# Capturing Expert Arguments from Medical Adjudication Discussions in a Machine-readable Format

Mike Schaekermann
University of Waterloo
Waterloo, Canada
mschaeke@uwaterloo.ca

Graeme Beaton
University of Waterloo
Waterloo, Canada
graeme.beaton@edu.uwaterloo.ca

Minahz Habib
University of Toronto
Toronto, Canada
minahz.habib@mail.utoronto.ca

Andrew Lim
University of Toronto
Toronto, Canada
andrew.lim@utoronto.ca

Kate Larson
University of Waterloo
Waterloo, Canada
kate.larson@uwaterloo.ca

Edith Law
University of Waterloo
Waterloo, Canada
edith.law@uwaterloo.ca

## ABSTRACT

Group-based discussion among human graders can be a useful tool to capture sources of disagreement in ambiguous classification tasks and to adjudicate any resolvable disagreements. Existing workflows for panel-based adjudication, however, capture graders' arguments and rationales in a free-form, unstructured format, limiting the potential for automatic analysis of the discussion contents. We designed and implemented a structured adjudication system that collects graders' arguments in a machine-readable format without limiting graders' abilities to provide free-form justifications for their classification decisions. Our system enables graders to cite instructions from a set of labeling guidelines, specified in the form of discrete classification rules and conditions that need to be met in order for each rule to be applicable. In the present work, we outline the process of designing and implementing this adjudication system, and report preliminary findings from deploying our system in the context of medical time series analysis for sleep stage classification.

## CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing systems and tools**; **Human computer interaction (HCI)**; *Collaborative interaction*; *Empirical studies in collaborative and social computing*.

## KEYWORDS

Ambiguity; Disagreement; Adjudication; Medical time series

**ACM Reference Format:**
Mike Schaekermann, Graeme Beaton, Minahz Habib, Andrew Lim, Kate Larson, and Edith Law. 2019. Capturing Expert Arguments from Medical Adjudication Discussions in a Machine-readable Format. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308560.3317085


*WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA*

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317085

## 1 INTRODUCTION

A common requirement in supervised machine learning is that objects can be *unambiguously* classified into categories. In practice, however, there exist many classification tasks that are inherently ambiguous and the reasons why domain experts may be in disagreement over the correct way to classify an object may vary from task to task and from data object to data object. Several researchers have recognized this problem and come up with different solutions to handle it. One main distinction between these different works can be made around the question of whether expert disagreement is a problem to be resolved or whether disagreement is treated as a signal that is leveraged in some useful way. Our work is situated along the latter line of research. In particular, we propose that a key component to trusted and explainable artificial intelligence (AI) systems is to understand and capture the logical arguments and the various pieces of evidence that lead to divergent interpretations among experts. The overall goal is to get one step closer to endowing AI systems with the ability to provide argument-based explanations about (potentially ambiguous) classification decisions to their end users. Extending prior work on the design of systems for real-time group deliberation among remote human annotators [20, 22] and observational studies of in-person adjudication among expert annotators [21], in this work, we propose a general approach for capturing experts' rationale for individual classification decisions in a structured, guideline-centric format–with the goal of capturing sources of ambiguity and the content of evidence-driven adjudication discussions in a machine-readable format. The remainder of this paper covers related work, briefly introduces the reader to the application domain of sleep stage classification, details the proposed solution and preliminary findings from pilot experiments, and concludes with a discussion of use cases for our approach.

## 2 RELATED WORK

### 2.1 Ambiguity and Inter-rater Disagreement

Ambiguity is an issue of central importance in the field of epistemology, where an openness to multiple interpretations complicates the justification of knowledge. In practice, ambiguity gives rise to inter-rater disagreement in expert domains when there is lack of consensus on a single interpretation of a subjective case. Both ambiguity and expert disagreement have received extensive coverage in the epistemological literature [2, 7, 13, 24]. In a recent discussion of

