Towards Explainable AI in Nuclear: Introducing Ad Hoc Model Explainability

Alex Xu[1], Nataly Panczyk [2,*], Majdi I. Radaideh[2]

[1]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109;
[2]Department of Nuclear Engineering and Radiological Sciences, University of Michigan, Ann Arbor, MI 48109

ABSTRACT

The purpose of this work is to demonstrate ad hoc explainability methods for artificial intelligence (AI) applications to the nuclear industry. We specifically consider using a feedforward neural network (FNN) to predict critical heat flux (CHF) using 60 years worth of data collected from a variety of experimental setups. The FNN uses combinations of six different features in the dataset to predict the onset of CHF, a safety critical parameter in a nuclear reactor. After assessing the state-of-the-art in employing explainability methods for various AI models, we chose Shapley Additive Explanations (SHAP) for its consistency and model-agnostic nature. We then use three flavors of the SHAP methodology (Exact SHAP, DeepLIFT, and Integrated Gradients) to rank the feature importances of the model. We found heated channel length, followed by mass flux and inlet temperature to be the most impactful on CHF prediction when considering a combination of five of the six input features. We validated these results using mean decrease in impurity and permutation importance metrics from a random forest model trained on the same data, and found an identical feature importance ranking. We also experienced a shortcoming in SHAP—it fails to consider feature correlations in determining their importance rankings. Future work should consider implementing explainability methods within the model, not just global metrics as shown in this work, for improved reliability.

Keywords: Explainable AI, Critical Heat Flux, SHAP, DeepLIFT, Neural Networks

1. INTRODUCTION

In a nuclear reactor, the critical heat flux (CHF) is a limit to the amount of heat that can reliably be carried away before the heat transfer coefficient of the coolant significantly drops, risking system damage. A reactor operating above this threshold poses a major threat to people and the environment. On the other extreme, a heat flux well below this threshold indicates a reactor is operating at a power level that is economically disadvantageous. Accurately predicting CHF promises both safe and economical reactors. While this quantity depends on a variety of known physical parameters, it does not have closed form solution. This scenario elucidates an opportunity for artificial intelligence (AI) and machine learning (ML) to inform CHF predictions in nuclear power applications, but without metrics to guarantee the trustworthiness of such an algorithm, AI's application to a safety critical parameter, such as CHF, is impossible. These metrics must capture both interpretability (understanding *how* the model came to a conclusion) and explainability (understanding *why* the model came to a conclusion) to maximize the trustworthiness of a model. Though the state-of-the-art is far from meeting both of these demands, this work seeks to explore AI explainability methods to start making AI models more transparent and available to sensitive industries like nuclear.

*npanczyk@umich.edu

This work begins by reviewing current progress in the explainability space and selecting methods to acquire our explainability metrics, all variants of classical Shapley Additive Explanations (SHAP) (2). Then, we describe the research setup, considering the dataset, feature selection process, and modeling process (3). We then outline the theory behind and application of the explainable AI methods chosen in 2 (4). We evaluate these results by comparing black-box models like neural networks to a model with a known alternative explainability methods, a random forest, to gauge the accuracy of our methods (5). Finally, we conclude this paper by presenting an opportunity for future work by identifying persisting gaps in the research space (6).

2. THE STATE-OF-THE-ART

A big wave of AI popularity is necessarily followed by a (smaller) wave of AI explainability methods. While using AI to summarize some meeting minutes or develop a recipe carries relatively low risk, using AI to predict safety critical parameters of a nuclear power plant does not. This sentiment is highlighted in the *U.S. Nuclear Regulatory Commission's (NRC) Artificial Intelligence Strategic Plan* for fiscal years 2023-2027 [1]. This document motivates increased understanding around black-box AI, like through the explainability methods described in this paper. Specifically, this research targets the NRC's second goal as described [1], which aims to "establish an organizational framework to review AI applications." Without well-designed methods to explain and interpret these models, AI cannot be regulated by the U.S. NRC, and if AI cannot be regulated, its use cases and potential for impact are diminished. In this work, explainability is not related to model uncertainty, which is typically studied and of importance in nuclear applications for design and safety analysis [2, 3].

