Expert Disagreement in Sequential Labeling: A Case Study on Adjudication in Medical Time Series Analysis

Mike Schaekermann and **Edith Law**
Human-Computer Interaction Lab
School of Computer Science
University of Waterloo

Kate Larson
Artificial Intelligence Group
School of Computer Science
University of Waterloo

Andrew Lim
Division of Neurology
Sunnybrook Health Sciences Centre
University of Toronto

Abstract

Low inter-rater agreement is typical in various expert domains that rely in part on subjective evaluation criteria. Prior work has predominantly focused on expert disagreement with respect to individual cases in isolation. In this work, we report results from a case study on expert disagreement in sequential labeling tasks where the interpretation of one case can affect the interpretation of subsequent or previous cases. Three board-certified sleep technologists participated in face-to-face adjudication sessions to resolve disagreement in the context of sleep stage classification. We collected 1,920 independent scoring decisions from each expert on the same dataset of eight 2-hour long multimodal medical time series recordings. From all disagreement cases (29% of the dataset), a representative subset of 30 cases was selected for adjudication and expert discussions were analyzed for sources of disagreement. We present our findings from this case study and discuss future application scenarios of expert discussions for the training of non-expert crowdworkers.

Introduction

One of the most common use cases for crowdsourcing is the classification of objects into categories. While crowdsourced classification tasks traditionally focused on problems not requiring domain expertise, recent work suggests that crowdsourcing can also be effective for expert-level classification. Examples of such expert tasks from the medical domain include the identification of low-level patterns in sleep-related biosignals [Warby et al.2014], the annotation of retinal images [Mitry et al.2016], and medical relation extraction [Dumitrache, Aroyo, and Welty2016].

In many mission-critical expert domains including the interpretation of medical data, low inter-rater agreement rates are the norm [Rajpurkar et al.2017, Krause et al.2018, Guan et al.2018, Penzel, Zhang, and Fietze2013]. Expert disagreement, however, poses fundamental challenges to quality control procedures in crowdsourcing, and to the use of data labels in supervised machine learning, as it is not immediately obvious how cases at the inter-subjective decision boundary should be disambiguated if multiple equally-qualified domain experts exhibit genuine disagreement.

Prior work has predominantly paid attention to the nature, sources and resolvability of expert disagreement on individual classification tasks in isolation [Beatty and Moore2010, Garbayo2014, Mumpower and Stewart1996, Solomon2007]. Many interpretation tasks, however, are sequential in nature, i.e., the interpretation of one case affects the interpretation of subsequent or previous cases. For example, in text translation, the semantic interpretation of one phrase or sentence can affect the translation of subsequent or previous phrases or sentences. Heidegger called this reciprocity of text and context the *hermeneutic circle*. Overall, sequential labeling makes up a large and diverse class of problems from numerous expert domains.

In this work, we present findings from a case study on expert disagreement in the context of sleep stage classification, the expert task of mapping a sequence of fixed-length pages of continuous multimodal medical time series (*polysomnogram*, see Figure 1) to a sequence of discrete sleep stages (*hypnogram*). Prior work has established that inter-rater agreement in sleep staging averages around 82.6% [Rosenberg and van Hout2013]. The objective of this case study is to identify various sources of expert disagreement in sleep stage classification and to investigate if and to what extent disagreement may be specific to the sequential nature of the labeling task and underlying data.

To answer these questions, we collected 1,920 independent sleep scoring decisions from a committee of three board-certified sleep technologists. We then selected a representative subset of the resulting disagreement cases which were resolved through in-person adjudication among the members of the expert committee.

The rest of this paper describes the related work, then details our study for collecting and analyzing the expert deliberation data, and concludes with a discussion of application scenarios for the training of non-expert crowdworkers.

