HOKKAIDO UNIVERSITY

Title	Dealing with limited data and uncertainty in deep learning using Kolmogorov-Arnold Neural Network
Author(s)	Yang, Zhen; Patelli, Edoardo
Relation	Proceedings of the 10th International Symposium on Reliability Engineering and Risk Management (ISRERM 2026), June 28 - July 1, 2026, Sapporo, Japan
Citation	Proceedings of the 10th International Symposium on Reliability Engineering and Risk Management (ISRERM 2026)
Issue Date	2026-07-01
Doc URL	https://hdl.handle.net/2115/100454
Type	conference paper
File Information	C000130_1.pdf

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

10th International Symposium on Reliability Engineering and Risk Management (*ISRERM 2026*),
28 June-01 July 2026, Sapporo, Japan
Kitahara, T., Miyamori, Y. & Beer, M. (editors)

Dealing with limited data and uncertainty in deep learning using Kolmogorov–Arnold Neural Network

Zhen Yang[1], Edoardo Patelli[2,*]

[1] *Civil and Environmental Engineering, University of Strathclyde, Glasgow, UK.*
E-mail: z.yang.2021@uni.strath.ac.uk
[2,*] *Civil and Environmental Engineering, University of Strathclyde, Glasgow, UK*
E-mail: Edoardo.patelli@strath.ac.uk

Accurate modelling of complex physical systems under limited data availability and inherent uncertainty remains a major challenge in the deep learning field. Conventional deep learning models rely heavily on large datasets and easily overfit when data is limited. Physics-informed neural networks (PINNs) address this issue by incorporating physical rules, but their performance is often limited by rigid loss balancing and is difficult to further improve. Recently, Kolmogorov–Arnold Networks (KANs) have shown improved expressiveness and interpretability. However, the reliance on B-spline bases introduces high computational cost and limits generalization, and uncertainty cannot be quantified. We propose a novel model based on the KAN structure that integrates Bayesian inference, physical constraint regularization, attention mechanism and Chebyshev polynomial-based functional representations. The capability and flexibility of the proposed approach are demonstrated by solving a benchmark problem from manufacture experiment from NASA Prognostic Data Repository. Comparative results with baseline models further demonstrate the advantages. The findings indicate that the proposed approach provides a robust and effective solution for dealing with small dataset in complex engineering systems.

Keywords: Bayesian inference, Physics-Informed Neural Networks, Kolmogorov-Arnold Networks, Uncertainty.

1 Introduction

Accurate modelling of complex physical systems under limited data availability remains a major challenge in the deep learning field due to the high dimensionality of physical parameters and the nonlinear correlations among them. Under such circumstances, the rapid development of machine learning methods has brought new data- driven paradigms, in which neural networks emerge as popular tools for prediction tasks [9]. While neural networks possess strong representation capability of automatically learning from feature relationships, conventional architectures such as Multiple Layer Preceptors (MLP) generally provide only approximate solutions and heavily rely on large volumes of high-quality data to achieve reliable and high-quality predictive performance. Without enough data, neural network training is generally unstable, leading to overfitting and poor generalization performance. This problem is particularly obvious in complex physical systems and manufacturing processes, where data may be missing or degraded due to sensor inaccuracies, equipment failures, or operational interruptions.[5]. Consequently, there is a strong need to develop neural network models capable of dealing with uncertainty and achieving reliable performance under limited data training, e.g. [12].

