A Conceptual Study on Surrogate Modeling of Nuclear Power Plant using KAN

Young Ho Chae [a*], Seo Ryong Koo [a]
[a]*Korea Atomic Energy Research Institute, 111 Daedeok-daero 989 beon-gil, Daejeon, 34057*
[*]*Corresponding author: yhchae@kaeri.re.kr*

Keywords : Surrogate Modeling, Kolmogorov-Arnold Network, Nuclear Power Plant, Physics-Informed

1. Introduction

Dynamic Probabilistic Risk Assessment (DPRA) has become an essential methodology for evaluating time-dependent accident scenarios in nuclear power plants (NPPs). Unlike traditional static PRA, DPRA captures the temporal evolution of accident sequences, control system responses, and operator actions, providing realistic representations of plant behavior under various failure conditions. However, the computational burden of high-fidelity thermal-hydraulic codes such as MARS, RELAP5, and TRACE presents a significant bottleneck. A single transient simulation can require several hours, making Monte Carlo-based uncertainty quantification computationally prohibitive for real-time risk assessment applications.

This computational challenge has motivated extensive research into surrogate modeling techniques. Traditional approaches have employed Gaussian Process regression and Polynomial Chaos Expansion to approximate complex simulator outputs. More recently, deep learning methods have shown promising results in capturing nonlinear NPP dynamics. For instance, Antonello et al. [1] developed a neural network-based surrogate model. Similarly, Lu et al. [2] applied neural networks to model KLT-40S Nuclear reactor.

Despite these advances, a critical limitation remains: the lack of interpretability in black-box models poses significant challenges for regulatory acceptance in safety-critical applications. While Physics-Informed Neural Networks (PINNs) incorporate domain knowledge into the learning process, they often impose overly restrictive constraints that limit flexibility when dealing with complex multi-physics phenomena and discontinuous events.

To address this interpretability gap, this paper proposes a novel surrogate modeling framework based on Kolmogorov-Arnold Networks (KAN). Unlike traditional neural networks that use fixed activation functions, KANs employ learnable univariate functions represented as B-splines, motivated by the Kolmogorov-Arnold representation theorem. This architectural choice enables the discovery of symbolic relationships between variables, potentially revealing underlying physical laws through network pruning and simplification.

Our proposed framework adopts a snapshot-based learning approach, where the KAN model is trained on pairs of input conditions (initial plant state and malfunction specifications) and corresponding thermal-hydraulic parameters from MARS simulations. The framework incorporates specialized techniques for handling discontinuous phenomena through input augmentation with derivative terms and physics-informed regularization. This conceptual study presents the theoretical foundation and implementation strategy for applying KAN to NPP surrogate modeling, addressing both the computational efficiency requirements of DPRA and the interpretability demands of nuclear safety analysis.

2. Kolmogorov-Arnold Networks

2.1 Mathematical Foundation

Kolmogorov-Arnold Networks [3] (KANs) represent a fundamentally different approach to neural network architecture, inspired by the Kolmogorov-Arnold representation theorem. This theorem, proven by Andrey Kolmogorov and Vladimir Arnold in the 1950s, states that any multivariate continuous function $f: [0,1]^n \to \mathbb{R}$ can be expressed as a finite composition of continuous univariate functions and addition:

$$f(x_1, \ldots, x_n) = \sum_{q=0}^{2n} \Phi_q \left(\sum_{p=1}^{n} \phi_{q,p}\,(x_p) \right)$$

where $\phi_{q,p}: [0,1] \to \mathbb{R}$ and $\Phi_q: \mathbb{R} \to \mathbb{R}$ are continuous univariate functions. This decomposition suggests that complex multivariate relationships can be represented through combinations of simpler one-dimensional transformations, providing the theoretical foundation for KAN architectures.

2.2 Network Architecture

Unlike Multi-Layer Perceptrons (MLPs) that employ fixed activation functions (ReLU, sigmoid, tanh) with learnable weights, KANs place learnable activation functions on network edges while eliminating traditional weight matrices. In a standard MLP, the transformation between layers is expressed as:

$$\mathbf{x}^{(l+1)} = \sigma(\mathbf{W}^{(l)}\mathbf{x}^{(l)} + \mathbf{b}^{(l)})$$

where $\mathbf{W}^{(l)}$ represents the weight matrix, $\mathbf{b}^{(l)}$ the bias vector, and σ a fixed activation function.

