Optimal Feature Subset Selection and Dimensionality Reduction for Pulse Waveform Discrimination

Authors: Ding Tingmeng, Jiang Xiaofei, Jiang Yuhang, Yang Luying, Jiang Xiaofei

Date: 2024-05-31T00:00:00+00:00

Abstract

With the widespread application of machine learning in neutron-gamma (n-γ) discrimination, feature subset selection in pulse shape discrimination has become a problem worthy of attention. Empirical methods, Random Forest classification, and Logistic regression feature selection algorithms have comprehensively refined the feature subset selection methodology, while Kernel Principal Component Analysis (KPCA) further reduces the dimensionality of the feature subset. Experimental results demonstrate that feature selection algorithms exhibit suboptimal performance on weak nuclear signals, with error rates reaching over 30%. The selection range of feature subsets in empirical methods is crucial; the error rate for feature subset "1-62" reaches 49.096%, significantly higher than the approximately 1% error rate for feature subsets derived from the pulse tail. The optimal feature subset does not completely coincide with the sampling points corresponding to the tail integral, but the difference is minor; the sampling points corresponding to the tail integral can be approximated as the optimal feature subset. Through investigating currently representative feature selection algorithms such as Random Forest classification and Logistic regression, along with meticulous empirical methods, the results of this paper possess universality and provide further theoretical support for feature subset selection.

Full Text

Study on Optimal Feature Subset Selection and Dimensionality Reduction in Pulse Shape Discrimination

DING Tingmeng[1], JIANG Yuhang[1], YANG Luying[1], JIANG Xiaofei[1]

[1](College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China)

Abstract

[**Background**] With the widespread application of machine learning in neutron-gamma (n-γ) discrimination, feature subset selection in pulse waveform discrimination has become a notable issue. [**Purpose**] By investigating representative feature selection algorithms such as Random Forest classification and Logistic regression, as well as detailed empirical methods, the results of this paper are universally applicable, providing further theoretical support for the selection of feature subsets. [**Methods**] Empirical methods, Random Forest classification, and Logistic regression feature selection algorithms have comprehensively improved the methods of feature subset selection, while Kernel Principal Component Analysis (KPCA) further reduces the dimensionality of feature subsets. [**Results**] Experimental results indicate that feature selection algorithms perform poorly in weak nuclear signals, with error rates exceeding 30%. The selection range of feature subsets in empirical methods is crucial, with error rates reaching 49.096% for feature subset "1-62", significantly higher than the approximately 1% error rate from features originating from the pulse tail. [**Conclusions**] The optimal feature subset does not entirely overlap with the sampling points corresponding to the tail integral, but the difference is minor, suggesting that the sampling points corresponding to the tail integral can be approximated as the optimal feature subset.

Keywords: Pulse shape discrimination, Feature subset, Feature selection, Dimensionality reduction

Introduction

In most neutron fields, γ-rays are always produced alongside neutrons. Currently, commonly used scintillation detectors are sensitive to both neutrons and γ-rays, making neutron-gamma (n-γ) discrimination one of the key challenges in neutron detection.

The Charge Comparison Method (CCM) is a classical n-γ discrimination method that is computationally simple and exhibits excellent discrimination performance in high-energy domains, leading to its widespread application in pulse discrimination firmware. However, CCM cannot discriminate low-energy pulses. With technological advancements, machine learning methods, leveraging their outstanding performance in classification and regression problems, have been applied to the n-γ discrimination field by many researchers [1][2][3].