In recent years, Physical informed neural networks (PINNs) have attracted close attention. PINNs are networks developed based on multilayer perceptron with introducing the constraints of objective physical laws to guide models to learn during the training stages. PINNs were first proposed to solve partial differential equations and later used in different tasks including as prediction models with physical regularization [1]. The physical constraints guarantee physical consistency for predictions, thereby providing significant advantages for modelling engineering tasks, particularly in guiding model learning when available training data is limited. However, PINNs suffer of rigid loss balance and poor capabilities in capturing dynamic physical processes [2]. In contrast, the Kolmogorov–Arnold Networks (KANs), first proposed in 2024 [3], have shown expressive capability and numerical stability by replacing linear transformations with learnable univariate functional mappings by decomposing high-dimensional functions into combinations of low-dimensional functions. The structure allows KANs to approximate complex nonlinear relationships more efficiently and interpretably, making them another promising foundation for physics-informed learning frameworks. Typically, KANs adopt B-splines as basic functions to capture complex nonlinear relationships and enables smooth functional representation although requiring substantial computational resources [11]. The unique architecture design of KANs provides enhanced interpretability which helps reduce model uncertainty compared to conventional architecture. This makes KANs a promising direction for dealing with uncertainty and small-data challenges in physics-related modelling tasks.

In recent years, researchers have proposed an advanced KAN-based architecture network known as Attention-Enhanced and Chebyshev Polynomial-Based Physics-Informed Kolmogorov-Arnold Networks (AC-PKAN) [4]. AC-PKAN is a physics-informed deep learning framework that integrates attention mechanisms with functional KAN representations. However, due to its design for solving partial differential equations, AC-PKAN is not directly suitable

for data-driven prediction tasks. The original AC-PKAN framework is formulated as a deterministic model and lacks the capability to quantify uncertainty. This limitation becomes especially critical in limited-data scenarios, where the uncertainty estimation and robust generalization are essential for reliable modelling of complex physical systems. Thus, to address these limitations, we propose 1) a novel data-driven approach that combines a Chebyshev polynomial-based KAN architecture with 2) a Bayesian output layer and an attention-enhanced layer for prediction tasks and 3) we reformulate the physical component into a prediction-oriented physical loss function, rather than relying on the original PDE-solving formulation. This allows our proposed model to capture and quantify uncertainty, thereby enhancing prediction robustness and reliability.

2 Methodology

Conventional architectures approximate complex high-dimensional functions through deep nonlinear stacking and huge amount of parameters training:

$$\mathrm{MLP}(x) = (W_3 \cdot \sigma_2 \cdot W_2 \cdot \sigma_1 \cdot W_1)(x) \tag{1}$$

while the KAN decomposes high-dimensional functions into structured compositions of low-dimensional functions

$$\mathrm{KAN}(x) = (\Phi_3 \cdot \Phi_2 \cdot \Phi_1)(x). \tag{2}$$

This distinction arises from different theoretical foundations. Conventional MLP models are designed based on the Universal Approximation Theorem (UAT). This guarantees that large neural networks can approximate arbitrary continuous functions through nonlinear transformations expressed:

$$f(x) \approx \sum_{i=1}^{N(C)} a_i \sigma(w_i \cdot x + b_i) \tag{3}$$

In contrast, KAN is derived from the Kolmogorov–Arnold Representation Theorem (KART), which provides a constructive representation of multivariate functions as explicit compositions of univariate functions (see Eq. (4))

$$f(x) = \sum_{q=1}^{2n+1} \Phi_q \left[\sum_{p=1}^{n} \varphi_{q,p}(x_p)\right] \tag{4}$$

More specifically, UAT emphasizes approximation capability through parameter scaling, while KART establishes an explicit structural decomposition of functions. Functional decomposition mechanism of KAN reduces the effective dimensionality of the function representation, and constraint the model's hypothesis space. The relatively modest parameter requirements also enable KAN to better control the network's parameter count, and remain smaller than the training dataset size under small datasets training, thereby preventing overfitting. This characteristic gives KAN the advantage over MLP of dealing with limited datasets.

