# XGBoost-Based Single-Pulse Signal Recognition (Postprint)

**Authors:** Ling Yu, Zhang Jinqu, Li Xiangru, Li Hui

**Date:** 2021-06-18T09:43:08+00:00

## Abstract

Pulsar search represents a significant research direction in radio astronomy. With the continuous construction and development of large radio telescopes, the volume of collected data is experiencing exponential growth, posing a substantial challenge for the timely and accurate identification of pulsar signals from rapidly acquired massive datasets. This study utilizes observation data from the LOFAR Tie-Array Survey project as a case study, designs ten feature variables for single-pulse signal recognition, investigates the application of XGBoost combined with wrapper feature selection method in single-pulse signal identification, and comparatively analyzes the experimental performance of models including GBDT, AdaBoost, Random Forest, and BP neural network for single-pulse signal recognition. Experimental results demonstrate that XGBoost combined with wrapper feature selection method exhibits superior comprehensive performance in single-pulse signal recognition, achieving the lowest misclassification rate while simultaneously attaining the highest precision, recall, and F1-score values—approximately 1–2 percentage points higher than other models on average. Regarding feature selection, nine features are identified as optimal. The feature variables and recognition methodology designed in this research can provide methodological and technical references for pulsar search initiatives in China that primarily focus on signal detection from FAST.

## Full Text

## Preamble

**Research on Single-Pulse Signal Recognition Based on XGBoost**

Yu Ling, Jinqu Zhang*, Xiangru Li, Hui Li
School of Computer Science, South China Normal University, Guangzhou, Guangdong, 510631, China

## Abstract

Pulsar searching is a crucial research direction in radio astronomy. With the continuous construction and development of large-scale radio telescopes, the volume of collected data is growing exponentially, making it a tremendous challenge to accurately identify pulsar signals from rapidly acquired massive datasets in a timely manner. This paper takes observation data from the LOFAR Tied-Array All-Sky Survey (LOTAAS) as an example, designs ten feature variables for single-pulse signal recognition, and further investigates the application of XGBoost combined with wrapper-based feature selection in single-pulse signal identification. The experimental performance of XGBoost is compared with GBDT, AdaBoost, Random Forest, and BP neural network models for single-pulse signal recognition. Experimental results demonstrate that XGBoost combined with wrapper feature selection offers comprehensive advantages in single-pulse signal recognition, achieving the lowest misclassification rate while attaining the highest precision, recall, and F1-score values—approximately 1 to 2 percentage points higher than other models on average. In terms of feature selection, nine features were identified as optimal. The feature variables and recognition methods designed in this study can provide methodological and technical references for pulsar searching efforts in China, particularly those utilizing FAST-detected signals.

**Keywords:** Single-pulse; XGBoost; Feature selection; Wrapper method

## 1. Introduction

Pulsars are rapidly rotating neutron stars that emit electromagnetic pulse signals continuously, hence their name. The discovery of pulsars represents one of the most significant astronomical findings of the 1960s [**?**]. Research on pulsars provides crucial information for studying neutron stars with extreme physical properties and greatly advances developments in physics, astronomy, navigation, and time measurement [**?**][**?**]. Based on the periodic characteristics of pulsar signals, the primary method for detecting pulsar signals involves converting time-domain signals to frequency-domain signals using Fast Fourier Transform combined with fast folding algorithms [**?**]. As pulse signal mining has progressed, two types of astronomical pulse signals lacking periodic characteristics have been discovered in recent years: those from Rotating Radio Transients (RRATs) and Fast Radio Bursts (FRBs) [**?**][**?**][**?**]. RRAT signals are emitted very sporadically and intermittently in time, making them undetectable through traditional periodic searches. FRBs consist of extragalactic radio burst signals; although extremely rare periodic phenomena have been found, they remain primarily characterized by their lack of periodicity. Due to their fleeting nature, the pulse signals from these two astronomical phenomena are referred to as single-pulse celestial signals. Single-pulse signal searching is not only a valuable complement to periodic signal search methods but also the primary detection approach for RRATs and FRBs [**?**].