Unsupervised learning algorithms in machine learning perform clustering based on data distribution in feature space without relying on pre-labeled samples and possess certain capabilities for identifying anomalous pulse events, making them suitable for mixed neutron-gamma fields. The Gaussian Mixture Model (GMM), commonly used in unsupervised learning, demonstrates good performance in n-γ

discrimination [4][5]. However, direct application of GMM clustering to high-dimensional data suffers from the "curse of dimensionality." A nuclear pulse signal contains hundreds of sampling points, and using such high-dimensional data directly for GMM clustering results in not only large computational overhead but also poor clustering performance. To reduce data dimensionality, the current common approach is to empirically select sampling points with large pulse differences as a feature subset, then apply dimensionality reduction algorithms to obtain new features. Liu et al. (2023) used 14 sampling points from the pulse tail as a feature subset and then applied Kernel Principal Component Analysis (KPCA) to extract 4 new features [6][7]. Hu Wanping (2024) used a KPCA-MPA-ELM model to improve n-γ discrimination accuracy [8], where the feature subset consisted of 80 sampling points from the pulse tail and KPCA was used for feature dimensionality reduction.

Empirically selected feature subsets often exhibit large fluctuations, and the specific selection criteria for sampling points within feature subsets require further investigation. Additionally, feature selection algorithms represent an important method for obtaining feature subsets [9][10], yet research on feature selection algorithms in n-γ discrimination remains scarce. If feature selection algorithms can be applied to pulse features, the methods for obtaining feature subsets in n-γ discrimination can be further improved. Feature subsets generally contain more than ten features, and KPCA can reduce the dimensionality of feature subsets to further extract features.

Data and features determine the upper limit of machine learning, while models and algorithms merely approximate this limit. Selecting feature subsets and reducing dimensionality help obtain more information from raw data, thereby yielding more accurate results. This paper first employs three methods to obtain feature subsets: empirical methods that select sampling points with large pulse differences, Random Forest classification [11] feature selection algorithm, and Logistic regression [12] feature selection algorithm. KPCA is then used to further reduce the dimensionality of feature subsets, and finally, more refined empirical feature subsets are used to determine the optimal feature subset selection range.

1 Methods and Principles

In machine learning, high-dimensional data is not suitable for direct input. To minimize feature dimensionality as much as possible, this paper obtains feature subsets through empirical selection of sampling points with large pulse differences and feature selection algorithms such as Random Forest classification and Logistic regression. The feature subsets are then further reduced to obtain several optimal low-dimensional features (Figure 1 [Figure 1: see original paper]).

Figure 1 [Figure 1: see original paper] Flowchart of impulse signal feature selection and dimensionality reduction

1.1 Construction of Feature Subsets When dealing with high-dimensional data, it is necessary to eliminate redundant items as much as possible and select a portion of features from the original pulse sampling points to form a feature subset. Feature subsets can be obtained through empirical selection of sampling points with large pulse differences or through feature selection algorithms such as Random Forest classification and Logistic regression.

The pulse tail exhibits the largest difference between neutron and gamma-ray pulses, and features within the feature subset should originate from sampling points in the pulse tail, though the number of sampling points is not uniform. The conventional approach for constructing feature subsets is to empirically select sampling points with large pulse differences—typically dozens of points from the pulse tail—though this method relies on "experience" and estimation.

Feature selection algorithms reduce dimensionality by removing irrelevant, redundant, or noisy features and selecting a small subset from the original features. Both Random Forest classification and Logistic regression belong to feature selection methods. Random Forest is a classifier algorithm model containing multiple decision trees, with each tree composed of root nodes, internal nodes, and leaf nodes. Leaf nodes represent classification results, while root and internal nodes represent decision criteria. Logistic regression is a binary classification algorithm that predicts the probability of classification variables through linear combinations of multiple independent variables.

To enhance the reliability of Random Forest feature selection algorithm results, we gradually increase the size of original features, taking 34 sampling points from the pulse tail, 62 sampling points from the non-baseline portion of the pulse, and 120 sampling points including partial baseline as original features. By examining the differences between the features included in the feature subset and the sampling points from the pulse tail, we can evaluate the ability of the Random Forest classification feature selection algorithm to eliminate redundant items. Random Forest classification and Logistic regression each select a feature subset from the original features. If the feature selection algorithms are reliable, the feature subsets obtained by Random Forest classification and Logistic regression should be consistent. If there are features with significantly higher importance in the nuclear pulse sampling points, the baseline should not affect the feature selection results, and including baseline sampling points in the original features can assess the stability of the feature selection algorithms.