However, practical training time of KAN appears lengthy while maintaining accuracy [10]. This limitation is related to the use of B-Spline of order *k* function as the basis function that is recursively defined as

$$B_{i,k}(x) = \frac{x - t_i}{t_{i+k-1} - t_i} B_{i,k-1}(x) + \frac{t_{i+k} - x}{t_{i+k} - t_{i+1}} B_{i+1,k-1}(x) \tag{5}$$

where $\{t_i\}$ denotes the knot sequence. The recursive evaluation of spline basis functions requires repeated local computations during forward and backward propagation. Such process will increase computational overhead and lead to longer training time compared to simple activation functions in MLPs. Furthermore, the original KAN architecture, as a deterministic model, is unable to capture either data uncertainty or model uncertainty and associated reliable uncertainty estimations for its predictions. Although KAN generally requires fewer trainable parameters than conventional MLPs, the compact representation may increase the risk of overconfident predictions in limited-data scenarios. This stresses the importance of embedding uncertainty quantification structure into model design. Another limitation is that KAN does not incorporate prior physical knowledge, making it unable to generate physically consistent predictions when applied to complex physical process-related problems. In addition, KAN lacks adaptive feature importance mechanisms, which limit the ability to dynamically adjust the contributions of input features during training. As a result, part of the network parameters may be used for learning irrelevant features for the outputs, resulting in computational resources and training time waste.

In our model designing, we use Chebyshev Polynomial as the basis functions of our KAN:

$$T_{n+1}(x) = 2xT_{n+1}(x) - T_{n-1}(x) \tag{6}$$

$$f(x) = \sum_{k=0}^{N} C_k T_k(x) \tag{7}$$

The Chebyshev Polynomial based KAN represents functional mappings through polynomial combinations, which helps significantly reduce storage requirements and computational complexity compared to B-spline based KAN. In addition, B-spline basis functions typically require dense data distribution to achieve stable fitting performance, while

Chebyshev polynomials only require sufficient domain coverage. Since limited datasets generally provide domain coverage instead of sufficient sampling density, Chebyshev polynomials are more suitable as basis functions in KAN for limited-data prediction tasks.

We also introduce a probabilistic model. Instead of using deterministic output layer, we design the output layer to be a Bayesian output layer which can provide probabilistic outputs as

$$y = W_{out} \cdot f_{acpkan}(x) + b \tag{8}$$

where $W_{out} \sim N(\mu_w, {\sigma_w}^2)$ and $b \sim N(\mu_b, {\sigma_b}^2)$, this leads to the output $y \sim N(\mu(x), \sigma^2(x))$. Correspondingly, the loss function of our KAN is formulated to be negative log-likelihood (NLL) objective to maintain consistency with the probabilistic output layer and to align with the uncertainty quantification framework

$$L_{nll} = \frac{(y-\mu)^2}{\sigma^2} + log\sigma^2 \tag{9}$$

Because the output weights and biases are parameterized as probability distributions, the NLL objective guarantees the coherent probabilistic inference of the model parameters. In our design, we choose not to introduce fully Bayesian KAN where probability distributions are assigned to all the weights. This design choice avoids the substantial computational overhead and optimization difficulty associated with fully Bayesian parameterization while still enabling effective uncertainty quantification.

The physical prior information is integrated in our model to guide the training process. We embed the physical rules as constraints into our model. The physical constraints work as physical residual forms in part of the loss function. The loss function in model is defined as

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_{\text{data}}}_{\text{Data fitting}} + \lambda \underbrace{\mathcal{L}_{\text{physics}}}_{\text{Physical consistency}} \tag{10}$$

In the total loss function, L_{data} is the part for data fitting, also it is L_{nll} we mentioend at Eq. (9). For physical loss part, we set an adjustable parameter L that is used for balancing relative contribution of data loss and physical loss to work in same magnitude. The physical loss can be expressed by the general term as

$$\mathcal{L}_{\text{physics}} = \frac{1}{N} \sum_{i=1}^{N} \left\| f_{\text{phys}}\left(x_i, \mu(x_i)\right) \right\|^2 \tag{11}$$

where f_{phys}denotes the physical consistency function derived from prior physical knowledge.