While the necessity for explainable AI methods is evident for the nuclear industry, other industries are marching towards this goal as well. In [4], the authors provide a comprehensive review of explainable AI methods and applications to describe the state-of-the-art as of 2023. Their review found the highest proportion of explainable AI research in the health care domain, which much like nuclear, has high sensitivity and thus high demand for explainability. The authors of [4] describe some benefits of the SHAP methods we employ in this work, including being model-agnostic and able to provide some of the most consistent results among explainability methods, with its major cost being computational expense [4]. This is mitigated in our work by using derivatives of exact SHAP, which are explained in Section 4.

The need for explainable AI in nuclear has already been explored in some extent by [5], [6], and [7]. In [5], the authors use SHAP to explain nuclear power plant fault diagnoses determined by both a convolutional neural network (CNN) and a recurrent neural network (RNN). In [6], the authors apply explainable AI methods via LightGBM and SHAP to develop a more reliable diagnostic assistant for nuclear power plant operators. The inclusion of LightGBM, which is a variant of a decision tree model instead of a feedforward neural network, adds interpretability to this study, in addition to explainability. In [7], the authors consider a specific case of using AI for predictive (instead of preventative) maintenance in nuclear power plants. They outline the demand for explainability methods for any kind of AI that will approach this problem, and weigh the characteristic tradeoffs of using AI: performance versus explainability [7]. Our work hopes to minimize the tradeoffs described by [7] by applying SHAP methods, as shown by [5] and [6] for a CNN/RNN and LightGBM, respectively, to a feedforward neural network (FNN) instead.

3. RESEARCH SETUP

3.1. Dataset and Feature Selection

Critical heat flux (CHF) in nuclear reactors refers to the maximum heat flux at the surface of a reactor fuel element where the cooling liquid (typically water) can effectively remove heat by nucleate boiling. Beyond

this point, the heat transfer mechanism deteriorates significantly due to the formation of a vapor film on the heated surface, which insulates the fuel element and prevents efficient heat removal. This phenomenon can lead to a sudden increase in surface temperature, potentially causing fuel damage or failure. CHF is a crucial safety parameter in reactor design and operation because it determines the thermal limits to prevent overheating and ensure the safe operation of the reactor.

Numerous experiments have been conducted globally to measure CHF for vertical water-cooled tubes, forming the basis of CHF lookup tables. The OECD Nuclear Energy Agency in collaboration with the U.S. NRC made available experimental CHF data collected over 60 years that included various methods for determining CHF, such as visual observations, physical burnout, changes in resistance in the test sections, and thermocouple readings [8, 9]. We used this dataset [8, 9], which contains 21,453 CHF data points and includes the following features:

- Geometry of the tube: Test section diameter (D), heated length (L)
- Boundary conditions: Pressure (P), mass flux (G), inlet temperature (T)
- Measured parameters: Outlet quality (X_e) and critical heat flux (CHF)

The parameters span the ranges shown in Table I. In training our ML models, we selected the five geometric and boundary features as our default set of inputs and aimed to predict the CHF, i.e. $CHF = F(D, L, P, G, T)$. Another common formulation uses X_e instead of T due to the moderately high correlation between T and P, as shown in the correlation matrix in Figure 1. We train an alternate model using these features as well, i.e. $CHF = F(D, L, P, G, X)$. The full feature space would include all six input parameters to predict the CHF, i.e., $CHF = F(D, L, P, G, T, X)$.

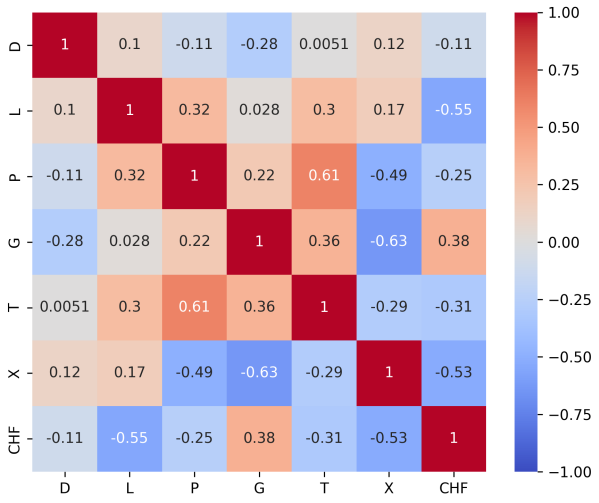

Figure 1. Correlation matrix between CHF and the full set of input features available in the dataset.