Related Work

Ambiguity and Sources of Inter-rater Disagreement

Ambiguity, the *quality of being open to more than one interpretation*, and the phenomenon of expert disagreement are central to the justification of knowledge, and have been extensively discussed in the epistemic literature [Beatty and

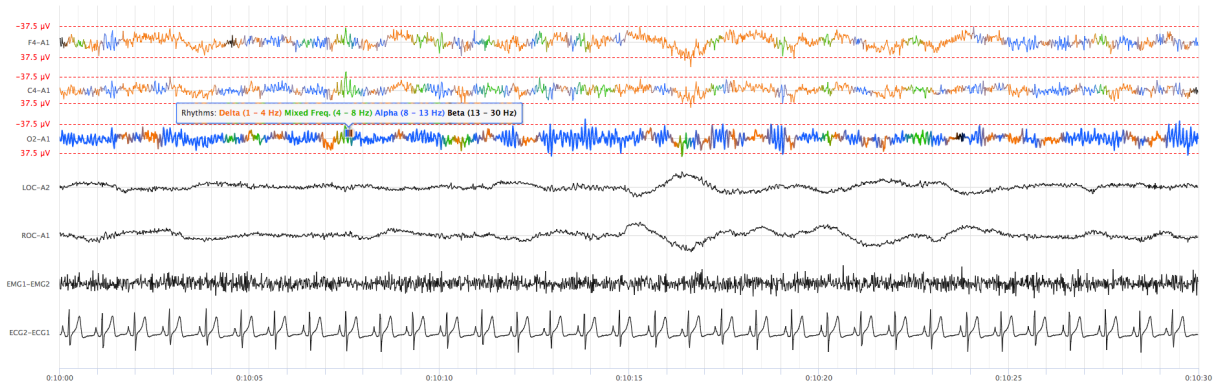

Figure 1: Visualization of one 30-second epoch of biosignal data to be scored into one of five sleep stages

Moore2010, Garbayo2014, Mumpower and Stewart1996, Solomon2007]. An early theoretical account identified three types of expert disagreement [Mumpower and Stewart1996]: *personality-based* disagreement that arises due to incompetence, venality or ideology of experts, *judgment-based disagreement* that arises due to missing information, or *structural disagreement* that arises due to experts adopting different problem definitions or organizing principles. Garbayo [Garbayo2014], on the other hand, distinguished *verbal* disagreement, i.e., discrepancy in terminology leading to misunderstanding among experts, from a form of *legitimate* disagreement, arising when experts have access to the same evidence, but diverge in interpretations.

Recent work in the field of human-computer interaction (HCI) has explored the issue of disagreement in the context of crowdsourcing tasks. Gurari and Graumen [Gurari and Grauman2017] analyzed visual question answering tasks and found that disagreement can be attributed to ambiguous and subjective questions, insufficient or ambiguous visual evidence, differing levels of annotator expertise, and vocabulary mismatch. Chang et al. [Chang, Amershi, and Kamar2017] found that workers disagree because of incomplete or ambiguous category definitions, and proposed to elicit help from the crowd for refinement. Kairam and Heer [Kairam and Heer2016] introduced a clustering-based technique to identify subgroups of workers with diverging, but equally valid interpretations of the same (entity annotation) task. Their work shows that disagreement can depend on how conservatively or liberally workers interpret category definitions.

Our study revolves around the task of biomedical time series classification, a field with typically low inter-scorer reliability. For example, Rosenberg and van Hout [Rosenberg and van Hout2013] conducted a large-scale study on inter-scorer reliability in sleep stage classification and found that average expert agreement is as low as 82.6%. In a comment on this study, Penzel et al. [Penzel, Zhang, and Fietze2013] explained that systematic studies on the inter-rater reliability of sleep automatically bring up the question of truth, claiming that the "true" state (i.e., sleep stage) is unknown and can only be approximated through aggregation of expert opinions.

Group Deliberation as a Method for Disambiguation

Group deliberation is an interactive form of decision making among humans which typically involves group members with conflicting beliefs who try to reach consensus on a given question by presenting arguments, weighing evidence and reconsidering individual positions.

Several works explored factors that influence the process and outcomes of group deliberation. Nemeth [Nemeth1977] found that when jurors are required to reach a unanimous decision, there is more conflict, more changes in assessments, and higher confidence in the final verdict reported by members of the group. Solomon [Solomon2006] sees conflict as an important feature of any effective deliberation system. He argues that dissent is both necessary and useful—as *"dissenting positions are associated with particular data or insights that would be otherwise lost in consensus formation"*—and criticizes procedures that push deliberators to reach consensus. Instead, he advocates for a structured deliberation procedure that avoids the undesired effects of *groupthink* [Janis1972]—the tendency to agree with the group by suppressing dissent and appraisal of alternatives—by actively encouraging dissent, organizing independent subgroups to deliberate on the same problem, and ensuring diversity of group membership.

