A Comparative Performance Study of Deep Learning-Based Surrogate Models for APR1400 Core Design: Prediction of Boron Concentration and Power Peaking Factor (Fq)

Yongrae Kim[a*], Jinsun Kim[a], Gonghoon Bae[a], Changhee Lee[a], Seongho Park[a], Jonghoon Kim[a], Junggyu Lee[a]
[a]*KEPCO Nuclear Fuel, 242, Daedeok-daero 989 beon-gil, Yuseong-gu, Daejeon 34057, Republic of Korea*
**Corresponding author: yongraekim@knfc.co.kr*

***Keywords :** Core Design, Deep Learning, Surrogate Model, Boron Concentration, Power Peaking Factor (Fq)*

1. Introduction

In the core design process of pressurized water reactors (PWRs), the search for loading patterns is a representative combinatorial optimization problem that requires repeated depletion calculations, and has been pointed out as a major bottleneck in core design optimization [1]. This optimization problem has traditionally been addressed using meta-heuristic techniques such as genetic algorithms (GA) and particle swarm optimization (PSO) [2], and more recently, active attempts have been reported to combine machine learning techniques with meta-heuristics to improve search efficiency [3]. Meanwhile, the critical boron concentration and the Power Peaking Factor (Fq) are key indicators used to assess core safety and operating margin, and are computed at every burnup step using detailed core analysis codes. However, because such detailed analyses are computationally expensive, the need for surrogate models that satisfy both computational speed and accuracy has been raised at the optimization stage, where numerous loading pattern candidates must be evaluated rapidly.

Recently, numerous studies have reported the use of deep learning techniques to predict boron concentration and power distribution from core state variables [4]. In particular, a prior study [5] that predicted the power peaking factor (Fq) and boron concentration from the infinite multiplication factor (k-inf) and burnup for a core analogous to the Angra-1 plant addresses a problem setting similar to that of the present study, and was therefore referenced for comparison. MLP, CNN [6], and Transformer [7] architectures are known for learning nonlinear relationships through fully connected layers, extracting spatial features, and capturing sequential dependencies, respectively. Accordingly, these three architectures were selected for comparison in this study.

The objective of this study is to develop MLP-, CNN-, and Transformer-based surrogate models that predict the boron concentration and Fq of the APR1400 core from core state variables (k-inf, burnup), and to quantitatively compare and analyze the prediction performance of the three models for both output variables. Rather than validating the effectiveness of a specific improvement strategy, this paper focuses on sharing the experimental results obtained at the current stage and analyzing the observed phenomena — in particular, the training instability observed in the Fq prediction process — in order to suggest directions for future research.

2. Methods and Results

2.1 Dataset Construction

In this study, loading patterns randomly configured to reflect the experience of core designers were used, and a depletion calculation was performed for each loading pattern using the core analysis code ASTRA. The APR1400 core was modeled as a quarter (1/4) core using its symmetry and represented as a 9×9 assembly grid, and the infinite multiplication factor (k-inf) and burnup of each fuel assembly location on this grid were extracted at each burnup step. Each loading pattern consists of 23 burnup steps, yielding a total of 22,977 data points (999 loading patterns × 23 burnup steps).

Because data from consecutive burnup steps within the same loading pattern can be strongly correlated, splitting the dataset randomly at the individual (step) level risks data leakage. To prevent this, the dataset was split at the loading-pattern (case) level rather than at the individual data level, using a train : validation : test ratio of 7 : 1.5 : 1.5. Table I summarizes the number of loading patterns and the resulting data counts for each split.

Table I: Dataset Composition (Train/Validation/Test)

Dataset	Loading Patterns	Steps	Data Count
Train	699	23	16,077
Validation	150	23	3,450
Test	150	23	3,450
Total	999	23	22,977

The input variables (k-inf, burnup) were normalized before being fed into the models. To prevent information from the validation and test sets from leaking into training, the mean and standard deviation used for normalization were computed using only the training set. Grid positions in the 9×9 layout that do not correspond to an actual fuel assembly were excluded from this calculation so that they would not distort the statistics. The output variables (critical boron

concentration, Fq) were likewise normalized using Z-score normalization based on the mean and standard deviation of the training set.