Table I. Parameter ranges for the CHF dataset used in this study

Feature	D (mm)	L (m)	P (MPa)	G (kg/m^{2}s)	T (°C)	X (-)	CHF (kW/m^2)
Minimum	2.39	0.07	0.10	17.7	9.0	−0.445	130
Maximum	16.0	15.0	20.0	7712	353.6	0.986	13345
Mean	8.88	2.63	9.86	1992	197.2	0.352	1761

3.2. Modeling

In prior research, we explored the aforementioned three CHF cases (full set, no inlet temperature, no quality) using an automatic ML (autoML) framework called pyMAISE [10]. This autoML tool tested a variety of ML models, including feedforward neural networks (FNN) for these three combinations of the CHF dataset. We found that using all six inputs yielded the most accurate CHF prediction, while dropping inlet temperature (T_{in}) from the input features yielded the least accurate CHF prediction. We summarize these results in Table II. As metrics for the models, we report MAPE (mean absolute percentage error), root mean squared percentage error (RMSPE), and coefficient of determination (R^2). We calculated these values using the test split of the full dataset, so the data were unseen to the model during FNN training. This paper picks up as an additional evaluation of these models by attempting to rank the importance of each feature in the FNN's prediction of CHF. We will henceforth focus on explainable AI methods to understand various feature importance rankings methods for FNN predictions using these models.

Table II. FNN test metrics for the CHF dataset with different feature combinations (see [10] for full analysis).

Input Features	**D, L, P, G, *T***	**D, L, P, G, *X***	**D, L, P, G, *T, X***
MAPE	3.08%	8.65%	1.34%
RMSPE	4.73%	13.19%	2.36%
R^2	0.9963	0.9790	0.9994

Aside from the FNN, our pyMAISE analysis included a variety of classical ML methods for the three cases listed in Table II, including a random forest, gradient boosting, K-nearest neighbors, support vector machines with radial basis kernels, Gaussian processes, and Adaboost. These methods performed worse than FNN for all cases except for $CHF = F(D, L, P, G, X)$, where gradient boosting and random forest models slightly outperformed the FNN with R^2 scores of 0.988 and 0.984, respectively, compared to 0.981 for FNN.

Because the FNNs of all three feature combinations outperformed the CHF lookup table approach, which had a 36% MAPE [9], we will continue this paper with the model $CHF = F(D, L, P, G, T)$. This model does not assume both T and X_e measurements are available for CHF, so it is more practical for future systems, and it demonstrated a relatively lower error than the $CHF = F(D, L, P, G, X)$ case. This choice is made to respect the space limit of this paper, but the method and analysis can be generalized to other cases.

4. METHODOLOGY

This section is dedicated to the ad-hoc explainable AI methods used in this work, which include classic Shapley Additive Explanations (SHAP), Integrated Gradients (IG), and Deep Learning Important FeaTures (DeepLIFT).

Let us define $x = [x_1, x_2, \ldots, x_d]$ as a sample instance with d features for a black-box model f. Let $f(x)$ be the output of the black-box model for that input instance (x). The goal is to explain the model's prediction $f(x)$ by attributing contributions to each feature x_i.

4.1. Classical SHAP

SHAP is a unified framework in explainable AI that provides interpretable feature attributions for model predictions. It is based on cooperative game theory, which allocates the contribution of each player (i.e., model feature) fairly based on their participation in a "game" (i.e., model prediction). SHAP considers all possible subsets of features and calculates the contribution of each feature towards the final prediction,

which may be positive or negative. The Shapley value ϕ_i for feature i is calculated as

$$\phi_i = \sum_{S \subseteq \{1,2,\ldots,d\} \setminus \{i\}} \frac{|S|!(d - |S| - 1)!}{d!} \left[f(S \cup \{i\}) - f(S) \right] \tag{1}$$

where d is the number of features in the model, $\frac{|S|!(d-|S|-1)!}{d!}$ is the weight for each subset S, which accounts for the permutations of the feature set, $f(S)$ represents the model's prediction using only the features in the subset S, and $f(S \cup \{i\}) - f(S)$ represents the marginal contribution of player i to the team S, i.e., it measures how much additional value feature i brings to the model prediction. The runtime of SHAP is $O(d \cdot 2^d)$. With $d = 5$ or even 6 in this study, it is reasonable to run Exact SHAP, but when d becomes large, approximations like Kernel SHAP are used instead [11].