In addition, a Channel Attention mechanism is incorporated to introduce adaptive feature weighting which can be formulated as

$$\alpha = \sigma(W_2 \delta(W_1 x)), x' = \alpha \odot x \tag{12}$$

where a denotes the channel attention weights, $\sigma(\cdot)$ represents the sigmoid activation function δ $(\cdot)$ denotes a nonlinear transformation, and $\odot$ indicates element-wise multiplication. Through the mechanism, our model can dynamically adjust feature contributions and improve learning stability under limited-data scenarios.

3. Experiments

To validate the effectiveness of our proposed model, a prediction task is set up using the data set from real experiments provided by the UC Berkeley Emergent Space Tensegrities (BEST) Lab [7]. A series of milling experiments were performed under different cutting conditions, including spindle speed, feed rate, and depth of cut, also measuring the corresponding flank wear (VB) of the milling insert. The purpose of the milling experiment is to study tool wear under various cutting conditions, the experimental machine is Matsuura MC-510V machining center, the tool is 70 mm face mill for heavy roughing with KC710 inserts, and the workpiece materials is cast iron and steel. For the total experiments. There are 167 runs grouped into 16 cases belonging to 2 different material type, different cases have different Cutting conditions. The number of runs depended on the degree of flank wear that was measured between runs at irregular intervals up to a wear limit. The data feature details are shown in Table 1.

Table 1. Data feature names and description from [7]. In bold the used input features and VB is the considered output.

Field name	Description
case	Case number (1–16)
run	Counter for experimental runs in each case
VB	Flank wear (VB), measured after runs; not recorded after every run
time	Duration of experiment (restarts for each case)
DOC	Depth of cut (constant within each case)
feed	Feed rate (constant within each case)
material	Material type (constant within each case)
smcAC	AC spindle motor current
smcDC	DC spindle motor current
vib_table	Table vibration
vib_spindle	Spindle vibration
AE_table	Acoustic emission measured at the table
AE_spindle	Acoustic emission measured at the spindle

The original dataset contains high-frequency sensor signals collected during each run. As each signal consists of 9000 sampling points, directly employing the raw time-series data would lead to extremely high input dimensionality. To construct a compact prediction dataset, a statistical feature extraction procedure was applied in our design. For each run, the mean value of every sensor signal was calculated over the sampling duration, resulting in low-dimensional scalar descriptors that preserve the overall signal magnitude characteristics. The processed dataset was subsequently organized in tabular form for model training and evaluation. In this study, we focus on case 1 condition, which has the fewest available data points – 13 samples. The input variables selected are shown in bold in Table 1 and the output is the Flank wear (VB). We conducted comparative experiments with MLP, Bayesian Neural Network (BNN), and conventional KAN models. An ablation study was further performed to assess the impact of channel attention, Chebyshev polynomial basis, and physics-informed regularization. Due to the extremely limited number of samples (13 runs), Leave-One-Out Cross-Validation (LOOCV) was adopted. Model performance was evaluated using Mean Average Error (MAE) and R^2, while probabilistic models were additionally assessed using NLL, Prediction Interval Coverage Probability of 95% (PICP@95%) and Mean Prediction Interval Width of 95% (MPIW@95%). Our network architecture uses one hidden layer with 4 nodes (i.e. 7–4–1 architecture) with first-order Chebyshev polynomials, and all baseline models were configured with comparable parameter scales and consistent training settings to ensure fair comparison.

