

## Article

# An Interpretable Time Series Data Prediction Framework for Severe Accidents in Nuclear Power Plants

Yongjie Fu <sup>1,2</sup>, Dazhi Zhang <sup>3</sup>, Yunlong Xiao <sup>4</sup>, Zhihui Wang <sup>3</sup> and Huabing Zhou <sup>1,2,\*</sup> 

<sup>1</sup> College of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, China; fuyongjie1217@163.com

<sup>2</sup> Hubei Provincial Key Laboratory of Intelligent Robot, Wuhan Institute of Technology, Wuhan 430205, China

<sup>3</sup> CNNC Key Laboratory on Nuclear Industry Simulation, China Nuclear Power Operation Technology Corporation, Ltd., Wuhan 430040, China; zhangdz02@cnnp.com.cn (D.Z.); ablezhh9703@163.com (Z.W.)

<sup>4</sup> China Nuclear Power Operation Technology Corporation, Ltd., Wuhan 430040, China; xiaoyl01@cnnp.com.cn

\* Correspondence: zhouhuabing@gmail.com

**Abstract:** Accurately predicting severe accident data in nuclear power plants is of utmost importance for ensuring their safety and reliability. However, existing methods often lack interpretability, thereby limiting their utility in decision making. In this paper, we present an interpretable framework, called GRUS, for forecasting severe accident data in nuclear power plants. Our approach combines the GRU model with SHAP analysis, enabling accurate predictions and offering valuable insights into the underlying mechanisms. To begin, we preprocess the data and extract temporal features. Subsequently, we employ the GRU model to generate preliminary predictions. To enhance the interpretability of our framework, we leverage SHAP analysis to assess the contributions of different features and develop a deeper understanding of their impact on the predictions. Finally, we retrain the GRU model using the selected dataset. Through extensive experimentation utilizing breach data from MSLB accidents and LOCAs, we demonstrate the superior performance of our GRUS framework compared to the mainstream GRU, LSTM, and ARIMAX models. Our framework effectively forecasts trends in core parameters during severe accidents, thereby bolstering decision-making capabilities and enabling more effective emergency response strategies in nuclear power plants.

**Keywords:** time series prediction; GRU; SHAP; MSLB; LOCA



**Citation:** Fu, Y.; Zhang, D.; Xiao, Y.; Wang, Z.; Zhou, H. An Interpretable Time Series Data Prediction

Framework for Severe Accidents in Nuclear Power Plants. *Entropy* **2023**, *25*, 1160. <https://doi.org/10.3390/e25081160>

Academic Editor: Christian H. Weiss

Received: 21 June 2023

Revised: 25 July 2023

Accepted: 31 July 2023

Published: 2 August 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

With the rapid growth in the social economy, traditional thermal power generation fails to meet societal needs and results in irreversible environmental pollution. The large demand for energy in various countries makes the vigorous development of nuclear energy inevitable [1]. Additionally, promoting the development of nuclear energy is one of the essential measures for China to adjust its energy structure and achieve the goal of being “dual carbon” [2]. However, in the nuclear energy field, nuclear power plants’ operational safety is highly critical, and any mistake may lead to catastrophic consequences. Therefore, to ensure the safe and reliable operation of nuclear power plants, the diagnosis and prediction of nuclear power operating conditions must be accurate and efficient, and it is essential to study the prediction of severe nuclear power accidents [3].

A loss of coolant accident (LOCA), main steam line break (MSLB) accident, and steam generator tube rupture (SGTR) accident are critical accidents that pose safety threats to nuclear power plants. Predicting the progression of these accidents is a crucial emergency measure to assess safety risks in advance and enable timely and effective countermeasures. However, their prediction is challenging due to the transient nature and complex characteristics of nuclear power accidents. The data involved are a nonstationary time series with nonlinear and noisy characteristics, further complicated by the influence of temporal and characteristic dimensions. Thus, developing a forecasting model that considers complex

systems, nonlinear parameters, and multiple operating nodes remains a key research focus in time series data prediction [4].

To enhance forecast accuracy, researchers initially employed traditional models such as autoregressive integral moving average (ARIMA) [5] and support vector regression (SVR) [6,7] to explore nuclear power forecasting. In the following years, artificial neural networks (ANNs) [8], convolutional neural networks (CNNs) [9], recurrent neural networks (RNNs), long short-term memory networks (LSTMs), gated recurrent units (GRUs) [10], etc., along with their variants, were employed to improve prediction accuracy.

Improvements in model prediction accuracy often come at the expense of increased model complexity. However, increased complexity can result in reduced explanatory power of the prediction outcomes, compromising the credibility and practicality of the model. The highly nonlinear and abstract characteristics of deep learning models pose challenges for intuitively comprehending the decision-making process. The models are regarded as “black box” models due to the opaque nature of the models’ internals. Consequently, existing deep learning methods suffer from a lack of interpretability. This falls short of meeting the rigorous safety standards required in nuclear power plants. Therefore, in terms of practical engineering deployment, pursuing model interpretability represents a critical challenge for artificial intelligence models at this stage [11].

To tackle these challenges, we integrated the GRU and SHAP methodologies to construct a framework called GRUS. This framework aims to forecast and elucidate the trends of severe accidents in nuclear power plants, and it establishes a trusted relationship between users and models, effectively mitigating potential threats of models in the deployment and application of nuclear power. The key contributions of this study can be summarized as follows.

(1) We propose a hybrid framework called GRUS for predicting the trends of severe accidents in nuclear power plants, achieving a higher prediction accuracy. Furthermore, the interpretability analysis of the prediction model provides a more reliable and comprehensive reference and practical value.

(2) We pioneer the utilization of the SHAP algorithm for the interpretability analysis of nuclear power severe accident time series prediction. This approach leads to a deeper understanding of key features that have a significant impact on predictive models, thereby improving the effectiveness and efficiency of forecasting.

(3) We conduct a comparative analysis, evaluating the performance of our proposed GRUS model against well-established GRU, LSTM, and classic ARIMAX models. The results conclusively demonstrate the superiority of our model over these state-of-the-art comparison models.

## 2. Related Work

A comprehensive review of relevant research indicates that the development of artificial intelligence (AI) has facilitated its utilization in the simulation and analysis of operational data, and has received increasing attention from the nuclear industry over the past few decades, resulting in successful applications in nuclear power plants. Despite the increasing application of AI models, interpretability remains a nascent research field, with further development required. Specifically, interpretability analysis is of paramount importance to ensure the safe and effective deployment of these models. Therefore, there is a pressing need for future research in nuclear power plants to focus on enhancing the accuracy of models and integrating interpretability analysis.

### 2.1. Artificial Intelligence Method

Researchers in the past have utilized conventional statistical modeling methods to forecast the functioning of nuclear power plants. For instance, Wu et al. [5] applied the ARIMA model to predict the levels of radioactive substances in the vicinity of nuclear power plants. Moshkbar et al. [12] developed a robust accident recognition system that integrates the ARIMA model and the EBP algorithm for identifying and monitoring the

transient processes during nuclear power plant accidents. With the advancements in artificial intelligence, researchers have started employing neural networks to model and predict the operational data of nuclear power plants. Santhosh et al. [13] proposed a data-driven nuclear reactor accident state identification method based on the reactor process parameters using an ANN, which can effectively diagnose potentially threatening accident situations. Kim et al. [14] proposed a method based on a fuzzy neural network (FNN) to predict the leakage flow rate of an LOCA. Do et al. [15] used a deep neural network (DNN) to achieve the water level prediction of a nuclear reactor pressure vessel. Tian et al. [16] proposed an approach for detecting an LOCA in nuclear power plant condition monitoring using a neural network architecture optimization method based on the constrained random search algorithm. Sun et al. [17] employed principal component analysis (PCA) to identify an SGTR occurring in the steam generator tubes of a small modular reactor and to quantify the size of the fracture. In addition, Wang et al. [18] introduced the fault identification and diagnosis method based on kernel principal component analysis (KPCA) and similarity clustering for sensor anomaly diagnosis in nuclear power plants. Yang et al. [19] used the RELAP5/MOD3.3 code to simulate the MSLB of Gen III reactors with passive safety functions, and proposed an optimal estimation plus uncertainty analysis method, which can analyze and evaluate the MSLB event in Gen III reactors. Xiang et al. [20] introduced an unsupervised representation clustering method based on deep learning for the automatic transient pattern recognition of nuclear power plant condition monitoring data.

A CNN is a variation of a DNN that is commonly employed in various fields such as natural language processing and image processing. Because of its ability to extract valuable features from data properties, a CNN exhibits high robustness for large-scale and complicated datasets. Yao et al. [21] applied a small-batch-size processing CNN to fault diagnosis problems in nuclear energy production, including detecting fault types such as cracks, sensor failures, pipeline corrosion, etc. Lin et al. [22] converted the input sensor measurement time series data into a picture form, and classified and reconstructed the signal based on a CNN to deal with the problem of sensor failure in nuclear power plants. He et al. [23] proposed a data-driven adaptive fault diagnosis method based on non-dominated sorting genetic algorithm II (NSGAI) and a CNN to improve the accuracy and efficiency of fault diagnosis in nuclear power systems.