In their study of communication technology and its effects on group decision making, Kiesler and Sproull [Kiesler and Sproull1992] found that time limits imposed on deliberation tend to decrease the number of arguments exchanged and to polarize discussions. The authors suggest the use of voting techniques or explicit decision rules to structure the deliberation timeline.

Liu et al. [Liu et al.2018] proposed a method for visualizing disagreement across multiple subjective decision criteria, and demonstrated that highlighting individual points of disagreement accelerates consensus formation in a comparative hiring task. Participants reported varying reasons for shifting opinions, from genuine consensus to appeasement, reinforcing the significance of *groupthink* in the orchestration of deliberation procedures. Navajas et al. [Navajas et al.2018] studied the effectiveness of in-person group delib-

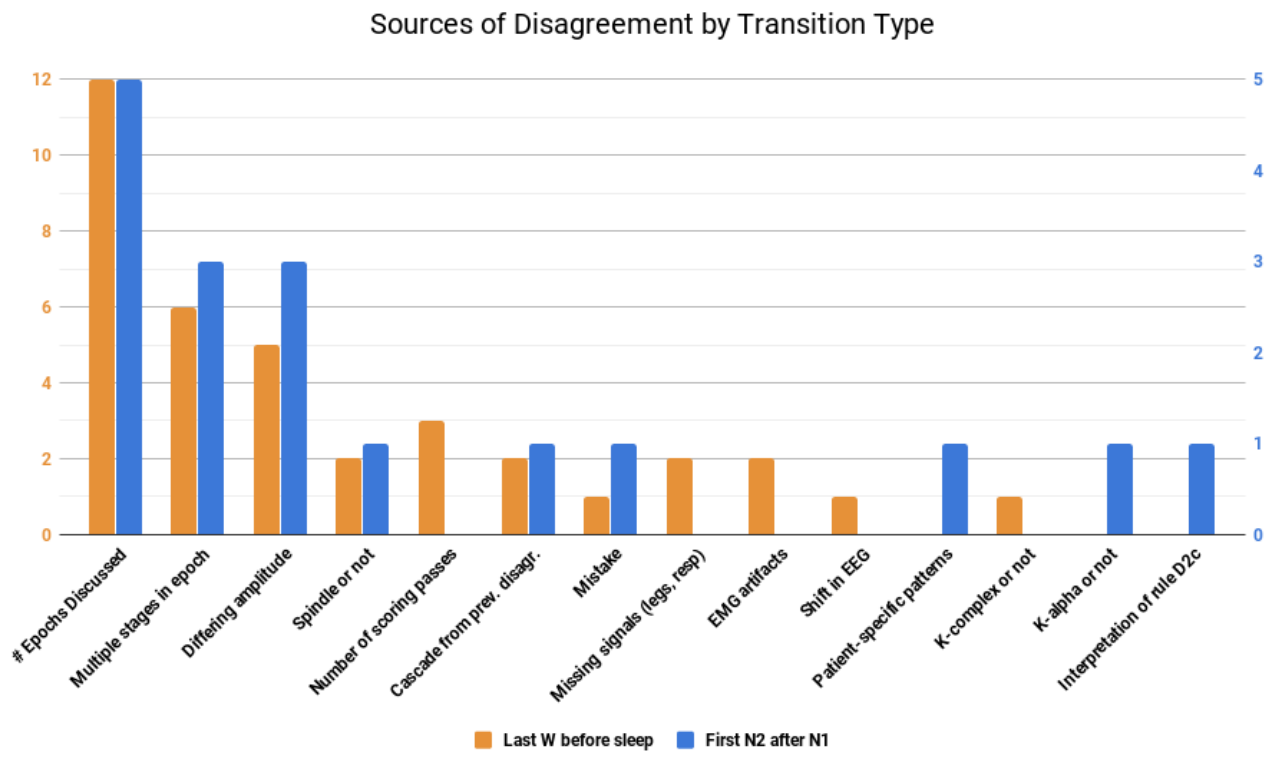

Figure 2: Sources of disagreement by transition type. The vertical axis plots the number of times a particular source of disagreement was mentioned in an expert discussion about a case from one of two transition types: Last Wake before sleep onset and transitions from N1 sleep to N2 sleep. Note the two vertical axes, one for each transition type, are re-scaled to facilitate a visual comparison of both distributions relative to the number of epochs discussed (# Epochs Discussed) for each transition type. Expert discussions could mention more than one source of disagreement.

eration in the context of general-knowledge questions (e.g., what is the height of the Eiffel Tower?) and reported that averaging consensus decisions yielded substantially higher accuracy than averaging individual independent responses.

