

PNNL-33400

# Method for Generating Expert Derived Confidence Scores

September 2022

Corey K Fallon  
Tim Yin  
Alexander A Anderson

## DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Battelle Memorial Institute, nor any of their employees, makes **any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights.** Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST NATIONAL LABORATORY  
*operated by*  
BATTELLE  
*for the*  
UNITED STATES DEPARTMENT OF ENERGY  
*under Contract DE-AC05-76RL01830*

Printed in the United States of America

Available to DOE and DOE contractors from  
the Office of Scientific and Technical Information,  
P.O. Box 62, Oak Ridge, TN 37831-0062

[www.osti.gov](http://www.osti.gov)

ph: (865) 576-8401

fox: (865) 576-5728

email: [reports@osti.gov](mailto:reports@osti.gov)

Available to the public from the National Technical Information Service  
5301 Shawnee Rd., Alexandria, VA 22312

ph: (800) 553-NTIS (6847)

or (703) 605-6000

email: [info@ntis.gov](mailto:info@ntis.gov)

Online ordering: <http://www.ntis.gov>

# **Method for Generating Expert Derived Confidence Scores**

September 2022

Corey K Fallon  
Tim Yin  
Alexander A Anderson

Prepared for  
the U.S. Department of Energy  
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory  
Richland, Washington 99354

## Acknowledgments

This research was supported by the Mathematics for Artificial Reasoning in Science (MARS) Initiative, under the Laboratory Directed Research and Development (LDRD) Program at Pacific Northwest National Laboratory (PNNL). PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL01830.

## Contents

Acknowledgments.....	ii
1.0 Introduction.....	1
2.0 Machine Learning Classifier .....	2
3.0 Data .....	3
3.1 Data Overview .....	3
3.2 Data Selection .....	3
4.0 Generating Human Derived Confidence Scores and Qualitative Descriptions .....	5
4.1 Subject Matter Expert .....	5
4.2 Learning Phase.....	5
4.3 Scoring Phase .....	9
5.0 Evaluation .....	12
6.0 Discussion.....	14
6.1 Modifications to the Learning Phase .....	14
6.2 Insights from the SME's Learning Process.....	14
6.3 Combining SME and ML Confidence .....	15
6.4 Plans for FY23 .....	15
7.0 References.....	16
Appendix A – Event 1 of EDC Scoring Questionnaire .....	1

## Introduction

Machine Learning (ML) classifiers may help support power grid operators by identifying important events on the grid. However, even the most higher performing ML will occasionally misclassify events (Zhang, Liao & Bellamy, 2020). For example, imagery data with small distortions due to variability in format and image compression may cause a classifier to misclassify these images (Zheng, Song, Leung & Goodfellow, 2016). Power grid operators understand the classifier's potential for error, but they may be less clear about when these errors are likely to occur and conversely when the tool's classification can be trusted with an accurate recommendation.

To provide some guidance, developers can display a confidence score associated with each classification to communicate the tool's confidence in its decision. Although confidence scores may serve as a useful guide for some classifications, for other events, the tool's confidence may not be a good indicator of its own performance. For some classifiers, confidence scores can 'deviate substantially from their true outcome probabilities' (Zhang et al., 2020, p. 298). In addition, these confidence scores suffer from the same lack of transparency as the underlying classifications. The operator does not understand the underlying analytical processes that led to both the classifications and associated confidence scores. This lack of understanding may lead to automation bias if the operator replaces their own assessment of the situation with an automated system's incorrect recommendation (Mosier, Skitka, Burdick & Heers, 1996). Wickens, Clegg, Vieane and Sebok (2015) found that automation bias experienced during a process control simulation degraded participant's ability to diagnose faults in the automation and increased operator workload. On the other hand, lack of understanding may lead to algorithm aversion which is a tendency to avoid recommendations offered by algorithms in favor of human judgment despite the potential performance benefits of relying on such technology (Dietvorst, Simmons & Massey, 2015).

With our FY22 funds, we executed a pilot demonstration of our methodology for developing an expert derived confidence score and associated qualitative descriptions to accompany an ML classifier's classification decision. The score and qualitative description are intended to help operators calibrate their trust and reliance on the tool (i.e., know when to accept or reject an ML classifier's classification decisions). This methodology provides a way for operators to receive guidance from a domain expert who also has a clear mental model of the error boundaries of the ML classifier.

The pilot demonstration involved a Learning Phase where our domain expert learned the error boundaries of an ML model for classifying power systems data. This phase was followed by a Scoring Phase. In this Phase our expert rated their confidence in the model's ability to accurately classify a subset of events in the training data. Next, we assessed the predictive power of the expert derived confidence scores compared to the ML's traditional uncertainty quantification score.