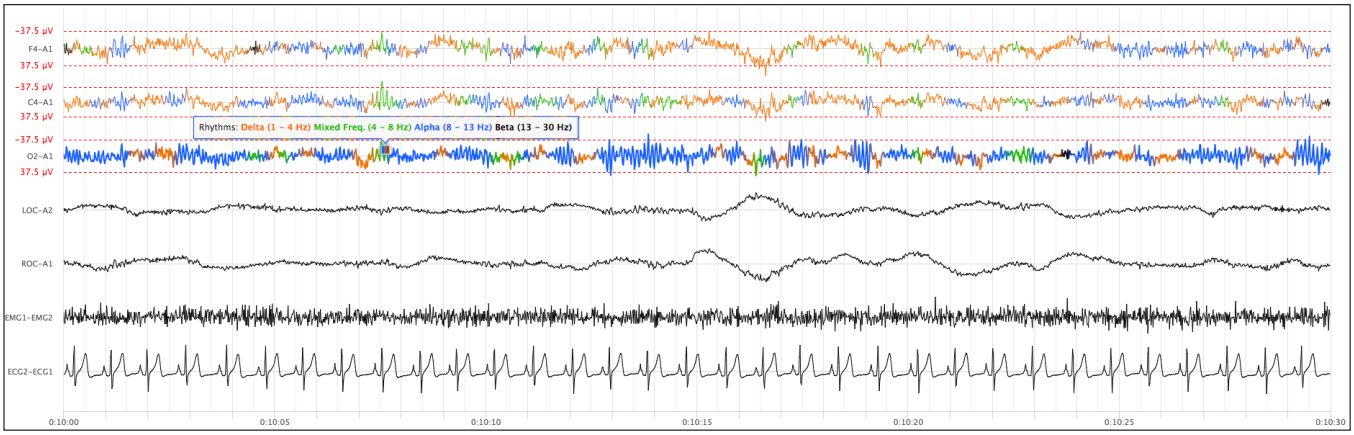


**Figure 1: Visualization of one 30-second epoch of biosignal data to be scored into one of five stages of sleep.**

the issue, Garbayo (2014) distinguished legitimate disagreement—where experts disagree despite access to the same evidence— from verbal disagreement, where experts misunderstand each other due to differences in terminology or semantics [13]. In an earlier theoretical account, Mumpower and Stewart (1996) delineate three forms of expert disagreement: (1) personality-based disagreement, beget by ideology, venality, or incompetence of the experts themselves, (2) judgement based disagreement, where information gaps exist, or (3) structural disagreement generated by different problem definitions or organizing principles held by experts [13].

In the clinical domain, expert disagreement is prevalent in diagnostic tasks that rely on visual analysis of subjective criteria. EEG interpretation is one such example. For instance, in the context of epilepsy diagnosis, interictal epileptiform discharges (IEDs) are key distinguishing features within an EEG—however, neurologists will often disagree over whether a particular waveform constitutes an IED. In a study by Bagheri et al. (2017), it was found that inter-rater agreement rates surrounding IED detection could be predicted based on particular wavelet features, given that the sample of experts was large enough [1]. In the context of sleep stage classification, which also depends on the identification of transient and infrequent features (i.e., sleep spindles and K-complexes), the average agreement rate among experts is 82.6% [19].

## 2.2 Adjudication in Medical Data Analysis

Disambiguating edge cases generated by inter-rater disagreement in the medical domain is a matter of contention in the contemporary literature. Majority-vote techniques have been criticized for their tendency to promote artificial consensus over valuable data or insights that might be had from group discussion and deliberation [23]. Indeed, group deliberation—where group members who hold conflicting beliefs present arguments and weigh evidence in light of their individual positions in order to reach a decision—has been shown to be a useful and productive technique for aggregating expert opinions and reaching consensus.

A study by Krause et al. (2017) found that in-person, group deliberation resulted in significantly higher recall among experts in diagnosing eye disease from images of the fundus, when compared to the majority vote technique [10]. Furthermore, it was demonstrated in the same study that group deliberation, when performed on just a small portion of a dataset, can be used to train the hyperparameters of deep learning models for more effective automated analysis. Guan et al. (2018) later used the same consensus data set to train multiple, grader-specific machine learning models, and showed that the aggregate performance of these models could beat out a single-prediction model trained with majority labels [8].

Adjudicated diagnoses have also proved valuable as reference standards for training machine learning models. In work done by Rajpurkar et al. (2017), cardiologists engaged in group deliberation to generate an adjudicated electrocardiogram (ECG) data set in the context of arrhythmia detection. This consensus validation data set was then used as a benchmark for a convolutional neural network, which was found to outperform individual cardiologists in ECG classification when trained solely on independent data labels [16].

Where sleep staging is concerned, it has been argued that group deliberation, also referred to as "consensus-scoring" or adjudication, is an optimal method of training human sleep scorers [15].

## 2.3 Computational Models of Argumentation

Argumentation is an approach to reasoning focused not only on the conclusions reached, but also on the data and the inference steps involved in inferring conclusions from the data. Argumentation has a considerable history in the field of computer science, including the problem of understanding common patterns of argumentative discourse in human decision making (e.g., [6, 25]) mapping natural language to a more formal, machine readable representation of argumentative discourse (e.g., [3–6, 11, 12, 18]), and using formal representations of arguments to generate new conclusions for previously unseen queries (e.g., [14, 17]).

