Uncertainty Quantification of Machine Learning to Establish AI Trustworthiness in Nuclear Engineering Applications

Xu Wu
xwu27@ncsu.edu

Assistant Professor
Department of Nuclear Engineering
North Carolina State University

Data Science and Artificial Intelligence Regulatory Applications Workshops
Workshop #4: AI Characteristics for Regulatory Consideration
Panel Session on "AI Safety, Security and Explainability"
The U.S. Nuclear Regulatory Commission (NRC)

September 19, 2023

Sources of uncertainties in physical modeling & simulation

Parameter Uncertainty

unknown exact values of the model input parameters, and/or randomness

Experimental/Data Uncertainty

noise or error in the measurement and/or data processing process

Numerical Uncertainty

numerical approximation errors due to e.g., insufficient convergence, mesh

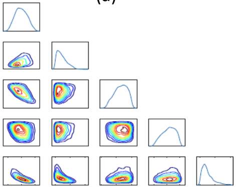
Model Uncertainty (Bias, Discrepancy)

missing, inaccurate and/or incomplete underlying physics in the computer model

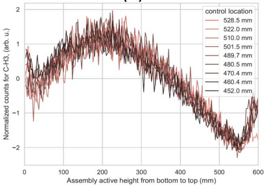
Code/Interpolation Uncertainty

emulation of expensive computer codes using data-driven surrogate model

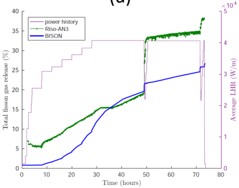
(a)



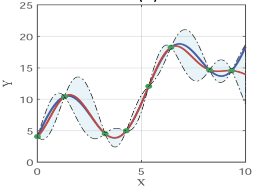
(b)



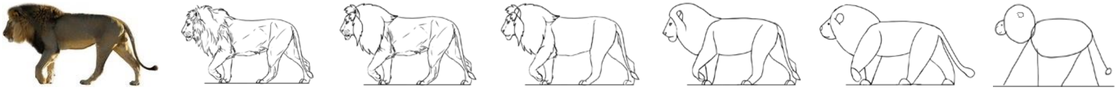
(d)



(e)



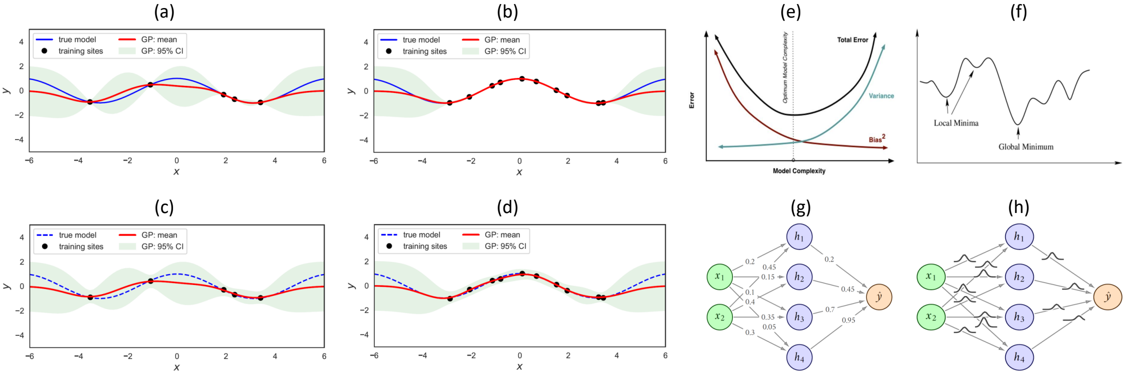
(c)



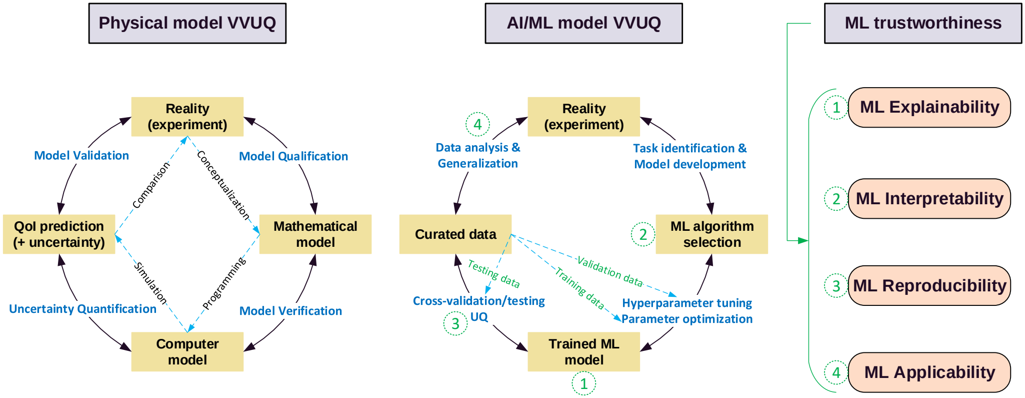
Observation | Good model | Model 1 | Model 2 | Model 3 | Model 4 | Model 5