In terms of the real-time detection and trend prediction of nuclear power plant operating conditions, an RNN is good at processing time series data, and has been successfully employed to sequence data modeling in previous explorations. Şeker et al. [24] presented the study of an RNN for the condition monitoring of nuclear power plants and rotating machinery. LSTM as a variant of traditional RNNs, has attracted much attention due to its high-precision prediction of time series data. Radaideh et al. [25] used an expert system based on a DNN and LSTM to accurately predict an LOCA in nuclear power plants. Nguyen et al. [26] proposed a multi-step time series signal prediction method based on ensemble empirical mode decomposition (EEMD) and LSTM, which was applied to data analysis in nuclear power plants. She et al. [27] constructed a combination of a CNN and LSTM named ConvLSTM for fault diagnosis and the post-accident prediction of an LOCA. Zhang et al. [28] proposed a fault diagnosis model based on a CNN-LSTM neural network optimized by the sparrow search algorithm, which achieved accurate and real-time fault diagnosis and prediction in nuclear power plants. Gong et al. [29] introduced a sigmoid-based LSTM method for the time series prediction of an LOCA. Lei et al. [30] developed a prediction model for key parameters affected by sensor failures during nuclear power plant accidents by using LSTM and historical key parameter operation sequences. The proposed model successfully predicted parameters including the pressurizer pressure, pressurizer water level, and steam generator water level, etc. in the case of an LOCA and SGTR. In comparison with the LSTM model, the GRU model simplifies the design of the gating mechanism, has a faster training speed and fewer parameters, and can effectively process long sequence data. As a result, it exhibits a superior performance in applications. Zhang et al. [31] proposed a GRU-based method for reconstructing the pressurizer water

level, which can monitor and predict the pressurizer water level of marine pressurized water reactors. Wang et al. [32] proposed an advanced fault diagnosis method for nuclear power plants based on a convolutional gated recurrent unit (CGRU) and enhanced particle swarm optimization (EPSO). Fukun et al. [33] introduced an LOCA accident state prediction method for nuclear power plants based on a GRU-CNN, which can effectively predict the LOCA accident status. He et al. [34] used LSTM and GRU algorithms to construct a U-tube steam generator water level prediction model, which can more accurately predict the change trend of the water level.

Through a comprehensive examination of the recent literature in the field of nuclear power plants, this study chose the current mainstream GRU model as the primary method for accurate nuclear power accident prediction.

## 2.2. Interpretability Method

AI methods have demonstrated remarkable success in analyzing predictive models of structured datasets. However, transparency and interpretability are essential requirements for model development. In nuclear power scenarios and other industrial situations, most models are regarded as “black box” models due to their limited transparency. For instance, the construction of a model for predicting the development trend of a variable in an LOCA relies on historically related parameter values to obtain a future predicted value for the variable. However, the model’s internal prediction mechanism cannot be explained, ultimately reducing the credibility and practicability of the model. To address these limitations and ensure safe engineering deployment when applying AI models in nuclear power plants, explainable artificial intelligence (XAI) has gained increasing attention in recent research endeavors. In this context, XAI is concerned with the development of methods that aim to enhance the models’ interpretability and facilitate operators’ comprehension. Kim et al. [35] systematically summarized the application and problems associated with XAI technology in the domain of operator assistance systems, and provided valuable references to guide future research in the field of nuclear energy. Mortenson et al. [36] conducted a comprehensive review of the application of AI in the nuclear power field, focusing on the analysis of research related to the concept of interpretability in AI systems. The interpretability and transparency of models are particularly important in critical infrastructure domains such as nuclear power, where both staff and the general public demand high levels of reliability from these systems and facilities. Park et al. [37] presented an intelligent diagnosis approach for nuclear power plants based on a GRU-AE autoencoder that extracts and reconstructs the data, followed by processing and analysis of the data using a LightGBM decision tree classifier to achieve the fault classification and diagnosis. Notably, the SHAP algorithm was employed to explain LightGBM model results, thereby enhancing model reliability. Shin et al. [38] used a CNN for the detection and diagnosis of abnormal events in nuclear power plants, and introduced several interpretability analysis methods, including saliency mapping, guided gradient-weighted class activation mapping (Grad-CAM), and SHAP, to enhance the system’s interpretability.

To sum up, practical applications of models in nuclear power and other industrial fields necessitate the incorporation of interpretability. The interpretability of models not only improves the robustness by identifying potential adversarial perturbations that impact prediction results, but also assists decision makers to understand the rationale behind model predictions and verify the potential causal relationships between variables. This study chose the SHAP method as an important method for interpretability analysis.

## 3. Methods

This section commences with an introduction of the GRU model as the fundamental model, while simultaneously introducing the concept of XAI to provide explainable analysis for models. Finally, the GRUS framework proposed in this study and the corresponding evaluation metrics are introduced.

### 3.1. Gated Recurrent Unit (GRU)

GRU model was proposed by Cho et al. [39] for time series that permits the incorporation of different time scales while simplifying calculations. GRU is a type of RNN architecture that can achieve comparable performance to other RNNs, yet is computed more efficiently. GRU has a simpler structure than other RNNs and consists of only two gating mechanisms, namely the reset gate and the update gate, which regulate the retention or deletion of the information at each time step. The update gate manages the amount of data that historical memory information can preserve until the current moment, while the reset gate determines the extent of historical information to forget. Thus the structure of GRU enables it to adaptively capture dependencies from an extensive sequence of data without discarding information from early sections of the sequence. The process for each GRU unit updates its parameters according to the formula below:

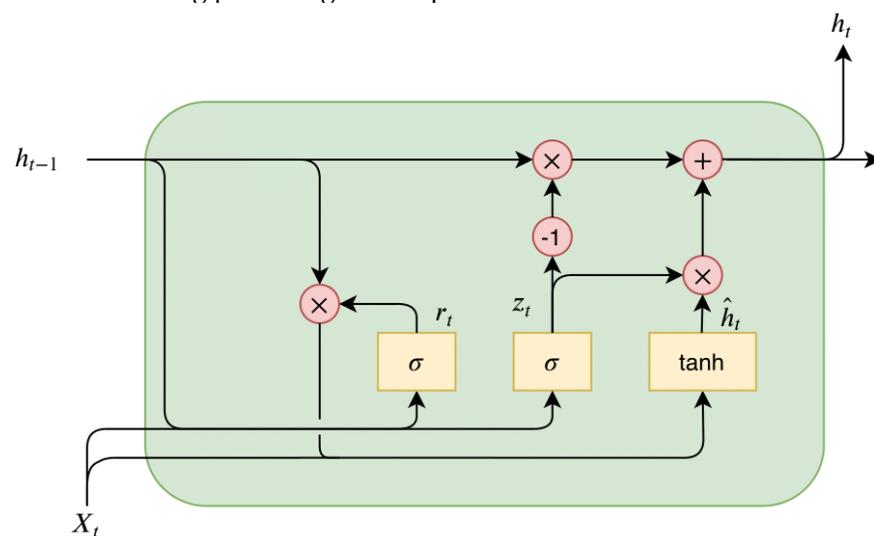
$$z_t = \sigma(W_z \cdot [h_{t-1}, X_t] + b_z), \quad (1)$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, X_t] + b_r), \quad (2)$$

$$\hat{h}_t = \tanh(W_h \cdot [r_t * h_{t-1}, X_t] + b_h), \quad (3)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \hat{h}_t, \quad (4)$$

where  $W_z$ ,  $W_r$ , and  $W_h$  are weight parameters,  $b_z$ ,  $b_r$ , and  $b_h$  are biases.  $\sigma$  is the sigmoid function,  $X_t$  is input time step,  $h_{t-1}$  is the previous hidden layer state,  $h_t$  is the output.  $r_t$  represents the reset gate,  $z_t$  represents the update gate. As with any other neural network, the gates in GRU are associated with weights and biases whose values are learned during the model training phase. Figure 1 depicts the internal network structure of GRU.



**Figure 1.** The internal structure of one neural unit of GRU.

### 3.2. Explainable AI (XAI)

XAI has emerged as a prominent trend in AI. The purpose of an XAI system is to make the behavior of models more intelligible to humans by providing explanations [40]. XAI models are commonly classified into two types: global interpretability and local interpretability [41]. Global interpretability enables analysts to investigate the structure and parameters of a complex model and comprehend how the model works globally. On the other hand, local interpretability examines the model's individual predictions locally and attempts to comprehend why the model made a particular decision. The interpretability of a predictive model often depends on estimating the contribution of

individual characteristics (i.e., independent variables) to the model's results. This study introduces the SHAP methodology as an interpretability analysis approach for time series prediction models.