Consensus Scoring in Medical Data Analysis

Group deliberation has also been proposed as a technique for disambiguating edge cases in the interpretation of medical data. Rajpurkar et al. [Rajpurkar et al.2017] employed group deliberation among cardiologists to generate a high-quality validation data set in the context of arrythmia detection from electrocardiograms (ECGs). Their work revealed that a convolutional neural network trained on independent labels (i.e., labels collected without deliberation) exceeded the classification performance of individual cardiologists when benchmarked against the consensus validation set.

Krause et al. [Krause et al.2018] compared majority vote to in-person deliberation as techniques for aggregating expert opinions for diagnosing eye diseases from photos of the eyeground. Compared to majority vote, in-person deliberation yielded substantially higher recall, suggesting the potential of group deliberation for mitigating underdiagnosis of diabetic retinopathy and diabetic macular edema. Krause et al. also showed that performing group deliberation on a small portion of the entire data set can make tuning of hyperparameters for deep learning models more effective. The same consensus data set was later used by Guan et al. [Guan et al.2018] to validate the classification performance of a novel machine learning approach involving the training of multiple grader-specific models. They demonstrated that training and aggregating separate grader-specific models can be more effective than training a single prediction model on majority labels.

In the context of sleep stage classification, Penzel et al. [Penzel, Zhang, and Fietze2013] refer to the concept of in-person group deliberation as *consensus scoring*, concluding that an *"optimal training for [...] sleep scorers is participation in consensus scoring rounds"*. In this work, we translate this idea to the non-expert domain by augmenting training procedures for crowd workers by using edge-case examples and the associated expert discussion dialogues to teach disambiguation skills to human learners in the context of sleep stage classification.