Sources of uncertainties in data-driven Machine Learning models

Data Noise	Data Coverage	Extrapolation	Imperfect Model	Training
noises in training data from either physical simulation models or experiments	few and/or gappy data that has incomplete coverage of training domain	generalization to the extrapolated domains outside of the training domain	ML model architecture is not properly defined, e.g., model is too simple or too complex	random initialization, convergence to local minima, hyper-parameter tuning, etc



- Application-agnostic algorithms, or those designed for more traditional ML applications such as computer vision and natural language processing, cannot typically be directly applied to scientific data in nuclear applications and require non-trivial, task-specific modifications.
- **Low-consequence error-tolerant settings** → **high-consequence nuclear systems**, need to establish ML trustworthiness, including accuracy, robustness (reproducibility, applicability), algorithmic fairness, algorithmic transparency (explainability, interpretability), and privacy.



- Explainability[1,2]: the ability to ensure that algorithmic decisions as well as any data driving those decisions can be explained to end-users and stakeholders in non-technical terms. It helps stakeholders and decision makers to understand ML solutions by "opening the black-box".
- Interpretability[3]: the degree that an ML model obeys structural knowledge of the domain, such as monotonicity, causality, structural constraints, additivity, or physical constraints that come from domain knowledge.
- Reproducibility[4,5]: the ability of being able to replicate the ML model from data processing to model design, reporting, model analysis, or evaluation to successful deployment.
- Applicability[6]: the usability of ML for new scenarios such as unseen domains.
- Other definitions: NIST framework on AI trustworthiness[7,8] consists of nine factors: accuracy, reliability, resiliency, objectivity, security, explainability, safety, accountability and privacy.

[1] Barocas, S., Friedler, S., Hardt, M., Kroll, J., et al. (2018). The FAT-ML Workshop Series on Fairness, Accountability, and Transparency in Machine Learning.

[2] Gunning, D., Vorm, E., Wang, J. Y., & Turek, M. (2021). DARPA's explainable AI (XAI) program: A retrospective. Applied AI Letters, 2(4), e61.

[3] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

[4] Heil, B. J., Hoffman, M. M., Markowetz, F., Lee, S. I., et al. (2021). Reproducibility standards for machine learning in the life sciences. Nature Methods, 18(10), 1132-1135.

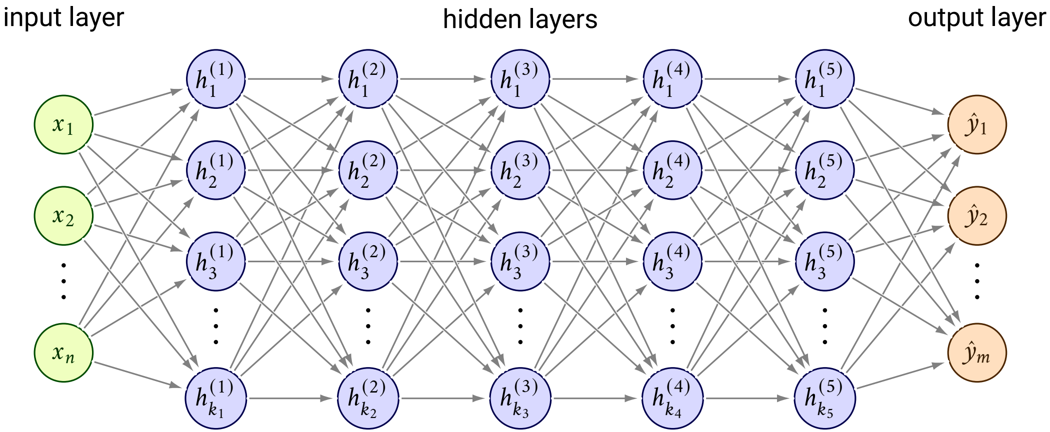
[5] Beam, Andrew L., Arjun K. Manrai, and Marzyeh Ghassemi. "Challenges to the reproducibility of machine learning models in health care." Jama 323.4 (2020): 305-306.

[6] Wang, J., Lan, C., Liu, C., Ouyang, Y., Zeng, W., & Qin, T. (2021). Generalizing to unseen domains: A survey on domain generalization. arXiv preprint arXiv:2103.03097.