# 3 APPLICATION DOMAIN

We leverage biomedical time series classification, a field with typically low inter-scorer reliability, as an application domain for embedding our work. In particular, we use examples from sleep stage classification, the expert task of mapping a sequence of fixed-length pages (typically 30 seconds) of continuous multimodal medical time

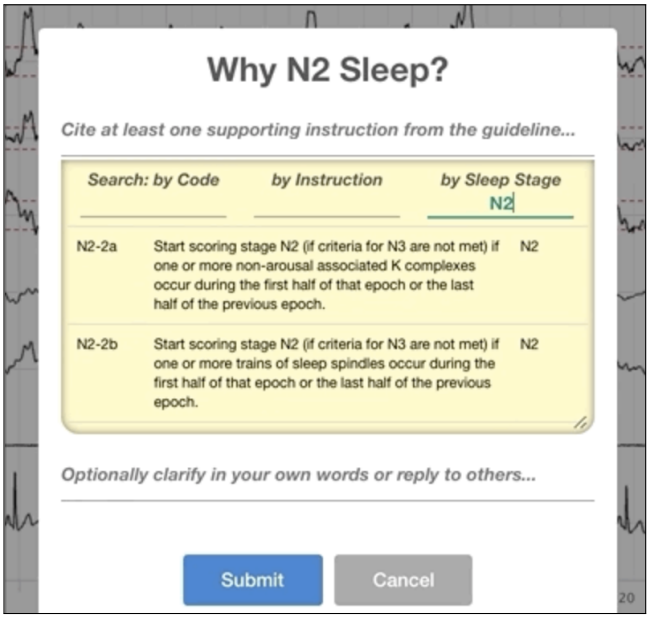


**Figure 2: Rationale form for expert graders to cite guideline instructions in support of their classification decision.**

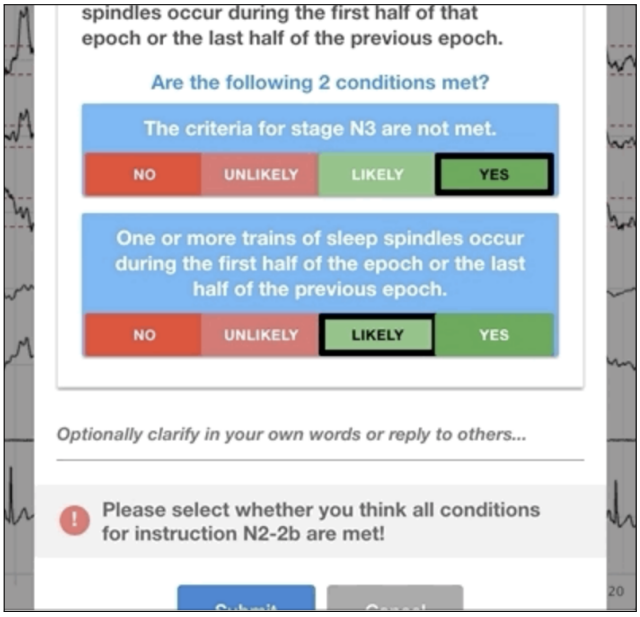


**Figure 3: Expert graders specify their level of confidence for individual conditions required for a cited instruction.**

series (*polysomnogram*, see Figure 1) to a sequence of discrete sleep stages (*hypnogram*). Each fixed-length page of time series (epoch) is classified into one of five different stages of sleep—Wake, NREM1, NREM2, NREM3 or REM sleep—based on the stage comprising the greatest portion of the epoch. Rosenberg and van Hout [19] conducted a study on inter-scorer reliability in sleep stage classification, finding that expert agreement averages around 82.6%.

## 4 PROPOSED SOLUTION

We propose to augment the traditional process of collecting ground truth labels for supervised learning (i.e., querying one or more experts for the "correct" label to a given input example), by introducing an extra step to elicit the reasons for certain classification decisions (i.e., *rationale*) in a structured form. In particular, we propose to collect expert rationale in the form of propositional logic where experts specify the evidence and inference rules they used to arrive at their classification decisions. In this section, we guide the reader through the input and output of our proposed approach and the intermediate steps required to transform the input to the output. We illustrate our explanations with examples from sleep stage classification.

**Input.** Our system will take as input a pool of *human experts* and a set of *data objects* (e.g., images, text documents, medical time series) to be classified into one of several preset *categories*. In the case of sleep stage classification, a set of pages of physiological time series (see Figure 1) is classified into one of 5 stages of sleep by a pool of sleep technologists.