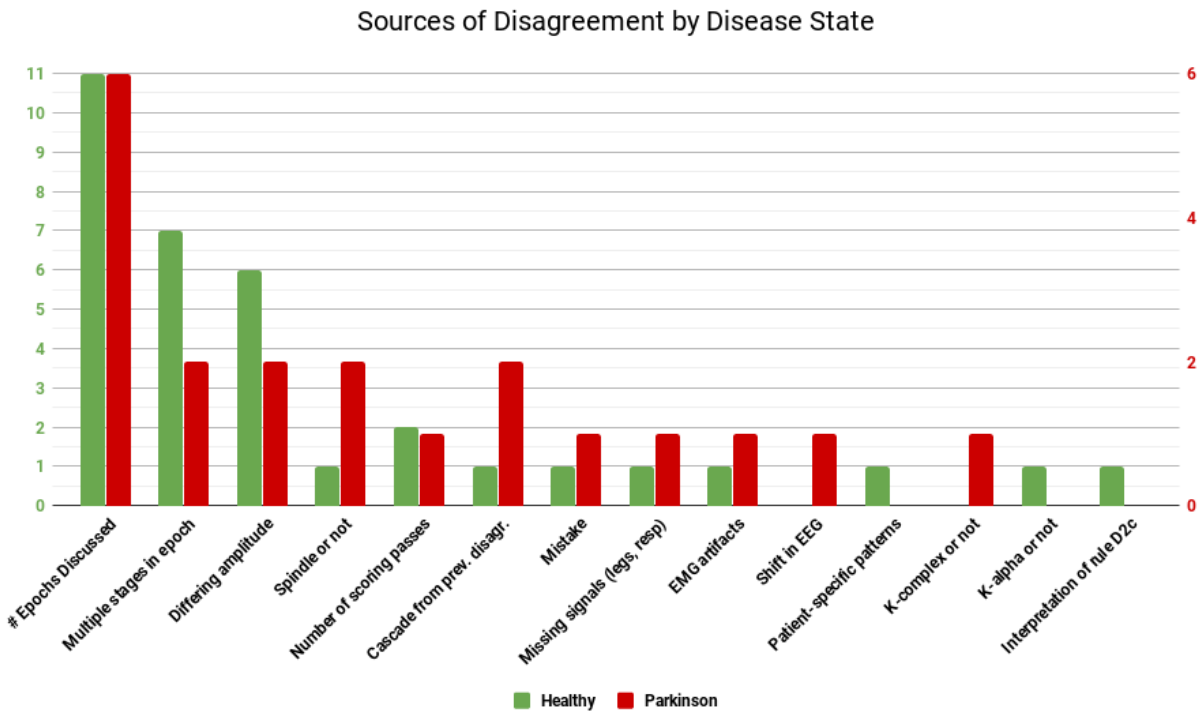

Figure 3: Sources of disagreement by disease state. The vertical axis plots the number of times a particular source of disagreement was mentioned in an expert discussion about a case from one of two disease states: Healthy and Parkinson's Disease. Note the two vertical axes, one for each disease state, are re-scaled to facilitate a visual comparison of both distributions relative to the number of epochs discussed (# Epochs Discussed) for each disease state. Expert discussions could mention more than one source of disagreement.

Expert Deliberation Data Set

An in-person deliberation study was conducted with an expert committee of three board-certified sleep technologists at Sunnybrook Health Sciences Centre in Toronto to investigate the extent and potential sources of inter-rater disagreement in sleep stage classification, and the effectiveness of group deliberation as a method for consensus formation.

Data Set

We prepared a data set of eight 2-hour-long PSG recording fragments. Each 2-hour-long fragment contained a sequence of 240 30-second epochs of biosignal data, resulting in 1,920 (240 x 8) epochs for the entire data set. Half of the fragments were from healthy subjects, the other half from patients with Parkinson's disease. Both parts of the data set (Healthy and Parkinson) contained examples of different transition types. We included examples from four different transition types identified by Rosenberg et al. [Rosenberg and van Hout2013] as regions with typically low inter-rater agreement: the last epoch of stage Wake before sleep onset, the first epoch of stage N2 after stage N1, the first epoch of stage REM after stage N2, and transitions between stages N2 and N3.

Procedure

The full data set was first scored independently by each sleep technologist, resulting in 5,760 individual scoring decisions, three for each of the 1,920 epochs. We then identified all epochs with disagreement among scorers and selected a subset of 30 epochs for in-person group deliberation. The selected disagreement epochs represented both disease states and all four transition types. All 30 epochs were discussed in person by the three scorers using a graphical scoring interface to facilitate detailed discussions about patterns present in the time series data. The experts participants were not explicitly required to reach unanimous consensus, and could instead choose to declare a case as *irresolvable*. We did not impose an explicit voting scheme or limit the amount of time available per discussion, but instead left the discussion dynamics open until all experts either agreed on one sleep stage or declared a case as irresolvable. Unanimous decisions were reached for all 30 epochs through a process of verbal argumentation and re-interpretation of the patterns shown in the biosignal data The *irresolvable* option was never used. Discussions were recorded (screen capture and audio), transcribed and qualitatively coded for the different sources of disagreement.