1.2 Dimensionality Reduction Kernel Principal Component Analysis (KPCA) is a fundamental feature extraction method that maps high-dimensional data onto low-dimensional orthogonal features, which are called principal components. The number of features within a feature subset remains relatively high, and KPCA maps these features into new principal components to achieve dimensionality reduction. At this point, only a few principal components are needed to obtain a high cumulative variance contribution rate.

1.3 Qtail and Qtotal As shown in Figure 2 [Figure 2: see original paper], Qtail and Qtotal represent the total charge integral and tail integral of the pulse, respectively. CCM uses the ratio of these two values as the discrimination factor. Qtail and Qtotal can be regarded as two independent features extracted from non-baseline sampling points of the pulse and can serve as features for GMM clustering to achieve better classification results than CCM.

Figure 2 [Figure 2: see original paper] Diagram of the tail integral Qtail and the total integral Qtotal

1.4 GMM Clustering The Gaussian Mixture Model (GMM) is a probabilistic model used to describe datasets composed of multiple Gaussian distributions. For each Gaussian component, its probability density function is:

(1.1)

where n is the data dimension, is the n-dimensional mean vector, and Σ is the n×n covariance matrix. Clearly, the Gaussian distribution is determined by and Σ. The initial pulse has 248 features, and such high dimensionality leads to the "curse of dimensionality." To reduce the number of pulse features, feature extraction or selection must first be performed.

Ignoring pulse pileup, in n-γ discrimination, the model contains only two components: neutrons and gamma rays. For a Gaussian mixture distribution with two components, its probability density is:

(1.2)

where αi is called the "mixing coefficient," representing the probability of selecting the i-th Gaussian mixture component, with $\alpha i > 0$ and $\Sigma \alpha i = 1$. The model parameters αi, i, and Σi need to be solved iteratively through the EM algorithm. Each iteration of the EM algorithm includes two steps: the E-step, which estimates the expectation of hidden variables based on current parameters; and the M-step, which updates model parameters using maximum likelihood estimation based on the E-step results.

When different features are input into GMM clustering, the clustering results will vary accordingly. This paper aims to explore optimal methods for obtaining feature subsets, and the quality of different feature subsets can be evaluated by comparing the GMM clustering results obtained from them.

2 Results and Discussion

The experimental flowchart is shown in Figure 3 [Figure 3: see original paper]. The neutron source is a 241Am-Be source, the detector is an organic liquid scintillator detector EJ-301, and the digitizer is DT5730B. The detector acquires current pulses, which are digitized to obtain raw data. The raw data undergo preprocessing steps including smoothing filtering, normalization, and baseline restoration before being stored in a computer [13]. The 60,000 preprocessed

pulses are divided into two parts: 30,000 pulses are used for GMM clustering to obtain a reliable training set, and the other 30,000 pulses are used for testing to compare the performance differences of different algorithms.

Figure 3 [**Figure 3: see original paper**] Experimental flowchart

2.1 Feature Subset Construction The pulse tail integral corresponds to 34 sampling points, and the non-baseline portion of the pulse includes 62 sampling points. The portion with the largest difference between neutron and gamma-ray pulses is the 34 sampling points corresponding to the tail integral, and feature subsets can be obtained through empirical selection of sampling points with large pulse differences.

In addition to empirical feature subset selection, we also use feature selection algorithms to improve the methods for obtaining feature subsets in pulse shape discrimination. Random Forest classification is an important method for obtaining feature subsets in feature engineering, and we adopt 5-fold cross-validation with subset size ranging from 1 to 13.

Random Forest classification and Logistic regression rely on prior knowledge. We take pulses with probabilities greater than 99% from the GMM clustering results as the training set to find feature subsets. GMM clustering using Qtail and Qtotal as features yields classification results consistent with the classical CCM in the 100-2100 keV energy range, with a pulse classification accuracy 5.52% higher than CCM in the 0-100 keV range. The GMM clustering results are probability values. After excluding low-probability events (classification probability $< 99\%$), the remaining pulses form a training set of size 26,261.

