Chapter 11
Statistical Framework for Deep Learning Model Comparison and Evaluation

Samuel Myren, Nidhi Parikh, Rosalyn Rael, Garrison Flynn, Dave Higdon, and Emily Casleton

Abstract Deep learning for structural health monitoring (SHM) has enabled researchers to extract information from big data acquired from a sensor network or images to detect, identify, and characterize events of interest. However, choosing the best model from a set of candidates for a given problem remains challenging due to the need for a robust evaluation framework that considers model performance and the uncertainty in the employed data environments. We contribute such a framework for deep learning model evaluation and comparison that is rooted in statistical foundations. Given a performance metric of interest and a set of competing models or training approaches, our framework estimates the distribution of performance metrics and segregates uncertainty arising from the training data choice and the model fitting process (i.e. weight initialization and stochastic gradient descent). We demonstrate this framework by training a deep learning architecture using ground vibration data to detect events under a set of training conditions to quantify the effects of pre-training, fine-tuning, and out-of-distribution training approaches. Additionally, we evaluate these models across varying amounts of training data to mimic real world scenarios where data labeling is expensive/limited and the model is deployed in an environment slightly different from the training environment. In this demonstration, we exemplify how easy it can be to select the incorrect model if the models are only trained once and that model initialization can unexpectedly influence performance uncertainty. Our rigorous and generalizable framework which is supported through a real example provides value to the SHM researchers by enabling them to confidently compare various models and rank them according to their priorities.

Keywords Uncertainty Quantification · Foundation Models · Learning Efficiency · Artificial Intelligence · Signal Processing

Introduction

Equipped with technological advances in artificial intelligence (AI) and big data from advanced sensors, engineers and practitioners in structural health monitoring (SHM) are actively researching and incorporating traditional deep learning (DL) models to accomplish tasks within their workflows (Flah et al., 2021). In recent years, these advances have culminated with foundation models (FMs) (Bommasani et al., 2022; Harsuko & Alkhalifah, 2022), large models pre-trained on vast amounts of data that can be fine-tuned using smaller, labeled datasets to accomplish a variety of specific tasks. FMs take traditional DL a step further by promising to accomplish multiple tasks simultaneously, even potentially replacing the workflow altogether. However, for the engineering and scientific domains such as SHM and geosciences, we lack robust evaluation frameworks for even the simpler DL models, which ill-prepares us to evaluate FMs upon their inevitable adoption.

We address this challenge for the signal processing community for current DL models and future FMs by identifying and jointly incorporating three crucial aspects into an evaluation framework: performance uncertainty, learning efficiency, and overlap between training and test datasets. The first evaluation aspect, performance uncertainty (see Bouthillier et al., 2021), measures the variability in the model's performance arising from two sources of variation: the stochastic training process and the random training data sample. Characterizing the variation of a model's performance helps practitioners qualify expectations and guards against a model developer erroneously claiming a technological advancement. The second

Samuel Myren · Nidhi Parikh · Rosalyn Rael · Garrison Flynn · Emily Casleton
Los Alamos National Laboratory, Los Alamos, NM, 87545
e-mail: samuelmyren@gmail.com; nidhip@lanl.gov; rrael@lanl.gov; garrison@lanl.gov; ecasleton@lanl.gov

Dave Higdon
Department of Statistics, Virginia Tech, Blacksburg VA, 24061
e-mail: dhigdon@vt.edu

Roland Platz et al. (eds.), *Model Validation and Uncertainty Quantification, Vol. 3*, Conference Proceedings of the Society for Experimental Mechanics Series, https://doi.org/10.13052/97887-438-0148-1_11

evaluation aspect, learning efficiency (for reference see Hlynsson et al., 2019), measures the model's capability to perform under varying sizes of training data. Learning efficiency is fundamental to FM evaluation because a user needs to understand the cost to benefit ratio of obtaining more labeled data during the fine-tuning stage. It also is imperative to current DL methods to determine which model can do more with less. The third evaluation aspect is the effect on model performance resulting from overlap in the training and test data (e.g., Elangovan et al., 2021), which directly targets the FM's pre-training on big data phase where the test dataset may overlap or be partially/fully contained in the pre-training dataset.