In contrast, a KAN layer performs the transformation:

$$x_j^{(l+1)} = \sum_{i=1}^{n_l} \phi_{i,j}^{(l)} (x_i^{(l)})$$

where $\phi_{i,j}^{(l)}$ are learnable univariate functions connecting neuron i in layer l to neuron j in layer $l+1$. Each function $\phi_{i,j}$ is parameterized using B-spline basis functions:

$$\phi_{i,j}(x) = \sum_{k=1}^{K} c_{i,j,k} \cdot B_k(x)$$

where $c_{i,j,k}$ are learnable coefficients and $B_k(x)$ are B-spline basis functions defined on a grid. The B-spline representation provides several advantages: local support for efficient computation, smooth derivatives for stable training, and adaptive grid refinement for capturing fine-scale features.

2.3 Training and Optimization

The training process for KANs involves optimizing the B-spline coefficients to minimize a task-specific loss function. For regression tasks in surrogate modeling, the primary loss component is the mean squared error:

$$\mathcal{L}_{data} = \frac{1}{N} \sum_{i=1}^{N} \| \mathbf{y}_i - \hat{\mathbf{y}}_i \|^2$$

To promote interpretability and prevent overfitting, additional regularization terms are incorporated:

$$\mathcal{L}_{total} = \mathcal{L}_{data} + \lambda_1 \mathcal{L}_{smooth} + \lambda_2 \mathcal{L}_{sparse}$$

The smoothness penalty $\mathcal{L}_{smooth}$ encourages simpler activation functions by penalizing the second derivatives:

$$\mathcal{L}_{smooth} = \sum_{i,j} \int |\phi_{i,j}''(x)|^2 dx$$

The sparsity penalty $\mathcal{L}_{sparse}$ promotes network pruning by encouraging small magnitudes for less important connections:

$$\mathcal{L}_{sparse} = \sum_{i,j} \| \phi_{i,j} \|_1$$

2.4 Interpretability Mechanisms

The interpretability of KANs arises from three key mechanisms:

1. Symbolic Extraction: After training, the learned B-spline functions can be analyzed to identify their mathematical form. Simple functions like linear relationships, quadratic terms, or exponential decay can be recognized through pattern matching or symbolic regression techniques.

2. Network Pruning: Connections with small activation function magnitudes can be removed without significant accuracy loss, revealing the essential computational graph. The pruning threshold ϵ is chosen such that:

$$|\phi_{i,j}(x)| < \epsilon \quad \forall x \in [x_{min}, x_{max}]$$

$\Rightarrow$ remove connection

3. Dimensional Analysis: The univariate nature of activation functions allows direct examination of how each input variable transforms through the network, enabling physical interpretation of learned relationships.

3. Proposed KAN-based Framework for NPP Surrogate Modeling

3.1 Framework Overview

The proposed framework integrates Kolmogorov-Arnold Networks with domain-specific techniques to create interpretable surrogate models for nuclear power plant thermal-hydraulic analysis (Fig.1). The framework consists of four main components: (1) Data Generation, (2) Data preprocessing, (3) KAN training phase, (4) Interpretability analysis

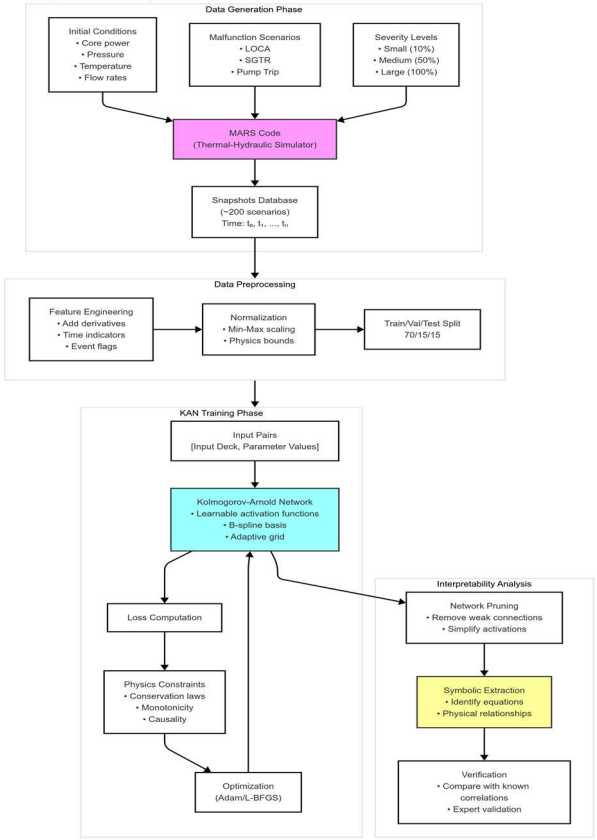

Fig. 1. KAN-based Surrogate modeling framework

3.2 Data Generation Strategy

3.2.1 Scenario Matrix Design

The training data is generated through systematic MARS code simulations covering diverse operational and accident conditions. The scenario matrix is constructed using three dimensions:

Initial Conditions: Plant operational states are sampled across the normal operations

Malfunction Types: Representative accident scenarios are selected based on design basis events and beyond design basis considerations:

- ✓ Loss of Coolant Accident (LOCA) with varying break locations
- ✓ Steam Generator Tube Rupture (SGTR)
- ✓ Reactor Coolant Pump trip
- ✓ etc.