**Output.** For each input data object, our method will output a distribution of classification labels, one from each expert. For those data objects that led to some disagreement among the experts throughout the labeling process, the system will also output each individual expert's rationale for their final classification decision in the form of propositional logic. In other words, for ambiguous cases, the system will list the inference rule(s) each expert used to arrive at a certain classification decision as well as the expert's confidence levels for the evidence criteria that need to be met in order for the chosen inference rule(s) to be applicable. An example for sleep stage classification may look as follows:

- **Expert A:**
  - **Classification:** Wake
  - **Rule:** "*[W-3a] Score epochs without alpha rhythm as stage W if eye blinks are present.*"
  - **Confidence for Evidence Criteria:**
    - Alpha rhythm absent: **Yes**
    - Eye blinks present: **Likely**
- **Expert B:**
  - **Classification:** NREM1 Sleep
  - **Rule:** "*[N1-2] In patients who generate alpha rhythm, score stage N1 if the alpha rhythm is extenuated and replaced by low-amplitude, mixed-frequency activity for more than half of the epoch.*"
  - **Confidence for Evidence Criteria:**
    - Patient generates alpha rhythm: **Likely**
    - Alpha rhythm is extenuated and replaced by low-amplitude, mixed-frequency activity for more than half of the epoch: **Yes**

### 4.1 Capturing Structured Rationale

Our solution requires two components to capture rationale in the format above:

(1) Rule-based representation of the classification guidelines
(2) User interface to collect rationale in structured form

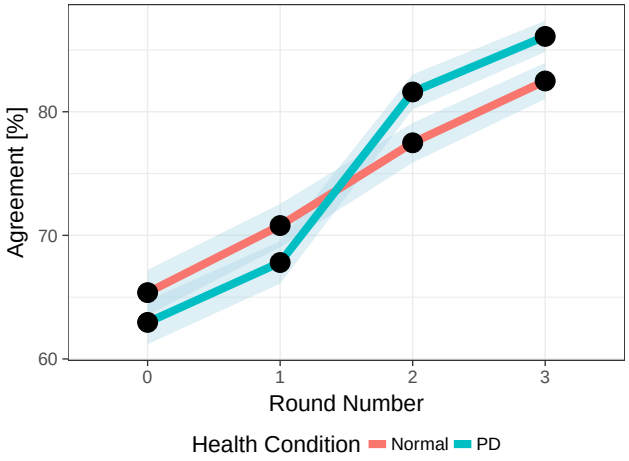


**Figure 4: Agreement rate by adjudication round number and patient's health condition. Transparent overlays indicate 95 percent confidence intervals.**

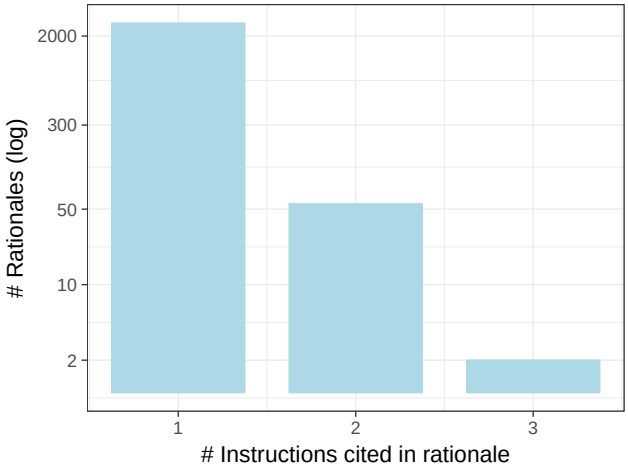


**Figure 5: Number of rationales citing one, two and three guideline instructions.**

**Rule-based representation of the classification guidelines.** The first step in collecting rationale in a structured form is to define the set of possible inference rules for classifying objects, and a set of evidence criteria (i.e., Boolean propositions) that need to be met in order for a given rule to be applicable. For our application domain of sleep stage classification, we adapted the official *AASM Manual for the Scoring of Sleep and Associated Events* [9] into a set of 36 individual inference rules (8 for Wake, 10 for NREM1 Sleep, 9 for NREM2 Sleep, 3 for NREM3 Sleep, and 6 for REM Sleep) with 52 unique evidence criteria.

**User interface to collect rationale in structured form.** The resulting rule-based representation of the classification guidelines needs to be exposed through a user interface enabling experts to specify individual inference rules and to indicate the extent to which they believe that the evidence criteria required for the selected rules are met. Figure 2 illustrates our implementation of such an interface. In our example, the interface starts out as an empty rationale form with two input fields—one to select discrete inference rules, and one to optionally explain more in one's own words. The first input field will automatically suggest possible inference rules based on the current classification decision (e.g., Wake) and the keywords typed into the input field. Once the user selects an inference rule, the rationale form automatically lists the evidence criteria that need to be true in order for the rule to be applicable (Figure 3), prompting the user to indicate the extent to which they believe each condition is met, one of: No, Unlikely, Likely, Yes. The interface produces warnings for invalid inputs (e.g., selecting rules while indicating that their conditions are not met) to ensure the user has selected at least one inference rule in support of their classification decision and specified their confidence level for each of the evidence criteria, before submitting the rationale and proceeding to the next disagreement case.