Currently, single-pulse signal recognition methods are mainly divided into heuristic threshold search algorithms and machine learning algorithms. Heuristic search utilizes problem-specific heuristic information to guide the search and discover targets, reducing problem complexity and improving computational efficiency by narrowing the search scope. These methods are primarily based on the single-pulse signal classification framework proposed by Cordes and McLaughlin [**?**]. This framework divides single-pulse signal extraction into four steps: dedispersion, matched filtering, thresholding, and judgment, to determine the existence of single-pulse signals in detected signals. For example, Deneva et al. used a clustering algorithm to filter suspected pulse events with signal-to-noise ratios above a certain threshold as single-pulse candidates [**?**]. Karako-Argaman et al. grouped pulse events based on dispersion measure (DM) and signal time, then determined whether pulse signals had peak occurrences based on the maximum signal-to-noise ratio in adjacent groups, creating diagnostic plots for manual inspection [**?**]. Ryan et al. further proposed a simple recursive peak identification algorithm that uses the slope of fitted lines for Dispersed Pulse Groups (DPGs) to identify large slope trends in DPGs and thereby judge single-pulse event candidates [**?**]. While these methods have some effectiveness in detecting pulse signals, they primarily rely on threshold segmentation to extract pulsar signals, with features derived from the strongest pulse signals in groups, resulting in limited accuracy and requiring substantial manual participation, making them unsuitable for large-scale, massive data processing.

In recent years, with the development of sensor technology and the advancement of large-scale radio surveys, machine learning has become an important approach for pulsar signal recognition [**?**]. Machine learning methods involve statistical analysis of known pulsar signal features to establish learning models, which are then used to judge unknown pulse signals. This approach typically requires four steps: (1) establishing a benchmark dataset; (2) feature extraction; (3) model training and evaluation; and (4) model application. McFadden et al., in summarizing the application of machine learning in pulse signal screening, noted that existing machine learning algorithms are primarily used for periodic pulse signal searching [**?**]. For instance, Artificial Neural Network (ANN) algorithms [**?**]-[**?**] and pattern recognition algorithms [**?**] have been applied in periodic pulse signal searches. Although machine learning has seen much exploration in periodic pulse signals, its application in single-pulse signal recognition has only just begun and is gradually gaining attention. In terms of machine learning applications for single-pulse signal recognition, Eatough et al., building upon heuristic threshold search algorithms, selected 12 features including signal-to-noise ratio and pulse width as inputs for a three-layer artificial neural network, pioneering the use of machine learning for single-pulse signal screening [**?**]. Ryan et al. utilized datasets observed by the Green Bank Telescope, extracted 16 features from pulse number-DM plots and signal-to-noise ratio-DM plots, and compared SVM, ANN, RULE, and decision tree methods, concluding that a random forest ensemble tree classifier provided the best overall performance in terms of recall and precision [**?**]. Michilli et al., using the LOTAAS dataset, selected five impor-

tant metrics for single-pulse signal classification based on information gain from each feature: peak detection window width, average pulse DM, pulse signal-to-noise ratio, excess kurtosis of window width distribution curve, and excess kurtosis of signal-to-noise ratio distribution curve (further explanations of these metrics are provided in Section 2.3 of this paper) [**?**]. This work, after comparing several different machine learning algorithms, concluded that the method based on Gaussian-Hellinger fast decision trees exhibited the best performance in single-pulse signal classification.

Based on previous research, decision tree-based methods are considered among the best-performing approaches. However, parameter estimation for decision tree models has mostly employed small-scale random sampling methods, which cannot guarantee the optimality of final classification results. In recent years, the machine learning field has improved and enhanced decision tree models, particularly gradient boosting-based GBDT and XGBoost algorithms, which have been widely applied in many domains [**?**]. Therefore, this paper aims to investigate the performance analysis of XGBoost combined with wrapper-based feature selection for single-pulse signal recognition. The second part of the paper introduces the dataset used for the research, the third part details the XGBoost algorithm principles, the fourth part presents experimental results and comparative analysis discussions, and the final section provides conclusions and future prospects.

## 2.1 Data Source

A high-quality benchmark dataset is fundamental for machine learning training applications and research. However, given the massive volume of pulse signals, annotating them is a task that cannot be completed in the short term. Therefore, in this paper, we directly use the annotated dataset from Michilli' s work [**?**] for model research. This dataset originates from the LOFAR Tied-Array All-Sky Survey (LOTAAS) project. The Low Frequency Array (LOFAR), led by the Netherlands Institute for Radio Astronomy, is a large radio telescope composed of thousands of antennas grouped and distributed across observation stations in the Netherlands and other European countries. It operates at the lowest frequency bands with high resolution and high sensitivity for extensive and in-depth pulsar research [**?**]. The LOTAAS project utilizes 12 sub-stations for observations, generating 222 simultaneous radio data streams for each sky pointing, with each observation lasting one hour, a time resolution of 0.492 milliseconds for recorded data, and the capability to receive 16.9 TB of raw data per hour [**?**]. The dataset used in this study' s experiments was extracted from various LOTAAS observations.

