Assessment of Practical Performance Gains in AI-based CHF Prediction

Mooneon Lee*, Juhyung Lee, Dae-Hyun Hwang, Hyouk Kwon
Korea Atomic Energy Research Institute (KAERI), 111, Daedeok-daero 989beon-gil, Yuseong-gu, Daejeon
**Corresponding author: melee@kaeri.re.kr*

***Keywords :** Critical Heat Flux, Artificial Intelligence, Generalization Performance, Lookup-Table

1. Introduction

In water-cooled nuclear reactors, critical heat flux (CHF) is a key design parameter that directly affects safety margins and economic efficiency. Due to the physical complexity of the boiling phenomenon, the accurate prediction of CHF remains a major engineering challenge. Conventional data-driven CHF prediction approaches used in reactor design are typically based on empirical correlations and the Look-up Table (LUT) method[1].

Recently, the prediction of CHF using artificial intelligence (AI) has been actively investigated, largely accelerated by the public release of the 2006 CHF LUT database and the OECD/NEA benchmark program[2]. Most recent AI-based studies report significant performance improvements, often achieving error reductions several times greater than the predictive performance of the LUT[3-5]. However, these overwhelming gains are frequently the result of an unfair comparison stemming from a lack of domain knowledge regarding input parameter structures, as well as the inherent vulnerability of deep learning models to overfitting.

Therefore, this study proposes a robust framework to fairly compare and evaluate the accuracy of AI-based CHF prediction against the conventional LUT. Through this framework, we quantitatively assess the practical and objective performance improvements that can be achieved by adopting AI methodologies.

2. Domain Knowledge : Input parameter effect

In flow boiling CHF experiments within circular tubes, tube diameter (D), heated length (L), system pressure (P), mass flux (G), and inlet subcooling (dHin) are generally used as independent variables, with CHF measured as the outcome. The dependent variable, critical quality (Xcr), is calculated using the heat balance equation.

Input parameter combinations for prediction models are primarily classified into three types: 1) Local conditions (D, P, G, Xcr), which use only parameters at the local CHF occurrence location, 2) Exit conditions (D, P, G, Xcr, L); and 3) Inlet conditions (D, P, G, dHin, L). The amount and reliability of information vary significantly depending on these input conditions, directly determining the achievable prediction accuracy. The LUT method can be applied in two forms: the Direct Substitution Method (DSM) using local conditions, and the Heat Balance Method (HBM) using inlet conditions. The prediction error between these two methods can differ by up to a factor of six[1].

Despite this, many recent studies[3-5] have trained their models using information-rich inlet or exit conditions, yet compared their results against the LUT (DSM) which relies on information-poor local conditions, thereby claiming overwhelming performance gains. To prevent this overestimation, this study strictly restricted the input conditions of the AI model to match those of the LUT (Inlet vs. Inlet, Local vs. Local), eliminating the unfair advantage caused by mismatched input variables.

3. Method

To minimize the structural bias of the AI model and independently evaluate the impact of the optimization pipeline, a Multilayer Perceptron (MLP) architecture was adopted. To prevent an exponential increase in the hyperparameter search space, a Block-wise MLP architecture—combining a linear layer, ReLU activation, and dropout into a single block—was utilized.

Model optimization was conducted by gradually expanding the structural size (depth and width) of the network. Within each stage, Bayesian Optimization and Grid Search were sequentially applied, followed by batch size optimization. Furthermore, to analyze the impact of data splitting strategies on model performance, two cross-validation (CV) techniques were employed: 1) the conventional Random Split 5-Fold CV, and 2) the Source-based Group K-Fold CV, which splits data by literature reference to control for facility bias.

4. Effect of Input Conditions and Model Size

Initially, the predictive performance (RMSPE) was analyzed across different model sizes under the conventional Random Split condition. As the number of parameters increased, the prediction accuracy continuously improved and eventually converged. When the AI model was sufficiently large, it demonstrated significantly improved predictive performance compared to the LUT using identical input conditions.

However, the "several times error reduction" claimed in previous literature was drastically reduced when the input conditions were strictly controlled. This proves that a substantial portion of the dramatic performance improvements reported in existing studies was an unfair benefit derived from the difference in the amount of input information (Inlet/Exit vs. Local), rather than the superiority of Deep Learning Model. Nevertheless,

within the Random Split environment, the large AI models still recorded an RMSPE approximately 30~40% lower than that of the LUT.