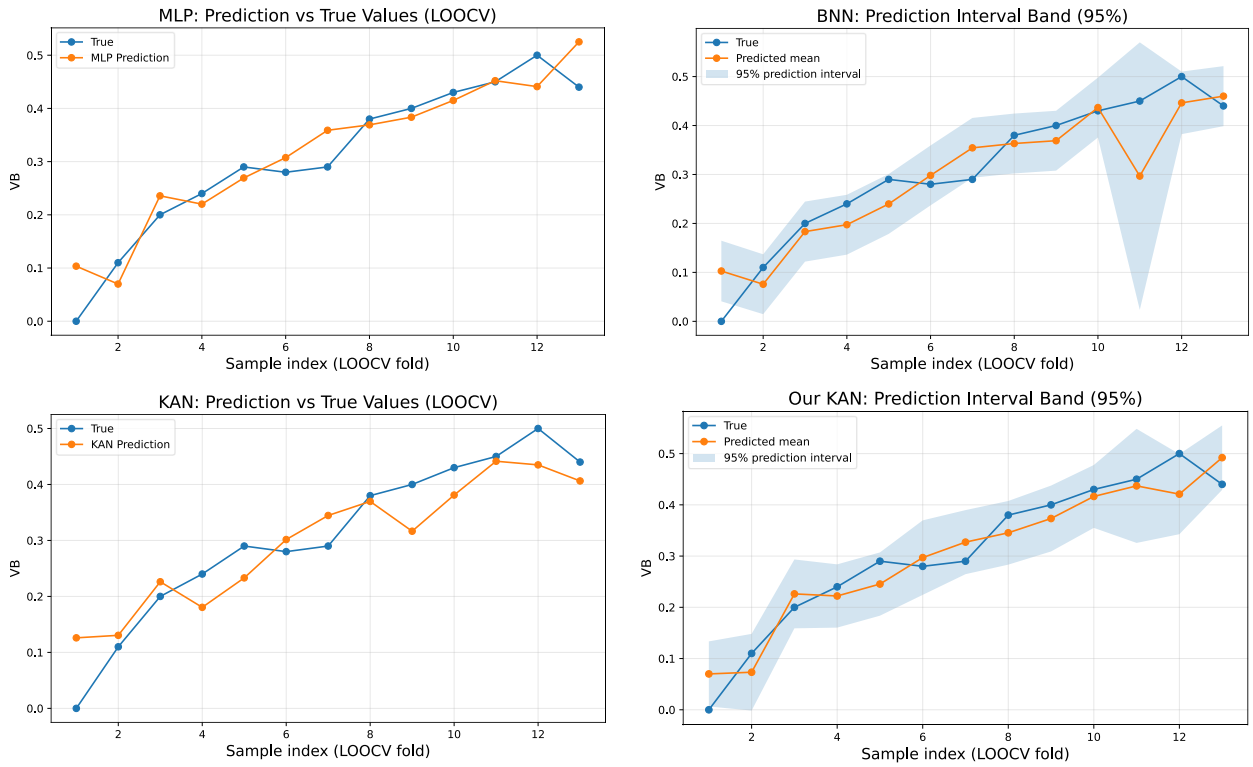

Fig 1. Model Prediction vs true value for MLP (top-left), BNN (top-right), KAN (left-bottom) and the proposed KAN Model (right-bottom)

Table 2. Model Evaluation of MLP, BNN, KAN and Our KAN Model

Model	MAE	R^2	NLL	PICP@95%	MPIW@95%
MLP	0.0388	0.8773	–	–	–
BNN	0.0470	0.8080	-1.3987	84.62%	0.1556
KAN	0.0473	0.8347	–	–	–
Our KAN model	0.0361	0.9130	-1.7135	84.62%	0.1391

Table 3. Ablation Study of our KAN model

Our KAN Model	MAE	R^2	NLL	PICP@95%	MPIW@95%	Running time
Full Model (All Component)	0.0361	0.9130	-1.7135	84.62	0.1391	59s
Without Attention layers	0.0434	0.8729	-1.2824	84.62	0.1581	24s
B-spline replacing Chebyshev	0.0440	0.8352	-1.0916	84.62	0.1406	120s
Without base function	0.0726	-0.3857	-0.8511	92.31	0.2618	31s

All the above experimental results are conducted without using any physics-informed constraints. As shown in Table 2, the MLP slightly outperforms the BNN in predictive accuracy. This may be because Bayesian models introduce additional distribution parameters to estimate uncertainty, which makes parameter estimation more difficult when only limited training samples are available. Nevertheless, BNN still provides valuable uncertainty quantification information compared to deterministic models. The conventional KAN model performs slightly worse than MLP, suggesting that the original KAN structure may not be enough to analyze the complex multi-sensor input conditions in this case. As shown in Fig 1 and Table 2, the proposed KAN model achieves the best predictive performance among the compared baseline models with the lowest MAE (0.0361) and the highest R^2 (0.9130) which enabling more effective feature utilization and improved prediction accuracy.