4.2. Deep Learning Important Features (DeepLIFT)

DeepLIFT (Deep Learning Important Features) is an attribution method designed to determine how changes in input from a baseline (or reference) state to the actual state affect the output of a neural network [12]. Unlike traditional gradients that can be highly sensitive to small changes in inputs, DeepLIFT uses "reference differences" to assess the relative importance of each input feature. The method employs "multipliers" to propagate the contributions through each neuron in the network, ensuring that the sum of attributions at one layer corresponds to the sum at the subsequent layer. Like IG, DeepLIFT offers completeness where the sum of attributions for all inputs matches the difference between the network's output for the given input and the baseline output. In addition, DeepLIFT has the ability to handle non-linearities—it is capable of providing meaningful attributions even when the gradients are zero [12]. This feature helps overcome certain limitations inherent to gradient-based methods, particularly in regions where activation functions like ReLU are flat. The DeepLIFT contribution for a feature x_i is computed as:

$$\text{DeepLIFT}_i(x) = (x_i - x_i') \cdot \text{Multiplier}_i \tag{2}$$

where the multiplier is defined as:

$$\text{Multiplier}_i = \frac{\Delta f}{\Delta x_i} = \frac{f(x) - f(x')}{x_i - x_i'} \tag{3}$$

where x_i is the actual input value of feature i, x_i' is the reference (baseline) input value of feature i, $f(x)$ is the output of the neural network for the actual input x, $f(x')$ is the output of the neural network for the baseline input x', Δf is the change in the network output, Δx_i is the change in the input feature i, and Multiplier_i is a term representing the contribution of the change in input feature i to the change in the network output.

4.3. Integrated Gradients (IG)

Integrated Gradients (IG) is an attribution technique based on gradients that mitigates some shortcomings of traditional gradient methods, such as the "gradient saturation" issues, where minimal gradients make it difficult to evaluate the significance of input features [13]. IG computes feature importance by integrating the gradients of the model's output concerning the input along a path from a baseline (often zero) to the actual input. IG offers two main features: (1) completeness where the total attributions equal the difference between the function's output at the given input and its output at the baseline, and (2) symmetry where if two input features equally impact the output, they receive identical attributions. The IG for a feature x_i is computed as:

$$\mathrm{IG}_i(x) = (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} \, d\alpha \tag{4}$$

where x is the actual input vector, x' is the baseline input vector (often chosen as a zero vector or some neutral input), x_i is the i-th feature of the actual input, x_i' is the i-th feature of the baseline input, f is the function representing the model output, α is a scaling parameter that ranges from 0 to 1, used to interpolate between the baseline and the actual input, and $\frac{\partial f(x)}{\partial x_i}$ is the gradient of the model output f with respect to the input feature x_i. Lastly, it is worth highlighting that we have used the implementation of the SHAP package for classical SHAP, DeepLIFT, and integrated gradients in this work [11].

4.4. Explaining Random Forests

As indicated in [10], random forests (RF) can also be a reliable machine learning model for CHF prediction with comparable performance to FNN. Therefore, RF can be used to validate other explainable methods for FNN. For RF, two methods can be used: **Permutation Importances (PI)** and **Mean Decrease in impurity (MDI)**. In PI, for each feature x_i, its values in the dataset are randomly permuted, and the RF model is then evaluated on this shuffled dataset; this procedure is repeated many times to reduce variance. A large decrease in model accuracy indicates a high feature importance, and vice versa. This method tends to be more reliable but also more computationally expensive than MDI, which measures the importance by examining how much each feature reduces impurity across all the decision trees in the forest. Impurity refers to a measure of the homogeneity of the output within the nodes. When a decision tree splits on a feature (x_i), the impurity of the resulting child nodes is generally lower than that of the parent node. The difference in impurity before and after the split is known as the decrease in impurity. The higher the MDI, the higher the importance of the feature.

5. RESULTS

The results in this section focus on the case where CHF is predicted by five input features: $CHF = F(D, L, P, G, T)$, but will also show an alternative model: $CHF = F(D, L, P, G, X)$ for comparison. Future work can generalize the methodology to the other cases but the conclusions about feature importance may differ. The results shown in this section for all methods, with the exception of random forests, include two-component figures. The left-hand side component of these figures shows a "beeswarm" plot. The beeswarm plot contains point-wise information about the impact on model output by feature variations and given SHAP method. Each dot on the beeswarm plot shows a tested feature value (from low to high, indicated by its color), and its horizontal position shows the overall impact that adjustment had on the output (heat flux at which CHF occurred). Each SHAP method repeats this procedure for each input feature many times. Through this iteration through possible feature values, the beeswarm summary plot is compiled. The right-hand side component shows a bar plot that is a higher level summary of the information contained in the beeswarm plot to make feature importance rankings obvious to the viewer. Each bar shows the mean of the absolute value of all the SHAP values (dots in the beeswarm plot) for each feature, thus demonstrating a feature's total magnitude of impact.