As of conducting this research, easily accessible foundation models were unavailable for the SHM community, so we demonstrate our evaluation framework by employing traditional DL methods borrowed from the seismic community. We focus on methods of interest to the SHM community by using a model trained to identify and temporally locate events in ground vibration data. We apply our framework to compare four competing models under various training methods including pre-training, fine-tuning, and trained-from-scratch. To address the three crucial evaluation aspects, we employ rigorous, thoughtful partitioning of the training, validation, and test data splits and we utilize random data selection and deep ensembles. We then compare the models with uncertainty across a variety of performance metrics and discuss the results.

Background

To demonstrate the evaluation framework, we borrow a deep learning algorithm from the seismic community, PhaseNet (Zhu & Beroza, 2018). PhaseNet is a U-Net type architecture (Ronneberger et al., 2015) designed to identify characteristic signatures in tri-axial ground acceleration data (a waveform) that propagates through the Earth from an event's (i.e. earthquake or blast) hypocenter to the station housing the seismometer. Prominent signatures include the primary (P) compressional wave and the secondary (S) longitudinal wave. The process of phase picking involves manually inspecting waveforms and identifying the time of arrival of the P/S-waves. Identification of these phases at various stations allows seismologists to locate and characterize an event. For demonstration purposes, we focus exclusively on identifying P-waves, called P-picking.

Data

In this work, we use two publicly available, noncontinuous, labeled datasets. The first, the STanford EArthquake Dataset, STEAD (Mousavi et al., 2019), comprises ∼1.3 million waveforms including earthquakes and noise from sources and stations distributed across the globe. The second, the Italian seismic dataset for machine learning, INSTANCE (Michelini et al., 2021), also comprises ∼1.3 million earthquake and noise waveforms, but are regionally constrained near Italy. To measure the effect of overlap in training and test data, we spin off a third dataset we call STEAD Masked, which comprises all of STEAD data but removes the sources and stations that overlap with the Italian region. An example of a waveform featuring a labeled P-wave is shown in Fig. 1.

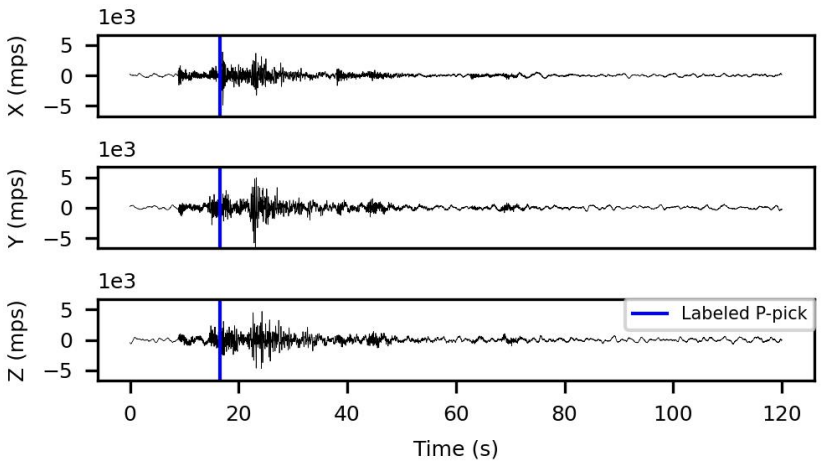

Fig. 1 Example of a triaxial waveform of seismic data from the INSTANCE dataset with labeled P-pick (blue line).