2.2 Model Architecture and Training Configuration

MLP, CNN, and Transformer architectures were each used to build a dedicated boron-concentration prediction model and a dedicated Fq prediction model. The detailed configuration of each model is as follows, and the architecture of each model is illustrated in Fig. 1.

MLP: consists of four fully-connected layers (three hidden layers and one output layer), with hidden layer sizes of 256, 128, and 64, respectively, and uses the ReLU activation function. The 9×9×2 input (k-inf, burnup) is flattened into a 162-dimensional vector (Fig. 1(a)).

CNN: takes the 9×9 grid as a 2-channel (k-inf, burnup) map and consists of two convolutional layers (3×3 kernel, 32 channels, padding=1) with no pooling layers, followed by a two-layer fully-connected head (sizes 64 and **1**), using the ReLU activation function (Fig. 1(b)).

Transformer: consists of two encoder layers with d_model=32, 4 attention heads, and a feed-forward network (FFN) dimension of 128 (dropout rate 0.1). The 9×9 grid is divided into patches of size 1, yielding 81 tokens, to which a learned positional embedding is applied; the final output is obtained via mean pooling (Fig. 1(c)).

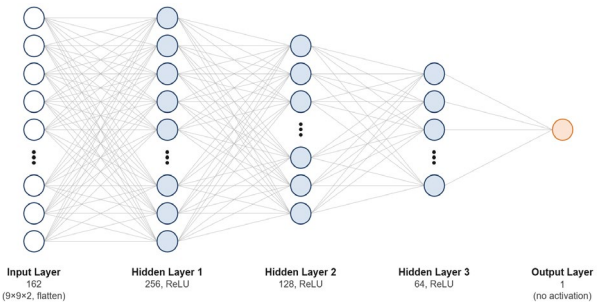


Fig. 1(a). Architecture of the MLP model.

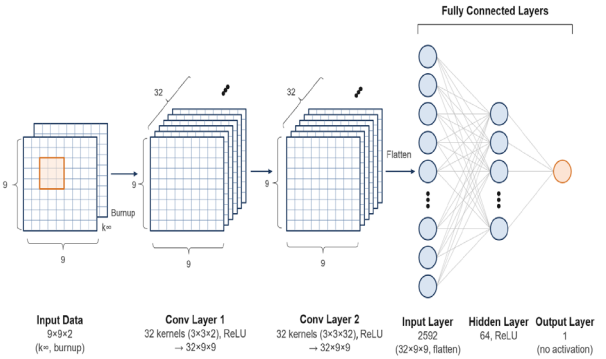


Fig. 1(b). Architecture of the CNN model.

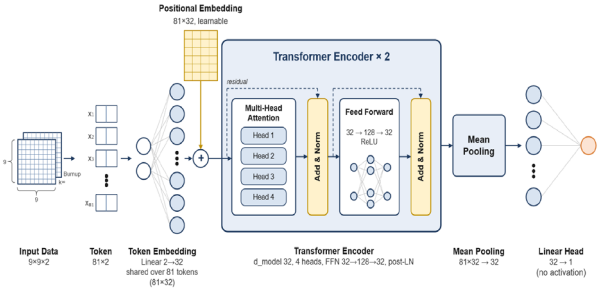


Fig. 1(c). Architecture of the Transformer model.

The loss function for all models is the Mean Squared Error (MSE), defined as in Eq. (1) (N is the number of data, y_i is the true value, and $\hat{y}_i$ is the predicted value).

$$MSE = (1/N)\,\Sigma\,(y_i - \hat{y}_i)^2 \qquad (1)$$

The Adam optimizer was used without weight decay. The learning rate was set to 1e-3, the batch size to 64, and the total number of training epochs to 500. Instead of early stopping, the weights corresponding to the minimum validation loss observed over the full 500-epoch training run were adopted as the final model. To assess the dependence of the results on the random seed, each model was trained three times with different random seeds (0, 1, and 2), which govern weight initialization, mini-batch ordering, and dropout, while the train/validation/test split was kept fixed; deterministic GPU operations were enforced so that runs with the same seed are exactly reproducible. The numbers of trainable parameters are 82,945 (MLP), 175,873 (CNN), and 28,129 (Transformer); the model capacities were not matched, and the Transformer has the fewest parameters.