To begin, we directly computed the importances for each feature using Exact SHAP. We show these results in Figure 2. As shown in Figure 2, the overall feature importance ranking is heated channel length (L), mass flux (G), inlet temperature (T_{in}), heated channel diameter (D), then pressure (P).

Figure 3 shows the explainability results using the DeepLIFT method. The overall feature ranking here matches that of Figure 2, but with slightly lower overall magnitudes of impact, as shown by the shifting of the y-scale.

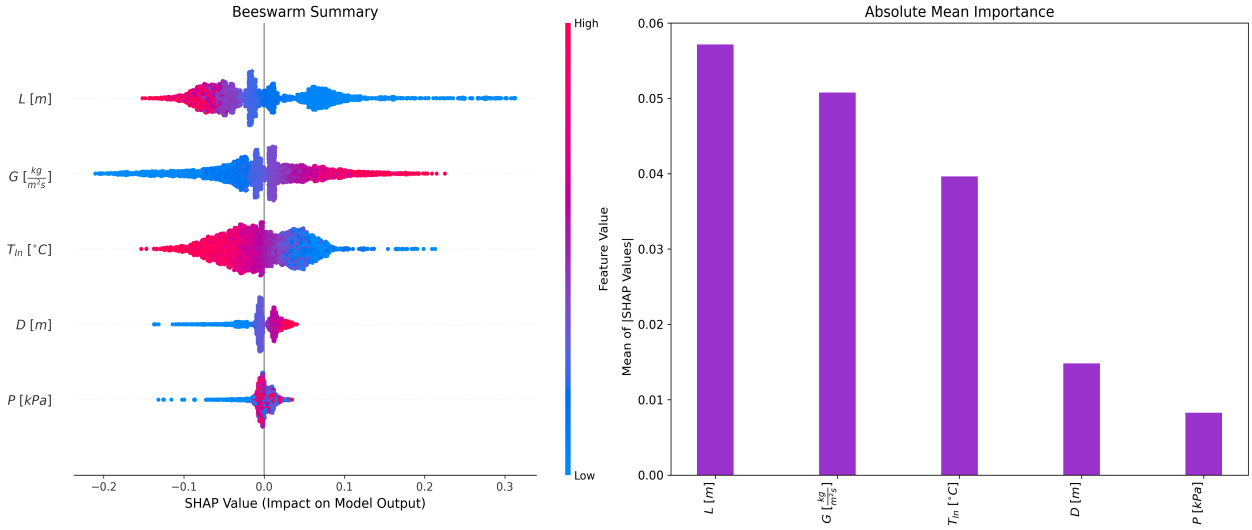

Figure 2. Explainability results using Exact SHAP for the best CHF FNN model with all features *except* outlet quality (X_e).