## Machine Learning Classifier

The machine learning algorithm chosen for the pilot demonstration was the support-vector machine (SVM). SVM is a conventional machine learning algorithm that is successful in many applications (Cristianini 1992). A SVM forms hyperplanes to separate the training data into different categories and to predict new data according to such separations in the feature space. SVM is usually good at handling high-dimensional data when using radial basis function kernel, which is the kernel of our choice. During the model training process, repeated cross validation has been applied to fine-tune the hyper-parameters of SVM to avoid over-fitting.

The uncertainty quantification was provided by the softmax equation in the caret package in R (Kuhn 2008). The softmax equation was selected based on previous research. In past work the softmax equation generated scores that most closely aligned with the current model's performance when compared to other equations for generating uncertainty.

## Data

### Data Overview

The initial data set in our study included the real-world phasor measurement unit (PMU) data of power system events from the Eastern Interconnection in the US. The events are classified by a power system engineering expert into two categories, namely generator tripping events and other frequency events as ground truth for the following analysis, since the generator tripping events are of much more interest than the other types of events within the data set. The time range of this data set is from July to December 2018. The data set contains 42 generator tripping events and 165 other frequency events. Each event is captured by a one-minute window of PMU data of two types of channels, the frequency channel and the phase angle difference (PAD) channel. This data set is used for training the machine learning classifier.

PMU data are time series data and must be distilled into several features that the model can be trained on. The features used are the 16 signatures developed in work by Amidan (2005), as well as six frequency multi-channel features and six PAD multi-channel features. The detail descriptions of all the features can be found in (Follum 2020).

### Data Selection

We selected a subset of 124 events from the training data. The events consisted of five types:

- False Positives (FP) – The ML classified these events as Generator Trips but they were actually Other Frequency events.
- Misses – The ML classified these events as Other Frequency events but they were actually Generator Trip events.
- Near Neighbor FP – Each FP had two near neighbor events (one Generator Trip and one Other Frequency event). These events were closest in latent space to the FP.
- Near Neighbor Miss – Each Miss had two near neighbor events (one Generator Trip and one Other Frequency event). These events were closest in latent space to the Miss.
- Exemplar – These events included only correctly classified events that were not near neighbors. Exemplars were also selected based on their similarity to each other. Events that were dissimilar from each other based on their distance in latent space and visual inspection were selected.

A primary goal of the method was for the SME to learn the performance boundaries of the ML classifier. The Learning Phase was designed to encourage learning by giving



SMEs the environment and structure to study the various types of misclassified events and contrasting them with correctly classified events. The research team needed to provide the SME with a representative sample of misclassified events and this focus on misclassified events became the primary factor determining the total number of events selected for the learning and scoring phases.

Our ML classifier was 85% reliable which provided us with far fewer misclassified events to choose from when compared to correctly classified events. Of the 28 FPs in the training data, we selected 20 FP events. Although 20 was not the total number of FP events, it represented the large (71%) representative majority of FP events in the data, and we believe reflected the diversity of FPs in the training data. The ML only missed 8 generator trip events. This small number made it manageable for the SME to work with the entire population of missed events. Therefore, we selected all 8 events from the training data.

Each misclassified event had a Near Neighbor FP and a Near Neighbor Miss selected for our method. Near Neighbor events were identified using the dynamic time warping method of calculating similarities between time series data (Keogh, 2005). This method can account for the different event starting times within the one-minute data window. By accounting for these differences, dynamic time warping provides a more accurate similarity measurement considering the various events within the data set compared to more traditional similarity measurements such as L1 or L2 norm. The Near Neighbors allowed the researcher to compare and contrast each misclassified event to similar events that were correctly classified.

All correctly classified events that were not identified as Near Neighbors were eligible for selection as Exemplar events. The research team selected 10 Exemplar events for each class (Generator Trip Events, Other Frequency Events). To guide selection, we calculated the similarity of the elidable Exemplar events using dynamic time warping and selected events with the highest degree of dissimilarity based on both dynamic time warping results and visual inspection. Dissimilarity was prioritized to provide the SME with a visually diverse sample of Exemplar events. All event types were divided equally into two groups and assigned to either the Learning or Scoring phases (see Table 1).

**Table 1.** Number of Events for the Learning and Scoring Phases by Class and Event Type

Type	Learning Phase		Scoring Phase		Total
	Gen	Other	Gen	Other	
FP		10		10	20
Miss	4		4		8
Near Neighbor FP	10	10	10	10	40
Near Neighbor Miss	4	4	4	4	16
Exemplar	10	10	10	10	40
Total	28	34	28	34	124

Note. *Gen* refers to the Generator Trip Event Class, *Other* refers to the Other Frequency Event Class

## Generating Human Derived Confidence Scores and Qualitative Descriptions

### Subject Matter Expert

One participant was our Subject Matter Expert (SME). This participant has a background in power systems analysis and real-time simulation of large-scale transmission and distribution networks using electromagnetic and long-term dynamic models. The participant was also previously a technical trainer for transmission system operators, balancing authorities, and reliability coordinators working in transmission control rooms worldwide.