## 2.2 Data Preprocessing

Pulsar searching generally requires four stages: radio signal data collection, dedispersion processing, periodic pulse or single-pulse searching, and manual dis-

crimination [**?**], with dispersion effects being one of the key distinctions between astrophysical signals and Radio Frequency Interference (RFI) signals [**?**]. When astrophysical signals reach Earth, they are affected by free electrons of varying densities in space, causing different frequency signals to experience different delay effects. Dispersion Measure (DM) quantifies the total number of free electrons along the signal propagation direction. Since the DM corresponding to a celestial signal is unknown beforehand, dedispersion processing requires testing with different DM values. Consequently, for a single-pulse signal, although it essentially corresponds to a unique DM, the dedispersion process generates many candidate pulse signals based on different DM values, and these candidate pulse signals corresponding to different DMs may still be detected as peak signals. Thus, a theoretically single pulse signal might be detected as multiple peak signals with very close DM values. Therefore, cluster analysis can be performed on the detected series of peak signals according to their corresponding DM values, with clustered peak signals forming a Dispersed Pulse Group (DPG). Subplot 1 in [Figure 1: see original paper] shows the signal-to-noise ratio distribution of pulse signals obtained at different DM values within a DPG. The recognition of single-pulse signals primarily involves identifying whether a DPG originates from a pulsar or RFI; if identified as a pulsar signal, feature map information is further output for manual verification.

The data used in this paper were processed within a DM range of 0 to 550 pc $cm^{-3}$, with calculations performed at intervals of 0.01 to 0.1 pc $cm^{-3}$. For data processed by DM, peak detection was conducted using rectangular windows of different lengths, and signals with signal-to-noise ratios greater than 5 were saved to form a signal event table, storing information including window width, DM, and signal time. Based on the proximity of signal time and DM values for each record in the signal event table, signal events were clustered and grouped, with signal events within 30 milliseconds in time and 2 pc $cm^{-3}$ in DM difference classified into a single DPG. [Figure 1: see original paper] shows the distribution of signal events for a DPG from pulsar B1133+16 and a DPG composed of RFI signals. The figure reveals significant differences in the morphology of signal-to-noise ratio distribution curves between pulsar and RFI DPGs, with noticeable differences in their window width distribution curves as well, which aids in the recognition of pulsar DPGs.

## 2.3 Data Feature Design

After filtering and peak detection screening, the signal event table contains approximately 3.74 million total records, forming 53,066 DPGs, of which 35,063 are RFI records and 18,003 are pulse records belonging to 47 known pulsars. Designing features for DPGs is crucial for correct classification. Referencing existing DPG feature application methods, this paper designs the following features:

1. **Dispersion Measure (DM)**: The integrated column density of free electrons between the pulsar and Earth along the signal propagation direction,

measured in pc $cm^{-3}$. The DM value for a DPG is taken as the DM value corresponding to the strongest signal event within it.

2. **Signal-to-Noise Ratio (S/N)**: The ratio of signal to noise, i.e., the ratio of the voltage value of the signal received by the radio telescope to the simultaneously recorded noise voltage. Higher S/N indicates stronger signals and weaker noise. S/N is the primary basis for judging pulse events, and a DPG's S/N is taken as the value corresponding to its strongest signal event.

3. **Window Width (Duration)**: The window width of the rectangular window function used for peak detection in time series signals, representing the time range of the window and serving as a computational parameter for peak extraction. A series of different window widths were used for peak detection, and different window widths may detect different peak results. A DPG' s window width is taken as the value corresponding to its strongest signal event.

4. **DM Extent (DM_E)**: The range of DM values for all signal events in a DPG, i.e., the coverage range of the curve in subplot 1 of [Figure 1: see original paper].

5. **Time Extent (Time_E)**: The time range for all signal events in a DPG, measured in seconds.

6. **Number of Events (N_{Events})**: The number of signal events contained in a DPG. Too few events indicate weak dispersion effects and likely represent non-pulsar signals.

7. **Average DM (aDM)**: The average dispersion of all signal events belonging to the same DPG.

8. **Average Time of Pulse (aTime)**: The average time of all signals forming a DPG. Since the LOTAAS project uses 12 sub-stations for simultaneous observations, generating 222 celestial radiation data streams for each sky pointing, these data undergo preliminary processing to form many different time series. The average pulse time helps determine whether pulse signals from different time series originate from the same celestial body. For pulsar signals, multiple sub-stations may observe them simultaneously, whereas RFI signals are often observed by only one sub-station.