2.3 Evaluation Metrics

Because the actual boron-concentration values can include values near zero, applying relative error can cause severe distortion (the relative error diverges sharply near zero); the absolute error (ppm) was therefore used as the evaluation metric. In contrast, since Fq values are sufficiently far from zero, the relative error (%) was used as the evaluation metric.

For the i-th test sample, the absolute error of the boron concentration (AE_i) and the relative error of Fq (RE_i) are defined as in Eqs. (2) and (3), respectively. The average (Avg) and maximum (Max) errors reported in Tables II–IV are the mean and maximum of these values over all N test samples. Each value is reported as the mean ± standard deviation over the three random seeds.

$$AE_i = |\hat{y}_i - y_i| \ \text{(ppm)} \qquad (2)$$
$$RE_i = |\hat{y}_i - y_i| \,/\, |y_i| \times 100\ (\%) \qquad (3)$$

In addition, because the boron concentration varies over a wide range, the absolute error was also expressed as a ratio to the range of the boron concentration in the

test dataset (RR_i), as defined in Eq. (4), where R is the difference between the maximum (1,513.7 ppm) and minimum (−203.6 ppm) values in the test dataset (R = 1,717.3 ppm). The results are summarized in Table III.

$$RR_i = AE_i / R \times 100\ (\%) \quad (4)$$

2.4 Results

Tables II–IV summarize the model-wise prediction errors on the test dataset (3,450 samples) as the mean ± standard deviation over the three random seeds. Fig. 2 shows the boron-concentration absolute error and Fig. 3 shows the Fq relative error for each test sample, with the results of the three models (seed 0) overlaid on a common axis for direct comparison.

Table II: Boron Absolute Error (Test Dataset, Mean ± Std over Three Seeds)

Model	Avg (ppm)	Max (ppm)
MLP	14.79 ± 0.62	79.67 ± 14.25
CNN	12.11 ± 1.12	74.82 ± 17.72
Transformer	3.64 ± 0.10	22.35 ± 2.27

Table III: Ratio of Boron Absolute Error to the Test-Set Range (Test Dataset, Mean ± Std over Three Seeds)

Model	Avg (%)	Max (%)
MLP	0.86 ± 0.04	4.64 ± 0.83
CNN	0.71 ± 0.07	4.36 ± 1.03
Transformer	0.21 ± 0.01	1.30 ± 0.13

Table IV: Fq Relative Error (Test Dataset, Mean ± Std over Three Seeds)

Model	Avg (%)	Max (%)
MLP	5.41 ± 0.16	29.35 ± 4.18
CNN	4.10 ± 0.09	22.44 ± 3.12
Transformer	5.66 ± 0.30	33.78 ± 3.16

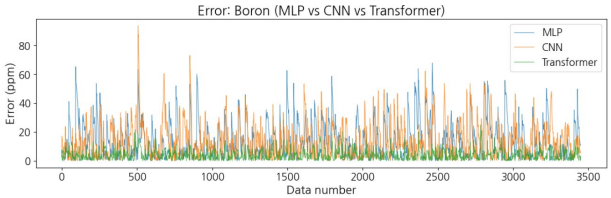


Fig. 2. Boron absolute error for each test sample (MLP, CNN, Transformer; seed0).

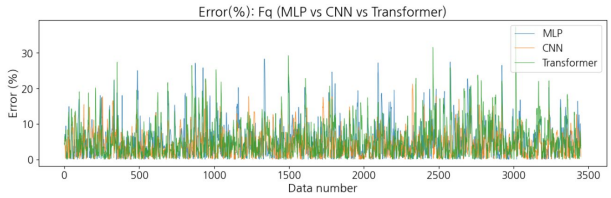


Fig. 3. Fq relative error for each test sample (MLP, CNN, Transformer; seed 0).

3. Discussion and Conclusion

For the boron-concentration prediction, all three models showed stable convergence of the train/validation loss after the early stage of training. In terms of absolute error, the average ranged from 3.64 to 14.79 ppm and the maximum from 22.35 to 79.67 ppm (means over three seeds), indicating a relatively low level of absolute error across all three models. Relative to the range of the boron concentration in the test dataset (1,717.3 ppm), the average errors correspond to 0.21-0.86% and the maximum errors to 1.30-4.64%, indicating that the errors of all three models are small compared with the overall variation range. Among the three models, the Transformer reduced the average error to about one-quarter to one-third of that of the other two models, showing a clear performance improvement, and this ranking was identical for all three seeds. The seed-to-seed standard deviation of the average error (0.10-1.12 ppm) was much smaller than the differences between the models, and because the Transformer has the fewest parameters, this improvement is attributed to its architecture rather than to model capacity.