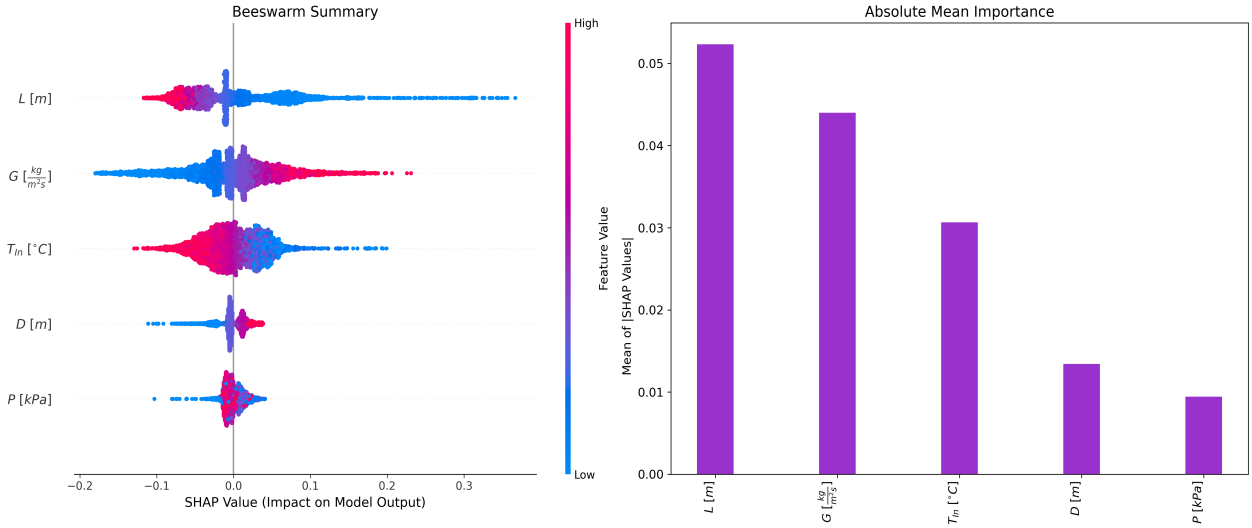

Figure 3. Explainability results using DeepLIFT for the best CHF FNN model with all features *except* outlet quality (X_e).

Figures 4 and 5 show the feature importance rankings using integrated gradients for the model, $CHF = F(D, L, P, G, T)$ and for the alternative model, $CHF = F(D, L, P, G, X)$, respectively. Figure 4 shows the same importance ranking as Figures 2 and 3, but with a higher overall magnitude of impact than both of these models, as shown, again, by the shift in the y-scale. These shifts in magnitude are symptoms of how each method approximates the exact (classic) SHAP value calculation. Nonetheless, the overall importance ranking is the most important takeaway, and all three methods agree for the $CHF = F(D, L, P, G, T)$ model.

Figures 4 and 5 show that replacing inlet temperature (T) for outlet quality (X_e) drastically changed the feature importance rankings. Outlet quality became the most important feature, followed by pressure, heated channel length, mass flux, and channel diameter. This result highlights a shortcoming of SHAP

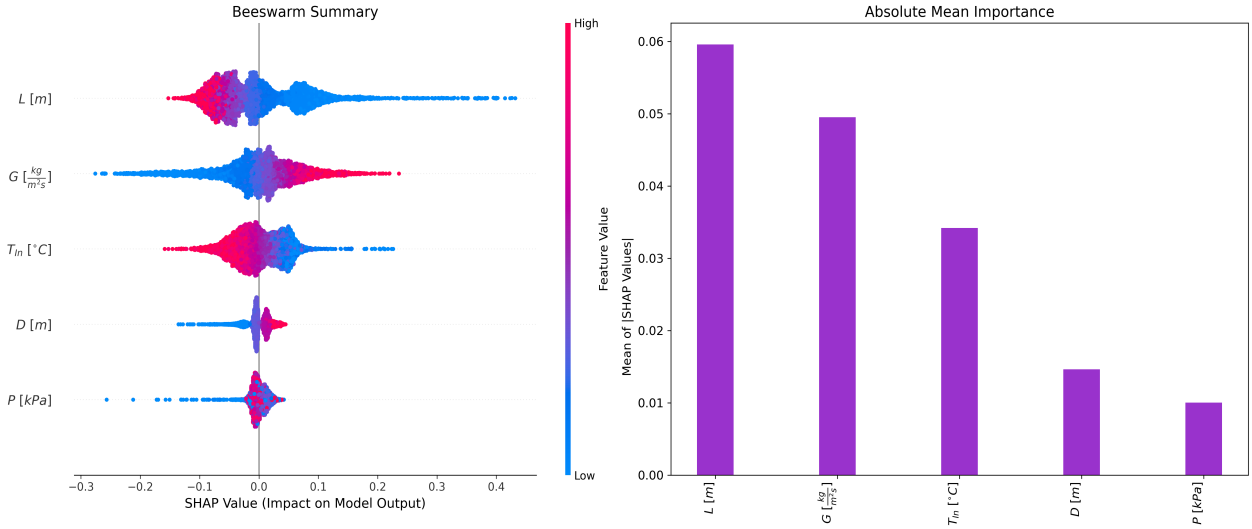

Figure 4. Explainability results using Integrated Gradients for the best CHF FNN model with all features *except* outlet quality (X_e).