[7] Stanton, B., & Jensen, T. (2021). Trust and artificial intelligence. Draft NIST Interagency/Internal Report (NISTIR) 8332, National Institute of Standards and Technology, Gaithersburg, MD, URL: https://doi.org/10.6028/NIST.IR.8332-draft.

[8] Phillips, P. J., Hahn, C. A., Fontana, P. C., Broniatowski, D. A., & Przybocki, M. A. (2020), Four Principles of Explainable Artificial Intelligence, NIST Interagency/Internal Report (NISTIR) 8312, National Institute of Standards and Technology, Gaithersburg, MD, https://doi.org/10.6028/NIST.IR.8312

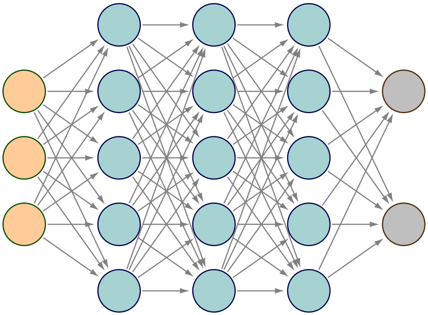
How to quantify the approximation/prediction uncertainties in deep neural networks[9]?



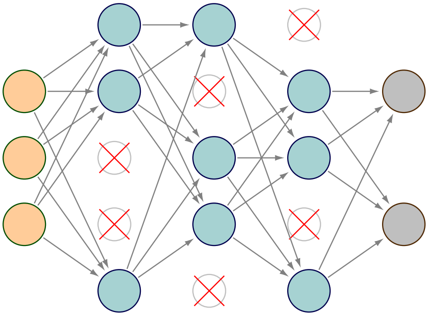
[9] Yaseen, M., & Wu, X. (2023). Quantification of Deep Neural Network Prediction Uncertainties for VVUQ of Machine Learning Models. Nuclear Science and Engineering, 197(5), 947-966.

Monte Carlo Dropout (MCD)[10] introduces randomness to prediction in addition to training

- The training step is performed in the regular way, using stochastic gradient descent methods and re-evaluating the dropout matrices before each learning step.
- At the prediction step, we again evaluate the dropout matrices before every forward pass resulting in random network outputs.



Regular DNN



DNN after dropout

[10] Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In international conference on machine learning (pp. 1050-1059).

Deep ensembles (DE)[11] changes network predictions as distributional parameters

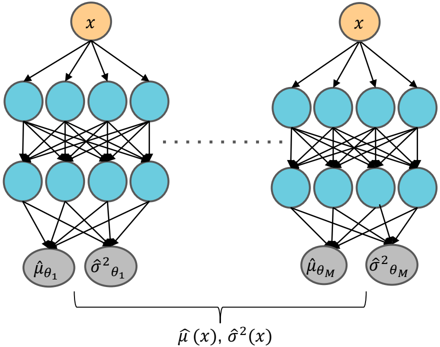
- DE assumes the data to have a given parameterized distribution (e.g., Gaussian) where the distribution parameters depend on the input.
- Use the negative log-likelihood function of the Gaussian distribution as the cost function:

$$\mathcal{L}_\theta(\mathbf{x}, y) = -\log \phi_\theta(y|\mathbf{x}) = \frac{\log \hat{\sigma}_\theta^2(\mathbf{x})}{2} + \frac{(y - \hat{\mu}_\theta(\mathbf{x}))^2}{2\hat{\sigma}_\theta^2(\mathbf{x})} + c$$

- With an ensemble of M neural networks, the joint Gaussian has mean and variance given by:

$$\hat{\mu}(\mathbf{x}) = \frac{1}{M}\sum_{i=1}^{M} \hat{\mu}_{\theta_i}(\mathbf{x})$$

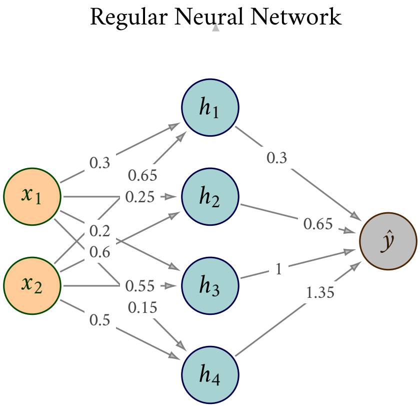
$$\hat{\sigma}^2(\mathbf{x}) = \frac{1}{M}\sum_{i=1}^{M} \left(\hat{\sigma}_{\theta_i}^2(\mathbf{x}) + \hat{\mu}_{\theta_i}^2(\mathbf{x})\right) - \hat{\mu}^2(\mathbf{x})$$

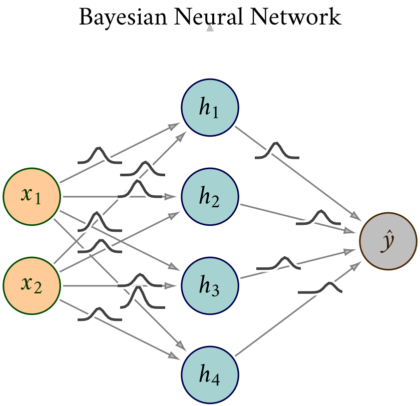


[11] Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in neural information processing systems (pp. 6402-6413).