SHAP is an explanatory model of additivity constructed by Lundberg [42] that draws inspiration from cooperative game theory. SHAP is used to explain individual and global predictions of the model. It aims to explain the contribution of each input feature to the prediction. Through the computation of Shapley values, which measures the marginal contribution of features towards predictions, SHAP offers crucial insights into the positive and negative impacts of feature values on estimation results at the sample level. The methodology also reveals the interactions between variables and how this relationship is reflected in the model. In addition, SHAP provides visualizations that aid in comprehending the importance of each feature in the model output. As a result, SHAP enhances the interpretability of the model, promoting trust in its predictions. The Shapley value of each feature is calculated as shown in the following equation:

$$\phi_{i,j} = \sum_{S \subseteq M \setminus \{x_{i,j}\}} \frac{|S|!(m - |S| - 1)!}{m!} [f_{x_i}(S \cup \{x_{i,j}\}) - f_{x_i}(S)], \quad (5)$$

where  $M$  is the set of all input features in the dataset with dimension  $m$ .  $S$  is a subset of  $M$  with a size of  $|S|$ .  $f_{x_i}(S)$  denotes the predicted value of the model for the sample  $x_i$  when only the feature set  $S$  is utilized. When  $S$  is an empty set,  $f_{x_i}(S)$  is referred to as the basic value  $\phi_0$ , which equals the average value of the predicted value of the model output among all samples.  $f_{x_i}(S \cup \{x_{i,j}\})$  is computed by adding the feature value  $x_{i,j}$  to the feature set  $S$ , which yields the average predicted value of the model for the sample  $x_i$  in all samples.  $\frac{|S|!(m - |S| - 1)!}{m!}$  represents the coefficient used to normalize the differences. As there are numerous potential values of  $S$ , it can be inferred from the formula that the Shapley value  $\phi_{i,j}$  of the eigenvalue  $x_{i,j}$  takes into account all possible combinations of  $S$ , along with the influences of all other features except  $x_{i,j}$ . When  $\phi_{i,j} > 0$ , it means that the corresponding feature has a positive effect in enhancing the predicted output, otherwise it reduces the contribution to the prediction. SHAP utilizes the feature attribution method to calculate the attribution value of each feature, which reflects its impact on the final prediction. A higher SHAP score indicates a greater contribution to the feature.

### 3.3. GRUS Framework

In this study, we propose an explainable framework GRUS based on the GRU model, which integrates the SHAP methods. GRUS utilizes the relevant features of severe nuclear power accidents and the information to predict the trend of key parameters. Simultaneously, it conducts explainable analysis to facilitate understanding of the model's prediction mechanism, identification of internal factors influencing nuclear power accidents, and improvement of trust in AI models for nuclear power scenarios among relevant personnel.

The GRUS framework mainly includes two parts: data processing and framework analysis. The data processing component consists of four sequential steps, i.e., feature engineering, data normalization, sliding update, and dataset splitting. The framework encompasses the GRU model for training and prediction, the SHAP method for interpretability analysis, model optimization, and enhancement of prediction results. To elaborate, the following provides a detailed description of the specific implementation process.

#### 3.3.1. Data Processing

- Feature Engineering

The aim of feature engineering is to enhance prediction accuracy and accelerate computations by reducing the number of features during model training. Due to the large number of features in the data, a feature correlation heat map is generated to comprehend the features. For feature analysis, Spearman, Pearson, and Kendall correlation coefficients (i.e., SCC, PCC, KCC) are commonly used to evaluate monotonicity, linearity, and depen-

dence of different state parameters, respectively. In this study, we employed Spearman correlation and Pearson correlation to analyze the relationships between features, aiming to select highly correlated features with the target features. By utilizing these two correlation methods, we ensured a comprehensive and precise analysis of the correlations, which facilitated the selection of input variables for our modeling process.

Additionally, the F-score can be employed to select highly relevant features for analysis. By perturbing a column of the feature matrix, the error score of the feature represented by the column to the target feature is regarded as the F-score. The F-score of each feature is calculated as a measure of its discriminatory strength, with higher scores indicating greater discriminative capability and more importance for the feature. All F-scores are computed for each feature to facilitate feature selection. By comparing feature importance, the most effective features can potentially be identified.

Finally, the features are filtered again by incorporating expert experience, resulting in the selection of key system parameters as modeling features.

- Data Normalization

Raw data may not be appropriate for model input if their values are too large or small, hindering the model’s ability to learn temporal correlations. Therefore, it is important to normalize input data by scaling it to a specified range. In our experiment, we adopted the MinMaxScaler method to normalize the data, which results in rapid convergence during model training.

- Sliding Update

Accurate time series data prediction often necessitates considering the influence of historical data on the succeeding moment. The sliding window technique is a valuable method for processing data to maximize the utilization of all data.

The sliding window technique retains a fixed time step and uses the initial  $k$  data points as inputs for the model. The  $k + 1$ th data point serves as the label, representing the actual value of the time series prediction for that data group. As new data arrive, the sampling window slides and updates, appending the latest data while removing historical data. This method is widely used for processing time series data because it efficiently captures temporal dependencies and improves forecasting accuracy. Its schematic diagram is shown in Figure 2.

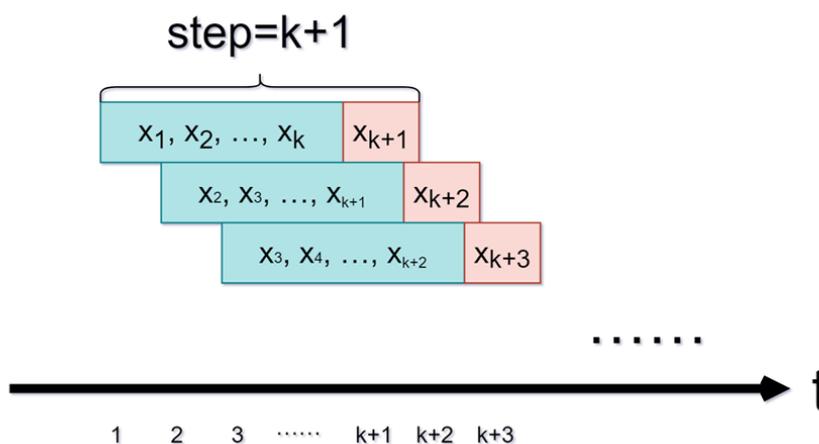


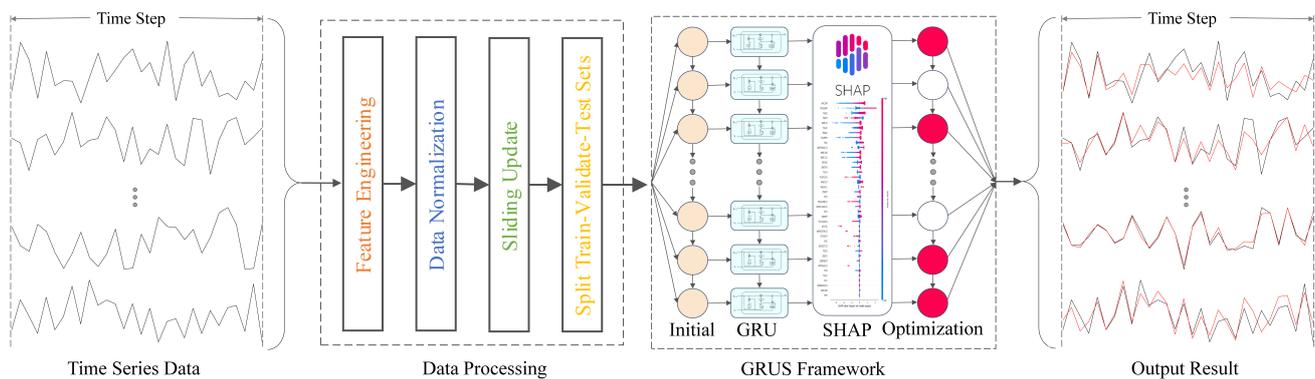
Figure 2. The sliding window technique.

- Splitting Dataset

The datasets are divided into sub-datasets of the training set, validation set, and test set, which are utilized for model training, model verification, and model testing, respectively. Based on the sub-datasets, we can obtain a well-trained GRUS to predict the time series trend of some key parameters in nuclear power accidents.

### 3.3.2. Framework Analysis

After data preprocessing, the GRUS framework is employed to predict time series data and conduct interpretability analysis. The GRU model is leveraged to predict time series data by capturing the nonlinearity of fluctuating features and controlling information flow through the network via gating mechanisms to achieve accurate forecasting. Subsequently, SHAP analysis is conducted to analyze the model's interpretability and comprehend its prediction mechanism, enabling optimization of the model at the feature level by selecting important relevant features in the interpretability analysis, so as to achieve the improvement in the prediction performance. Ultimately, prediction results are generated and corresponding evaluation metrics are utilized for evaluation and comparison. The flowchart of the GRUS framework is shown in Figure 3.