To measure the three performance aspects, we carefully construct data splits using a clustering approach. We leave STEAD and STEAD Masked as they are because we primarily use them for pre-training models. For INSTANCE, since we aim to measure learning efficiency which tests model performance at various data amounts, we cluster the data using a K-means clustering algorithm based on source and station latitude and longitude (see Fig. 2). We employ 20 clusters as this provided a balance between cluster size and geographical division to help limit data leakage between clusters to better measure data sampling uncertainty and learning efficiency.

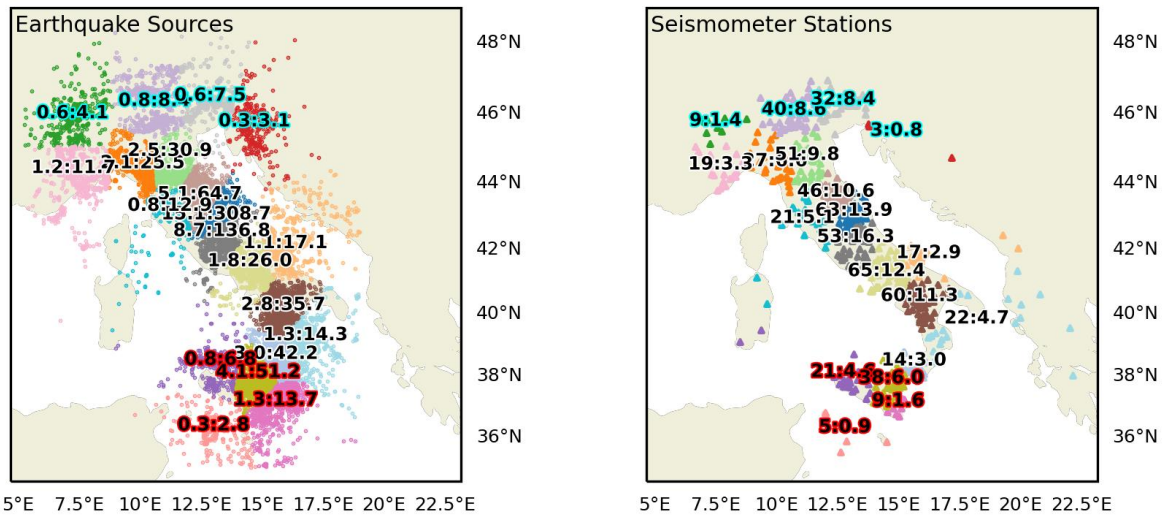

Fig. 2 INSTANCE sources (left, circles) and stations (right, triangles) are clustered using 20 clusters based on geographical locations. For the left panel (sources), the text shows the number of sources in thousands followed by the number of waveforms in thousands. For the right panel (stations), the text shows the number of stations and the number of noise waveforms in thousands. The text highlighting indicates the group: cyan is north test group, white is training/validation group, and red is south test group.

With the INSTANCE clusters, we create the training, validation, and test splits to accomplish our evaluation goals. We use the 4 northern most clusters and 4 southernmost clusters to comprise the test set, balancing the data within each region approximately equally through random source sampling resulting in about 50,000 earthquake waveforms and 5,000 noise waveforms. The remaining 12 central clusters form the training and validation pool of data. To make the validation set, we randomly choose 159 sources and 276 noise waveforms from each cluster, totaling 27,169 earthquake waveforms and 3,312 noise waveforms to form the validation set. The remaining data from the 12 central clusters comprises the training pool. When training models, the amount of training data is varied and randomly chosen from the quantity of clusters assigned. In general, we aim to keep an 80-20 training to validation ratio, so each training cluster comprises 636 sources (which on average corresponds to 9,705 waveforms) and 1,106 noise waveforms.

Methods

We apply the PhaseNet architecture (Zhu & Beroza, 2018) for all models which has been well researched for the phase picking domain. PhaseNet is a U-Net type architecture (Ronneberger et al., 2015) with identical input and output sizes of 3x3001. That is, the model accepts 30s tri-axial vibration data sampled at 100Hz and produces outputs of the same size. Thus, we train the model to predict the probability of 3 features as a function of time: one vector for predicting P-waves, one for S-waves, and one for noise. For demonstration, we discard the S-wave and noise predictions for all subsequent analysis.