From the ablation study results from Table 3, we explored how different components matter in our KAN model structure. With all components, our model can achieve the best performance. Without Attention layers, the training time reduces nearly half but the predictive performance becomes slightly worse. This suggests the use of attention mechanism introduces additional computational overhead but still improves the model's capability to adapt for learning the sparse sensor features, improving predictive performance under such limited information conditions. Replacing Chebyshev base to B-spline function leads to double training time and worse predictive performance, this observation is consistent with the design motivation discussed earlier in our methodology part, indicating that the Chebyshev-based representation provides a more efficient functional approximation under limited-data conditions. This result further highlights the importance of the Chebyshev basis in the proposed model. The ablation experiment without the base function proves that the kernel functional representation is a critical component of the proposed KAN framework. Once the base function is removed, the model suffers a severe degradation in both predictive accuracy and uncertainty quality. This indicates that the performance gain does not arise solely from auxiliary modules such as attention, but fundamentally from the Chebyshev-based functional representation itself.

Next, we conducted additional experiments to evaluate the effectiveness of physical knowledge to our model. Since most of the inputs are sensor signals, instead of enforcing complete strict physical equations, we introduced end-based constraints describing the relationship between time and tool wear VB. Two physical constraints are considered: monotonicity in time and VB smoothness. The monotonicity constraint reflects the physical knowledge that tool wear is non-decreasing with respect to time as follow:

$$L_{mono} = softplus(-(VB_{t+1} - VB_t)) \tag{14}$$

while the smoothness constraint assumes that the wear process evolves progressively without sudden jumps

$$L_{smooth} = \frac{1}{N}\sum(VB_{t+1} - 2VB_t + VB_{t-1})^2 \tag{15}$$

Both constraints are applied to BNN and our proposed model, with physical weights of 0.2 and 1, respectively. An annealing strategy is adopted during training for stabilize our training, which enables us to better compare the effects produced when physical constraints are appropriately applied within the model.

Table 4. Performance after incorporating physics-informed constraints

Model	MAE	R^2	NLL	PICP@95%	MPIW@95%
BNN	0.0470	0.8080	-1.3987	84.62%	0.1556
BNN + Physics	0.0443 (↓0.0027)	0.8125 (↑0.0045)	-1.4532 (↓0.0545)	92.31% (↑7.69%)	0.2167 (↑0.0611)
Our KAN model	0.0361	0.9130	-1.7135	84.62%	0.1391
Our KAN model + Physics	0.0355 (↓0.0006)	0.9159 (↑0.0029)	-1.7481 (↓0.0346)	100% (↑15.38%)	0.1378 (↓0.0013)

Incorporating physics-informed constraints leads to different effects on two models as shown in Table 4. For BNN, there are slight improvements observed in MAE and R^2. Although the coverage probability increases to 92.31%, this is achieved at the cost of a much wider prediction interval, indicating a more conservative uncertainty estimation. For our proposed model, the physical constraint further improves an already strong baseline, and it achieves PICP=100% while maintaining an even narrower MPIW, suggesting better calibrated and more efficient prediction intervals.

4. Conclusions

In this paper, we propose a novel Bayesian physics-informed Chebyshev-based KAN with an attention mechanism to address prediction challenges under extremely limited data conditions. This method was tested using a real milling dataset from NASA, focusing on predicting cutting insert flank wear with limited sensor and time data. The findings indicate that the proposed model surpasses established baseline models, such as MLP, BNN, and conventional KAN, in terms of both predictive precision and uncertainty quantification. Furthermore, the results reveal that the combination of functional representation and physics-informed learning can ensure the physical consistency and at the same time, improve the prediction accuracy.