### Learning Phase

The research team used the software platform MURAL to conduct the learning phase. MURAL is a digital white board that can be accessed by remote collaborators for synchronous and asynchronous collaboration and allows teammates to import and manipulate images. In the Learning Phase all five types of events were visually displayed on MURAL for a total of 62 events. Each event was labeled on MURAL according to its Class and Type and was assigned a unique ID number within its Type. The ID numbers of the misclassified events matched their near neighbor event ID numbers so that the SME could track these associations on MURAL. In the example below we see an FP event in the center of two Near Neighbor events (see Figure 1). The numeric ID '3' distinguishes this FP event from the other FP events in the Learning Phase. To the left and right of FP\_3 are its near neighbors which were also assigned the label '3' to indicate that they are near neighbors of FP\_3. Both start with 'FPNN' which stands for False Positive Near Neighbor. The two Near Neighbor events differ by Class. The Near Neighbor on the left is a generator trip event and is assigned the label 'Gen'. The Near Neighbor event on the right is an Other Frequency event and is assigned the label 'Other'. FP\_3 does not need an explicit class designation because all FPs belong to the Other Frequency event class.

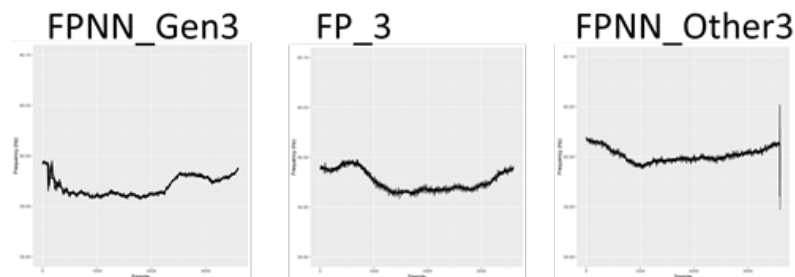


Figure 1. Example Labeled Events as they were displayed on MURAL during the Learning Phase.

Using both the visual profile of the time series data and their knowledge of model performance for each event (i.e., correctly classified or misclassified), the SME organized the events into categories that were meaningful to them and provided a label for each category. This exercise allowed the SME to see subtle distinctions between events that are typically misclassified and those that are similar, but correctly classified by the model (i.e., Near Neighbor events). The sorting task was intended to stimulate critical thinking and learning of the performance boundaries of the model by providing visual clues into why the model might be misclassifying particular events. By developing a rich mental model of the classifier the goal was for the SME to successfully predict when misclassification is likely to occur and provide an informed explanation for their prediction. Card sort tasks are commonly employed in human factors psychology as a technique for understanding humans' mental models (Smith-Jentsch, Campbell, Milanovich & Reynolds; 2001; Wright, et al. 2020).

*Step 1.* The sorting task was largely self-guided because the researchers did not want to constrain the SMEs mental model development. The SME in our pilot demonstration of the method choose to divide the sorting task into two steps. In the first step the SME classified the events using their domain knowledge independent of how the ML classified the events. The purpose of this step was to become familiar with the characteristics of generator trip and other frequency events unique to the particular data set. The SME used their knowledge of three specific domain characteristics to guide grouping of generator trip events. The domain characteristics include the following:

- The magnitude of the frequency decrease corresponds to the size of the unit that tripped. For example, loss of a large unit results in a large frequency drop and loss of a small unit causes smaller decreases in frequency.
- The Rate of Change of Frequency (ROCOF) during the event corresponds to the total spinning inertia of the system. Reduced inertia due to either high renewable penetration or fewer online generators during light load conditions results in faster decreases in frequency.
- The rate at which the system recovers after the event. This characteristic has no impact on the SME's determination of whether an event is generator trip. It is just an effect of how quickly turbine governors respond to the disturbance.

Knowledge of these characteristics helped the SME sort the Generator Trip Events into one of four categories:

- a. Events with "textbook" generator tripping characteristics including a large frequency drop and some oscillations due to control system overshoot.
- b. Events with very fast frequency drops (high ROCOF)
- c. Events with small frequency decrease associated with tripping of a small generator
- d. Events with quick system recovery after the event

Example events from each category can be seen in Figure 2.