9. **Kurtosis of S/N Distribution (KurtSigma)**: The kurtosis value of the S/N distribution curve for all signals forming a DPG, minus the kurtosis of a normal distribution (i.e., the kurtosis of the curves in the first two subplots of [Figure 1: see original paper] minus the kurtosis at normal distribution, where the normal distribution kurtosis coefficient is 3).

10. **Kurtosis of Duration Distribution (KurtDuration)**: The kurtosis value of the distribution curve of window width values used for peak detection for each event in a DPG, minus the kurtosis of a normal distribution

(i.e., the kurtosis of the distribution curves in the last two subplots of [Figure 1: see original paper] minus the kurtosis at normal distribution).

## 3.1 Wrapper Feature Selection

The purpose of feature selection is to remove irrelevant and redundant features for the current learning task, reduce learning difficulty, and promote understanding of features and problems. The key is to establish an evaluation criterion to distinguish which feature combinations aid recognition. To enhance the correlation between features and models and improve model performance, this paper employs the wrapper method for feature selection before recognition.

Wrapper feature selection methods are directly related to the chosen classifier for subsequent tasks, using the classifier' s performance as the evaluation criterion for feature subsets—meaning wrapper feature selection directly optimizes for the given learner ([Figure 2: see original paper]). Therefore, the feature subset determined by wrapper feature selection is most compatible with the currently selected classifier.

This paper implements wrapper feature selection using Recursive Feature Elimination (RFE). The classifier is trained on a given feature set, the least important feature is removed from the current feature set, and training continues on the new feature set. This process is recursively repeated until the desired number of features is reached, thereby determining the optimal feature subset. For a given classifier, the features in the final selected subset are the most important.

## 3.2 XGBoost Classification Learner

XGBoost is an ensemble learning algorithm that employs an ensemble strategy based on decision trees. XGBoost comprises a collection of iterative residual trees, using gradient boosting to continuously reduce the loss of generated decision trees. Each tree learns the residuals of all previous trees, with the final prediction result for a sample being the sum of predictions from all trees.

XGBoost uses a forward distribution algorithm to learn an additive model containing K trees:

$$\hat{y}_i = \sum_{t=1}^{K} f_t(x_i), \quad f_t \in F$$

where K is the total number of trees, $f_t$ represents the t-th tree, $x_i$ denotes the input sample, $\hat{y}_i$ is the prediction result, $f_t(x_i)$ is the prediction result of the t-th tree, and F represents the function space composed of decision trees.

To solve the entire decision tree function space, the objective function must be continuously optimized. XGBoost' s overall objective function can be expressed as [**?**]:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

where $l$ is the loss function representing the difference between the predicted value $\hat{y}_i$ and target value $y_i$, and $\Omega(f_t)$ is the regularization term for the t-th tree, used to constrain the complexity of decision trees—the higher the complexity, the larger the regularization term.

First, a greedy algorithm is used to find local optimal solutions:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$$

where $\hat{y}_i^{(t-1)}$ represents the prediction result of the i-th tree at iteration t-1. At each iteration, we seek $f_t$ that maximally reduces the loss function. The objective function can then be rewritten as:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

Second, the objective function is approximated using second-order Taylor expansion:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} [l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t)$$

where $g_i$ and $h_i$ represent the first and second derivatives of the loss function, respectively:

$$g_i = \partial_{\hat{y}^{i(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$$
$$h_i = \partial^2_{\hat{y}^{i(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$$

Removing the constant term $l(y_i, \hat{y}_i^{(t-1)})$ for iteration t, the regularization term in XGBoost measures tree complexity:

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2$$

where T represents the number of nodes in each tree, $w$ is the output score of each leaf node, and $\lambda, \gamma$ are constants. The objective function can be further expressed as:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n}[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)] + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2$$

If $q(x_i)$ maps sample $x_i$ to a leaf node, then:

$$f_t(x) = w_{q(x)}, \quad w \in \mathbb{R}^T$$

and defining the sample set on each leaf node j as $I_j = \{i|q(x_i) = j\}$, the objective function can be expressed as:

$$\mathcal{L}^{(t)} = \sum_{j=1}^{T}[(\sum_{i\in I_j} g_i)w_j + \frac{1}{2}(\sum_{i\in I_j} h_i + \lambda)w_j^2] + \gamma T$$

Finally, the objective function is optimized by calculating the leaf node output score $w_j$ that minimizes the objective function at iteration t. Taking the derivative with respect to $w_j$ and setting it to zero yields:

$$w_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda}$$

Substituting equation (10) into (9) gives the final optimized objective function:

$$\mathcal{L}^{(t)} = -\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{i\in I_j} g_i)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T$$

When selecting feature attributes for node splitting, XGBoost uses a greedy algorithm or approximate greedy algorithm to traverse all feature split points, calculating the gain in objective function values for each, and selects the optimal feature for splitting. Tree growth stops when the gain from new splitting is less than a set threshold or the maximum depth is reached. XGBoost performs second-order Taylor expansion on the cost function and introduces operations such as shrinkage, row sampling, and column sampling, providing excellent overfitting prevention, high computational efficiency, and strong generalization capabilities. For XGBoost implementation, the Python-based machine learning toolkit Scikit-learn can be used directly.

