Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.12 n.57, p. 930-940. January/February, 2026.
DOI: https://doi.org/10.5935/jetia. v12i57.3132

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

RANDOM FOREST GUIDED FEATURE WEIGHTED KNN MODEL FOR NUCLEAR REACTOR EVENT IDENTIFICATION

Arunprasath V[1], T V Santhosh[2], Gopika Vinod[3]

[1]Homi Bhabha National Institute, Mumbai, India
[2]Reactor Safety Division, Bhabha Atomic Research Centre, India
[3]Board of Research in Nuclear Sciences, India

[1]http://orcid.org/0009-0008-2401-4788, [2]http://orcid.org/0000-0003-3266-1324, [3]http://orcid.org/0000-0003-4460-7341

E-mail: *arunv@barc.gov.in

ARTICLE INFO

Article History
Received: December 13, 2025
Revised: January 10, 2026
Accepted: January 15, 2026
Published: February 28, 2026

Keywords:
Machine learning,
Random Forest,
KNN,
Event detection,
Nuclear reactor monitoring.

ABSTRACT

Accurate identification of developing events is crucial in achieving reactor operation safety, decision making and accident management. Machine learning techniques are increasingly used as alternative to conventional model based methods due to their data driven nature for abnormal event detection. Among these, the K-Nearest Neighbors (KNN) algorithm remains widely used across industrial diagnostic systems, including applications in nuclear reactor monitoring, owing to its simplicity and good performance in multi class classification tasks. However, a key limitation of standard KNN is that all input features are treated with equal importance, which can limit its accuracy. To address this limitation, this study proposes a Random Forest guided feature weighted KNN classifier for reactor event identification. In the proposed approach, feature importance extracted from a Random Forest model is utilized to strengthen the ability of standard KNN model. The method is evaluated on a five class nuclear reactor event dataset and is found to outperform the base KNN and RF models. The performance of the proposed model is also compared with other KNN - RF ensemble methods and is shown to perform the best among all with superior performance metrics.

I. INTRODUCTION

Early detection of abnormal conditions is very important to preserving reactor safety through timely intervention by operators. Conventionally, decisions about plant operation in the control room were based on alarms/trip systems, displays for monitoring plant conditions, and emergency operating procedures. These systems have limitations because they were developed from simplifications of fundamental physical laws and they operate in response to those events which are covered under such principles. Further, accurate modelling of nuclear reactors via first principles could be challenging due to the complicated process behavior, therefore it limits the uses of model-based diagnostics entirely. On the other hand, in contrast to first principle methods, data driven methods learn patterns from operational data and real time measurements and hence do not require detailed mathematical descriptions regarding reactor behavior. Since modern industrial plants including nuclear plants acquire large amounts of operational data, the information can be utilized to train machine leaning models to identify developing events in their earlier stages. Whenever an event occurs, operators in a nuclear reactor control room is expected to analyze many signals, alarms and trends simultaneously under huge mental pressure. Human error has been identified as an important contributor to nuclear accidents worldwide. Hence, machine learning techniques can be used to develop data driven systems learning from huge volumes of operational data to provide diagnostic support to the operators reducing human errors. Owing to their benefits in terms of safe and cost effective reactor operation, many techniques for the identification and diagnosis of events have been studied. In general, such approaches aim to represent relationships within reactor systems or derive informative features through pattern recognition strategies, enabling the detection of abnormal reactor conditions based on measured process data [1]. In recent years, there has been a growing shift toward the use of data driven methodologies for monitoring nuclear reactors.

Currently, a wide range of machine learning techniques—such as neural networks, support vector machines (SVMs), dimensionality reduction algorithms, ensemble learning frameworks, regression models, and others—have been adopted to forecast the operational behavior of nuclear power plants, as noted by Hu G et al. [2]. Moreover, machine learning tools, including advanced deep learning architectures, are increasingly applied for condition monitoring, equipment health evaluation, and remaining useful life (RUL) estimation within nuclear facilities [3], [4]. Studies illustrate the use of SVM based models [5], [6], where deviations in process variables are exploited to identify potential accident scenarios in reactors. Reference [7] proposes a diagnostic approach that not only detects faults but also quantifies their severity by combining Principal Component Analysis with neural network techniques. In [8], Santosh et al. developed a neural network driven framework to diagnose severe coolant loss events in nuclear power plants. Additional research has explored a variety of data driven solutions for assessing and managing plant wide and component level health conditions, as summarized in [9].

Reference [10] examines multiple classification techniques—such as PCA, SVM, decision trees, and hybrid schemes—for identifying equipment and actuator faults in PWR systems. Meanwhile, [11] reports that ensembles based on nearest neighbor classifiers achieved the highest accuracy for fault detection, whereas ensembles composed solely of decision trees performed comparatively poorly. Motivated by these works, our earlier work involved development of KNN model for the purpose of transient event identification in reactors [12]. The developed KNN model achieved an accuracy of 97% which is a hold out test data performance. In this work, a novel method is proposed for an improved KNN model for event identification with enhanced accuracy. The proposed model is a Random Forest guided feature weighted KNN classifier to improve the accuracy of classification. By integrating the local feature relevance derived from Random Forests with the neighbourhood based decision structure of KNN, the proposed model is shown to outperform the basic KNN model and other KNN-RF ensemble models. Different performance metrics of the model such as categorical accuracy, confusion matrix, class wise performance score are studied and are shown to provide the best results when compared with the other models.