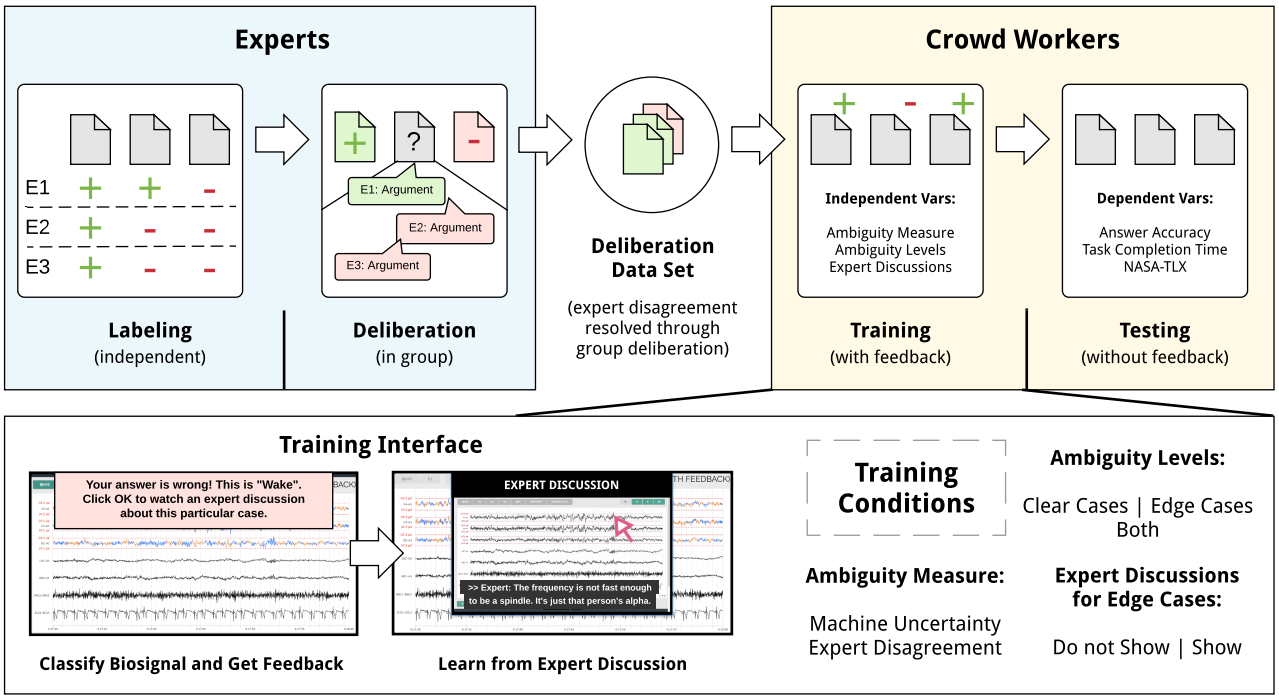

Figure 4: Application scenario of using expert discussions for improving example-based training for non-expert crowdworkers.

	Tech A	Tech B	Tech C	Majority	# Obs.
Tech B	0.71	—	—	—	
Tech C	0.71	0.68	—	—	N=1920
Majority	0.87	0.83	0.84	—	
Deliberation	0.63	0.50	0.02	**0.54**	N=30

Table 1: Pairwise agreement between all sleep technologists (Tech A, Tech B, Tech C), as well as the group labels as determined by majority vote and the deliberation process. Agreement is measured by Cohen's kappa.

Inter-rater Disagreement

We measured pairwise agreement between all scorers (Tech A, Tech B, Tech C), as well as the group labels as determined by majority vote (Majority) and the deliberation process (Deliberation). Agreement was measured by Cohen's kappa. Table 1 summarizes all agreement results. Pairwise agreement among scorers was moderate, ranging between 0.68 and 0.71 (N=1920). Agreement between individual scorers and the majority vote was high, between 0.84 and 0.87 (N=1920). For the epochs discussed in person, we measured pairwise agreement between the deliberation decision and individual scorers' decisions. Two of the three scorers showed weak agreement with deliberation outcomes (Cohen's kappa of 0.63 and 0.50, N=30), while the third scorer showed no systematic agreement with the deliberation outcomes (Cohen's kappa of 0.02, N=30). Agreement between the majority vote and deliberation decisions was low (Cohen's kappa of 0.54, N=30).

Sources of Disagreement

Initial qualititative coding of the expert discussions for 17 cases from two major transition types revealed a broad range of reasons why sleep technologists may disagree on the correct sleep stage label. Figures 2 and 3 compare the relative frequency of different sources of disagreement across two transition types (*Last W before sleep* and *First N2 after N1*) and across two disease states (*Healthy* and *Parkinson*) respectively. Overall, we identified two sources of disagreement which occurred with the highest frequency in both transition types and disease states. These were (a) the presence of multiple stages in one epoch causing disagreement about which stage was the dominant one, and (b) different configurations of the graphical scoring interface in terms of amplitude scaling causing divergent interpretations of visual patterns in the signal.

While these two sources of disagreement could persist on individual cases without the sequential context, we identified two other sources of disagreement that explicitly depend on the sequential nature of the labeling task and underlying data:

- **Number of scoring passes**:
  for 3 out of 30 adjudicated cases, experts explicitly mentioned that their scoring decision depended on the number of passes they had taken on a particular recording. In other words, experts indicated that their interpretation of biosignals is often updated once certain patient-specific patterns are observed towards the end of the recording. A subsequent re-interpretation (i.e., second scoring pass) would then allow experts to take into account observations they have made in the other parts of the data sequence in one of

the earlier scoring passes. Disagreement could therefore arise if one expert had only performed one initial pass whereas other experts may have performed two or more passes.

- **Cascade from previous disagreement**:
  3 out of 30 adjudicated cases could be resolved automatically once the disagreement on one of the close-by preceding cases had been resolved. This dynamic was observed since evidence for specific stages of sleep may sometimes be observed only at the transition point from one sleep stage to another. Consequently, disagreement may arise at a "critical" transition point and persist over multiple steps in the sequence. Once the disagreement at the transition point is resolved, the resolution can cascade to the subsequent steps until the next transition point.