## 5 PILOT EXPERIMENT

A pilot experiment was conducted to explore the usefulness of the proposed procedure and to demonstrate sample analyses made possible by the resulting structured adjudication data. For our pilot experiment, we sampled six EEG recordings from six unique patients, three with Parkinson's disease (PD) and three normal control subjects. As is the case for other neurological disorders, sleep studies from PD patients may exhibit slight differences in the expression of sleep-relevant features (e.g., sleep spindles and K-complexes), compared to healthy subjects, and may therefore lead to different disagreement patterns among domain experts. For annotation, we recruited 18 sleep technologists as expert graders forming six panels of three experts each. Each EEG recording was assigned to exactly one of the six expert panels, and each expert grader participated in exactly one panel. Graders first performed an initial independent round of scoring on their assigned recordings, followed by three rounds of adjudication, one round per grader in the panel. In each adjudication round, the active grader stepped through each individual epoch with any level of disagreement among panel members, re-scored the epoch and provided a rationale for their reclassification decision. In each adjudication round and for each disagreement epoch, the active grader was presented with the most recent grades from all three panel members, as well as the grades and rationales submitted during each of the preceding rounds. The three panels adjudicating recordings from PD patients used the structured, guideline-centric way of collecting rationale during adjudication discussions. In contrast, the three remaining panels adjudicating recordings from normal control subjects used the non-structured, free-form way of collecting rationale.

## 6 RESULTS

The output of our rationale form (i.e., sleep stage classification labels and their given expert rationales) provided the basis for a detailed set of quantitative results across multiple items of analysis.

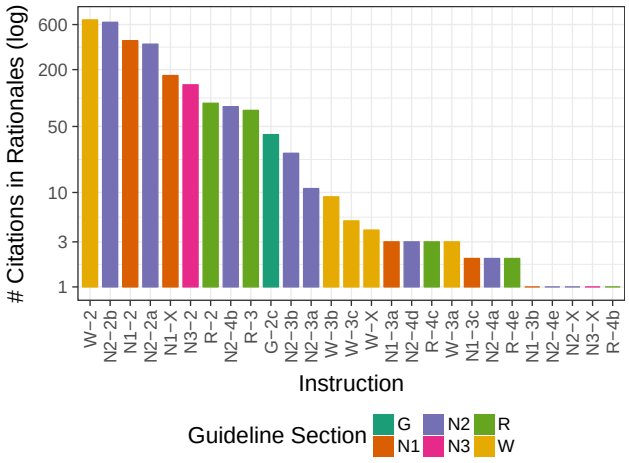


**Figure 6: Number of citations per guideline instruction. Bar colours indicate the guideline section for each instruction: General (G), NREM1 Sleep (N1), NREM2 Sleep (N2), NREM3 Sleep (N3), REM Sleep (R) and Wake (W).**

For ambiguous cases in which experts were required to provide a rationale and confidence levels for their assessments, data were collected on the number of citations each guideline instruction in the AASM scoring manual [9] received. As seen in Figure 6, select guideline sections received a disproportional number of citations compared to others, based on the sleep stage to which the classification label pertained. The guideline section listing rules for scoring NREM2 sleep received the greatest number of citations overall, while instruction W-2 received the highest number of citations for a given instruction. Certain guideline instructions, such as R-4b, received only a single citation across all ambiguous cases.

Despite a wide distribution in the subsections indexed by expert rationales, rarely did individual rationales cite more than one guideline instruction (see Figure 5). The significant majority of rationales (>2000) referenced only one instruction, while less than 3% of rationales given cited two. For all rationales collected, none cited more than three instructions in total.

Since adjudication decisions in sleep stage classification often hinge on the presence or absence of distinguishing features of an EEG waveform, references to feature types were also collected from expert rationale output. 15 basic features in total were mentioned across all rationales, again, with select features (i.e., alpha rhythm, train of sleep spindles, non-arousal associated K-complexes) receiving far more mentions than others. However, unlike the data for the number of citations per guideline instruction (Figure 6), there was a much flatter distribution in the number of mentions received by each feature type, with no substantial difference between the number of mentions across 7 of the 15 feature types, as shown in Figure 7.