Severity Levels: Each malfunction is simulated at multiple severity levels to capture the full spectrum of plant response.

3.2.2 Snapshot-based Data Structure

Rather than treating the problem as time-series prediction, the framework adopts a snapshot-based approach where each training sample consists of:

$$\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i) | i = 1, \ldots, N\}$$

where the input $\mathbf{x}_i$ comprises:

Input deck parameters: $\mathbf{d} \in \mathbb{R}^{n_d}$ (initial conditions, malfunction specifications)

Query point: $\mathbf{q} \in \mathbb{R}^{n_q}$ (time stamp, spatial location identifier)

The output $\mathbf{y}_i$ contains the corresponding thermal-hydraulic parameters:

- ✓ Temperature at specified locations
- ✓ Pressure in major components
- ✓ Mass flow rates
- ✓ …

3.3 Handling Discontinuous Phenomena

Nuclear power plant operations involve numerous discontinuous events that challenge standard neural network architectures. The framework implements three strategies to address these challenges:

3.3.1 Input Augmentation

Discontinuous variables are augmented with continuous auxiliary features to smooth the learning landscape:

$$\mathbf{x}_{aug} = [\mathbf{x}_{original}, \dot{\mathbf{x}}, \Delta t_{event}, e^{-t/\tau}]$$

where: - $\dot{\mathbf{x}}$ represents rate-of-change terms computed through finite differences - Δt_{event} indicates time elapsed since the last discrete event - $e^{-t/\tau}$ captures exponential decay of transient effects with appropriate time constant τ

3.3.2 Physics-Informed Regularization

Conservation laws are enforced through additional loss terms to maintain physical consistency across discontinuities:

$$\mathcal{L}_{physics} = \lambda_{mass}\mathcal{L}_{mass} + \lambda_{energy}\mathcal{L}_{energy} + \lambda_{momentum}\mathcal{L}_{momentum}$$

where each conservation loss penalizes violations of the respective physical principle.

3.4 Training Methodology

3.4.1 Adaptive Loss Weighting

The loss function weights are dynamically adjusted based on training progress:

$$\mathcal{L}_{total} = \mathcal{L}_{data} + \lambda_1(t)\mathcal{L}_{smoot} + \lambda_2(t)\mathcal{L}_{sparse} + \lambda_3(t)\mathcal{L}_{physics}$$

Early training emphasizes data fidelity ($\mathcal{L}_{data}$), while later stages increase regularization weights to promote interpretability.

3.4.2 Network Pruning and Simplification

Post-training pruning identifies and removes redundant connections:

1. Compute activation function importance: $I_{ij} = \max_x |\phi_{ij}(x)|$
2. Remove connections where $I_{ij} < \epsilon_{prune}$
3. Retrain briefly to compensate for removed connections
4. Apply symbolic regression to simplify remaining activation functions

4. Conclusions and Future Works

This paper presented a conceptual framework for developing interpretable surrogate models of nuclear power plant thermal-hydraulic behavior using Kolmogorov-Arnold Networks. The proposed approach addresses the fundamental challenge in Dynamic Probabilistic Risk Assessment: achieving computational efficiency while maintaining interpretability for safety-critical applications.

The successful implementation of this framework could significantly enhance nuclear safety analysis capabilities. Real-time DPRA would enable continuous risk assessment during evolving scenarios, supporting operator decision-making during abnormal conditions. Furthermore, the ability to extract symbolic relationships may reveal new insights into complex thermal-hydraulic phenomena.

Future research should focus on experimental validation against actual MARS simulations, beginning

with simplified components before progressing to full plant models. Critical areas for investigation include: developing rigorous uncertainty quantification methods for risk-informed applications; exploring multi-fidelity approaches integrating various simulation levels; implementing online learning capabilities for adaptive model updating.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (No. RS-2022-00144150)

REFERENCES

[1] Antonello, Federico, Jacopo Buongiorno, and Enrico Zio. "Physics informed neural networks for surrogate modeling of accidental scenarios in nuclear power plants." Nuclear Engineering and Technology 55.9 (2023): 3409-3416.
[2] Lu, Qi, et al. "Prediction method for thermal-hydraulic parameters of nuclear reactor system based on deep learning algorithm." Applied Thermal Engineering 196 (2021): 117272..
[3] Liu, Ziming, et al. "Kan: Kolmogorov-arnold networks." arXiv preprint arXiv:2404.19756 (2024).