In contrast, while the average relative error for the Fq prediction appears low (4.10-5.66%), the maximum relative error (22.44-33.78%) is considerably larger than the average, indicating that the models effectively failed to predict Fq in a non-negligible number of cases. This is likely because the input variables currently used (k-inf, burnup) do not have a sufficient correlation with Fq, and the average error alone should not be interpreted as indicating good prediction performance. Furthermore, the small performance difference among the MLP, CNN, and Transformer models — despite their different architectures — suggests that the bottleneck to further performance improvement lies not in model capacity but in the type of input data used. The CNN showed the lowest average Fq error for all three seeds, but the differences among the models were small, and the seed-averaged maximum errors of all models exceeded 20%. Meanwhile, the Transformer, which clearly outperformed the other models for boron, showed no such advantage for Fq. This is likely because the boron concentration is a core-wide quantity that can be sufficiently described by the 2D inputs, whereas Fq is a local quantity that also depends on the axial power distribution, which is not contained in the 2D inputs. Examining the train/validation loss curves, the MLP exhibited mild overfitting, whereas the CNN and Transformer repeatedly showed large oscillations in validation loss throughout training; this too appeared consistently regardless of model architecture, further supporting the possibility that the current input variables alone cannot sufficiently explain Fq (e.g., spatial information such as assembly-wise power distribution).

In summary, the surrogate models developed in this study achieved a practically useful level of prediction accuracy for boron concentration. For Fq, however, the average error alone is not sufficient to judge performance; considering the maximum error and the

training instability together, the current input variables appear insufficient to predict Fq adequately. In particular, the fact that the three models achieved similar performance despite their different architectures suggests that the bottleneck to improvement lies in the type of input data rather than in the model itself. Rather than verifying the effectiveness of a specific improvement strategy for this phenomenon, this study aimed to organize and share the experimental results and observed phenomena obtained thus far. In future work, we plan to examine expanding the input variables (e.g., utilizing assembly-wise power distribution information) and redefining the output variable (e.g., comparing the prediction performance of Fq versus the radial peaking factor Fr). In addition, we plan to extend the surrogate modeling approach presented in this study to the core design of Small Modular Reactors(SMRs).

ACKNOWLEDGEMENTS

This work was supported by the Innovative Small Modular Reactor Development Agency grant funded by the Korea Government(MCEE) (No. RS-2024-00407975).

REFERENCES

[1] P.J. Turinsky, Nuclear Fuel Management Optimization: A Work in Progress, Nuclear Technology, Vol.151, No.1, pp.3-8, 2005.
[2] Z. Li, J. Wang, M. Ding, A Review on Optimization Methods for Nuclear Reactor Fuel Reloading Analysis, Nuclear Engineering and Design, Vol.397, 111950, 2022.
[3] M. Karimi-Mamaghan, M. Mohammadi, P. Meyer, A.M. Karimi-Mamaghan, E.-G. Talbi, Machine Learning at the Service of Meta-heuristics for Solving Combinatorial Optimization Problems: A State-of-the-Art, European Journal of Operational Research, Vol.296, No.2, pp.393-422, 2022.
[4] J.L. Francois, A. Castillo, J.J. Ortiz-Servin, C. Martin del Campo, J.L. Montes-Tadeo, R. Perusquia, Artificial Intelligence Techniques in In-Core Nuclear Fuel Management: Evolution and Contributions from Mexican Research, SSRN preprint, 2026. https://ssrn.com/abstract=6359343
[5] C.O. Pastrana Orejuela, F.I. Pinillos Nieto, V.H. Cabral Pinheiro, M.C. Santos, A.X. Da Silva, A.M.M. Lima, Deep Learning Techniques for Optimization of Fuel Reloading in Pressurized Water Reactor (PWR) Type Nuclear Reactors, Progress in Nuclear Energy, Vol.190, 105952, 2026.
[6] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-Based Learning Applied to Document Recognition, Proceedings of the IEEE, Vol.86, No.11, pp.2278-2324, 1998.
[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, Advances in Neural Information Processing Systems (NeurIPS), Vol.30, 2017.