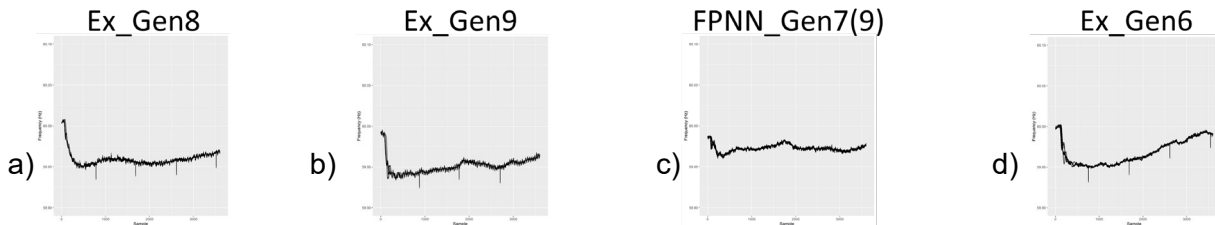


Figure 2. Example events from each Generator Trip class category created in step 1 of the learning phase

In the first step, Other Frequency Events were also organized into groups based on distinct features that differentiated these events from Generator Trip Events. These features include the following

Lack of change in frequency

- Linear decreases in frequency
- Increases in frequency
- Decreases in frequency that are too slow (low ROCOF)
- Decreases in frequency that have the wrong shape
- Harmonic distortion (possibly caused by inverter-based renewable generation)

Knowledge of these characteristics helped the SME sort the other frequency events into one of four categories:

- a. Events with no change in frequency
- b. Events that increase in frequency
- c. Events that have both incorrect concavity and harmonic distortion
- d. Events with a very slow decrease in frequency (very low ROCOF)

Examples of these characteristics are shown in Figure 3.

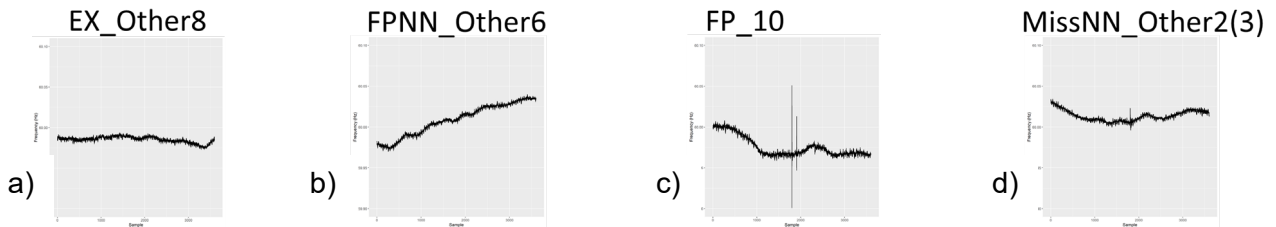


Figure 3. Example events from each other frequency event class category created in step 1 of the learning phase

*Step 2.* In the second step the SME set out to identify characteristics that were associated with the ML’s classification decisions (i.e., correct or incorrectly classified events). This step focused primarily on those event characteristics that appeared to lead to misclassifications (i.e., Misses or FPs). During a post learning phase debrief the SME stated there was no apparent pattern to the correctly classified images. Therefore, he developed an opinion that the classifier would identify an event correctly if the event did not contain any features associated with misclassifications.

The SME observed no characteristic pattern among Miss events other than the similarity of the last 600 data points to the other frequency event nearest neighbor. However, the SME did not believe this similarity provided insights into the classification decisions. FP events were grouped by a set of visual characteristics (some of which were informed by the SMEs domain expertise). Characteristics that appeared to be associated with false positives included the following:

- “Bumps” in the middle of the data series
- Harmonic distortion
- Large spikes in frequency

Events that had these characteristics were grouped together on the MURAL board. However, it is important to note that these groupings contained both correctly classified events as well as FPs. The groupings associated with harmonic distortion and large spikes can be seen in Figure 4.