These two sources of disagreement co-occurred once, meaning that 5 out of 30 adjudicated cases (17%) were associated with sources of disagreement that depend on the sequential nature of the labeling task and underlying data.

Discussion

In this work, we provided an initial investigation of expert disagreement in the context of sequential labeling tasks, studying the effectiveness of in-person adjudication for resolving disagreement and for analyzing information about the original source of disagreement.

Our results suggest that majority vote is not necessarily a good proxy for group deliberation decisions in sleep staging. This finding provides some confidence in the usefulness of expert discussions for the purpose of resolving disagreement cases. Beyond that, our qualitative analysis of expert discussion dialogues uncovered a diverse set of different reasons why domain experts disagree in the context of sleep stage classification, most of which go beyond the notion of mere input mistakes.

Perhaps most importantly, we identified two sources of disagreement with a clear connection to the sequential nature of the labeling task and underlying data. This observation provides some support for our hypothesis that the reciprocity of data and context in sequential labeling may lead to unique forms of expert disagreement that are characteristic for sequential labeling tasks, where the interpretation of one case affects the interpretation of subsequent or previous cases. One exciting avenue for future research is the problem of whether it is possible to detect the "critical" tasks that might set up a cascade of disagreement and potentially incorrect labels. Successful detection of such "critical" tasks would allow for a more cost-effective use of expert resources by focusing disambiguation procedures on those cases and saving expert resources on other cases that may be automatically resolved as a consequence.

Picking up on Penzel et al.'s comment on the nature of "truth" in sleep staging [Penzel, Zhang, and Fietze2013], some of the inherent difficulty may arise because there is some degree of both temporal and spatial continuity at transitions between states. In other words, despite the fact that any single neuron or cortical circuit may be thought of as existing in one state or another at any given moment, it is possible for local assemblies of neurons to take some time to transition from one state to another, and also that distant assemblies of neurons in different parts of the brain can exist in different states at the same time. These transitions may take minutes [Wright Jr, Badia, and Wauquier1995, Saper et al.2010] which encompasses several 30-second epochs. Thus, we hypothesize that some of the ambiguity stems from the need to force transitional states into one sleep stage category or another.

We posit that expert disagreement in complex tasks can be used as a signal to identify ambiguous edge cases, and as a driver for eliciting conclusive expert discussions to disambiguate such edge cases. For future work, we propose the idea that example-based training procedures for non-expert crowdworkers may benefit from the presentation of edge cases and their associated expert discussions. While expert disagreement may be one signal for the identification of edge cases, other techniques for the automatic selection of edge case examples, e.g., based on measures of machine uncertainty, have been proposed in prior work [Kim, Park, and Lee2016]. We believe expert disagreement and the associated expert discussions open up interesting opportunities for optimizing example-based training procedures for human learners, e.g., to improve disambiguation skills and depth of understanding.

Figure 4 illustrates a high-level overview of some of these future directions. In summary, we hope to conduct future research on augmenting example-based training procedures for non-expert crowdworkers using edge-cases and their associated expert discussions to help human learners develop more accurate classification strategies for expert-level exhibiting a certain amount of ambiguity tasks.

Another promising avenue for future work will be to explore the minimum "bandwidth" and effective protocols of communication between experts needed to result in successful disambiguation in the context of sequential labeling settings like the one presented in this work. Comparisons may include different styles of expert communication ranging from online text-based asynchronous approaches, to in-person verbal real-time communication.

Conclusion

In this work, we reported results from a case study on expert disagreement in sequential labeling tasks where the interpretation of one case can affect the interpretation of subsequent or previous cases. Three board-certified sleep technologists scored 1,920 cases in a sequential 5-class labeling task. Out of all disagreement cases, 30 cases were discussed and resolved through face-to-face adjudication. We identified various sources of disagreement that are specific to the sequential nature of the underlying data and labeling procedure. Our work concluded with a discussion of promising application scenarios of expert discussions for the training of non-expert crowdworkers that we hope to explore in future work.

References

Beatty, J., and Moore, A. 2010. Should We Aim for Consensus? *Episteme* 7(3):198214.

Chang, J. C.; Amershi, S.; and Kamar, E. 2017. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In *Proceedings of the 2017 SIGCHI Conference on Human Factors in Computing Systems - CHI '17*, 2334–2346. ACM.