II. PROBLEM STATEMENT AND DATASET

II.1 EVENT DETECTION

Systems that support timely event detection are crucial to maintaining the safe operation of any nuclear reactor. Research or test reactors, in particular, consist of highly intricate subsystems designed both for operational safety and for carrying out diverse experimental activities. This paper focuses on identifying abnormal events in such test reactors using a data driven strategy built upon a novel machine learning framework. In these facilities, the reactor core is usually located within a large pool filled with demineralized light water, which serves simultaneously as coolant and moderator. Under normal operating conditions with offsite electrical power available, primary cooling pumps ensure forced circulation around the core. Heated coolant is routed through heat exchangers, where its thermal energy is transferred to the secondary cooling loop. The secondary loop employs its own set of pumps that drive water toward cooling towers for ultimate heat removal.

If offsite power is lost, core cooling relies on natural circulation within the pool, enabled by passive valves engineered to open automatically and permit flow. Throughout operation, vital plant variables—including process parameters and equipment status indicators—are continuously monitored and archived. Because of their extensive instrumentation, these reactors generate and log large volumes of process data [13], [14]. Within nuclear reactor safety design and assessment, Postulated Initiating Events (PIEs) play a critical role. These events represent potential triggers that may drive the reactor away from its normal operational state, possibly progressing into anticipated operational occurrences or even accident scenarios. Rapid identification and mitigation of such events are essential, making the precision of detection systems or models highly important. For this reason, the present study approaches the challenge through the application of an appropriate and innovative machine learning methodology.

II.2 DATASET

Training effective machine learning models relies on the availability of credible and well defined datasets. In this work, the training data are generated by simulating key PIE scenarios using thermal hydraulic codes capable of reproducing the dynamic behavior of accident conditions in light water reactors, including research reactors [15], [16]. Such events represent transients which are departures from normal conditions that may occur due to abnormal system behavior. These transients can stem from a variety of sources, including equipment malfunctions, operator mistakes, and both internal and external hazards. For the reactor examined in this study, the relevant events were selected in accordance with IAEA SSR-3, the international standard that outlines essential safety requirements for research reactors [17]. Four significant events typical of test research reactors were chosen, and simulated data from these scenarios were used for model development. Table 1 summarizes the set of events considered in this work. Each scenario is generated through a detailed event sequence analysis, which produces time dependent trends for the plant parameters affected during the transient. For the four selected events, 16 operational variables were identified as influential. These 16 measurements serve as the feature set used to train and construct the machine learning models. Consequently, the dataset is a labeled one comprising five classes, including a baseline class that represents normal operation with no event. The feature set includes both continuous variables such as reactor power, coolant flow rate, and heat exchanger temperatures, and categorical indicators, including various reactor trip signals, alarm activations, and the status of passive cooling valves, among others.

Table 1: Events modeled for detection.

S.No.	Event
1	Failure of offsite power event
2	Failure of pump flow event
3	Loss of coolant event
4	Failure of heat sink event

Source: Authors, (2026).

III. METHODOLOGY

III.1 K-NEAREST NEIGHBORS CLASSIFIER

One of the simplest and most understandable supervised learning methods for classification is the K-Nearest Neighbors (KNN) algorithm. The main principle is that in the feature space, data points belonging to the same class typically appear near one another. Therefore, we may use the labels of the closest samples in the training set to estimate the class of a new, unseen sample. Considering training dataset D,

$$\{(x_1, y_1), (x_2, y_2), \dots (x_N, y_N)\}, \tag{1}$$

where $x_1 \epsilon \ \mathbb{R}^d$ represents a d dimensional feature vector and y_i is the corresponding class label, the goal is to assign class to a new sample x. The KNN classifier performs this in the following steps:

1. Measure the distance between the test sample x and all samples in the training set.
2. Identify the nearest K samples, referred to as the "neighbors."
3. Perform majority voting among these Kneighbors.
4. The class receiving the highest number of votes is assigned to x.

Mathematically, this can be expressed as:

$$\hat{y}(x) = mode\{y(1), y(2), \dots, y(K)\}, \tag{2}$$

where $y(i)$ denotes the class label of the i-th nerest neighbor.

Distance metric plays a crucial role in KNN. Commonly used metrics include:

- Euclidean distance

$$\mathrm{d_E(x, x_i)} = \sqrt{\sum_{\mathrm{j=1}}^{\mathrm{d}} \left(\mathrm{x_j - x_{i,j}}\right)^2} \tag{3}$$

- Manhattan distance

$$d_M(x, x_i) = \sum_{j=1}^{d} \left|x_j - x_{i,j}\right| \tag{4}$$

The decision boundary formed by KNN is non linear and adapts to the data distribution. This characteristic makes it suitable for complex classification tasks such as nuclear event identification, where the feature relationships are often non linear. The choice of Ksignificantly affects the classifier's performance. A small K makes the classifier sensitive to noise and outliers, potentially overfitting. A larger K smooths the decision boundary and improves generalization but may overlook small scale structures in the data [18]. KNN is a relatively simple to use and straightforward algorithm. Since there is no training phase, it is adaptable to incremental data. KNN is one of the most suitable models to handle multiclass problems. However, KNN suffers from the curse of dimensionality and is sensitive to duplicate or noisy features [19]. Without specific weighting, contribution of each feature is treated equally by the algorithm. These limitations are especially pertinent to complicated industrial datasets, such nuclear reactor transients, where some features could be more important than others. This insight naturally emphasizes the necessity for an improved classifier that guides the decision making process by incorporating extra knowledge, such as feature importance information.