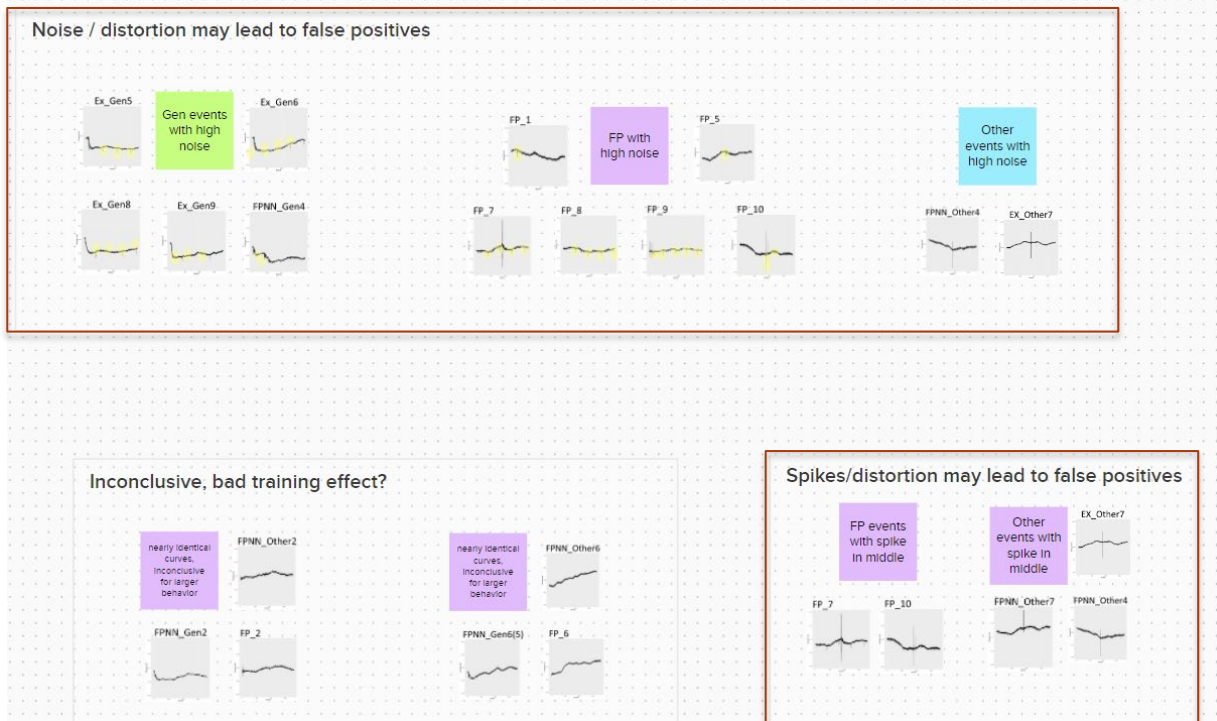


Figure 4. Portion of the MURAL board that shows the SME’s categorization of FP events into a Noise/Distortion group and a Spikes group (both highlighted in red).

## Scoring Phase

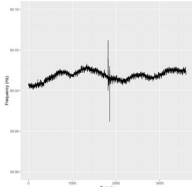
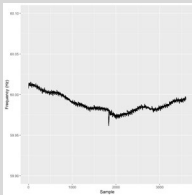
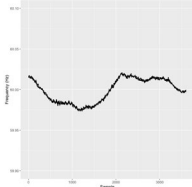
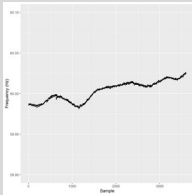
**Questionnaire.** The research team developed a questionnaire to capture SME confidence scores and explanations for each event in the Scoring Phase data set (see Table 1). SME scores and descriptions were recorded on a questionnaire posted to the project’s Microsoft Teams page where it was accessed and completed by the SME. Each page of the questionnaire displayed a different event from the dataset and, unlike events in the Learning Phase, these events were *not* labeled to indicate their class or type. For each event, participants were asked to identify the event as either a Generator Trip or Other Frequency Event. Participants were also asked to provide a likelihood rating from 0 to 1 that the ML will correctly classify the event. This likelihood rating served as the participant’s confidence score for that event. For the final item the participant was asked to provide the reason for their likelihood score in one to two sentences. The presentation order of events in the questionnaire was randomized. See Appendix A for a sample event from the questionnaire.

**SME Scoring Strategy.** The SME spent 3 hours completing the questionnaire. In the Learning Phase debrief the SME reported that the number of events for each type in this Phase heavily influenced his perception of the overall reliability of the ML classifier. In the Learning Phase the ML correctly classified 24 of the 28 Generator Trip events (i.e., 86%) and correctly classified 24 of 34 Other Frequency events (i.e., 71%). The SME used these percentages to shape his confidence in the ML. He reported that overall he was

90% confident that Generator Trip events would be correctly classified and 70% confident that Other Frequency events would be correctly classified. In this instance the SMEs perception of overall reliability was fairly close to the actual 85% reliability. Future work should continue to match the proportion of misclassified to correctly classified events in the Learning Phase to overall classifier reliability since this proportion may influence perception of ML reliability.

The SME started with these initial estimates of the classifier's overall reliability as his baseline confidence for scoring each event. He then evaluated the event's characteristics in search of possible indicators of ML misclassification identified in the Learning Phase. The presence or absence of these indicators caused the SME to adjust his likelihood rating of the event from his baseline estimate. The SME kept the MURAL board open on a second external monitor as a reference when evaluating each event in the Scoring Phase. For some events, part of the evaluation included not just whether the event had a characteristic associated with a misclassification, but also the ratio of correct to incorrectly classified events with that particular characteristic. This quantitative approach can be seen in several of the SME's reasons provided for various confidence scores (see Table 2).