To exhibit the versatility of the evaluation framework and to measure the effect of overlapping training and test data, we construct 7 competing models (see Table 1) for comparison using various training schemes. The first two models act as a baseline and are trained and validated on STEAD and STEAD Masked but tested on INSTANCE. As such, we refer to these models as out-of-distribution (OOD) models (OOD and OOD Masked, respectively). The next 4 models utilize transfer learning (TL) and exemplify a situation where we have a pre-trained model available and some data available for further training. We refer to these models as TL Free, TL Free Masked, TL Frozen, and TL Frozen Masked. The Masked

label indicates if the model is initialized using weights from the STEAD or STEAD Masked OOD models, and the Free versus Frozen label indicates whether half of the model's weights have been fixed to mimic a fine-tuning approach common for foundation models. When freezing weights, we fix the first half of the model: all down sampling layers at their initialized values. The final model, Standard, represents the standard approach to deep learning, where the weights are randomly initialized, and the model is trained using INSTANCE data. All models are tested on the INSTANCE test regions.

Table 1 Details of the 7 competing models are displayed. Each row represents a model, and the columns designate the model name, how the model was initialized (pre-trained or randomly initiated), whether the weights are free or fixed, the training and validation datasets, and the quantity of cluster levels that each model was trained under.

Model Name	Initialization	Weights	Training Set	Validation Set	Quantity of Clusters
OOD	Random	Free	STEAD	STEAD	NA
OOD Masked	Random	Free	STEAD Masked	STEAD Masked	NA
Standard	Random	Free	INSTANCE	INSTANCE	{1,2,6,9,12}
TL Free	OOD	Free	INSTANCE	INSTANCE	{1,2,6,9,12}
TL Free Masked	OOD Masked	Free	INSTANCE	INSTANCE	{1,2,6,9,12}
TL Frozen	OOD	Half Frozen	INSTANCE	INSTANCE	{1,2,6,9,12}
TL Frozen Masked	OOD Masked	Half Frozen	INSTANCE	INSTANCE	{1,2,6,9,12}

To evaluate learning efficiency, we train each model under varying quantities of training clusters with levels 1, 3, 6, 9 and 12 clusters. We measure performance uncertainty arising from data sampling by training each model 12 times under each level, randomly choosing the cluster set that comprises the quantity of clusters each time. Then, to quantify training uncertainty, we hold the data fixed and utilize adaptations of deep ensembles (Ganaie et al., 2022; Huang et al., 2017) to retrain each model under 4 different weight initializations while keeping the training data fixed. Note that the two OOD models are exempt from this training scheme, as their training data remains fixed. However, to supply multiple initialization points for the other models, we retrain the two OOD models using deep ensembles 48 times each. For ease of discussion, we refer to a single "model instance" as the unique combination of model, quantity of clusters, cluster set, and initialization. For example, the TL Free model trained with 3 clusters comprising the randomly chosen cluster set of green, blue, and yellow (in reference Figure 1) and the 3rd initialization represents a single model instance. Thus, we obtain 1,200 model instances from the 5 non-OOD models and 96 model instances from the 2 OOD models, totaling 1,296 model instances.

Evaluation

For this work, we focus on metrics that summarize the model performance in terms of identifying and localizing the P-wave within each earthquake waveform. The model produces a prediction of the probability of a P-wave for each time point in the waveform. We implement a user-defined threshold (see Fig. 3) where if the prediction exceeds the threshold, we obtain a predicted pick. We then classify the picks for each earthquake waveform in the test dataset. We obtain a true positive if the predicted P-pick is within 0.3s of the labeled P-pick, false positives for picks extra picks and those outside of 0.3s of the labeled P-pick, and false negatives if no predictions are made. We do not collect true negatives as they are relatively noninformative for this domain. We designate the total number of true positives, false positives, and false negatives as TP, FP, and FN and compute metrics including $Precision = \frac{TP}{TP+FP}$, $Recall = \frac{TP}{TP+FN}$, and $F1 = \frac{2TP}{2TP+FN+FP}$.