III.2 RANDOM FOREST CLASSIFIER

Random Forest is an ensemble learning method that constructs a collection of decision trees and aggregates their outputs to improve classification reliability. The core idea is that multiple weak learners, when combined in an intelligent manner, can form a strong and robust classifier. Random Forests are particularly effective for complex, nonlinear datasets such as those encountered in nuclear reactor monitoring, where different features may influence the event dynamics in varying ways [20]. The concepts of random feature selection and bootstrap aggregation (bagging) are used to construct a Random Forest. A bootstrap sample taken from the original dataset is used to train each tree in the forest. Furthermore, only a randomly chosen subset of features is taken into account for determining the optimal split during each tree's growth. Because the trees are diverse due to these two types of randomness, there is less chance of associated errors [21]. Let the training dataset be

$$\{(x_1, y_1), (x_2, y_2), \dots (x_N, y_N)\} \tag{5}$$

Where $x_1 \epsilon \ \mathbb{R}^d$ represents a d dimensional feature vector and y_i is the corresponding class label. For each tree T_b, a bootstrap sample $D_b$is drawn by sampling N times with replacement from the original dataset D. Some samples may appear multiple times, while others may be excluded (out of bag samples). At each internal node of tree T_b, instead of evaluating all features, only a randomly selected subset $F_b \subseteq \{1,2, \dots, d\}$ is considered for splitting. Because of this, there is reduced correlation between trees and it helps capture different structural relationships in the data. Each decision node chooses the feature and threshold that offers the greatest reduction in impurity. One of the most used impurity measure, Gini impurity is defined for a dataset D as:

$$Gini(D) = 1 - \sum_{k=1}^{C} p_k^2 \tag{6}$$

Where p_k is the proportion of samples belonging to class k, and $C$is the total number of classes. If a split on feature $f$divides the data into subsets D_{left} and D_{right}, the impurity reduction is

$$\Delta Gini(f) = Gini(D) - \left\{\frac{|\ D_{left}|}{|D|}Gini(D_{left}) + \frac{|\ D_{left}|}{|D|}Gini(D_{right})\right\} \tag{7}$$

The feature and threshold giving the maximum ΔGini are selected for that node [22]. Once all trees are trained, classification of a new sample x is performed by passing it through each tree. Each tree produces an output $T_b(x)$, and the forest prediction is obtained through majority voting:

$$\hat{y}(x) = mode\{T_1(x), T_2(x), \dots, T_3(x)\} \tag{8}$$

Where B is the total number of trees. The misclassifications of each tree are often uncorrelated because they are trained on distinct subsets of data and characteristics. The ensemble greatly lowers variance and enhances generalization when combined. The capacity of Random Forests to produce a quantitative assessment of feature relevance is a key feature. The impurity reduction ascribed to each feature across all trees is averaged to calculate this.

III.3 RF–KNN HYBRID LEARNING APPROACHES

Random Forest and KNN possess complementary characteristics. While Random Forest is a powerful ensemble learner capable of handling nonlinear relationships and providing feature relevance information, KNN is a simple yet effective local classifier that operates based on distance relationships. Several hybrid strategies have been proposed in literature to utilize the strengths of both models. The most commonly adopted approaches are summarized below.

III.3.1 Voting Based RF–KNN Ensemble

In voting based ensembles, RF and KNN act as independent base classifiers. For a given test sample x, each classifier produces a class prediction, and the final output is determined based on majority voting [23]. If $\hat{y}_{\text{RF}}(x)$ and $\hat{y}_{\text{KNN}}(x)$ represent the predictions from RF and KNN respectively, the ensemble decision is given by

$$\hat{y}(x) = mode\ \{\ \hat{y}_{\text{RF}}(x), \hat{y}_{\text{KNN}}(x)\} \tag{9}$$

In soft voting, the class probabilities are averaged:

$$P(\ c \mid x\ ) = \frac{1}{2}\left(P_{RF}(\ c \mid x\ ) + P_{RF}(\ c \mid x\ )\right) \tag{10}$$

Although this approach often improves robustness, it does not exploit the internal knowledge of either model beyond their final predictions.

III.3.2 Stacking Based RF–KNN Ensemble

In stacking, RF and KNN serve as base learners, while a separate meta classifier is trained to combine their outputs [24]. Let the predictions of RF and KNN for a sample x be concatenated as:

$$z(x) = [\hat{y}_{\text{RF}}(x), \hat{y}_{\text{KNN}}(x)] \tag{11}$$

A meta classifier $f_m(\cdot)$is then trained such that:

$$\hat{y}(x) = f_m(x) \tag{12}$$

Stacking enables the learning of nonlinear relationships between base classifier outputs, but it introduces additional model complexity and it requires careful cross validation in order to avoid overfitting.

III.3.3 Feature Level Fusion of RF and KNN

In feature level fusion, transformed feature representations obtained from RF are concatenated with the original input features and given to a KNN classifier or another learning model [23]. If x is the original feature vector and $h_{\text{RF}}(x)$ represents the RF induced feature transformation, the fused feature vector is:

$$x_{fusion} = [x, h_{\text{RF}}(x)\ ] \tag{13}$$

KNN is then applied in this expanded feature space. This approach increases representational power but also increases dimensionality and computational cost. While the above strategies combine RF and KNN at the decision level or feature level, they do not explicitly utilize the feature importance knowledge generated by RF to guide the distance computation in KNN, which forms the central idea of the proposed model.

III.3.4 Proposed RF Guided Feature Weighted KNN Model

The proposed model integrates Random Forest and KNN in a more informed manner by embedding feature importance knowledge from RF directly into the KNN distance metric. Unlike conventional fusion strategies, here RF does not merely act as an independent classifier but provides guidance to KNN in the form of feature relevance. In standard KNN, all features contribute equally to the distance computation. However, in nuclear reactor systems, different parameters contribute unequally to various transient events. Noisy or weakly relevant features can distort distance measurements and lead to incorrect neighbor selection. Therefore, introducing feature weight awareness into KNN is essential to improve classification reliability. Random Forest offers a natural framework for estimating feature relevance based on impurity reduction [25]. In this work, this information is exploited to construct a feature weighted distance metric for KNN.