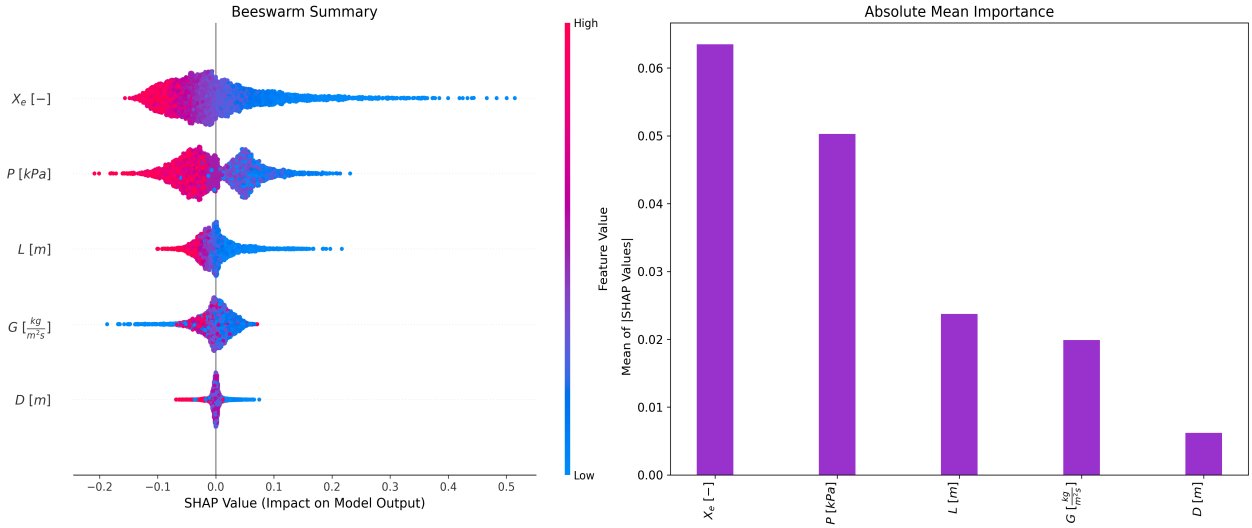

Figure 5. Explainability results using Integrated Gradients for the best alternative CHF FNN model: all features *except* inlet temperature (T_{in}).

explainability methods—that it is unable to consider correlations between features when ranking feature importances on model output. We can attribute the shifting in parameter rankings in this example to the moderately strong correlation (0.61) between T and P as described in Section 2. In other words, SHAP can treat a feature to be less important if another strongly correlated feature is present.

These results indicate that ad hoc explainability methods, such as SHAP, tend to reflect statistical properties of the AI/ML model rather than uncovering the underlying physics of the model. While these explanations may sometimes align with physical principles, when multiple features interact, the complexity of their interactions may become too challenging for SHAP methods to fully disentangle. This complexity does not always reflect a physical process but instead may highlight a well-known phenomenon in probability and

statistics called “Simpson’s paradox”.

5.1. Random Forest Explainability

To validate the various SHAP results, we employed a mean decrease in impurity (MDI) and permutation importance (PI) algorithm on a random forest model ($R^2 = 0.993$) that performed nearly as well as our FNN ($R^2 = 0.997$) [10]. These feature rankings, found with an entirely different method, generally agree on a feature importance ranking of L, G, T, D, P for our model considering all features except outlet quality (X_e). This provides some level of confidence in the accuracy of the SHAP methodology, though the industry will require further metrics on model interpretability in addition to explainability to have a comprehensive understanding, and thereby, trust, of an FNN’s prediction.

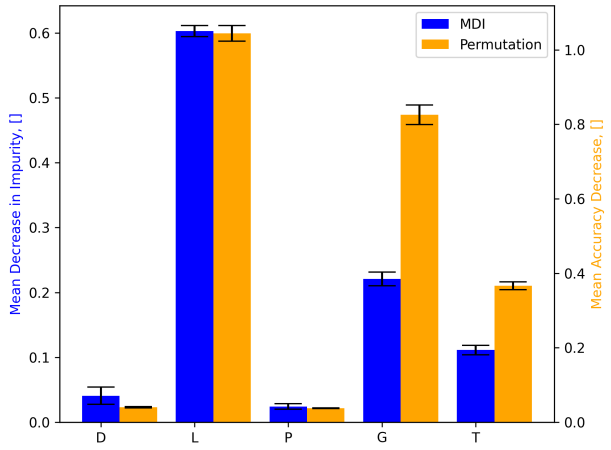

Figure 6. Feature importances using permutation importances (PI) and mean decrease in impurity (MDI) for a random forest model will all features except outlet quality (X_e).