With metrics of interest in hand, we present a statistical model to condense the information from the 1,296 model instances into a digestible form for comparison. Given a metric that summarizes the performance, designated y_{madi}, where m indexes one of the 7 models, a indexes the quantity of clusters, d indexes the cluster set, and i indexes each of the 4 initializations, we present the statistical model

$$y_{madi} = \gamma + \mu_m + \alpha_a + \theta_{ma} + \epsilon^{data}_{mad} + \epsilon^{train}_{madi} \, .$$

Since the OOD models are not trained over varying quantities of clusters and cluster sets, the statistical model simplifies by dropping the indices a and d and associated terms. In this model, γ is the grand mean, μ_m is the effect of the modeling approach, α_a is the effect of the quantity of clusters (i.e. data amount), and θ_{ma} is the interaction term. Furthermore, we assume $\epsilon^{data}_{mad} \sim N\left(0, \, \sigma^{2data}_{ma}\right)$ and $\epsilon^{train}_{madi} \sim N\left(0, \, \sigma^{2train}_{ma}\right)$ where the variance arising from the data and the training are separable and independent terms based on the training design.

Through this model, we estimate the effects of each model and its relationship with the quantity of clusters to evaluate the learning efficiency of each model. Furthermore, we can exploit the normal assumption to obtain confidence intervals

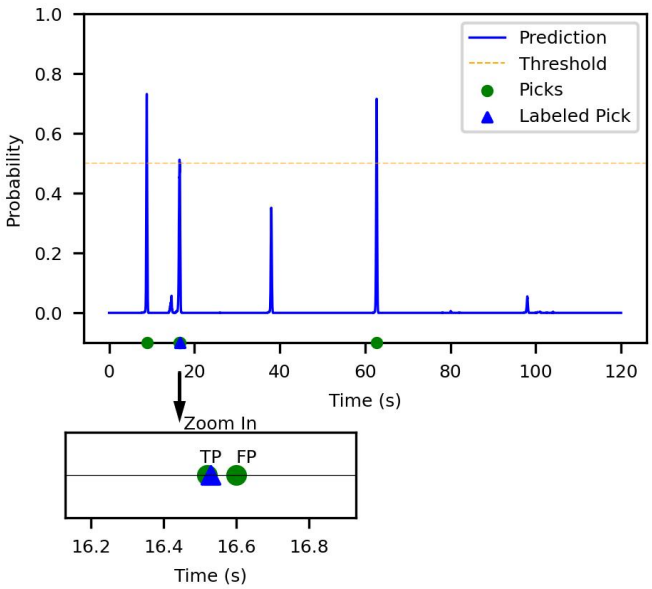

Fig. 3 An example of the model's prediction is shown (blue) for a single waveform. The labeled P-pick (blue triangle) is shown along with other predicted picks made by the model (green circles). This prediction results in 1 true positive and 3 false positives. The inset plot on the bottom zooms in to the region very near the true positive to show how multiple picks are made.

around the effects. Finally, we can estimate each variance term and form confidence intervals to determine which model exhibits minimal variability.