III.3.4.1 Feature Importance Extraction using Random Forest

A Random Forest consisting of Bdecision trees is trained on D. For each feature f_j, the importance is computed by aggregating the Gini impurity reduction over all the trees:

$$I(f_j) = \frac{1}{B} \sum_{b=1}^{B} \sum_{n \in T_{b(f_j)}} \Delta Gini_{b,n} \tag{14}$$

where $\Delta\mathrm{Gini}_{b,n}$ denotes the impurity reduction achieved by node n of tree T_b when splitting on feature f_j. The resulting importance vector is:

$$\boldsymbol{I} = [I(f_1), I(f_2), \dots, I(f_d)] \tag{15}$$

This vector is normalized such that:

$$w_j = \frac{I(f_j)}{\sum_{k=1}^{d} I(f_k)} \tag{16}$$

The normalized vector

$$W = [w_1, w_2, \dots, w_d] \tag{17}$$

Represents the relative importance of each feature.

III.3.4.2 Feature Weighted Distance Metric for KNN

The conventional Euclidean distance between two samples x and $x_i$is defined in equation 3. In the proposed model, this is modified into a feature weighted Euclidean distance:

$$d_w(x, x_i) = \sqrt{\sum_{j=1}^{d} w_j (x_j - x_{i,j})^2} \tag{18}$$

Where w_j represents the importance weight of the j-th feature obtained from Random Forest. This formulation ensures that features with higher diagnostic relevance have a stronger influence on the neighbor selection process, while less informative features are automatically suppressed.

III.3.4.3 RF Guided KNN Classification Procedure

For a test sample x:

1. Compute weighted distances $d_w(x, x_i)$ to all training samples.
2. Identify the Knearest neighbors based on these distances.
3. Assign x to the class receiving the majority vote among its neighbors:

$$\hat{y}(x) = mode\,\{y_{(1)}, y_{(2)}, \dots, y_{(k)}\} \tag{19}$$

The operational sequence of the model is given in figure 1. This structure highlights how Random Forest acts as a knowledge provider, guiding the neighbor selection process of KNN.

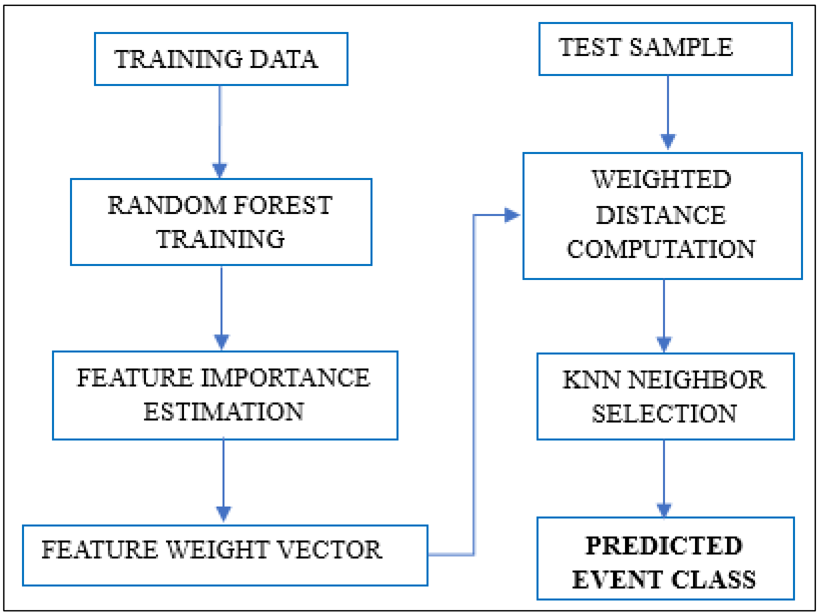

Figure 1: Operational sequence of the proposed RF guided feature weighted KNN model.
Source: Authors, (2026).

III.3.5 Data Preprocessing and Modeling Strategy

III.3.5.1 Data Preprocessing and preparation

All input features are normalized using z-score normalization to eliminate scale dominance:

$$x' = \frac{x-\mu}{\sigma} \tag{20}$$

where μ and σ denote the mean and standard deviation of a feature respectively.
The dataset is divided into three mutually exclusive subsets:

- **Training set** – used to train the Random Forest and KNN
- **Validation set** – used for hyperparameter optimization
- **Test set** – used for final performance evaluation

This separation ensures unbiased performance estimation.

III.3.5.2 Modeling and Hyperparameter Configuration

The performance of KNN is highly sensitive to the choice of K and the distance metric. To obtain the optimal configuration, a grid search combined with cross validation is employed on the training validation set.

The following hyperparameter ranges are explored:

- $K \in \{1,2,\ldots,30\}$
- Distance metric ∈{Euclidean, Manhattan}

For each parameter combination, cross validation accuracy is computed, and the configuration that yields the maximum average accuracy is selected. The optimal hyperparameters obtained for this study are: $K = 27, Distance\ Metric =$ Euclidean. The Random Forest model is trained to extract feature importance information required for the proposed RF guided weighted KNN framework. Optimal hyperparameters are obtained after performing grid search combined with cross validation. The following hyperparameter ranges are explored:

- No. of trees $N \in \{10,30,\ldots,200\}$

Other parameters of the model are chosen as

- Splitting criterion : Gini
- Maximum number of features at each split : $\sqrt{d}$, where d is the number of features.
- Maximum tree depth : None
- Minimum samples to split an internal node: 2
- Minimum samples in a leaf node: 1

The optimal number of trees obtained from hyperparameter tuning is N =70. After completing the independent modeling stages of KNN and Random Forest, both models are integrated by incorporating the feature importance weights in to the Euclidean distance calculation for the KNN classifier. Since the proposed RF guided weighting modifies only the distance metric and not the neighborhood structure, the optimal value K = 27, obtained from the standard KNN optimization, was retained for the weighted KNN model. After hyperparameter selection, the final RF guided weighted KNN model is trained using the combined training and validation data. The trained model is then evaluated on an independent hold out test dataset to assess its generalization capability.