When Random Forest classification selects feature subsets from the 34 sampling points of the pulse tail, the resulting feature subset contains only two features, but their importance scores are 0.091 and 0.087, respectively, indicating very low feature importance. The pulse tail exhibits the largest difference between neutron and gamma-ray pulses, and most features in the optimal feature subset should originate from sampling points in the pulse tail. To evaluate the reliability of feature selection methods, we expand the original features to 62 sampling points from the non-baseline portion of the pulse. By examining the differences between the features included in the optimal subset and the sampling points from the pulse tail, we can assess the method's ability to eliminate redundant items.

Figure 4 [Figure 4: see original paper] shows performance metrics for different feature subset sizes, including Root Mean Square Error (RMSE), R-squared coefficient, and Mean Absolute Error (MAE), which measure prediction accuracy and stability of models with different feature subset sizes.

Figures 4(a) and 4(b) are bar charts and line plots showing R-squared, RMSE, and MAE versus subset count. As the number of features in the subset increases from 1 to 13, R-squared shows an upward trend while RMSE and MAE show

downward trends. When the feature count exceeds 7, the rates of change for R-squared, RMSE, and MAE increase sharply. After the feature count exceeds 10, R-squared growth slows down while RMSE and MAE decline more slowly. The line plots clearly show that all three curves flatten after the feature count exceeds 10. Based on the combined results from bar charts and line plots, the optimal feature subset selected by Random Forest from 62 features contains 10 features, which are the 23rd, 22nd, 24th, 6th, 16th, 9th, 32nd, 21st, 23rd, 35th, 15th, and 48th sampling points. The pulse tail exhibits the largest difference between neutron and gamma-ray pulses, and most features in the optimal feature subset should originate from sampling points in the pulse tail. However, this subset has low overlap with the pulse falling edge sampling points, with only 3 sampling points from the pulse tail.

Since the baseline is not completely zero after baseline restoration and exhibits fluctuations, it is necessary to investigate the influence of pulse baseline on feature selection algorithms by including some baseline sampling points in the feature selection algorithm. When the original feature count is 120, the feature selection results are shown in Figure 5 [Figure 5: see original paper]. Figures 5(a) and 5(b) are bar charts and line plots showing R-squared, RMSE, and MAE versus subset count. After the feature count exceeds 6, R-squared growth slows down while RMSE and MAE decline more slowly. In the line plots of R-squared, RMSE, and MAE versus subset count, all three curves flatten after the feature count exceeds 6. Based on the combined results, the optimal feature subset selected by Random Forest from 120 features contains 6 features, which differs from the feature subset selected from 62 sampling points. Nuclear signals are extremely weak, and we cannot completely remove all noise, so the baseline cannot be exactly zero. When using the Random Forest classification algorithm for feature selection, the baseline significantly interferes with the results, indicating poor anti-interference capability of the feature selection algorithm.

Figure 5 [Figure 5: see original paper] Performance metrics versus feature subset size (120 sampling points as original features)

Random Forest classification and Logistic regression are both important feature selection algorithms. In addition to using the Random Forest classification model for feature selection, we can also use the Logistic regression model from the "leaps" package to select feature subsets from 62 tail sampling points. The Logistic regression model fitting effect is evaluated using four parameters: Residual Sum of Squares (RSS), Adjusted R-Squared (Adjusted R^2), Mallow's Cp (CP), and Bayesian Information Criterion (BIC).

Figure 6 [Figure 6: see original paper] shows performance metrics for different feature subset sizes in the Logistic regression model. When the feature subset contains 11 features, both CP and BIC reach their minimum values while Adjusted R^2 reaches its maximum, and RSS no longer shows a significant downward trend after the feature count exceeds 11. Based on this figure, the optimal feature subset selected by the Logistic regression model contains 11 features. The different results from Random Forest classification and Logistic regression

indicate that feature selection algorithms struggle to obtain reliable and stable feature subsets.