Results

With the 1,296 trained model instances and metrics calculated, we cast the information into the statistical model to obtain mean performance and 90% confidence intervals for each metric. The results are summarized in Fig. 4. For the simpler metrics (total TP, FP, and FN), we see the largest gains across all models occurring between 1 and 3 quantities of clusters and that increasing the quantity of clusters beyond 3 results in diminishing improvements. In comparing the 7 models directly, we identify the general trend that the Standard (green) and two TL w/ Free weights (blue) outperform the other 4 models. However, if we were to rank the models based on performance, our rankings would change depending on the quantities of clusters as exhibited by non-parallel lines. For example, the Standard (green) model does worse than the TL w/ Free Weights (blue) models at 1 cluster, but quickly outperforms as more data is added. Similarly, the 2 TL w/ Frozen Weights (orange) models perform even worse than the 2 OOD models at 1 cluster but perform better as more data is added.

This last result, that the TL w/Frozen Weights (orange) perform worse for small data is certainly a surprising result. We may expect that by freezing half the weights and performing some re-training, we see some performance improvement as the weights are fine-tuned from their pre-trained states. However, this hypothesis is ill supported in the TP panel alone. Fortunately, the other metrics provide some context. We see that the orange models also have few FPs. This indicates that the orange models are predicting more conservatively in general, which explains the results in the precision, recall, and F1 panels. We see that precision for the orange models resides in between the OOD models and the rest of the models. This makes sense because the TL w/ Frozen Weights are effectively a hybrid among all the modeling approaches. These results convey the importance of continuing to evaluate current DL/AI models and future FMs against simpler training approaches and that model rankings can change depending on the amount of data being employed.

To determine the effect of overlap between training and testing data, we compare the solid lines (non-Masked) to their dashed lines (Masked) models. We see that the non-Masked models typically outperform their Masked counterparts, although the effect is small and inconsistent. The exceptions to this trend are the OOD models (red) for FP and recall, where the masking appears to improve performance. This case is likely explained by how the masked models predict with lower confidence in general and thereby results in fewer FPs. Thus, these results are as expected: removing training data that

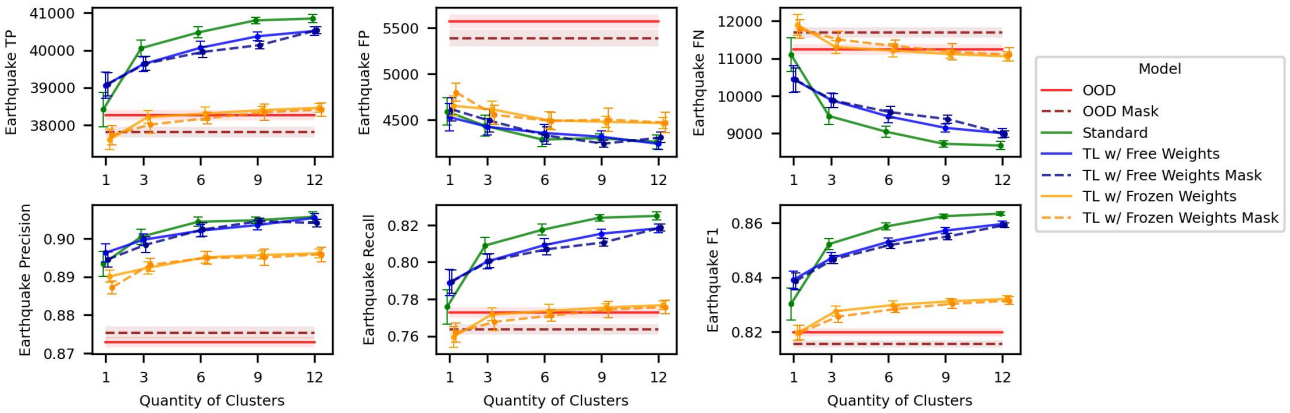

Fig. 4 Results of the model performance for each of the 7 competing models across different quantities of clusters. Each panel is a different metric, including total TP, FP, and FN, and precision, recall, and F1. The OOD models (red and dark red) are constant because they are exempt from the training design. Ribbons and vertical bars represent 90% confidence intervals.

overlaps with the test data reduces model performance. This is an important finding for evaluation FMs because it tells us that we must hold out full datasets from the pretraining phase of all the FMs to obtain fair evaluation.