IV. PERFORMANCE METRICS

The performance of the proposed RF guided feature weighted KNN classifier is evaluated using several standard classification metrics, namely accuracy, confusion matrix, precision, recall, and F1-score [26]. Since the considered problem involves multi class nuclear reactor event identification, it is essential to assess not only the overall classification accuracy but also the class wise prediction behavior. Each metric used in this study is briefly described below. The confusion matrix is the fundamental tool used to visualize the classification performance. For a multi class problem with Cclasses, the confusion matrix is a $C \times C$matrix, where the element M_{ij}denotes the number of samples belonging to class $i$that are predicted as class j. For a single class k, the following quantities are defined:

- **True Positives (TP)**: samples correctly classified as class k
- **False Positives (FP)**: samples incorrectly classified as class k
- **False Negatives (FN)**: samples of class k misclassified as other classes
- **True Negatives (TN)**: samples correctly classified as not belonging to class k

The confusion matrix provides detailed insight into which events are correctly detected and which are confused with one another. A typical confusion matrix for a 2 class problem is shown below:

$$\begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix} \tag{21}$$

Accuracy represents the proportion of total correctly classified samples to the total number of samples:

$$Accuracy = \sum_{i=1}^{C} M_{ii} \tag{22}$$

where:

- M_{ii} represents the diagonal elements of the confusion matrix (correct classifications),
- C is the number of classes,
- N is the total number of samples.

Although accuracy offers a broad indicator of performance, it does not show how effectively individual classes are recognized, particularly in situations when class distributions are unbalanced. For safety critical applications like nuclear reactor event identification, precision alone is therefore insufficient. Precision quantifies the percentage of samples that are correctly predicted to belong to a particular class. For class k:

$$Precision_k = \frac{TP_k}{TP_k + FP_k} \tag{23}$$

Precision is particularly important when false alarms are costly. In reactor monitoring, high precision ensures that when a specific transient is predicted, the likelihood of it being a false detection is low. Recall (also known as sensitivity or true positive rate) measures how many of the actual samples of a class are correctly identified:

$$Recall_k = \frac{TP_k}{TP_k + FN_k} \tag{24}$$

In safety related systems, recall is crucial since it shows how well the model can identify real anomalous events. A low recall suggests that the system might miss some crucial transients, which could result in dangerous situations. The F1-score is a fair performance indicator for every class and is the harmonic mean of precision and recall. F1 score is defined as:

$$F1_k = 2 * \frac{Precision_k * Recall_k}{Precision_k + Recall_k} \tag{25}$$

V. RESULTS AND DISCUSSIONS

This section presents the experimental results obtained for the standard KNN classifier, Random Forest classifier, various RF–KNN hybrid ensemble methods, and the proposed RF guided feature weighted KNN model. Initially, the individual base models were trained and optimized independently. Subsequently, different ensemble strategies, namely voting ensemble, stacking ensemble, and feature level fusion, were implemented and evaluated.

Finally, the proposed RF guided feature weighted KNN model was trained using the optimized hyperparameters and tested on a hold out dataset. The results are analyzed using accuracy trends, feature importance visualization, comparative performance plots, and the normalized confusion matrix. Figure 1 illustrates the variation of KNN classification accuracy as a function of the number of neighbors K. It can be observed that the accuracy increases rapidly for small values of K, indicating improved generalization as noise sensitivity reduces. Beyond this region, the performance gradually stabilizes as K increases. The accuracy reaches a peak and remains nearly constant in the range $K = 25$ to $K = 30$. indicating that the classifier has achieved a good balance between bias and variance. Based on this trend, $K = 27$ is selected as the optimal value for the KNN classifier and is used for all the hybrid models. Figure 2 shows the classification accuracy of the Random Forest classifier for different tree counts. As the number of trees increases, the performance first improves because to better ensemble averaging and reduced variation. However, adding more trees doesn't really improve the situation until the accuracy hits saturation. The plot clearly shows that the best accuracy is obtained with 70 trees, and the accuracy stabilizes after roughly 100 trees. Beyond this, raising the ensemble size primarily leads to increased computing costs and overfitting. This shows that a moderate number of trees is sufficient for precise feature extraction and classification in the given situation.

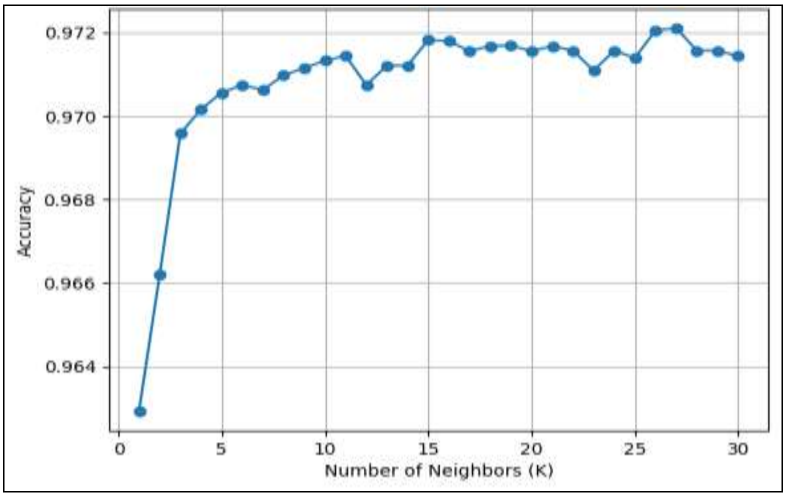

Figure 1: Accuracy Vs K (No. of Neighbors) plot of KNN classifier model.
Source: Authors, (2026).