**Figure 3.** The flow-chart of the GRUS framework.

### 3.4. Metrics

In order to evaluate the performance of the prediction model, this study employed three commonly utilized evaluation metrics in regression modeling, namely, mean square error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). Smaller MSE and MAE values indicate higher prediction precision, the range of MAPE is  $[0, +\infty)$ , 0% means a perfect model, and 100% means an inferior model. The formulas for these metrics are as follows:

$$MSE = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2, \quad (6)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i|, \quad (7)$$

$$MAPE = \frac{\sum_{i=1}^N |\hat{y}_i - y_i| / y_i}{N} \times 100\%, \quad (8)$$

where  $N$  represents the number of samples in the dataset,  $y_i$  represents the actual value,  $\hat{y}_i$  represents the predicted value generated by the evaluated model.

## 4. Experiment and Results

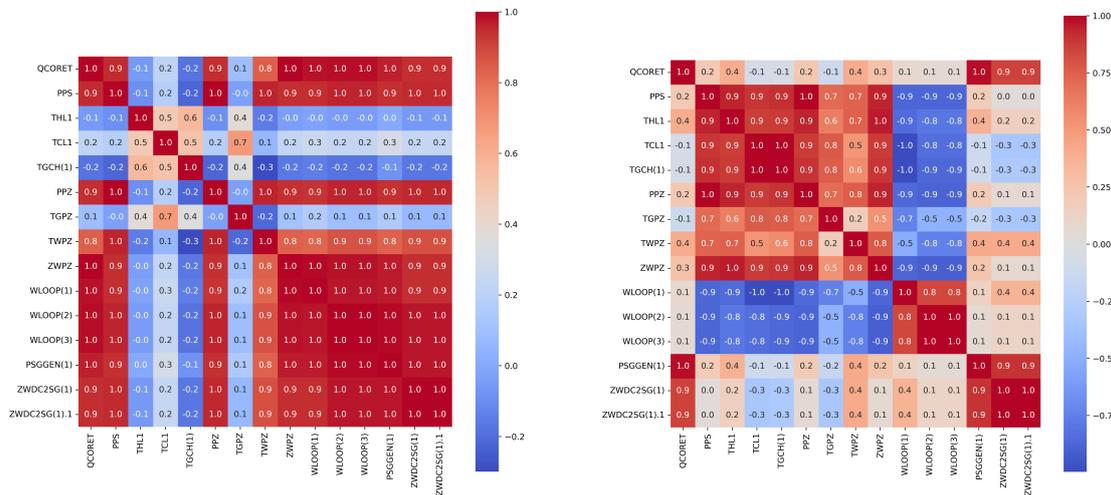
This section starts with a detailed description of the experimental datasets utilized in the study. Then, the data are subjected to rigorous processing for subsequent analyses. Additionally, this section introduces the relevant experimental settings, which include the both qualitative and quantitative experiments to evaluate the performance of the proposed GRUS framework. Finally, the experimental outcomes are meticulously presented and elaborately analyzed in detail to facilitate a comprehensive understanding of the research findings.

### 4.1. Data Source

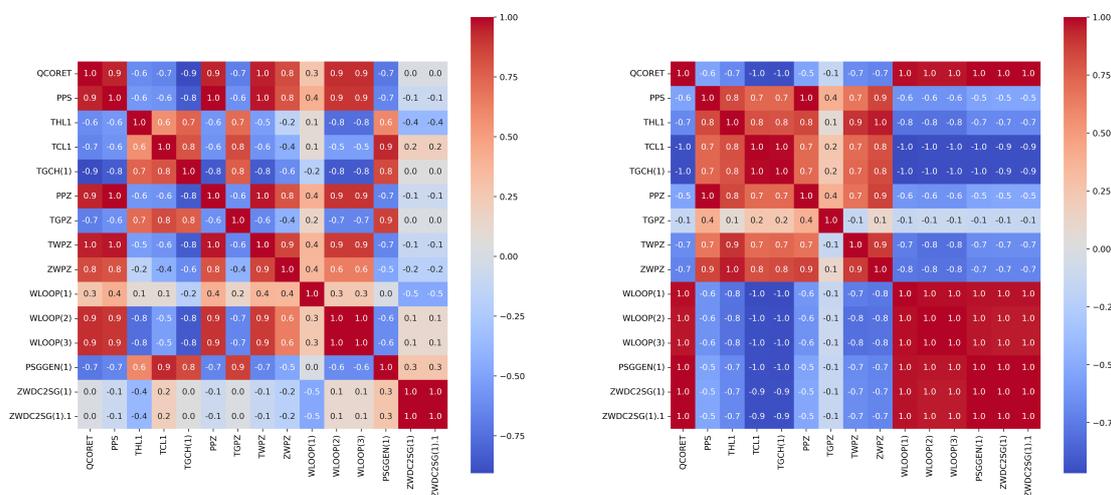
The experiment used the Modular Accident Analysis Program version 5 (MAAP5), an industrial-grade nuclear power simulation platform, to simulate severe nuclear power accidents. As an integrated system model, MAAP5 can simulate the response of the reactor core, reactor coolant system, and containment under conditions of core damage. The experiment simulated two typical accident scenarios of an LOCA and MSLB at 100% reactor power, with the accident pipe rupture size set to 0.5 m<sup>2</sup>. The resulting datasets contained 116 features divided into 10 categories, including the main coolant system, passive safety system, steam generator parameters, confinement parameters, primary coolant system parameters, and essential main system signals, among others.

### 4.2. Data Processing and Analysis

In order to further screen important feature parameters, we calculated and drew the PCC and SCC heatmap between all features in both types of accidents using feature engineering techniques. Then we were able to select the relevant features with the total power of the reactor (QCORET) based on expert experience. Figures 4 and 5 show the heatmaps for the Pearson and Spearman correlation analyses of some features, respectively.



**Figure 4.** The Pearson correlation coefficients heat-map of some features of the LOCA (left panel) and the heat-map of some features of the MSLB accident (right panel).



**Figure 5.** The Spearman correlation coefficients heat-map of some features of the LOCA (left panel) and the heat-map of some features of the MSLB accident (right panel).

Additionally, feature selection was conducted using the F-Score to identify high-relevance features specific to QCORET for further screening of the features. Figure 6 depicts the F-Scores of some features.

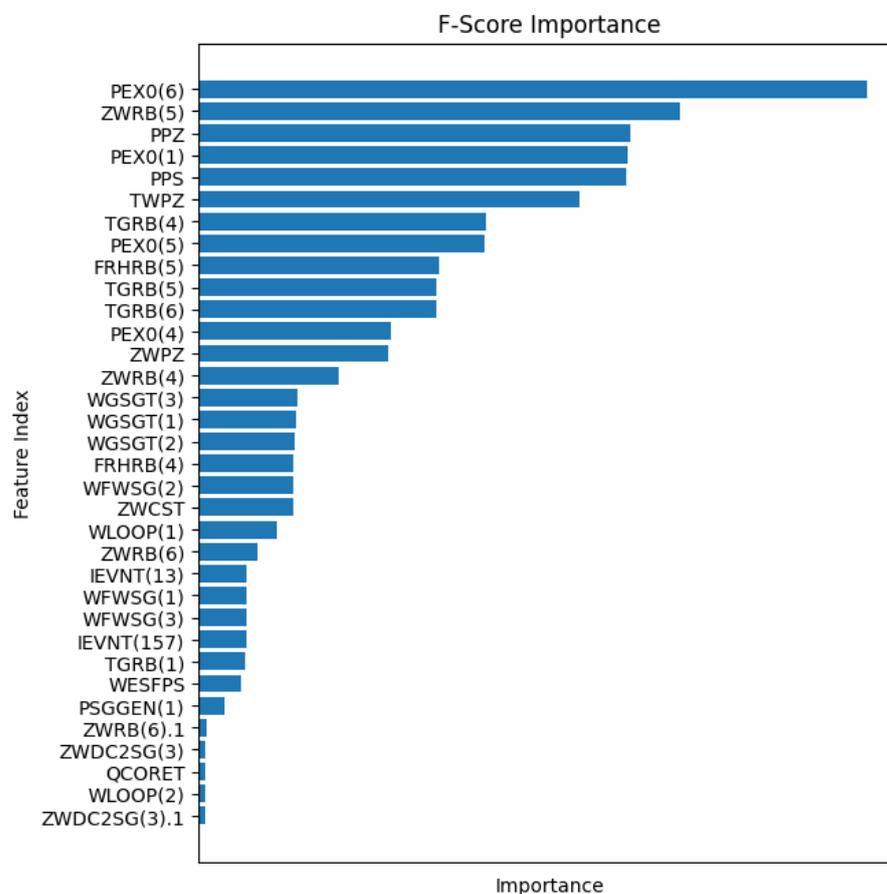


Figure 6. The F-Score of some features.