In it important to notice that the proposed approach can be further improved. For instance, the current Bayesian structure is implemented only at the output layer, which limits the expressiveness of the uncertainty modelling. Extending the probabilistic structure to more layers may further enhance the uncertainty representation capability of the model for dealing with more uncertain cases. Applying of Chebyshev-based functional representation as a function base may lead to unstable approximations when the training data only covers a limited region of the input domain. Future work may explore hybrid functional representations that combine global polynomial bases with local basis functions to improve approximation representation stability. The current physical constraints are implemented in a simplified deterministic form. However, in practical engineering systems, physical correlations may involve stochastic variations, uncertain parameters, or regime-dependent behaviors. In such cases, more comprehensive probabilistic modelling techniques (such as p-boxes and interval [13]) will be required. In addition, we plan to assess the proposed approach on a wider range of applications across different domains and further refine the framework to make it a more robust and adaptable tool for prediction in diverse scenarios. Despite the above-mentioned limitations, the proposed approach shows clear advantages respect to current machine learning method trained on limited dataset typical in practical engineering applications.

Acknowledgment.
This work has been supported by the Strathclyde Centre for Doctoral Network in Digital technologies for Resilient and Sustainable Infrastructures and the Research Innovation Scotland - International Collaboration Fund

References

[1] Raissi M, Perdikaris P, Karniadakis G E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational physics, 2019, 378: 686-707.
[2] Cuomo S, Di Cola V S, Giampaolo F, et al. Scientific machine learning through physics–informed neural networks: Where we are and what's next. Journal of Scientific Computing, 2022, 92(3): 88.
[3] Liu Z, Wang Y, Vaidya S, et al. Kan: Kolmogorov-Arnold networks. arXiv preprint arXiv:2404.19756, 2024.
[4] Zhang H, Huang Z, Wang Y. AC-PKAN: Attention-enhanced and Chebyshev polynomial-based physics-informed kolmogorov-arnold networks. arXiv preprint arXiv:2505.08687, 2025.
[5] Qi G J, Luo J. Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(4): 2168-2187.
[6] Zhang H, Huang Z, Wang Y. Ac-pkan: Attention-enhanced and Chebyshev polynomial-based physics-informed kolmogorov-arnold networks. arXiv preprint arXiv:2505.08687, 2025.
[7] A. Agogino and K. Goebel (2007). BEST Lab, UC Berkeley. "Milling Data Set ", NASA Prognostics Data Repository, NASA Ames Research Center, Moffett Field, CA
[8] Yang Z, Chen Y, Patelli E. Verification of Bayesian physics-informed neural networks. 35th European Safety and Reliability Conference (ESREL 2025) and the 33rd Society for Risk Analysis Europe Conference (SRA-E 2025). Research Publishing Services, 2025: 314-321.
[9] Iten R, Metger T, Wilming H, et al. Discovering physical concepts with neural networks. Physical review letters, 2020, 124(1): 010508.
[10] Yu R, Yu W, Wang X. Kan or mlp: A fairer comparison. arXiv preprint arXiv:2407.16674, 2024.
[11] Pal A, Das D. Understanding the limitations of B-Spline KANs: Convergence dynamics and computational efficiency. Dimensions, 2024, 21: 22.
[12] Lye, A., N. Prinja, and E. Patelli. 2022. 'Probabilistic Artificial Intelligence Prediction of Material Properties for Nuclear Reactor Designs'. *32nd European Safety and Reliability Conference*, https://rpsonline.com.sg/rps2prod/esrel22-epro/html/S24-02-306.xml.
[13] Sadeghi, J., M. de Angelis, and E. Patelli. 2019. 'Efficient Training of Interval Neural Networks for Imprecise Training Data'. *Neural Networks* 118: 338–51. https://doi.org/10.1016/j.neunet.2019.07.005.