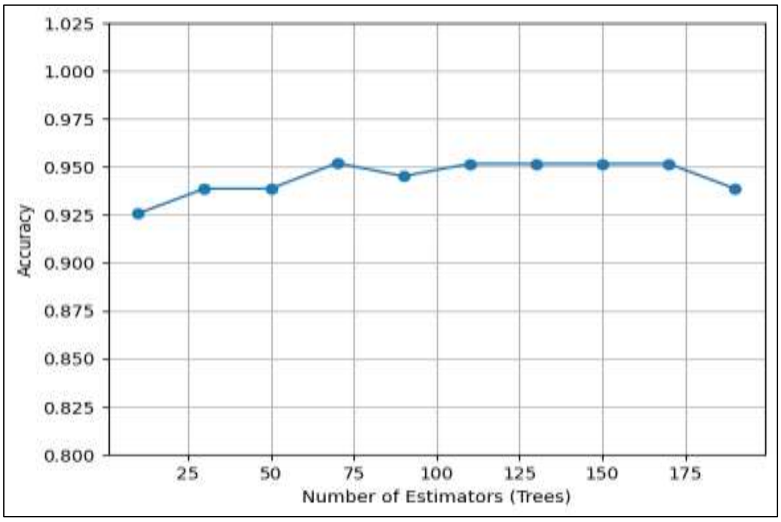

Figure 2: Accuracy Vs No. of estimators (trees) plot of Random Forest classifier model.
Source: Authors, (2026).

The trained Random Forest model's feature importance ratings are shown in Figure 3. It is evident that just a small number of features have a significant influence on the classification choice, whereas a number of features have comparatively little significance. The basic rationale for the suggested method—that considering all features equally in distance computation is not optimal for nuclear reactor event classification—is clearly supported by this observation. In the next KNN classification stage, the availability of such important information offers a solid foundation for the introduction of feature weighted distance computation. Performance metrics for several models are shown in Table 2, including the individual class F1 score and the total classification score for 1) Standard KNN 2) Random Forest 3) stacking ensemble 4) Voting ensemble 5) Feature level fusion 6) Proposed RF guided feature weighted KNN model. An overview of the models' comparative categorization accuracy is given in Figure 4. Strong baseline performance is demonstrated by the standard KNN classifier's high accuracy of 97.07%. At about 95.16%, the Random Forest classifier's accuracy is marginally lower. Although they offer a slight improvement over Random Forest, the stacking and voting ensembles are still not as good as the conventional KNN. Compared to the classical ensembles, the feature level fusion model produces a slight improvement. Nevertheless, the suggested RF guided feature weighted KNN model outperforms all individual models and hybrid baseline models with the maximum accuracy of 98.57%.

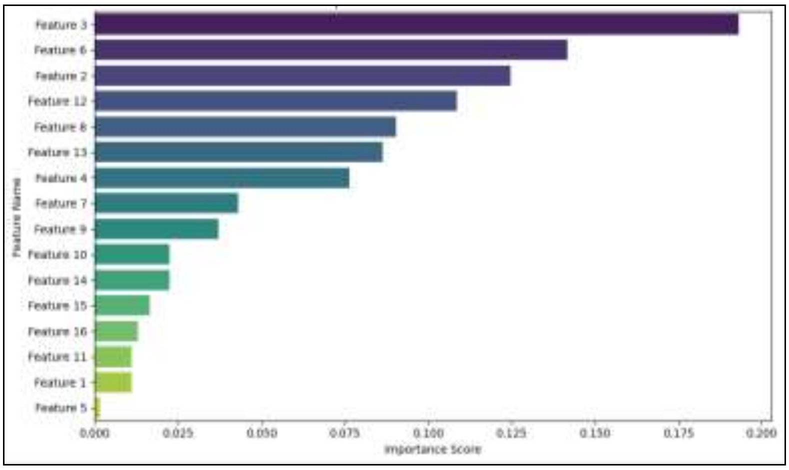

Figure 3: Feature importance plot of Random Forest Classifier model.
Source: Authors, (2026).

Table 2: Comparison of performance metrics of different models.

Model	**Categorical accuracy**	**F1-Score**				
		Class 1	**Class 2**	**Class 3**	**Class 4**	**Class 5**
KNN	97.07%	0.9	1	1	1	0.92
Random Forest	95.16%	0.88	1	1	1	0.91
KNN-RF Stack ensemble	95.86%	0.9	1	1	1	0.92
KNN-RF Voting ensemble	95.19%	0.88	1	1	1	0.91
KNN-RF feature level fusion	95.90%	0.89	1	1	1	0.92
Proposed RF guided feature weighted KNN model	98.57 %	0.96	1	1	1	0.97

Source: Authors, (2026).