Results from three adjudicated rounds of scoring with 6 panels of expert participants (3 experts per panel) showed a marked decrease in the number of controversial cases—and thus an increase in inter-rater agreement—between independent annotation and the end of adjudication, as seen in Figure 4. For the 3 panels that adjudicated EEG recordings from healthy controls, average agreement rate increased across three rounds of adjudication, and rose from ~66% to ~83%. These 3 panels provided rationales for disagreement cases through a non-structured, free-form interface. Among the three panels that adjudicated EEG recordings from Parkinsonian patients, average inter-rater agreement rate increased across three rounds of adjudication from ~63% to ~86%. Here, experts provided classification rationales for disagreement cases through our structured, guideline-centric interface. While overall change in average agreement rate did not substantially differ between free-form and structured adjudication, there were clear differences between the rates at which agreement rates increased across adjudication rounds in these two scenarios. Where free-form rationales were provided for disagreement cases between experts, average inter-rater agreement increased in a linear fashion between independent annotation and adjudication. For those expert panels that provided their rationales through the structured, guideline-centric interface, inter-rater agreement increased in a step-wise fashion, with the biggest jump in average agreement rate occurring between the first and second round of adjudication.

Furthermore, a stepwise logistic regression model was used to understand which feature types, when mentioned during a given adjudication round, were associated with the probability that the active adjudicator will change their classification decision in the same round. As outlined in Table 1, the likelihood that an expert would change their classification decision could be predicted based on which feature types were mentioned in their rationale. Graders mentioning arousals ($p < 0.05$) or low-amplitude mixed frequency activity (LAMF; $p < 0.001$) in their adjudication rationales were significantly more likely to stick with their classification decision than those not mentioning these features. Citing instructions pertaining to K-complexes ($p < 0.01$) or trains of sleep spindles ($p < 0.001$) was significantly associated with a change in the grader's classification decision in the same round. The logistic model selected additional feature types—low chin EMG tone, reading eye movements and rapid eye movements (REM)—contributing to the model fit without statistical significance.

## 7 DISCUSSION

The main contribution of our present work is a structured system and procedure for capturing expert rationale during adjudication of complex data sets–in this case, within the application domain of sleep stage classification. In addition to querying a group of experts for data labels (i.e., a sleep stage for a given epoch), we solicited rationales from these expert graders in the form of propositional logic (i.e., the reason for their classification decision in the form of a sleep scoring guideline), as well as their self-reported confidence levels for their evidence criteria. During structured adjudication, average inter-rater agreement rose by roughly 23%.

Applications of this system exist both within and outside of the present application domain. In the context of sleep stage classification, controversial cases surrounding classification decisions—especially those that remain after adjudication—may point to ambiguous instructions in the sleep scoring guidelines. With output from a structured rationale form, specific guideline sections and

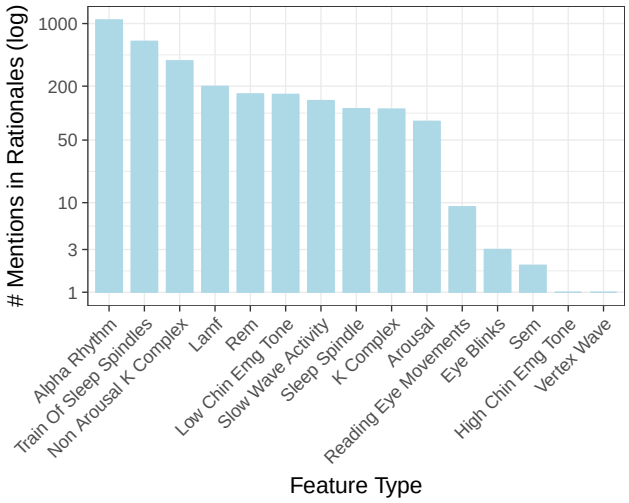


**Figure 7: Number of times each feature type was mentioned in a rationale.**

**Table 1: Logistic model for understanding the likelihood of a grader changing their decision in a given round, based on the feature types mentioned in their rationale.**

| | Model Parameters | | | |
|---|---|---|---|---|
| **Variable** | $\hat{\beta}$ | *Std. Error* | *t* | *p-value* |
| K Complex | 0.93 | 0.33 | 2.77 | ** |
| Train Of Sleep Spindles | 0.37 | 0.09 | 3.97 | *** |
| Arousal | -0.52 | 0.25 | -2.10 | * |
| Lamf | -1.72 | 0.51 | -3.36 | *** |
| Low Chin Emg Tone | 16.28 | 624.19 | 0.03 | |
| Reading Eye Movements | -14.22 | 294.25 | -0.05 | |
| Rem | -14.22 | 624.19 | -0.02 | |

instructions can be indexed for potential iteration. Likewise, particular feature types mentioned in given rationales may require more rigorous definitions.