Figure 6 [Figure 6: see original paper] Performance metrics for different feature subset sizes in the regression model

2.2 Optimal Feature Subset Feature subsets obtained through empirical selection, Random Forest classification, and Logistic regression are not identical. The pulse tail integral portion includes 34 sampling points (features), the feature subset selected by Random Forest classification contains 10 features, and the feature subset selected by Logistic regression contains 11 features. Compared to the 248 sampling points of the pulse, these feature subset sizes are significantly reduced. We first analyze GMM clustering performance using these feature subsets directly as features. To compare discrimination effects of different methods, we perform GMM clustering with different features and name different methods as "feature count + GMM (n-features GMM)." The 10-features GMM, 11-features GMM, 34-features GMM, and 62-features GMM methods use features corresponding to the Random Forest classification selected subset, Logistic regression selected subset, 34 sampling points from the pulse tail, and 62 sampling points from the non-baseline portion, respectively.

For pulse data obtained from the EJ-301 detector, neutrons and gamma rays are completely mixed in the lower energy domain and cannot be discriminated. To compare clustering performance differences under different features, pulses with reliable labels are essential. CCM is a classical method widely used in n-γ discrimination, where neutrons and gamma rays can be completely separated in the higher energy domain (100-2100 keV). CCM discrimination results in the high-energy domain are reliable, and comparing classification results of different methods for 100-2100 keV pulses can evaluate n-γ discrimination effectiveness.

To quantitatively analyze differences between discrimination results of different methods, we pairwise compare results from different methods to obtain a discrimination difference heatmap, as shown in Figure 7 [Figure 7: see original paper]. It is evident that the difference between 34-features GMM and CCM is the smallest, with only 1.36% difference, while differences between 10-features GMM (Random Forest classification), 11-features GMM (Logistic regression), and 62-features GMM (62 sampling points from non-baseline portion) and CCM all exceed 30%. Moreover, 10-features GMM, 11-features GMM, and 62-features GMM not only have low discrimination accuracy but also show enormous differences among their results.

When selecting feature subsets empirically, the sampling point selection range is extremely important. The error rate of 62-features GMM reaches 30.06%, far higher than that of 34-features GMM. On one hand, the features used in 62-features GMM are not from the portion with the largest pulse differences; on the other hand, the feature dimensionality of 62-features GMM remains relatively high, suffering from the "curse of dimensionality."

Another obvious fact is that feature selection algorithms perform extremely poorly in feature selection for pulse shape discrimination. The difference between Random Forest classification and Logistic regression, both belonging to feature selection algorithms, is 24.61%, with error rates exceeding 30%, indicating imprecise and unstable feature selection results. Nuclear pulses are extremely weak signals susceptible to noise, with large fluctuations in individual sampling points. Since there are no dominant sampling points, classification and regression-based feature selection methods yield different results with poor stability, even being affected by baseline sampling points.

Figure 7 [**Figure 7: see original paper**] Heatmap of classification result differences when performing GMM clustering with different feature subsets

Qtail and Qtotal reduce the impact of individual sampling point fluctuations on discrimination results through integration over sampling points. Principal component analysis computes new orthogonal features, whose explained variance is significantly higher than other features and plays a decisive role in discrimination results. To further explore optimal feature subset values, we reduce the dimensionality of empirically selected feature subsets for more detailed analysis.

2.3 KPCA Dimensionality Reduction The non-baseline portion of the pulse includes 62 sampling points, and the portion with the largest pulse difference—the pulse tail—has 34 sampling points. Both feature subsets are representative empirical selection methods.

Using KPCA to reduce dimensionality of the 62 non-baseline sampling points, we take the first three principal components as features for GMM clustering. It is difficult to accurately judge classification result accuracy in feature space. Figure 8 [Figure 8: see original paper] shows the distribution of clustering results in the Energy-PSD diagram, where squares represent neutrons and circles represent γ-rays, revealing numerous misclassifications.