6. CONCLUSION

This study marks our initial effort to incorporate ad hoc explainable AI methods within an automated AI/ML framework for the nuclear industry. Using techniques like DeepLIFT, Integrated Gradients (IG), SHAP, and random forests, we observed consistent feature ranking in predicting CHF, with inlet mass flux and channel length emerging as the primary influences. While these results are just a glimpse into understanding the full scope of *how* an AI/ML model makes a prediction, determining an accurate feature importance ranking for these models is a necessary first step in uncovering more detailed explainability. Future work will focus on a more granular explainability approach, examining the attribution of each layer and node in the neural network architecture to understand their specific impact on the output prediction.

ACKNOWLEDGEMENTS

This work is funded by the U.S. Nuclear Regulatory Commission’s University Nuclear Leadership Program for Research and Development (Award: 31310024M0013). Additionally, the second author (N. Panczyk) received sponsorship through the National Science Foundation’s Graduate Research Fellowship Program (Grant Number: DGE 2241144).

REFERENCES

[1] U.S. Nuclear Regulatory Commission. "NUREG-2261, "Artificial Intelligence Strategic Plan, Fiscal Years 2023-2027"."

[2] D. Price, M. I. Radaideh, D. O'Grady, and T. Kozlowski. "Advanced BWR criticality safety part II: Cask criticality, burnup credit, sensitivity, and uncertainty analyses." *Progress in Nuclear Energy*, **volume 115**, pp. 126–139 (2019).

[3] M. I. Radaideh, D. Price, and T. Kozlowski. "Criticality and uncertainty assessment of assembly misloading in BWR transportation cask." *Annals of Nuclear Energy*, **volume 113**, pp. 1–14 (2018).

[4] Saranya A. and Subhashini R. "A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends." *Decision Analytics Journal*, **volume 7**, p. 100230 (2023). URL https://www.sciencedirect.com/science/article/pii/S277266222300070X.

[5] J. Liu, Q. Zhang, and R. Macián-Juan. "Enhancing interpretability in neural networks for nuclear power plant fault diagnosis: A comprehensive analysis and improvement approach." *Progress in Nuclear Energy*, **volume 174**, p. 105287 (2024). URL https://www.sciencedirect.com/science/article/pii/S0149197024002373.

[6] J. H. Park, H. S. Jo, S. H. Lee, S. W. Oh, and M. G. Na. "A reliable intelligent diagnostic assistant for nuclear power plants using explainable artificial intelligence of GRU-AE, LightGBM and SHAP." *Nuclear Engineering and Technology*, **volume 54**(4), pp. 1271–1287 (2022). URL https://www.sciencedirect.com/science/article/pii/S1738573321006082.

[7] C. Walker, V. Agarwal, L. Lin, A. Hall, R. Hill, R. Boring, PhD, T. Mortenson, and N. Lybeck. "Explainable Artificial Intelligence Technology for Predictive Maintenance." Technical Report INL/RPT–23-74159-Rev000, 1998555 (2023). URL https://www.osti.gov/servlets/purl/1998555/.

[8] D. Groeneveld. "Critical Heat Flux Data Used to Generate the 2006 Groeneveld Lookup Tables." Technical Report NUREG/KM-0011, U.S. Nuclear Regulatory Commission (2019). URL https://www.nrc.gov/reading-rm/doc-collections/nuregs/knowledge/km0011/index.html.

[9] J.-M. L. Corre, G. Delipei, X. Wu, and X. Zhao. "Benchmark on Artificial Intelligence and Machine Learning for Scientific Computing in Nuclear Engineering. Phase 1: Critical Heat Flux Exercise Specifications." *NEA Working Papers* (2024).

[10] P. A. Myers, C. Craig, J. Cooper, V. Joynt, and M. I. Radaideh. "Pymaise: A Python Platform for Automatic Machine Learning Development and Benchmarking for Nuclear Engineering Applications." *Available at SSRN 4924326*.

[11] S. Lundberg. "A unified approach to interpreting model predictions." *arXiv preprint arXiv:170507874* (2017).

[12] A. Shrikumar, P. Greenside, and A. Kundaje. "Learning important features through propagating activation differences." In *International conference on machine learning*, pp. 3145–3153. PMlR (2017).

[13] M. Sundararajan, A. Taly, and Q. Yan. "Axiomatic attribution for deep networks." In *International conference on machine learning*, pp. 3319–3328. PMLR (2017).