Beyond the present application domain, we have shown that the use of statistical models like the one illustrated in Table 1 can help predict outcomes of structured adjudication, e.g., the likelihood that an individual grader will change their assessment based on particular evidence criteria. Similar models could be used to predict how many rounds of adjudication are necessary to resolve ambiguous cases. Of the total ~23% rise in average inter-rater agreement during our structured adjudication procedure, ~19% of that increase occurred between the first and second adjudication rounds. Additional analysis may reveal patterns as to which types of disagreements are resolved early vs. late.

The most notable limitation of work of this kind lies with the fact that adjudication is costly–especially in terms of time investment from expert graders. In addition, adjudication procedures like the one we deployed in this application domain depend on the existence of standardized grading guidelines (like the AASM sleep scoring manual [9]) which must be agreed upon by all expert graders participating. In this study, we mapped an existing standard scoring manual into a set of scoring instructions integrated into the adjudication interface. In cases where a single agreed-upon grading guideline does not exist in the community (e.g., interpretation of EEGs for epileptiform abnormalities), there is interesting potential for future work in the iterative development of scoring guidelines based on adjudication procedures that become increasingly structured, and thus less time-intensive, over time.

This study has laid the groundwork for future work in the area of analyzing the sources of inter-rater disagreement in labelling complex datasets, within the application domain of sleep stage classification. Throughout the continuation of this project, we intend to prepare and make publicly available a high-quality dataset of adjudicated human polysomnograms. The use cases for this data are severalfold. First, there is the potential to derive concrete suggestions for guideline development. Second, the data could be used to build machine learning models for predicting ambiguity and sources of disagreement for previously unseen data, an ability that would be helpful for guiding human labelling resources, and for establishing new pathways towards more informed and explainable concepts of uncertainty in machine learning. Third, these adjudication data sets can be used to train human graders in better disambiguating edge cases by leveraging structured information from adjudication rounds to target grader training towards different categories of ambiguity.

## 8 CONCLUSION

In this work, we introduced a novel perspective on the problem of handling expert disagreement in ambiguous classification tasks by proposing a structured procedure for collecting expert arguments put forward during panel-based adjudication in the form of propositional logic. We demonstrated the applicability of our approach in the context of medical time series analysis for sleep stage classification, and showcased how the data produced can facilitate detailed quantitative analyses of discussion contents and outcomes. Our solution has implications for the broader field of supervised learning from human-labeled data by translating the problem of trustworthy AI systems to that of trusted and explainable ground truth.

## ACKNOWLEDGMENTS

We thank Rui DeSousa for his invaluable help in recruiting participants for this study. This work was funded by NSERC CHRP (CHRP 478468-15) and CIHR CHRP (CPG-140200).

## REFERENCES

[1] Elham Bagheri, Justin Dauwels, Brian C. Dean, Chad G. Waters, M. Brandon Westover, and Jonathan J. Halford. 2017. Interictal epileptiform discharge characteristics underlying expert interrater agreement. *Clinical Neurophysiology* 128, 10 (10 2017), 1994–2005. https://doi.org/10.1016/j.clinph.2017.06.252

[2] John Beatty and Alfred Moore. 2010. Should We Aim for Consensus? *Episteme* 7, 3 (2010), 198âĂŞ214. https://doi.org/10.3366/E1742360010000948

[3] Floris Bex, Henry Prakken, Chris Reed, and Douglas Walton. 2003. Towards a Formal Account of Reasoning about Evidence: Argumentation Schemes and Generalisations. *Artificial Intelligence and Law* 11, 2/3 (2003), 125–165. https://doi.org/10.1023/B:ARTI.0000046007.11806.9a

[4] Katarzyna Budzynska, Mathilde Janier, Juyeon Kang, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yaskorska. 2014. Towards Argument Mining from Dialogue. In *Computational Models of Argument - Proceedings of {COMMA}*

*2014, Atholl Palace Hotel, Scottish Highlands, UK, September 9-12, 2014*. 185–196. https://doi.org/10.3233/978-1-61499-436-7-185

[5] Carlos Chesñevar, Jarred McGinnis, Sanjay Modgil, Iyad Rahwan, Chris Reed, Guillermo Simari, Matthew South, Gerard Vreeswijk, and Steven Willmott. 2006. Towards an argument interchange format. *The Knowledge Engineering Review* 21, 04 (12 2006), 293. https://doi.org/10.1017/S0269888906001044