Figure 8 [**Figure 8: see original paper**] GMM clustering result using the first three principal components after dimensionality reduction of non-baseline sampling points as raw features

Taking KPCA results from 14 randomly selected sampling points in the pulse tail as an example, the results are shown in Figure 9 [Figure 9: see original paper]. The explained variance of the first three principal components is 64.65%, 22.73%, and 3.35%, respectively, with the cumulative variance of the three exceeding 90% and the proportion of the first principal component being extremely high. Even with random sampling from the pulse tail, the first three principal components still have high explained variance.

Figure 9 [**Figure 9: see original paper**] Contribution rate and cumulative contribution rate of 14 principal component features from the pulse tail

To further improve empirical selection methods, we start from the peak position (the 28th sampling point) and take sampling points 28 to 38 as the first feature

subset, then incrementally add three sampling points up to the 62nd sampling point, resulting in 9 feature subsets ranging in size from 10 to 34. Additionally, we include sampling points before the 28th point to explore the impact of redundant items.

Table 1 shows dimensionality reduction results for partial feature subsets selected by empirical methods, where start-end sampling points indicate the selection range of the feature subset; cumulative variance of the first three principal components reflects KPCA results, with smaller feature subsets yielding higher cumulative variance; error rate is the misclassification rate after comparing GMM clustering results using three principal components with labels in the 100-2100 keV energy domain. Feature subsets "28-38," "28-50," and "28-62" all come from the 34 sampling points of the pulse tail, while subsets "25-62" and "1-62" include some non-baseline sampling points. It is evident that when feature subsets originate from pulse tail sampling points, error rates are around 1%. Feature subset "25-62" has the lowest error rate. The starting position of the pulse tail integral is determined based on optimal CCM discrimination results, which does not completely coincide with the optimal KPCA-GMM clustering result but shows minor differences. Feature subset "1-62" has an error rate as high as 49.096%, yielding poor results whether clustered directly or after KPCA. Combined with feature selection algorithm results, features in the subset must originate from sampling points with the largest pulse differences, which do not completely overlap with sampling points corresponding to the tail integral but differ only slightly, making tail integral sampling points a reasonable approximation of the optimal feature subset.

Table 1 Dimensionality reduction results of partial feature subsets selected by empirical selection methods

To obtain the optimal feature subset, this paper acquires feature subsets through empirical methods, Random Forest classification feature selection algorithm, and Logistic regression feature selection algorithm. Empirically selected feature subsets range from 10 to 62 sampling points, Random Forest classification yields a feature subset with 10 features, and Logistic regression yields a feature subset with 11 features. When selecting feature subsets empirically, the sampling point selection range is extremely important, with 62-features GMM having an error rate of 30.06%, far higher than 34-features GMM. Feature selection algorithms perform extremely poorly in pulse shape discrimination, with a 24.61% difference between Random Forest classification and Logistic regression and error rates exceeding 30%, indicating imprecise and unstable feature selection results.

Feature selection algorithms face three problems in feature selection: First, when original features are non-baseline sampling points (62 points), the selected feature subset has low overlap with tail sampling points, indicating low feature selection accuracy; second, when original features include some baseline sampling points (120 points), the feature subset differs from that selected from non-baseline sampling points, indicating poor method stability; finally, different results from Random Forest classification and Logistic regression also indicate

that feature selection algorithms struggle to obtain reliable and stable feature subsets. Nuclear signals are extremely weak, with individual sampling points having limited impact on pulse discrimination results and large numerical fluctuations. Without dominant sampling points, classification and regression-based feature selection methods yield different results with poor stability, even being affected by baseline sampling points. The optimal feature subset size obtained from 62 sampling points (10) differs from that obtained from 120 sampling points (6).

Principal component analysis computes new orthogonal features, with the first three principal components having significantly higher explained variance than other features. To obtain the optimal feature subset, we reduced the dimensionality of empirically selected feature subsets for more detailed analysis. When feature subsets originate from the pulse tail, error rates are around 1%. Feature subset "1-62" has an error rate of 49.096%, yielding poor results whether clustered directly or after KPCA. Feature subset "25-62" has the lowest error rate, indicating that the optimal feature subset does not completely coincide with sampling points corresponding to the tail integral but differs only slightly, making tail integral sampling points a reasonable approximation of the optimal feature subset.