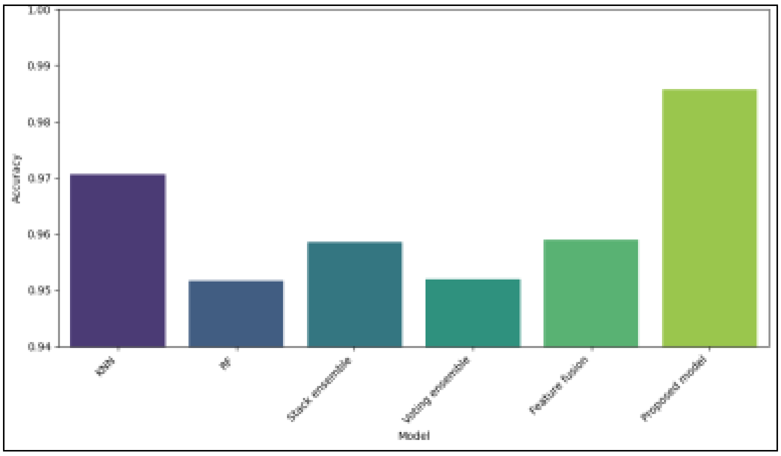

Figure 4: Classification accuracy of different models.
Source: Authors, (2026).

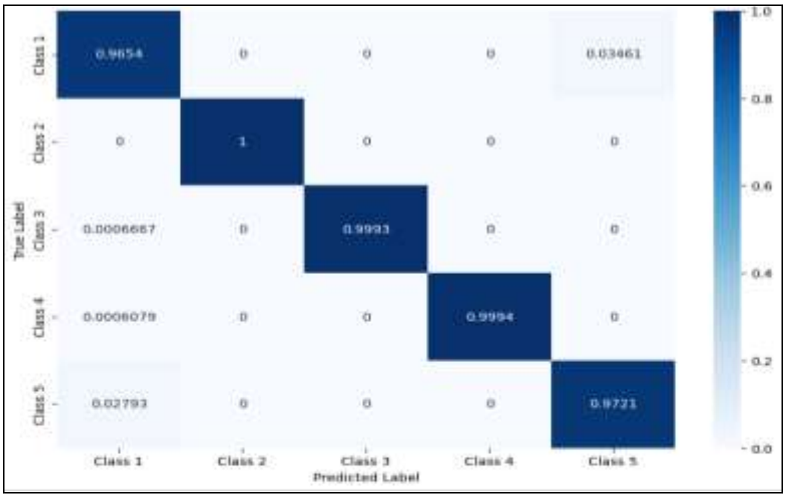

Figure 5: Normalized confusion matrix of the proposed RF guided feature weighted KNN model.
Source: Authors, (2026).

The proposed approach yields substantial improvements in the F1-scores of the relatively challenging event categories, namely Class 1 (No event) and Class 5 (Failure of heat sink event), where the F1-score improves from approximately 0.88–0.92 to 0.96 and 0.97, respectively. The remaining classes exhibit near perfect F1-scores across all models, indicating strong class separability. These improvements are consistent with the normalized confusion matrix in figure 5, which shows pronounced diagonal dominance and significantly reduced misclassification for the proposed model. This confirms that directly embedding Random Forest derived feature relevance into the KNN distance metric is more effective than combining classifiers only at the decision or feature fusion level.

V. CONCLUSIONS

In this work, a Random Forest guided feature weighted KNN classifier was developed to improve the performance of multi class reactor event classification. The key idea behind the proposed model is the integration of feature importance knowledge derived from Random Forest into the distance computation of KNN, thereby enabling the classifier to emphasize diagnostically significant reactor parameters while suppressing the influence of less informative features. The individual performances of KNN and Random Forest were first evaluated, followed by the implementation of several hybrid RF–KNN ensemble strategies, including stacking, voting, and feature level fusion. While these conventional ensembles offered marginal improvements, none of them surpassed the baseline KNN performance significantly. The proposed model, however, demonstrated a clear performance advantage. With the optimized hyperparameters and feature weighted distance formulation, the proposed RF guided KNN model achieved an overall classification accuracy of 98.57%, outperforming all baseline and ensemble models.

Class wise F1-scores further confirmed the robustness of the proposed method, especially for the more challenging event categories, where the F1-scores improved substantially compared to other approaches. The normalized confusion matrix illustrated strong diagonal dominance and minimal misclassification, validating the improved detection capability of the proposed model. These results collectively demonstrate that directly embedding Random Forest feature relevance into the KNN decision framework is a more effective strategy than combining the two models only at the decision or fusion stage. Overall, the proposed RF guided feature weighted KNN approach offers a simple and highly accurate method for nuclear reactor event identification. With improved performance, the work contributes to the growing field of safety in reactor operation and operator support, particularly where dependable event detection is critical.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Arunprasath V, T V Santhosh, Gopika Vinod.
Methodology: Arunprasath V, T V Santhosh, Gopika Vinod.
Investigation: Arunprasath V, T V Santhosh, Gopika Vinod.
Discussion of results: Arunprasath V, T V Santhosh, Gopika Vinod.
Writing – Original Draft: Arunprasath V, T V Santhosh, Gopika Vinod.
Writing – Review and Editing: Arunprasath V, T V Santhosh, Gopika Vinod.
Supervision: Arunprasath V, T V Santhosh, Gopika Vinod.
Approval of the final text: Arunprasath V, T V Santhosh, Gopika Vinod.

VII. ACKNOWLEDGMENTS

The authors greatly thank Mr.P.K.Guchhait, (Bhabha Atomic Research Centre, India) for his support in simulation and data generation.

VIII. REFERENCES

[1] Uhrig, R. E., Tsoukalas, L. H. "Soft computing technologies in nuclear engineering applications", Progress in Nuclear Energy, 34 (1) pp. 13–75, 1999. DOI:10.1016/S0149-1970(97)00109-1.