The importance of having error bars is worth emphasizing in Fig. 4. Consider if you were to train each of the 7 models only once under 1 cluster. Since each model instance that you train would come from the corresponding model's distribution of performance, you may mis-rank the models as exhibited by overlapping error bars. Thus, equipped with the uncertainty information, a practitioner when selecting the model can calibrate their expectations for a given model's performance.

In considering the performance error bars and ribbons in Fig. 4, we can decompose these intervals through estimating the two sources of uncertainty arising from training and data. These estimates for precision, along with 90% confidence intervals are shown in Fig. 5, however the uncertainty can be analyzed in this way for any metric. In the training variance, we see a generally flat trend as the quantity of clusters is increased. However, this trend is broken for the Standard (green) model where we see markedly higher training variability that decreases toward the trend as we move from 1 to 3 clusters. This tells us that for small data, we may avoid using the Standard model due to highly varying performance.

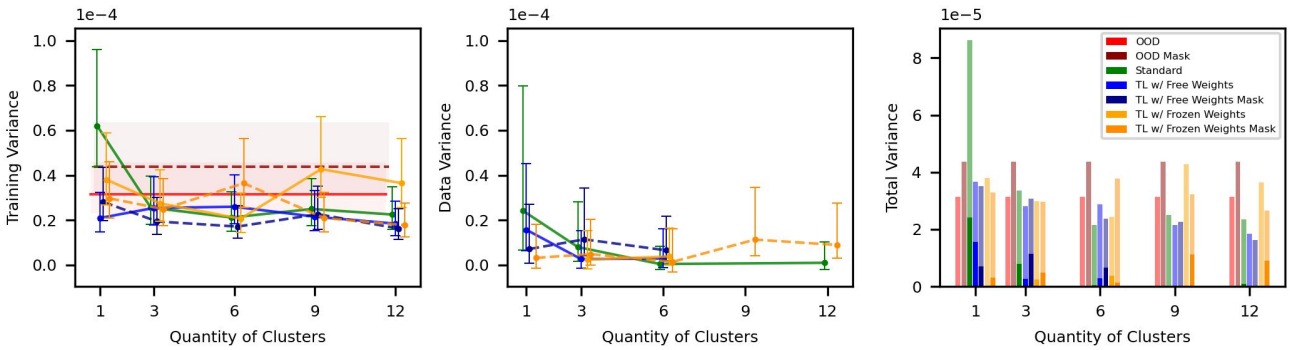

Fig. 5 Variance estimates are shown for the metric precision for each model as a function of quantities of clusters. Training variance (left panel) and data variance (middle panel) are shown along with ribbons or error bars representing 90% confidence intervals. The right panel shows the total variability as a stacked bar chart showing the relative contributions of the training variance (transparent) and data variance (opaque) to the total variance.

For the data variance (middle panel in Fig. 5), we see decreasing trends as more data is added. This is expected, because as we add more data, we better approximate the full dataset, thereby reducing variability. In addition, as exhibited by the right panel where the total variance is shown along with relative contributions of each source, the data variance contributes minimally relative to the training variance. Thus, in the middle panel, some of the curves abruptly stop at 6 clusters. This is because, as an artifact of the statistical framework, we obtain negative values for the data variance. Since negative variance is

theoretically impossible, we choose to remove that information from the plot, but it essentially means that the data variance is extremely small relative to the training variance. Overall, understanding which aspect in the modeling process is contributing to variance allows developers to design models more efficiently and allows practitioners to choose models that best fit their needs.

