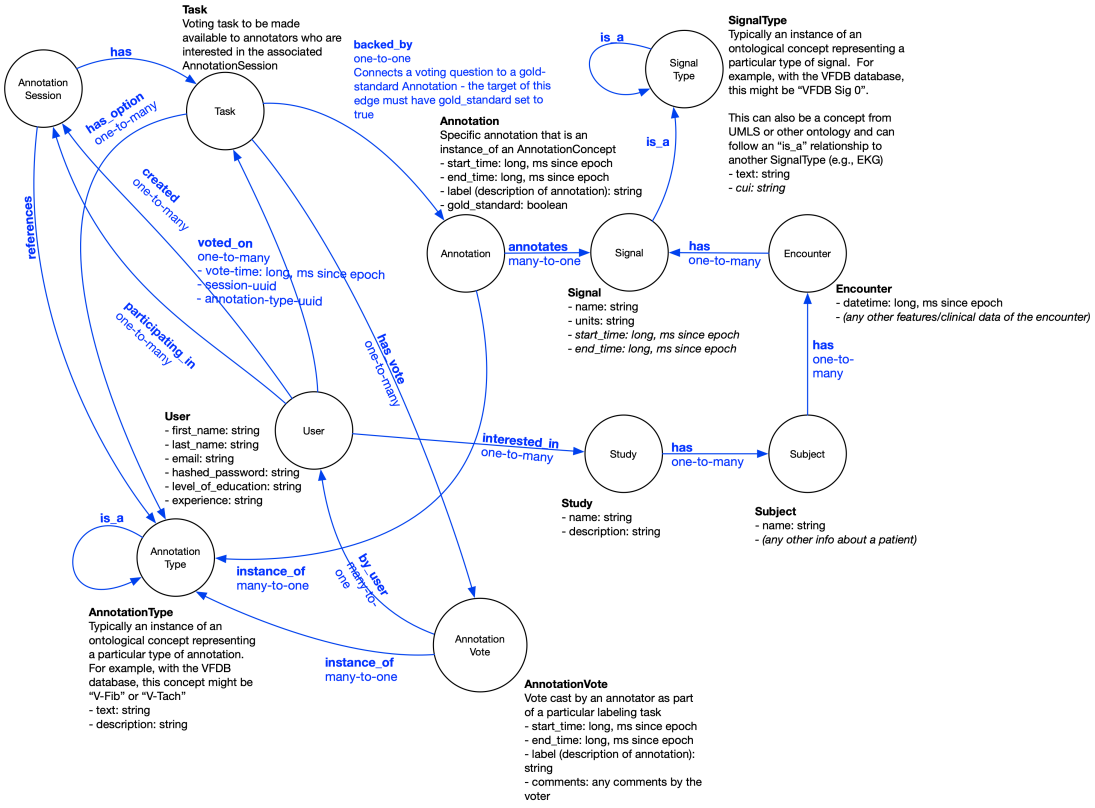

Figure 3. Schema for annotations and corresponding data

ACKNOWLEDGEMENTS

This work was partially funded by Google Cloud Platform research credits.

REFERENCES

[1] C. R. Newton and H. H. Garcia, "Epilepsy in poor regions of the world," *The Lancet*, vol. 380, no. 9848, pp. 1193–1201, 2012.

[2] J. M. Wilmshurst, G. L. Birbeck, and C. R. Newton, "Epilepsy is ubiquitous, but more devastating in the poorer regions of the world… or is it?" *Epilepsia*, vol. 55, no. 9, pp. 1322–1325, 2014.

[3] I. Obeid and J. Picone, "The temple university hospital EEG data corpus," *Frontiers in neuroscience*, vol. 10, p. 196, 2016.

[4] K. Gottlieb and F. Hussain, "Voting for Image Scoring and Assessment (VISA)-theory and application of a 2+ 1 reader algorithm to improve accuracy of imaging endpoints in clinical trials," *BMC medical imaging*, vol. 15, no. 1, p. 6, 2015.

[5] P. J. Boland, "Majority systems and the Condorcet jury theorem," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 38, no. 3, pp. 181–189, 1989.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[7] S. C. Warby, S. L. Wendt, P. Welinder, E. G. Munk, O. Carrillo, H. B. Sorensen, P. Jennum, P. E. Peppard, P. Perona, and E. Mignot, "Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods," *Nature methods*, vol. 11, no. 4, p. 385, 2014.

[8] J. Jing, A. Herlopian, I. Karakis, M. Ng, J. J. Halford, A. Lam, D. Maus, F. Chan, M. Dolatshahi, C. F. Muniz *et al.*, "Interrater reliability of experts in identifying interictal epileptiform discharges in electroencephalograms," *JAMA neurology*, vol. 77, no. 1, pp. 49–57, 2020.

[9] M. L. Scheuer, S. B. Wilson, A. Antony, G. Ghearing, A. Urban, and A. Bagić, "Seizure detection: Interreader agreement and detection algorithm assessments using a large dataset." *Journal of Clinical Neurophysiology: Official Publication of the American Electroencephalographic Society*, 2020.

[10]

[11] J. J. Halford, A. Arain, G. P. Kalamangalam, S. M. LaRoche, B. Leonardo, M. Basha, N. J. Azar, E. Kutluay, G. U. Martz, W. J. Bethany *et al.*, "Characteristics of EEG interpreters associated with higher interrater agreement," *Journal of clinical neurophysiology: official publication of the American Electroencephalographic Society*, vol. 34, no. 2, p. 168, 2017.

[12] J. J. Halford, W. B. Pressly, S. R. Benbadis, W. O. Tatum IV, R. P. Turner, A. Arain, P. B. Pritchard, J. C. Edwards, and B. C. Dean, "Web-based collection of expert opinion on routine scalp EEG: software development and interrater reliability," *Journal of Clinical Neurophysiology*, vol. 28, no. 2, pp. 178–184, 2011.

[13] A. Nguyen, W. Bosl, S. Herman, and T. Loddenkemper, "Crowdsourcing for research EEG annotation and accuracy estimation," *AMIA Informatics Summit*, 2018.

[14] L. J. Hirsch, S. M. LaRoche, N. Gaspard, E. Gerard, A. Svoronos, S. T. Herman, R. Mani, H. Arif, N. Jette, Y. Minazad, J. F. Kerrigan, P. Vespa, S. Hantus, J. Claassen, G. B. Young, E. So, P. W. Kaplan, M. R. Nuwer, N. B. Fountain, and F. W. Drislane, "American clinical neurophysiology society's standardized critical care EEG terminology: 2012 version," vol. 30, no. 1, p. 1.