Finally, feature screening based on expert knowledge was employed to further refine the selection, resulting in the identification of 25 key system parameters that were to be utilized as modeling features. The screening features are presented in Table 1.

Table 1. The screening features by feature engineering.

Feature Name	Feature Name	Feature Name	Feature Name
QCORET	TWPZ	ZWDC2SG(1)–W	ZWRB(4)
PPS	ZWPZ	ZWDC2SG(1)–N	FRHRB(5)
THL1	WLOOP(1)	WFWSG(1)	ZWRB(6)
TCL1	WLOOP(2)	WGS(1)	TWRB(6)
TGCH(1)	WLOOP(3)	PEX0(4)	ZWCST
PPZ	PSGGEN(1)	TGRB(4)	PACUM
TGPZ			

### 4.3. Experimental Setting

Following the application of data processing techniques, including filtering the original dataset and normalizing the data with a MinMaxScaler method, a sliding window mechanism with a step size of 50 was set to roll and split the data. By using these processed datasets, models were developed to predict the progression of two different types of severe nuclear power accidents, namely an MSLB and LOCA.

This study proposes a GRUS model that uses the historical data of an MSLB and LOCA to predict the respective development trends of the two nuclear power accidents. The

model is optimized by tuning hyperparameters, including the number of network layers, the number of neurons in each hidden layer, the learning rate of the optimizer, and the batch size used during training. Through the training and optimization of the GRUS model, the optimal parameters are obtained. The GRUS model in this study was constructed with a structure of 1 input layer, 3 hidden layers, and 1 output layer, where each hidden layer had 256 neurons. The optimal parameters are shown in Table 2.

**Table 2.** Optimal parameter settings for the GRUS framework.

Parameters	Values
Learning Rate	0.0001
Number of input layer units	25
Number of hidden layers	3
Number of neurons	256
Batch_size	32
Number of epochs	300
Optimizer	Adam
Dropout	0.2

There are two types of experiments, quantitative evaluation and qualitative evaluation. The quantitative evaluation experiment involves the comparison of the GRUS model framework with both ARIMAX, LSTM, and GRU, aimed at verifying its accuracy effect. In the qualitative evaluation experiment, given the lack of interpretability of the model, an interpretability analysis of SHAP in the GRUS framework is performed to shed light on the mechanism underlying the model's predictive performance. This approach facilitates a more thorough exploration of the model and improves its overall interpretability.

#### 4.4. Experimental Results

##### 4.4.1. Quantitative Evaluation

The experiment conducted a quantitative analysis to verify the accuracy of the GRUS framework on the time series prediction of nuclear power accident scenarios. The prediction performance of the GRUS was compared with that of the ARIMAX, LSTM, and GRU models using evaluation metrics such as MSE, MAE, and MAPE through quantitative analysis. The primary loop flow (i.e., WLOOP(1)), which is the most affected parameter during MSLB accidents, was chosen as the first key predictive feature, whereas pressurizer pressure (i.e., PPZ) was selected as the predictive feature for LOCAs. The experimental results are shown in Table 3, with the data in bold representing the optimal values after comparison.

**Table 3.** Performance comparison of the models' prediction results.

Scenarios	Model	MSE	MAE	MAPE
MSLB	ARIMAX	$4.6798 \times 10^{-3}$	$5.7732 \times 10^{-2}$	2.2553%
	LSTM	$1.7419 \times 10^{-4}$	$1.2245 \times 10^{-2}$	0.3077%
	GRU	$4.1627 \times 10^{-5}$	$5.0610 \times 10^{-3}$	0.0862%
	GRUS	<b><math>2.1881 \times 10^{-5}</math></b>	<b><math>3.2231 \times 10^{-3}</math></b>	<b>0.0587%</b>
LOCA	ARIMAX	$1.0805 \times 10^{-4}$	$6.6404 \times 10^{-3}$	0.2004%
	LSTM	$6.8415 \times 10^{-5}$	$6.8512 \times 10^{-3}$	0.5112%
	GRU	$3.7715 \times 10^{-5}$	$4.7580 \times 10^{-3}$	0.3553%
	GRUS	<b><math>8.6385 \times 10^{-6}</math></b>	<b><math>1.7704 \times 10^{-3}</math></b>	<b>0.0373%</b>

Table 3 presents the evaluation results of the GRUS framework, and the ARIMAX, LSTM, and GRU models for predicting the time series data related to LOCAs and MSLB accidents. The GRUS framework outperforms the other models in terms of all evaluation metrics, including MSE, MAE, and MAPE.

Figures 7 and 8 show the predicted results of the GRUS and other models taking a continuous 100 s of data after the accident. As can be observed from these figures, the predicted values of GRUS are much closer to the actual values, as indicated by the lower absolute differences between the predicted and actual values. Therefore, the GRUS framework can provide more accurate predictions and is a suitable choice for predicting severe nuclear power accidents due to its superior predictive capabilities.

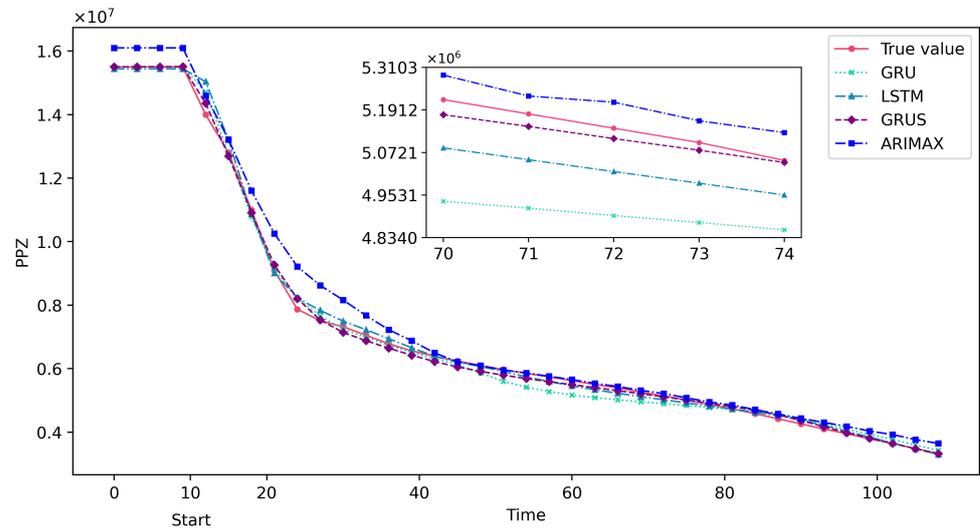


Figure 7. Partial prediction results of GRUS and other models on the test data in the LOCA.

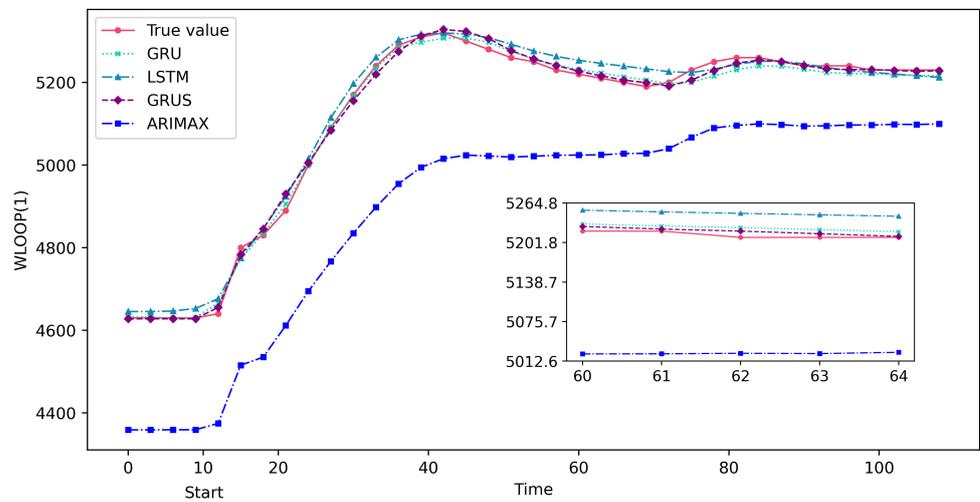


Figure 8. Partial prediction results of GRUS and other models on the test data in the MSLB accident.

#### 4.4.2. Qualitative Evaluation

The experiment conducted a qualitative analysis on the SHAP part of the GRUS framework. From the two perspectives of local and global interpretability, the SHAP index distribution of different features on a single sample and global samples was visualized to reveal the importance of the support of various features in the dataset to the prediction results. Additionally, an interpretability analysis was performed on the feature dependencies extracted from the prediction model to improve the prediction performance of the model.