**Table 2.** Examples of when the ratio of correct to incorrectly classified events for a particular characteristic influenced confidence scores.

Event Image	Event No.	Confidence Score	Justification Provided
	16	0.6	“The classifier is only 3/5 for events with a large spike / distortion in the middle with false positives on 2/5 events in the training set.”
	17	0.6	“In the training set, the classifier was only 7/12 correct for events with harmonic distortion. The overall linear increase/decrease is clearly an ‘other’ type event, but similar events have led to false positives.”
	41	0.7	“Frequency decrease is concave-down but otherwise looks like a gen trip, which may result in a false positive. The curve does not contain much noise or distortions, but the classifier is 12/17 for events with a bump in the middle.”
	42	0.8	“Frequency increased over the course of the event. The classifier identified 4/5 of these events correctly.”



## Evaluation

The SME's likelihood ratings provided in the Scoring Phase were treated as his confidence scores. We computed a series of logistic regressions to compare the predictive power of the SME's confidence scores to an ML derived uncertainty quantification score (i.e., softmax equation). In these analyses ML classification performance (i.e., correctly classified or incorrectly classified) was the dependent variable. The first regression included the ML derived uncertainty quantification score as the predictor. Results of the model show the ML derived uncertainty quantification score significantly predicted ML classification performance  $p=.002$  (see Table 3).

**Table 3.** Results for ML Derived Uncertainty Quantification Predictor

Predictor	Estimate	Std. Error	z	p
Intercept	-6.91	2.50	-2.77	0.006
ML	19.29	6.29	3.07	0.002

Next, the researchers computed a regression with our SME confidence scores as the predictor of ML classification performance. Results of the model show the SME confidence scores did not significantly predicted ML classification performance  $p=.29$  (see Table 4).

**Table 4.** Results for SME Confidence Scores Predictor

Predictor	Estimate	Std. Error	z	p
Intercept	3.77	2.48	1.52	0.129
SME	-3.156	3.019	-1.045	0.29

The researchers wanted to explore if combining the ML Uncertainty Score with SME confidence improved the predictive power of the Uncertainty Score. Although the SME confidence was not a significant predictor on its own perhaps it could be used to improve the ML Uncertainty Score. To explore this possibility the researchers computed the midpoint score between both predictors and included this new combined score as a predictor of ML classification performance. Results of the model show taking the average of both scores is a stronger predictor of ML classification performance ( $p<.001$ ) when compared to the ML Derived Uncertainty Quantification predictor alone ( $p=.002$ ) (see Table 5).

**Table 5.** Results for Combined Score Predictor

<b>Predictor</b>	<b>Estimate</b>	<b>Std. Error</b>	<b>z</b>	<b>p</b>
Intercept	-8.833	2.75	-3.21	0.001
ML+SME	16.245	4.568	3.556	<0.001

We were also interested in analyzing the confidence scores of a novice. Analyzing a novice's confidence scores may provide insight into the importance of prior domain expertise for providing accurate confidence scores. The PI of the project served this role as he did not develop the machine learning classifier and does not have expertise in power grid systems. Similar to the SME, the PI completed both the Learning and Scoring phases of the method and his results were analyzed. Table 6 shows that, like the SME, the PI's scores were not a significant predictor of ML classification performance,  $p=.55$ . However, unlike the SME scores, combining the novice scores with the ML derived uncertainty quantification scores did not improve prediction,  $p=.002$  (see Table 7).

**Table 6.** Results for Novice Predictor

<b>Predictor</b>	<b>Estimate</b>	<b>Std. Error</b>	<b>z</b>	<b>p</b>
Intercept	0.790	0.785	1.006	0.314
ML+N	0.669	1.118	0.598	0.550

**Table 7.** Results for Novice Predictor

<b>Predictor</b>	<b>Estimate</b>	<b>Std. Error</b>	<b>z</b>	<b>p</b>
Intercept	-2.570	1.222	-2.103	0.036
ML+N	6.923	2.286	3.028	0.002

## Discussion

Neither the SME nor the Novice scores significantly predicted ML classifier performance. These findings suggest that the Learning Phase was not successful in teaching participants the performance boundaries of the ML classifier. Interestingly, the combined ML Uncertainty Quantification scores and SME confidence did significantly improve predictions of ML performance and was actually a stronger predictor than the Uncertainty Quantification score alone. The results of this combined score suggests that there was still some value to incorporating expert ratings. In addition, although neither participant generated confidence values that significantly predicted ML performance, the SME's scores had a smaller 'p' value suggesting the SME's confidence was better calibrated to actual ML performance.