Dumitrache, A.; Aroyo, L.; and Welty, C. 2016. Crowdsourcing Ground Truth for Medical Relation Extraction. *Transactions on Interactive Intelligent Systems*.

Garbayo, L. 2014. Epistemic Considerations on Expert Disagreement, Normative Justification, and Inconsistency Regarding Multi-criteria Decision Making. *Constraint Programming and Decision Making* 539:35–45.

Guan, M.; Gulshan, V.; Dai, A.; and Hinton, G. 2018. Who said what: Modeling individual labelers improves classification. In *AAAI Conference on Artificial Intelligence*.

Gurari, D., and Grauman, K. 2017. CrowdVerge: Predicting If People Will Agree on the Answer to a Visual Question. In *Proceedings of the 2017 SIGCHI Conference on Human Factors in Computing Systems - CHI '17*, 3511–3522. ACM.

Janis, I. L. 1972. *Victims of groupthink: a psychological study of foreign-policy decisions and fiascoes.* Oxford, England: Houghton Mifflin.

Kairam, S., and Heer, J. 2016. Parting Crowds: Characterizing Divergent Interpretations in Crowdsourced Annotation Tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16*, 1635–1646. New York, New York, USA: ACM Press.

Kiesler, S., and Sproull, L. 1992. Group decision making and communication technology. *Organizational behavior and human decision processes* 52(1):96–123.

Kim, J.; Park, J.; and Lee, U. 2016. EcoMeal: A Smart Tray for Promoting Healthy Dietary Habits. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '16*, 2165–2170. New York, New York, USA: ACM Press.

Krause, J.; Gulshan, V.; Rahimy, E.; Karth, P.; Widner, K.; Corrado, G. S.; Peng, L.; and Webster, D. R. 2018. Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy. *Ophthalmology*.

Liu, W.; Xiao, S.; Browne, J. T.; Yang, M.; and Dow, S. P. 2018. ConsensUs: Supporting Multi-Criteria Group Decisions by Visualizing Points of Disagreement. *ACM Transactions on Social Computing* 1(1):4:1–4:26.

Mitry, D.; Zutis, K.; Dhillon, B.; Peto, T.; Hayat, S.; Khaw, K.-T.; Morgan, J. E.; Moncur, W.; Trucco, E.; and Foster, P. J. 2016. The Accuracy and Reliability of Crowdsource Annotations of Digital Retinal Images. *Translational Vision Science & Technology* 5(5):6.

Mumpower, J. L., and Stewart, T. R. 1996. Expert Judgement and Expert Disagreement. *Thinking and Reasoning* 2(23):191–21.

Navajas, J.; Niella, T.; Garbulsky, G.; Bahrami, B.; and Sigman, M. 2018. Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour*.

Nemeth, C. 1977. Interactions Between Jurors as a Function of Majority vs. Unanimity Decision Rules. *Journal of Applied Social Psychology* 7(1):38–56.

Penzel, T.; Zhang, X.; and Fietze, I. 2013. Inter-scorer reliability between sleep centers can teach us what to improve in the scoring rules. *Journal of Clinical Sleep Medicine* 9(1):81–87.

Rajpurkar, P.; Hannun, A. Y.; Haghpanahi, M.; Bourn, C.; and Ng, A. Y. 2017. Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks.

Rosenberg, R. S., and van Hout, S. 2013. The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring. *Journal of Clinical Sleep Medicine*.

Saper, C. B.; Fuller, P. M.; Pedersen, N. P.; Lu, J.; and Scammell, T. E. 2010. Sleep State Switching. *Neuron* 68(6):1023–1042.

Solomon, M. 2006. Groupthink versus the wisdom of crowds: The social epistemology of deliberation and dissent. *The Southern Journal of Philosophy* 44(S1):28–42.

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

Warby, S. C.; Wendt, S. L.; Welinder, P.; Munk, E. G. S.; Carrillo, O.; Sorensen, H. B. D.; Jennum, P.; Peppard, P. E.; Perona, P.; and Mignot, E. 2014. Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods. *Nature Methods* 11(4):385–392.

Wright Jr, K. P.; Badia, P.; and Wauquier, A. 1995. Topographical and temporal patterns of brain activity during the transition from wakefulness to sleep. *Sleep* 18(10):880–889.