- Global Interpretability

Taking the MSLB accident as an example, Figure 9 shows an analysis of feature importance using SHAP in the GRU model of the GRUS framework, which can understand

the importance of features in the neural network model. The first five important influencing factors were investigated, with WLOOP(1) identified as having the most significant impact on the model because the primary loop flow changes rapidly following the rupture of the main steam pipe. Other influential features include TCL1, TGCH(1), TGRB(4), and WLOOP(3), suggesting that parameters such as temperature also undergo significant changes. It can be seen that during the training process of the model, the neural network nodes focus on relevant feature nodes to realize accurate predictions of future results.

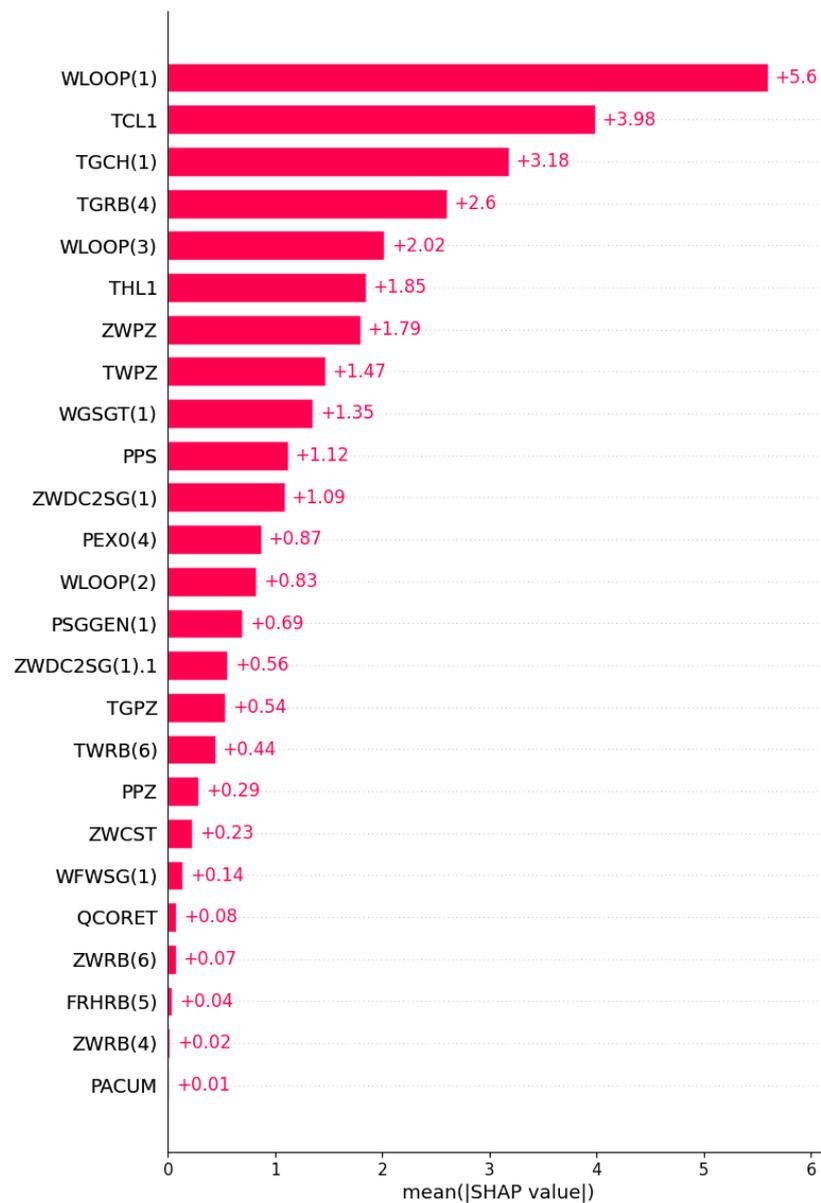


Figure 9. The importance of input features in the SHAP.

Figure 10 shows the distribution of the SHAP value of each influencing factor. Each data point on each feature stream in the figure represents each instance. The SHAP value of each feature determines the position of the point on the X-axis, the points along the feature rows are stacked to show the density, and the color indicates the value of the feature in the original dataset, from blue to red to indicate the change in the value from low to high. A positive SHAP value indicates that the contribution to the target variable is positive, and a negative SHAP value suggests that the contribution is negative. In addition, the variable’s value decreases as the points on the graph move closer to being

blue and increases as they move closer to being pink. For instance, if the air temperature in the upper compartment of the containment is raised (i.e., TGRB(4) is increased), the corresponding SHAP value of TGRB(4) also increases. This indicates that an increase in the air temperature of the upper compartment of the containment vessel caused by a rupture in the main steam pipe, produces a subsequent increase in the temperature of the coolant on the primary circuit side of the steam generator, thereby elevating the water vapor pressure inside the primary circuit. In order to safeguard the safety and stability of the reactor core, the flow rates and pressures within the system are precisely regulated via the control system. Consequently, the primary loop flow usually increases with the increase in the temperature in the upper compartment of the containment vessel, showing that the characteristic TGRB(4) is positively correlated with the predicted target WLOOP(1). However, the SHAP analysis suggests that WLOOP(1) is negatively correlated with the temperature of the cold pipe section (i.e., TCL1). Specifically, as the temperature of TCL1 decreases, its SHAP value will decrease, resulting in an increase in the predicted target. This can be explained by the fact that a decrease in the temperature of the cold pipe section results in an increase in the density of the coolant, leading to a reduction in the flow resistance and ultimately causing an elevation in the flow rate within the pipe. To sum up, the ability of the model to accurately predict outcomes can be attributed to its capacity for learning the interdependencies between features during the training process, allowing it to acquire a comprehensive understanding of feature interactions.

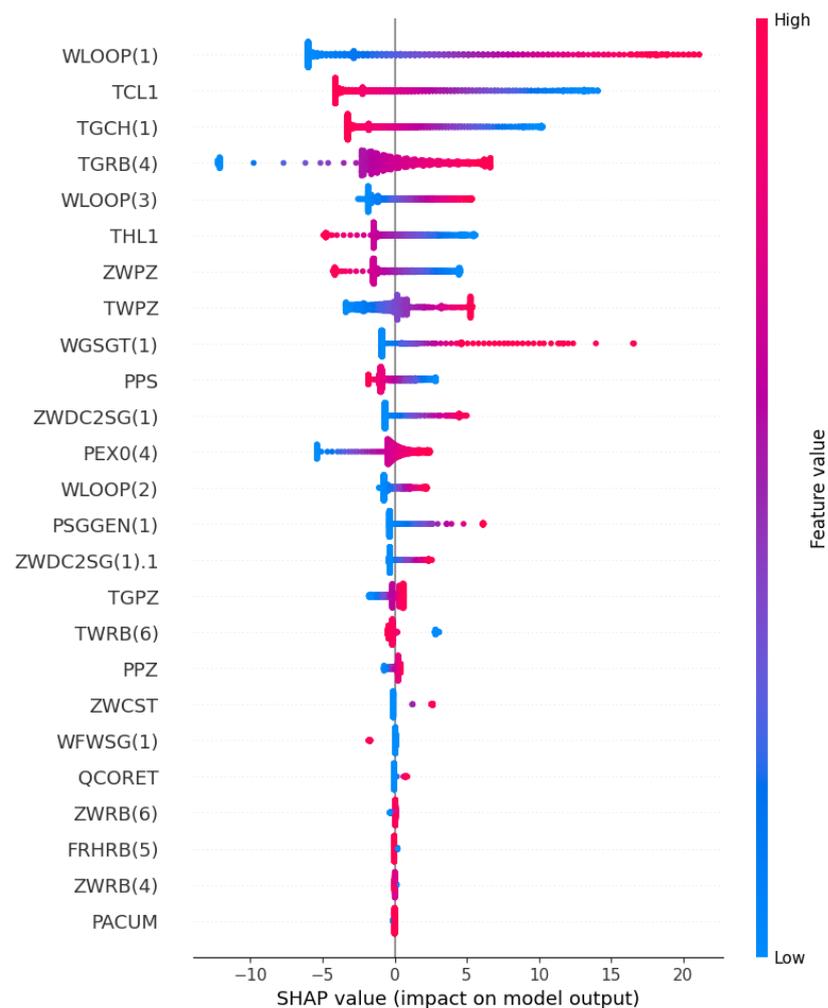
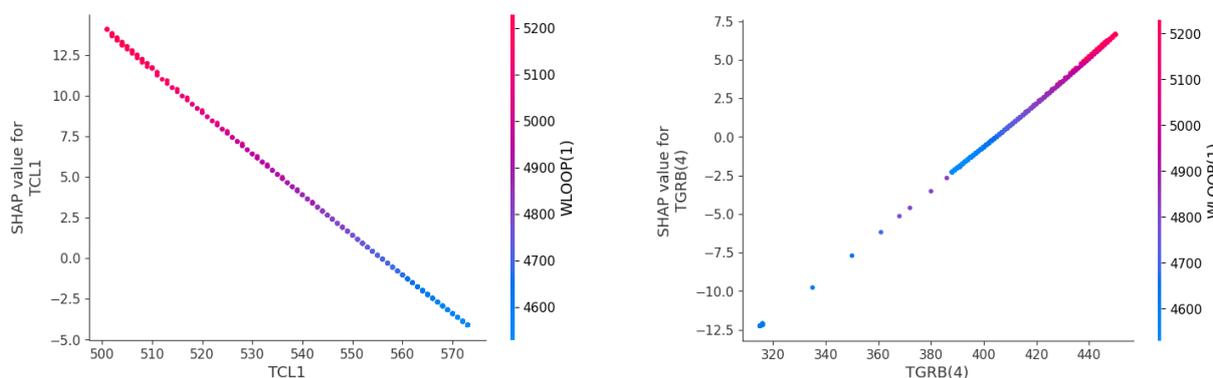