[2] Hu, G., Zhou, T., Liu, Q. " Data-Driven Machine Learning for Fault Detection and Diagnosis in Nuclear Power Plants: A Review", Frontiers in Energy Research, 9 (663296),2021. DOI: 10.3389/fenrg.2021.663296

[3] Yong X., Yunze C., Lin S., "Condition Assessment of Nuclear Power Plant Equipment Based on Machine Learning Methods: A Review", Nuclear Technology, 209(7), 929-962,2023. DOI:10.1080/00295450.2023.2169042

[4] Wang, F., Liu, A., Qu, C., Xiong, R., Chen, L., "A Deep-Learning Method for Remaining Useful Life Prediction of Power Machinery via Dual-Attention Mechanism", Sensors, 25(2), 497, 2025. DOI:10.3390/s25020497

[5] Gottlieb, C., Arzhanov, V., Gudowski W., et al. "Feasibility Study on Transient Identification in Nuclear Power Plants Using Support Vector Machines", Nuclear Technology, 155(1), pp. 67-77, 2006. DOI:10.13182/NT06-A3746

[6] Author,2024.

[7] Saeed, H.A., Peng, M.J.,Wang, H., et al. "Novel fault diagnosis scheme utilizing deep learning networks", Progress in Nuclear Energy, 103066,2020. DOI: 10.1016/j.pnucene.2019.103066

[8] Tian, D., Deng, J., Colantuono, G., Vinod, G., Santosh, T.V., & Gorse, C., "Identification of Loss of Coolant Accidents of Nuclear Power Plants Using Artificial Neural Networks", Nuclear Power Plant Life Management, Vienna, International Atomic Energy Agency (IAEA), 2021. DOI: 10.1016/j.ress.2008.08.005

[9] Gomez-Fernandez, M., Higley, K., Tokuhiro, A., Welter, K., Wong, W., & Yang, H., "Status of research and development of learning-based approaches in nuclear science and engineering: A review", Nuclear Engineering and Design, 359, 110479, 2020. https://doi.org/10.1016/j.nucengdes.2019.110479

[10] Ben Q., Jingang L., Jiejuan T., “Fault Diagnosis Techniques for Nuclear Power Plants: A Review from the Artificial Intelligence Perspective”, Energies, 16,1850,2023. https://doi.org/10.3390/en16041850

[11] Naimi, A., Deng, J., Doney, P., Sheikh-Akbari, A., Shimjith, S.R. and Arul, A.J., “Machine learning-based fault diagnosis for a PWR nuclear power plant”, IEEE Access, 10, pp.126001-126010, 2022.
https://doi.org/10.1109/ACCESS.2022.3225966

[12] Author, 2023.

[13] Mauro Cappelli, Instrumentation and Control Systems for Nuclear Power Plants, ed. 1, Woodhead Publishing Series in Energy, ed. 1, Mauro Cappelli, Woodhead Publishing, Cambridge,2023.

[14] IAEA, Modern Instrumentation and Control for Nuclear Power Plants: A Guidebook, Technical Reports Series No. 387, ed 1, IAEA, International Atomic Energy Agency, Vienna,1999.

[15] INL, RELAP5/Mod3.2 code manual, Vol. I-VII, Report NUREG/CR-5535, U.S. NRC, Washington D.C,1995.

[16] Guo, Y., Wang, G., et al., Thermal hydraulic analysis of loss of flow accident in the JRR-3M research reactor under the flow blockage transient.,Annals of Nuclear Energy, 118,147-153,2018. https://ui.adsabs.harvard.edu/link_gateway/2018AnNuE.118..147G/doi:10.1016/j.anucene.2018.04.014

[17] IAEA, Safety of research reactors, IAEA Specific Safety Requirements SSR-3, ed 1, International Atomic Energy Agency, Vienna,2016.

[18] Yang, L., Jin, R.: ‘Distance metric learning: A comprehensive survey’, Technical report, Department of Computer Science and Engineering, Michigan State University, 2007.

[19] I. A. Abu Amra and A. Y. A. Maghari, "Students performance prediction using KNN and Naïve Bayesian," 2017 8th International Conference on Information Technology (ICIT), Amman, Jordan, 2017, pp. 909-913, doi: 10.1109/ICITECH.2017.8079967

[20] Parmar, A., Katariya, R., Patel, V. (2019). A Review on Random Forest: An Ensemble Classifier. In: Hemanth, J., Fernando, X., Lafata, P., Baig, Z. (eds) International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018. ICICI 2018. Lecture Notes on Data Engineering and Communications Technologies, vol 26. Springer, Cham.,2018. https://doi.org/10.1007/978-3-030-03146-6_86

[21] Daho, Mostafa El Habib, Chikh, Mohammed Amine, Journal of Medical Imaging and Health Informatics, Volume 5, Number 3, pp. 539-544(6), June 2015.

[22] Tangirala S. Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. International Journal of Advanced Computer Science and Applications,11(2):612-9,Feb 2020. 10.14569/ijacsa.2020.0110277.

[23] Rojarath, Artittayapron, Wararat Songpan, and Chakrit Pong-Inwong. "Improved ensemble learning for classification techniques based on majority voting." In 2016 7th IEEE international conference on software engineering and service science (ICSESS), pp. 107-110. IEEE, 2016. 10.1109/ICSESS.2016.7883026.

[24] Mienye ID, Sun Y. A survey of ensemble learning: Concepts, algorithms, applications, and prospects. Ieee Access, Sep 16;10:99129-49,2022. 10.1109/ACCESS.2022.3207287.

[25] Rogers, Jeremy, and Steve Gunn. "Identifying feature relevance using a random forest." In International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection", pp. 173-184. Berlin,
Heidelberg: Springer Berlin Heidelberg, 2005. https://doi.org/10.1007/11752790_12

[26] Vujović, Željko. "Classification model evaluation metrics." International Journal of Advanced Computer Science and Applications 12, no. 6: 599-606,2021. 10.14569/IJACSA.2021.0120670.