### Modifications to the Learning Phase

The findings suggest a need to improve the Learning Phase. The current method relied on a card sorting approach that allowed the participant to organize various events into meaningful groups. This sorting task was designed to help the participant build a detailed mental model of the ML classifier's performance. Although the SME was able to construct a detailed mental model, the findings suggest that this sorting task was not sufficient for learning the performance boundaries of the ML.

We plan to revise this phase of the method to help the participant build a stronger association between the relevant event characteristics and the classifier decisions they influence. The research team plans to design an additional component of the Learning Phase focused on repeated performance assessments. After the initial sorting task, participants will have their knowledge tested by predicting the ML classifier's decisions on a subset of unlabeled learning phase data. The participant will receive feedback on their performance as learning researchers have long demonstrated the positive impact (both informational and motivational) of performance feedback on learning (Matthews et al., 2000; Wang, Zhang, He, 2022). Participants will continue regular assessments of their ability to accurately predict ML decisions until they have achieved a minimum standard of performance. Once this minimum standard is met participants will move to the Scoring Phase. Learning the performance boundaries of the ML is at least partly an experiential process that relies on associative learning (Evans & Stanovich, 2013). We believe repeated exposure to events and their associated ML classification decisions during performance assessment should help strengthen the associations needed to calibrate participant confidence scores with actual ML performance.

### Insights from the SME's Learning Process

The SME provided valuable insights into his process during the post Learning Phase debrief. It was evident from this debrief as well as from his responses in the Scoring Phase that he gleaned quantitative insights from the Learning Phase. For example, he used the ratio of Misses to correctly classified Generator Trip events and the ratio of FPs to correctly classified Other Frequency events presented in the Learning Phase to form an assessment of the classifier's baseline reliability. The SME was not instructed to use

the frequencies of correctly classified and misclassified events in this way. It is important to recognize that future participants may glean similar quantitative insights from the Learning Phase. To prevent incorrect assumptions about classifier reliability based on this quantitative approach, we must work to ensure the ratio of correctly classified and incorrectly classified events in the Learning Phase approximates model reliability.

### **Combining SME and ML Confidence**

Findings also suggest that there may be value in merging the human subjective evaluation of the ML's confidence with the ML's own uncertainty quantification. In our evaluation the combined score was the most significant predictor of ML performance. The research team plans to investigate these results to better understand how incorporating the non-significant SME scores improved prediction. In addition, the team would like to explore other approaches for computing a combined score such as a weighted average.

### **Plans for FY23**

Based on a projected budget of 200K we plan to develop a more complete methodology. These tasks include the following:

- Increase existing model complexity (increase number of classes and/or classify events at sub-class level)
- Develop simple interfaces for completing Learning and Scoring Phase exercises
- Improve the Learning Phase by including a component to assess participant model prediction performance and provide performance feedback
- Develop a novel uncertainty quantification score
- Select and Label data for Learning and Scoring Exercises for multiple classes
- Include multiple SMEs in the Learning and Scoring Phases
- Explore a weighted average approach for combining SME scores with ML Uncertainty Quantification
- Generalize scored events to entire dataset
- Evaluation including Human Subjects Testing to assess the effect of the associated qualitative descriptions on user prediction performance

## References

- Amidan, Brett G. and Ferryman, Thomas A. 2005. "Atypical event and typical pattern detection within complex systems," *2005 IEEE Aerospace Conference*, 2005, pp. 3620-3631, doi: 10.1109/AERO.2005.1559667.
- Cristianini, N., Ricci, E. (2008). Support Vector Machines. In: Kao, MY. (eds) *Encyclopedia of Algorithms*. Springer, Boston, MA. [https://doi.org/10.1007/978-0-387-30162-4\\_415](https://doi.org/10.1007/978-0-387-30162-4_415)
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Overcoming Algorithm Aversion. *People will use algorithms if they can (even slightly) modify them*, Philadelphia.
- Evans, J. B., & Stanovich, K. E., (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8, 223-241.
- Follum, James D., Betzold, Nick J., Yin, Tianzhixi, and Buckheit, John. 2020. "Event Screening Methods for the Eastern Interconnection Situational Awareness and Monitoring System (ESAMS)". United States. <https://doi.org/10.2172/1846589>. <https://www.osti.gov/servlets/purl/1846589>.
- Kuhn, Max. 2008. "Building Predictive Models in R Using the Caret Package". *Journal of Statistical Software* 28 (5):1-26. <https://doi.org/10.18637/jss.v028.i05>.
- Keogh, E., Ratanamahatana, C. Exact indexing of dynamic time warping. *Knowl Inf Syst* 7, 358–386 (2005). <https://doi.org/10.1007/s10115-004-0154-9>
- Matthews, G., Davies, D. R., Stammers, R. B., & Westerman, S. J. (2000). *Human performance: Cognition, stress, and individual differences*. Psychology Press.
- Mosier, K. L., Skitka, L. J., Burdick, M. D., & Heers, S. T. (1996, October). Automation bias, accountability, and verification behaviors. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 40, No. 4, pp. 204-208). Sage CA: Los Angeles, CA: SAGE Publications.
- Smith-Jentsch, K. A., Campbell, G. E., Milanovich, D. M., & Reynolds, A. M. (2001). Measuring teamwork mental models to support training needs assessment, development, and evaluation: Two empirical studies. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, 22(2), 179-194.
- Wang, X., Zhang, L., and He, T. (2022). Learning Performance Prediction-Based Personalized Feedback in Online Learning via Machine Learning. *Sustainability*, 14 (13): 7654. <https://doi.org/10.3390/su14137654>