Figure 10. The beeswarm plot of input features in the SHAP.

- Local Interpretability

Figure 11 shows a SHAP dependency plot, which illustrates the impact of the observation's variable values on the predicted result. The dependency plot enables a comprehensive description of both the major effects of individual observations and their interactions. The left figure in Figure 11 demonstrates that an increase in the temperature of the cold pipe section leads to a corresponding decrease in its SHAP value and contribution, resulting in a decrease in the estimated target WLOOP(1). It shows an inverse relationship between WLOOP(1) and TCL1, with a corresponding negative correlation between the two variables. The right figure depicts the connection between the air temperature in the top compartment of the containment and the primary loop flow rate. As TGRB(4) gradually increases, the positive correlation between TGRB(4) and WLOOP(1) becomes stronger.



**Figure 11.** The dependency plot of TCL1 analysis in the SHAP (left panel). The dependency plot of TGRB(4) analysis in the SHAP (right panel).

The model provides predictions based on the most significant chosen attributes, as demonstrated by the utilization of the SHAP approach for interpretability analysis. The identified relationship between the chosen input features and the predicted output variable enables robust conclusions to be drawn from the model's analysis and application. Furthermore, the model can be further optimized by removing less important features, improving the speed and efficiency of the prediction process.

## 5. Conclusions

In the event of a severe nuclear power accident, accurately predicting real-time operating conditions can transform nuclear power safety control from reactive measures to proactive intervention. Furthermore, incorporating interpretability analysis to comprehend the results of AI models enhances the trust of the operation and maintenance personnel, which is essential in providing effective decision support for nuclear power emergency response strategies.

This study proposes an interpretable hybrid framework that combines GRU and SHAP to address the challenges of accurately predicting key parameters in nuclear power accidents. By analyzing two types of accidents, an MSLB and an LOCA, we demonstrate the successful application of this model. The experiment carries out quantitative analysis and qualitative analysis. By comparing it with the ARIMAX, LSTM, and GRU models, it is shown that the GRUS framework is superior to other baseline models in nuclear power accident prediction. The evaluation of the two accident indexes results in MSEs of  $8.6385 \times 10^{-6}$  and  $2.1881 \times 10^{-5}$ , MAEs of  $1.7704 \times 10^{-3}$  and  $3.2231 \times 10^{-3}$ , and MAPEs of 0.0373% and 0.0587%. Simultaneously, the SHAP framework provides an explanation of the model and analyzes the influence of variables on different accidents, identifying WLOOP(1), TCL1, TGCH(1), and TGRB(4), etc. as indicators with significant impacts in the MSLB accident. Moreover, the correlation between TCL1 and TGRB(4) and the predicted target primary loop flow WLOOP(1) is analyzed. These results can enhance nuclear power

workers' understanding of the relationship between complex parameter indicators and facilitate decision making by optimizing resource utilization through the consideration of fewer features. Furthermore, these results enable the safety system to intervene proactively before the actual accident occurs, thereby mitigating the potential for more severe losses.

Several issues require further research beyond this work. In severe nuclear power accidents, it is necessary to utilize diverse breach accident data for analysis; as our experiment solely considers fixed breach data in two types of accidents, additional accident data can supplement future analyses. Additionally, real accident data from nuclear power plants can be utilized as a complement to this study if available. Furthermore, in terms of interpretability analysis, other explanation methods can be employed to model predictions to enhance the effectiveness of the model. Therefore, future research should explore and address these issues to augment nuclear power safety control measures.

**Author Contributions:** Conceptualization, Y.F. and H.Z.; methodology, Y.F. and H.Z.; software, Y.F. and D.Z.; validation, Y.F., D.Z. and H.Z.; formal analysis, Y.F.; investigation, Y.F. and H.Z.; resources, Y.F. and Z.W.; data curation, Y.F. and Z.W.; writing—original draft preparation, Y.F.; writing—review and editing, Y.F. and H.Z.; visualization, Y.F. and Z.W.; supervision, H.Z.; project administration, D.Z., Y.X. and H.Z.; funding acquisition, D.Z., Y.X. and H.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded in part by the National Natural Science Foundation of China under Grants 62171327, 62171328, and 62072350, and in part by the first batch of the application of the basic technology and science research foundation from the Hubei Nuclear Power Operation Engineering Technology Research Center under Grant B210610.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data are available from the corresponding author of this paper.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Lovering, J.R.; Yip, A.; Nordhaus, T. Historical construction costs of global nuclear power reactors. *Energy Policy* **2016**, *91*, 371–382. [CrossRef]
- Dai, H.; Su, Y.; Kuang, L.; Liu, J.; Gu, D.; Zou, C. Contemplation on China's energy-development strategies and initiatives in the context of its carbon neutrality goal. *Engineering* **2021**, *7*, 1684–1687. [CrossRef]
- Huang, Q.; Peng, S.; Deng, J.; Zeng, H.; Zhang, Z.; Liu, Y.; Yuan, P. A review of the application of artificial intelligence to nuclear reactors: Where we are and what's next. *Heliyon* **2023**, *9*, e13883. [CrossRef] [PubMed]
- Sikorska, J.Z.; Hodkiewicz, M.; Ma, L. Prognostic modelling options for remaining useful life estimation by industry. *Mech. Syst. Signal Process.* **2011**, *25*, 1803–1836. [CrossRef]
- Wu, Y.; Liu, J.; Cui, H.; Zhang, Q.; Shang, B.; Deng, J. Application of ARIMA models in the forecast of peripheral environmental radioactivity level of nuclear power plants. *Mod. Prev. Med.* **2014**, *41*, 1941–1943.
- Alamaniotis, M.; Tsoukalas, L.H.; Ikonomopoulos, A. On-line Surveillance of Nuclear Power Plant Peripheral Components Using Support Vector Regression. 2011. Available online: [https://inis.iaea.org/search/search.aspx?orig\\_q=RN:44079341](https://inis.iaea.org/search/search.aspx?orig_q=RN:44079341) (accessed on 20 March 2023).
- Yoo, K.H.; Back, J.H.; Na, M.G.; Kim, J.H.; Hur, S.; Kim, C.H. Prediction of golden time using SVR for recovering SIS under severe accidents. *Ann. Nucl. Energy* **2016**, *94*, 102–108. [CrossRef]
- Jiang, B.; Zhou, J.; Huang, X. Artificial Neural Networks in Condition Monitoring and Fault Diagnosis of Nuclear Power Plants: A Concise Review. In Proceedings of the International Conference on Nuclear Engineering: Virtual, Online, 4–5 August 2020; American Society of Mechanical Engineers: New York, NY, USA, 2020; Volume 83778, p. V002T08A032.
- Hu, G.; Zhou, T.; Liu, Q. Data-driven machine learning for fault detection and diagnosis in nuclear power plants: A review. *Front. Energy Res.* **2021**, *9*, 663296. [CrossRef]
- Song, H.; Liu, X.; Song, M. Comparative study of data-driven and model-driven approaches in prediction of nuclear power plants operating parameters. *Appl. Energy* **2023**, *341*, 121077. [CrossRef]
- Hong, S.R.; Hullman, J.; Bertini, E. Human factors in model interpretability: Industry practices, challenges, and needs. *Proc. ACM Hum.-Comput. Interact.* **2020**, *4*, 1–26. [CrossRef]
- Moshkbar-Bakhshayesh, K.; Ghofrani, M.B. Development of a robust identifier for NPPs transients combining ARIMA model and EBP algorithm. *IEEE Trans. Nucl. Sci.* **2014**, *61*, 2383–2391. [CrossRef]