Conclusion

In this work, we demonstrate a robust statistical framework that jointly incorporates three crucial evaluation aspects for current AI/DL models and future FMs, including performance uncertainty, learning efficiency, and overlap between training and test data. We demonstrate our evaluation framework using existing DL models from the seismic phase picking community by comparing 7 competing models under various training schemes, including a standard, out-of-distribution, pre-training, and fine-tuning model deployments. We design careful data splits and train each model under varying training sizes with multiple initializations to obtain segregated estimates of uncertainty arising from training process and data sampling. We summarize model results by computing metrics of interest and cast the information into a statistical model to summarize the results. We exemplify the importance of including uncertainty in an evaluation framework and how the model can interact strongly with the data amount. Furthermore, we find a non-negligible effect of overlap between training and test data. In general, these results show the need for such an evaluation framework to fairly evaluate models under a variety of contexts and to inform practitioners such that they can choose the best model for a given set of constraints.

Acknowledgments This work was performed under the US Department of Energy, National Nuclear Security Administration's Office of Defense Nuclear Nonproliferation Research and Development (NA-22) Steel Thread Venture. This work is approved for public release under LA-UR-24-30447.

References

1. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022). On the Opportunities and Risks of Foundation Models (No. arXiv:2108.07258). arXiv. http://arxiv.org/abs/2108.07258
2. Bouthillier, X., Delaunay, P., Bronzi, M., Trofimov, A., Nichyporuk, B., Szeto, J., Sepah, N., Raff, E., Madan, K., Voleti, V., Kahou, S. E., Michalski, V., Serdyuk, D., Arbel, T., Pal, C., Varoquaux, G., & Vincent, P. (2021). Accounting for Variance in Machine Learning Benchmarks (No. arXiv:2103.03098). arXiv. http://arxiv.org/abs/2103.03098
3. Elangovan, A., He, J., & Verspoor, K. (2021). Memorization vs. Generalization: Quantifying Data Leakage in NLP Performance Evaluation. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 1325–1335. https://doi.org/10.18653/v1/2021.eacl-main.113
4. Flah, M., Nunez, I., Ben Chaabene, W., & Nehdi, M. L. (2021). Machine Learning Algorithms in Civil Structural Health Monitoring: A Systematic Review. Archives of Computational Methods in Engineering, 28(4), 2621–2643. https://doi.org/10.1007/s11831-020-09471-9
5. Ganaie, M. A., Hu, M., Malik, A. K., Tanveer, M., & Suganthan, P. N. (2022). Ensemble deep learning: A review. Engineering Applications of Artificial Intelligence, 115, 105151. https://doi.org/10.1016/j.engappai.2022.105151
6. Harsuko, R., & Alkhalifah, T. (2022). StorSeismic: A new paradigm in deep learning for seismic processing. https://doi.org/10.48550/ARXIV.2205.00222
7. Hlynsson, H. D., Escalante-B., A. N., & Wiskott, L. (2019). Measuring the Data Efficiency of Deep Learning Methods. Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods, 691–698. https://doi.org/10.5220/0007456306910698
8. Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J. E., & Weinberger, K. Q. (2017). Snapshot Ensembles: Train 1, get M for free (No. arXiv:1704.00109). arXiv. http://arxiv.org/abs/1704.00109
9. Michelini, A., Cianetti, S., Gaviano, S., Giunchi, C., Jozinović, D., & Lauciani, V. (2021). INSTANCE – the Italian seismic dataset for machine learning. Earth System Science Data, 13(12), 5509–5544. https://doi.org/10.5194/essd-13-5509-2021
10. Mousavi, S. M., Sheng, Y., Zhu, W., & Beroza, G. C. (2019). STanford EArthquake Dataset (STEAD): A Global Data Set of Seismic Signals for AI. IEEE Access, 7, 179464–179476. https://doi.org/10.1109/ACCESS.2019.2947848
11. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation (No. arXiv:1505.04597). arXiv. http://arxiv.org/abs/1505.04597
12. Zhu, W., & Beroza, G. C. (2018). PhaseNet: A Deep-Neural-Network-Based Seismic Arrival Time Picking Method. Geophysical Journal International. https://doi.org/10.1093/gji/ggy423