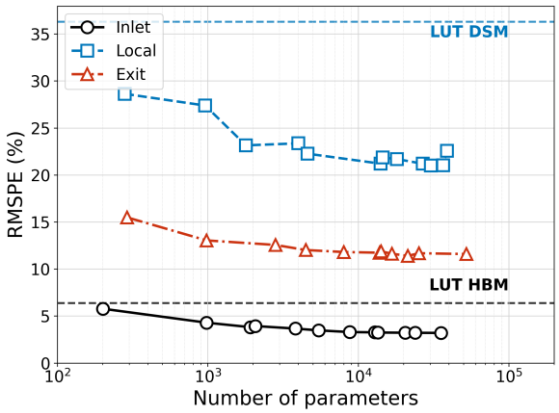

Fig. 1. Prediction performance (RMSPE) of AI models as a function of model parameters under random 5-fold cross-validation.

5. Effect of Systematic Bias

Generally, deep learning-based AI models can memorize minute features within data owing to their massive number of parameters. Due to the nature of CHF experimental data, data points produced from the same experimental facility often share completely identical tube diameters and heated lengths, with pressure and mass flux composed of repeated combinations. Consequently, large AI models are highly prone to overfitting, simply memorizing the systematic bias—such as specific measurement errors or heat loss characteristics inherent to a facility—rather than learning general physical relationships. The Random Split method neglects this data leakage, as data from the same source is mixed into both training and test sets.

To eliminate this memorization effect and evaluate the true generalization performance of the AI models, a Source-based Group 5-Fold CV was performed, which strictly separates data based on the experimental source. The results showed a completely different trend compared to the Random Split method. When the model predicted data from an experimental facility excluded from training, the error rate increased sharply, resulting in severe performance degradation comparable to that of the LUT.

Particularly, larger models with more parameters exhibited a greater decline in predictive performance under new experimental conditions. Conversely, optimal generalization performance was observed in small models with around 1,000 parameters. This suggests that excessive parameter expansion induces overfitting to experimental bias. To develop a robust model applicable to actual engineering fields, strict model size constraints and an appropriate evaluation framework tailored to data characteristics are essential.

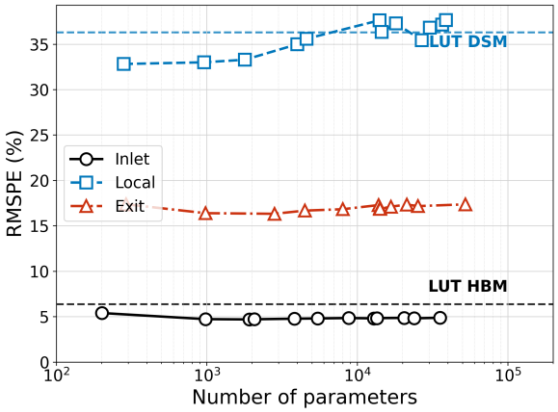

Fig. 2. Prediction performance (RMSPE) of AI models as a function of model parameters under source-based group 5-fold cross-validation.

6. Conclusion

While numerous recent studies have reported that AI models provide groundbreaking CHF prediction performance compared to the conventional LUT, our findings reveal that these results are largely due to overfitting caused by unfair input condition settings and random data splitting.

Under a fair evaluation framework that controls input variables and applies Source-based CV, the perceived superiority of AI was significantly offset. Excessively large AI models even demonstrated worse predictive capability than the LUT in unseen experimental environments (unseen sources).

In conclusion, to apply AI to fields requiring high reliability, such as nuclear reactor design, merely reducing numerical errors is insufficient. Physically valid variable control, model optimization for overfitting prevention, and conservative definitions of the applicability domain must precede deployment.

ACKNOWLEDGEMENTS

This work was supported by an Innovative Small Modular Reactor Development Agency grant funded by the Korean Government (Ministry of Science and ICT, MSIT) (No. RS-2023-00257680).

REFERENCES

[1] D. Groeneveld et al., “The 2006 CHF look-up table,” Nucl. Eng. Des., vol. 237, no. 15–17, pp. 1909–1922, 2007.
[2] J.-M. Le 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,” NEAWKP20231 NEA Work. Pap. OECD Publ., 2024.
[3] E. Helmryd Grosfilley, G. Robertson, J. Soibam, and J.-M. Le Corre, “Investigation of machine learning regression techniques to predict critical heat flux over a large parameter space,” Nucl. Technol., pp. 1–15, 2024.
[4] W. Zhou, S. Miwa, H. Wang, and K. Okamoto, “Assessment of the state-of-the-art AI methods for critical heat

flux prediction," Int. Commun. Heat Mass Transf., vol. 158, p. 107844, 2024.
[5] I. Ahmed, I. Gatti, and E. Zio, "Optimized ensemble of neural networks for the prediction of critical heat flux," Nucl. Eng. Des., vol. 439, p. 114111, 2025.