13. Santhosh, T.; Kumar, M.; Thangamani, I.; Srivastava, A.; Dutta, A.; Verma, V.; Mukhopadhyay, D.; Ganju, S.; Chatterjee, B.; Rao, V.; et al. A diagnostic system for identifying accident conditions in a nuclear reactor. *Nucl. Eng. Des.* **2011**, *241*, 177–184. [[CrossRef](#)]
14. Kim, D.Y.; Yoo, K.H.; Kim, J.H.; Na, M.G.; Hur, S.; Kim, C.H. Prediction of leak flow rate using fuzzy neural networks in severe post-LOCA circumstances. *IEEE Trans. Nucl. Sci.* **2014**, *61*, 3644–3652. [[CrossRef](#)]
15. Do Koo, Y.; Na, M.G.; Kim, K.S.; Kim, C.H. Prediction of nuclear reactor vessel water level using deep neural networks. In Proceedings of the 2018 International Conference on Electronics, Information, and Communication (ICEIC), Honolulu, HI, USA, 24–27 January 2018; IEEE: Piscataway, NJ, USA; 2018; pp. 1–3.
16. Tian, D.; Deng, J.; Vinod, G.; Santhosh, T. A constraint-based random search algorithm for optimizing neural network architectures and ensemble construction in detecting loss of coolant accidents in nuclear power plants. In Proceedings of the 2018 11th International Conference on Developments in eSystems Engineering (DeSE), Cambridge, UK, 2–5 September 2018; IEEE: Piscataway, NJ, USA; 2018; pp. 237–243.
17. Sun, P.; Cao, H.; Li, B. Fault Identification of a small modular reactor based on PCA method. In Proceedings of the International Conference on Nuclear Engineering (ICONE), Tsukuba, Ibaraki, Japan, 19–24 May 2019; The Japan Society of Mechanical Engineers, Shinjuku, Tokyo, Japan; 2019; p. 1254.
18. Wang, H.; Peng, M.J.; Yu, Y.; Saeed, H.; Hao, C.M.; Liu, Y.K. Fault identification and diagnosis based on KPCA and similarity clustering for nuclear power plants. *Ann. Nucl. Energy* **2021**, *150*, 107786. [[CrossRef](#)]
19. Yang, J.; Yang, Y.; Deng, C.; Ishii, M. Best Estimate Plus Uncertainty analysis of a large break LOCA on Generation III reactor with RELAP5. *Ann. Nucl. Energy* **2019**, *127*, 326–340. [[CrossRef](#)]
20. Li, X.; Fu, X.M.; Xiong, F.R.; Bai, X.M. Deep learning-based unsupervised representation clustering methodology for automatic nuclear reactor operating transient identification. *Knowl.-Based Syst.* **2020**, *204*, 106178. [[CrossRef](#)]
21. Yao, Y.; Wang, J.; Long, P.; Xie, M.; Wang, J. Small-batch-size convolutional neural network based fault diagnosis system for nuclear energy production safety with big-data environment. *Int. J. Energy Res.* **2020**, *44*, 5841–5855. [[CrossRef](#)]
22. Lin, T.H.; Wang, T.C.; Wu, S.C. Deep learning schemes for event identification and signal reconstruction in nuclear power plants with sensor faults. *Ann. Nucl. Energy* **2021**, *154*, 108113. [[CrossRef](#)]
23. He, C.; Ge, D.; Yang, M.; Yong, N.; Wang, J.; Yu, J. A data-driven adaptive fault diagnosis methodology for nuclear power systems based on NSGAI-CNN. *Ann. Nucl. Energy* **2021**, *159*, 108326. [[CrossRef](#)]
24. Şeker, S.; Ayaz, E.; Türkcan, E. Elman's recurrent neural network applications to condition monitoring in nuclear power plant and rotating machinery. *Eng. Appl. Artif. Intell.* **2003**, *16*, 647–656. [[CrossRef](#)]
25. Radaideh, M.I.; Pigg, C.; Kozłowski, T.; Deng, Y.; Qu, A. Neural-based time series forecasting of loss of coolant accidents in nuclear power plants. *Expert Syst. Appl.* **2020**, *160*, 113699. [[CrossRef](#)]
26. Nguyen, H.P.; Baraldi, P.; Zio, E. Ensemble empirical mode decomposition and long short-term memory neural network for multi-step predictions of time series signals in nuclear power plants. *Appl. Energy* **2021**, *283*, 116346. [[CrossRef](#)]
27. She, J.; Shi, T.; Xue, S.; Zhu, Y.; Lu, S.; Sun, P.; Cao, H. Diagnosis and prediction for loss of coolant accidents in nuclear power plants using deep learning methods. *Front. Energy Res.* **2021**, *9*, 665262. [[CrossRef](#)]
28. Zhang, C.; Chen, P.; Jiang, F.; Xie, J.; Yu, T. Fault Diagnosis of Nuclear Power Plant Based on Sparrow Search Algorithm Optimized CNN-LSTM Neural Network. *Energies* **2023**, *16*, 2934. [[CrossRef](#)]
29. Gong, S.; Yang, S.; She, J.; Li, W.; Lu, S. Multivariate time series prediction for loss of coolant accidents with a sigmoid-based LSTM. *Front. Energy Res.* **2022**, *10*, 294. [[CrossRef](#)]
30. Lei, J.; Ren, C.; Li, W.; Fu, L.; Li, Z.; Ni, Z.; Li, Y.; Liu, C.; Zhang, H.; Chen, Z.; et al. Prediction of crucial nuclear power plant parameters using long short-term memory neural networks. *Int. J. Energy Res.* **2022**, *46*, 21467–21479. [[CrossRef](#)]
31. Zhang, J.; Pan, Z.; Bai, W.; Zhou, X. Pressurizer Water Level Reconstruction for Nuclear Power Plant Based on GRU. In Proceedings of the 2018 Eighth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC), Harbin, China, 19–21 July 2018; IEEE: Piscataway, NJ, USA; 2018; pp. 1675–1679.
32. Wang, H.; Peng, M.J.; Ayodeji, A.; Xia, H.; Wang, X.K.; Li, Z.K. Advanced fault diagnosis method for nuclear power plant based on convolutional gated recurrent network and enhanced particle swarm optimization. *Ann. Nucl. Energy* **2021**, *151*, 107934. [[CrossRef](#)]
33. Fukun, C.; Xiaomeng, D.; Yicheng, L. GRU-CNN-Based Prediction of LOCA Accident Condition in Nuclear Power Plants. In Proceedings of the 2022 4th International Conference on System Reliability and Safety Engineering (SRSE), Guangzhou, China, 15–18 December 2022; IEEE: Piscataway, NJ, USA; 2022; pp. 174–179.
34. He, Q.; Li, B.; Peng, S. UTSG Water Level Prediction Model Based on LSTM and GRU Algorithms. In Proceedings of the International Conference on Nuclear Engineering: Virtual, Online, 8–12 August 2022 American Society of Mechanical Engineers: New York, NY, USA, 2022; Volume 86373, p. V003T03A015.
35. Kim, Y.G.; Park, D.S. Consideration on the use of Explainable AI in Operator Support System. In Proceedings of the KNS Autumn Meeting: Virtual, Online, 17–18 December 2020; pp. 13–14.
36. Mortenson, T.; Hill, R.; Hall, A. Considerations for Artificial Intelligence and Machine Learning in Nuclear Power: Literature Review. Available online: [https://scholar.google.com.hk/citations?view\\_op=view\\_citation&hl=zh-CN&user=bARXPC0AAAAJ&cstart=20&pagesize=80&citation\\_for\\_view=bARXPC0AAAAJ:kNdYIx-mwKoC](https://scholar.google.com.hk/citations?view_op=view_citation&hl=zh-CN&user=bARXPC0AAAAJ&cstart=20&pagesize=80&citation_for_view=bARXPC0AAAAJ:kNdYIx-mwKoC) (accessed on 22 March 2023).
37. Park, J.H.; Jo, H.S.; Lee, S.H.; Oh, S.W.; Na, M.G. A reliable intelligent diagnostic assistant for nuclear power plants using explainable artificial intelligence of GRU-AE, LightGBM and SHAP. *Nucl. Eng. Technol.* **2022**, *54*, 1271–1287. [[CrossRef](#)]

38. Shin, J.H.; Bae, J.; Kim, J.M.; Lee, S.J. An interpretable convolutional neural network for nuclear power plant abnormal events. *Appl. Soft Comput.* **2023**, *132*, 109792. [[CrossRef](#)]
39. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* **2014**, arXiv:1406.1078.
40. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [[CrossRef](#)]
41. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A survey of methods for explaining black box models. *ACM Comput. Surv.* **2018**, *51*, 1–42. [[CrossRef](#)]
42. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4768–4777.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.