Bayesian Neural Networks (BNNs)[12] - neural networks with distributions over parameters

- In BNNs, prior distributions are specified upon the parameters (weights, bias) of neural networks.
- Given the training data, the posterior distributions over the parameters are computed, which are used to quantify the predictive uncertainty.

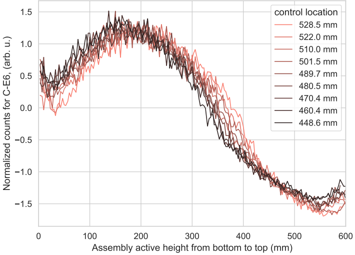


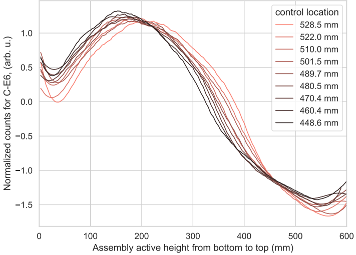


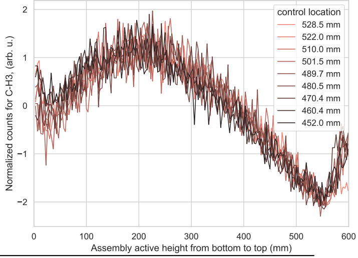
[12] Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural network. In International conference on machine learning (pp. 1613-1622). PMLR.

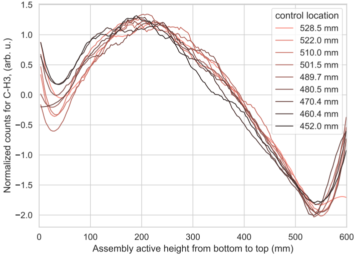
Example: using DNNs to predict the axial neutron flux profiles given the control rod bank position

- Training data[13]: copper-wire activation measurements and measured control bank positions obtained from the SAFARI-1 research reactor (South Africa) historical cycles.



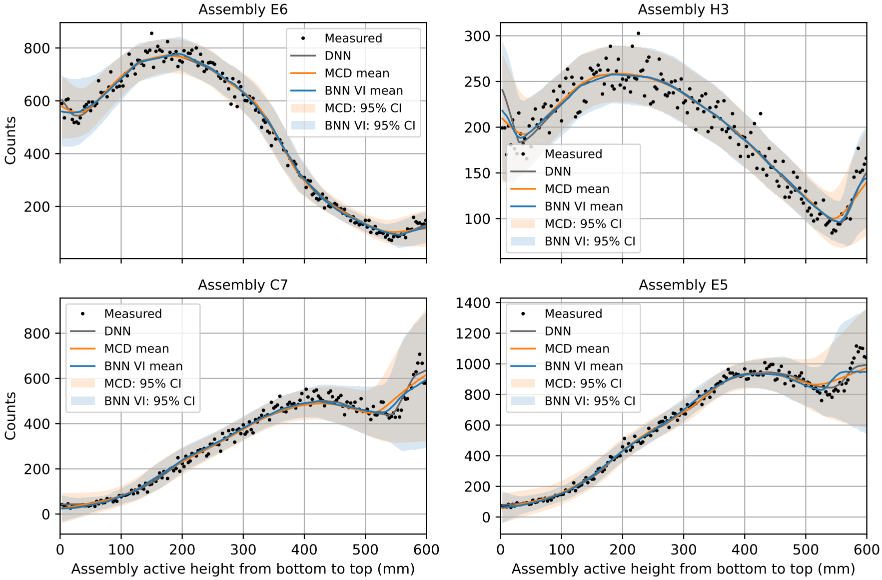






[13] Moloko, L. E., Bokov, P. M., Wu, X., & Ivanov, K. N. (2023). Prediction and uncertainty quantification of SAFARI-1 axial neutron flux profiles with neural networks. Annals of Nuclear Energy, 188, 109813.

Example: the DNN predictions are made on assemblies and cycles that are unseen during training



Thank you for your attention!

Questions and comments?