[6] Robin Cohen. 1987. Analyzing the Structure of Argumentative Discourse. *Comput. Linguist.* 13, 1-2 (1 1987), 11–24. http://dl.acm.org/citation.cfm?id=26386.26388

[7] Luciana Garbayo. 2014. Epistemic Considerations on Expert Disagreement, Normative Justification, and Inconsistency Regarding Multi-criteria Decision Making. *Constraint Programming and Decision Making* 539 (2014), 35–45. http://link.springer.com/10.1007/978-3-319-04280-0%5C_5

[8] Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton. 2018. Who said what: Modeling individual labelers improves classification. In *AAAI Conference on Artificial Intelligence*. https://arxiv.org/pdf/1703.08774.pdf

[9] Conrad Iber, Sonia Ancoli-Israel, Andrew L Cheeson Jr., and Stuart F Quan. 2007. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. American Academy of Sleep Medicine.

[10] Jonathan Krause, Varun Gulshan, Ehsan Rahimy, Peter Karth, Kasumi Widner, Greg S. Corrado, Lily Peng, and Dale R. Webster. 2018. Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy. *Ophthalmology* (3 2018). https://doi.org/10.1016/j.ophtha.2018.01.034

[11] John Lawrence and Chris Reed. 2015. Combining Argument Mining Techniques. In *Proceedings of the 2nd Workshop on Argumentation Mining at ACL 2015*. 127–136. https://doi.org/10.3115/v1/W15-0516

[12] John Lawrence and Chris Reed. 2016. Argument Mining using Argumentation Scheme Structures. *Proceedings of the 6th International Conference on Computational Models of Argument (COMMA 2016)* 0 (2016), 379 – 390. https://doi.org/10.3233/978-1-61499-686-6-379

[13] Jeryl L. Mumpower and Thomas R. Stewart. 1996. Expert Judgement and Expert Disagreement. *Thinking & Reasoning* 2, 2-3 (7 1996), 191–212. https://doi.org/10.1080/135467896394500

[14] Simon Parsons, Elizabeth Sklar, Jordan Salvit, Holly Wall, and Zimi Li. 2013. ArgTrust: Decision Making with Information from Sources of Varying Trustworthiness. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems (AAMAS '13)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1395–1396. http://dl.acm.org/citation.cfm?id=2484920.2485242

[15] Thomas Penzel, Xiaozhe Zhang, and Ingo Fietze. 2013. Inter-scorer reliability between sleep centers can teach us what to improve in the scoring rules. *Journal of Clinical Sleep Medicine* 9, 1 (2013), 81–87.

[16] Pranav Rajpurkar, Awni Y. Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y. Ng. 2017. Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks. (7 2017). http://arxiv.org/abs/1707.01836

[17] Chris Reed and Timothy Norman. 2004. *Argumentation Machines*. Argumentation Library, Vol. 9. Springer Netherlands, Dordrecht. https://doi.org/10.1007/978-94-017-0431-1

[18] Chris Reed and Doug Walton. 2005. Towards a Formal and Implemented Model of Argumentation Schemes in Agent Communication. 19–30. https://doi.org/10.1007/978-3-540-32261-0{_}2

[19] Richard S. Rosenberg and Steven van Hout. 2013. The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring. *Journal of Clinical Sleep Medicine* (1 2013). https://doi.org/10.5664/jcsm.2350

[20] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work. In *Proceedings of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW'18)*. New York City, NY. https://doi.org/10.1145/3274423

[21] Mike Schaekermann, Edith Law, Kate Larson, and Andrew Lim. 2018. Expert Disagreement in Sequential Labeling: A Case Study on Adjudication in Medical Time Series Analysis. In *1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing at HCOMP 2018*. Zurich, Switzerland.

[22] Mike Schaekermann, Edith Law, Alex C Williams, and William Callaghan. 2016. Resolvable vs. Irresolvable Ambiguity: A New Hybrid Framework for Dealing with Uncertain Ground Truth. In *1st Workshop on Human-Centered Machine Learning at SIGCHI 2016*. San Jose, CA.

[23] Miriam Solomon. 2006. Groupthink versus The Wisdom of Crowds : The Social Epistemology of Deliberation and Dissent. *The Southern Journal of Philosophy* 44, S1 (3 2006), 28–42. https://doi.org/10.1111/j.2041-6962.2006.tb00028.x

[24] Miriam Solomon. 2007. The social epistemology of NIH consensus conferences. In *Establishing medical reality*. Springer, 167–177.

[25] D Walton, C Reed, and F Macagno. 2008. *Argumentation Schemes*. Cambridge University Press. https://books.google.ca/books?id=qc3LCgAAQBAJ