Wickens, C. D., Clegg, B. A., Vieane, A. Z., & Sebok, A. L. (2015). Complacency and automation bias in the use of imperfect automation. *Human factors*, 57(5), 728-739.

Wright, M. C., Radcliffe, S., Janzen, S., Edworthy, J., Reese, T. J., & Segall, N. (2020). Organizing audible alarm sounds in the hospital: a card-sorting study. *IEEE transactions on human-machine systems*, 50(6), 623-627.

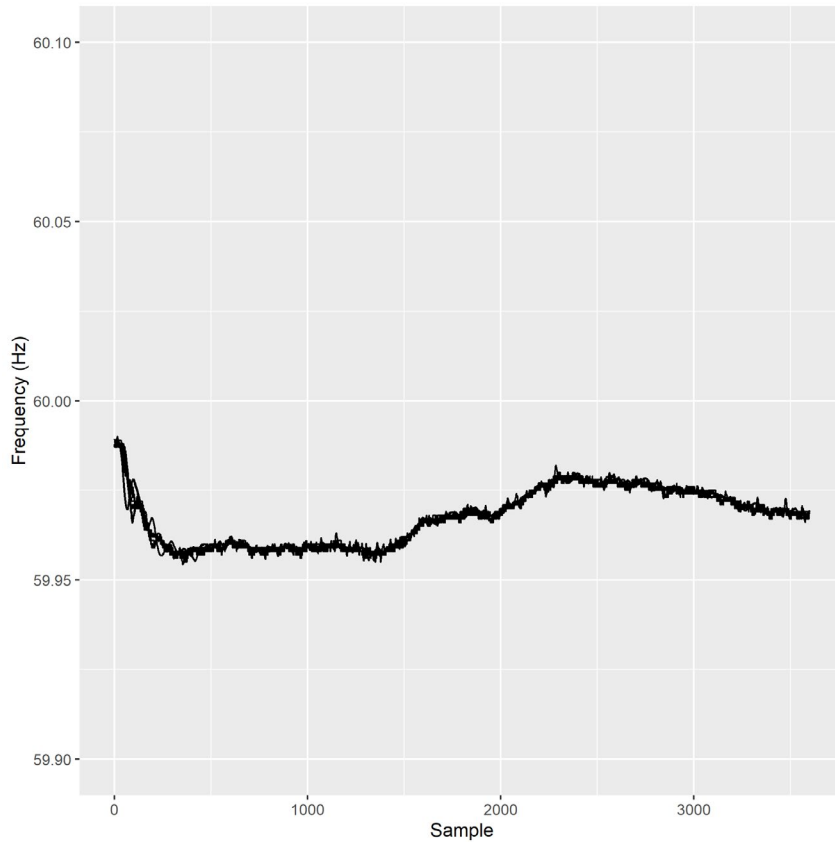
Zhang, Y., Liao, Q.V. & Bellamy, R.K. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-Assisted Decision Making. *In Conference on Fairness, Accountability and Transparency*, Barcelona, Spain, ACM.

Zheng, S., Song, Y., Leung, T., & Goodfellow, I. (2016). Improving the robustness of deep neural networks via stability training. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4480-4488).

## – Event 1 of EDC Scoring Questionnaire

### EDC Scoring Phase

1



Please Classify this event (circle the option below).

- Generator Trip
- Other Frequency Event

What is the likelihood the *Machine Learning Classifier* will correctly classify this event (please provide a number between 0 to 1)? \_\_\_\_\_

In one or two sentences please provide a reason for your likelihood score (i.e., your answer to question 2).

# **Pacific Northwest National Laboratory**

902 Battelle Boulevard  
P.O. Box 999  
Richland, WA 99354

1-888-375-PNNL (7665)

***[www.pnnl.gov](http://www.pnnl.gov)***