References

[1] Durbin M., Wonders M.A., Flaska M., et al. K-Nearest Neighbors regression for the discrimination of gamma rays and neutrons in organic scintillators, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 2021, 987:164826. DOI: 10.1016/j.nima.2020.164826

[2] Arahmane H., Hamzaoui E.M., Maissa Y. B., et al. Neutron-gamma discrimination method based on blind source separation and machine learning. *Nuclear Science and Techniques*, 2021,32: 18. DOI: 10.1007/s41365-021-00850-w.

[3] HUANG Kunxiang, ZHANG Jiangmei, WANG Jiaqi, et al. Study on n/γ Discrimination Method Based on GAF-CNN[J]. *Atomic Energy Science and Technology*, 2024,58(02):461-470. DOI: 10.7538/yzk.2023.youxian.0398.

[4] Andrew G., Qi C., Kaplan A.D., et al. Pulse pileup rejection methods using a two-component Gaussian Mixture Model for fast neutron detection with pulse shape discriminating scintillator[J]. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*,2021, 988: 164905. DOI: 10.1016/j.nima.2020.164905.

[5] Wang F P., Yang M H., Wang J Y., et al. A comparison of small-batch clustering and charge-comparison methods for n/γ discrimination using a liquid scintillation detector. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 2022,1028: 166379. DOI:10.1016/j.nima.2022.166379.

[6] Liu L F, Shao H. Study on neutron-gamma discrimination method based on the KPCA-GMM-ANN[J]. *Radiation Physics and Chemistry*, 2023, 203: 110602. DOI: 10.1016/j.radphyschem.2022.110602.

[7] Liu L F, Shao H. Study on neutron-gamma discrimination method based on the KPCA-GMM[J]. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 2023, 1056: 168604. DOI: 10.1016/j.nima.2023.168604.

[8] HU Wanping, ZHANG Guiyu, ZHANG Yunlong, et al. Neutron/gamma (n/γ) discrimination method based on KPCA-MPA-ELM[J]. *NUCLEAR TECHNIQUES*, 2024,47(04):75-84. DOI: 10.11889/j.0253-3219.2024.hjs.47.040403

[9] ZHANG Jin-qu, LING Yu, DU Ping, et al. Ensemble Feature Selection Method for Single Pulse Classification[J]. *Acta Astronomica Sinica*, 2023,64(05):59-69. DOI: 10.15940/j.cnki.0001-5245.2023.05.006.

[10] Ishwaran H., Malley J. D. Synthetic learning machines[J]. *Biodata Mining*,2014,7. DOI: 10.1186/s13040-014-0028-y.

[11] Vladimir S, Andy L, Christopher T, et al. A Classification and Regression Tool for Compound Classification and QSAR Modeling[J]. *Journal of Chemical Information and Computer Sciences*, 2003, 43 (6): 1947-1958. DOI: 10.1021/ci034160g

[12] Nuttanan W, Kang Y Y, Zhang F Q. Random feature selection using random subspace logistic regression[J]. *Expert Systems with Applications*, 2023, 217(119535): 0957-4174.DOI: 10.1016/j.eswa.2023.119535.

[13] Wang X X. The n-γ pulse waveform discrimination of scintillator detectors[D]. GuiYang: GuiZhou University,2023. DOI: 10.27047/d.cnki.ggudu.2023.001514.

Author Contributions

DING Tingmeng: Data processing, investigation, visualization, experimental results analysis, and paper writing; JIANG Xiaofei: Project management, resource provision, guidance, and review; JIANG Yuhang: Investigation; YANG Luying: Data management.

Funding

This work was supported by the National Natural Science Foundation of China (No. 12205062) and the Guizhou Provincial Science and Technology Program Project (Qiankehe LH Zi [2017]7225).

Note: Figure translations are in progress. See original paper for figures.