### 3.3 Feature Selection Evaluation Process

Based on the aforementioned theories and methods, this paper combines wrapper feature selection with the XGBoost algorithm. According to the input

dataset, a threshold is set to obtain the optimal feature subset under that threshold, which is then input into the XGBoost algorithm for classification to obtain results. The specific flowchart is shown in [Figure 3: see original paper].

To analyze the classification effectiveness of the current method, we evaluate model predictions using a confusion matrix. The dataset in this paper is divided into RFI DPGs and single-pulse DPGs. If a pulsar' s DPG is correctly identified as a single-pulse signal, it is called a True Positive (TP); if incorrectly classified as RFI, it is a False Negative (FN). Similarly, if RFI data is incorrectly classified as a single-pulse signal, it is a False Positive (FP); if correctly classified as RFI, it is a True Negative (TN). shows the confusion matrix for binary classification.

Common evaluation metrics for binary classification problems include accuracy, error rate, precision, recall, and F1-score [**?**]. Accuracy represents the proportion of correctly classified samples to total samples. However, when sample categories are imbalanced in the dataset, classifiers tend to judge samples as coming from the majority category, resulting in artificially high accuracy. Therefore, accuracy alone cannot objectively evaluate algorithm performance, and other metrics must be introduced. Precision represents the proportion of actual pulsar signals among samples predicted as pulsars. Recall represents the proportion of pulsar signals correctly identified as pulsar signals. Precision and recall have a trade-off relationship, while F1-score combines both metrics through harmonic averaging. A higher F1-score obtained by the current method indicates better overall performance.

## 4. Experiments and Analysis

The experimental dataset contains 18,003 DPGs from 47 known pulsars and 35,063 RFI DPGs. The specific procedure involves randomly partitioning the dataset 10 times using cross-validation for model training and evaluation, with 80% used for training and the remaining 20% for validation. To prevent data leakage in classification and ensure relative balance between single-pulse and RFI samples, the dataset was not randomly partitioned directly. Instead, DPGs belonging to the 47 known pulsars and RFI DPGs were randomly grouped 10 times separately, with 80% (i.e., DPGs from 38 known pulsars and 80% of RFI DPGs) used for training and the remaining DPGs from 9 known pulsars and 20% of RFI records used for validation.

For comparative analysis, this paper conducted experimental comparisons using XGBoost, GBDT, AdaBoost, Random Forest, and BP Neural Network (BPNN) models. To ensure comparable results, each method was optimized before comparison, with experimental results using the optimal parameters found. BPNN employed a three-layer architecture (10 nodes in the input layer, 56 nodes in the hidden layer, and 2 output nodes), a learning rate of 0.0015, cross-entropy loss function, and Adam optimizer. GBDT and Random Forest had maximum iterations of 100, maximum depth of 20, and learning rate of 2. AdaBoost had maximum iterations of 100. shows the average experimental results of the five

models across 10 random partitions on this dataset.

The wrapper feature selection process is closely integrated with the selected classifier, testing the classification performance of multiple feature subsets on the model through some feature search strategy. lists the classification evaluation results for different classifiers with their optimal feature combinations. The results show that all five models achieve relatively high precision and recall for classifying pulsar DPGs and RFI DPGs. Notably, XGBoost achieves the highest precision, recall, and F1-score among the five models, averaging 1 to 2 percentage points higher than other models. In terms of misclassification rate, GBDT has the highest while XGBoost has the lowest. Combining these metrics, XGBoost demonstrates comprehensive advantages in DPG classification and recognition for single-pulse signals.

Regarding feature usage, nine feature parameters—Duration, DM, S/N, DM_E, Time_E, N_{Events}, aDM, KurtDuration, and KurtSigma—were selected as optimal feature combinations by all five models. AdaBoost and BPNN further selected aTime as an optimal feature. The fact that aTime was not selected as an optimal feature by the other three models suggests its role in single-pulse recognition is not particularly significant.

For the XGBoost classifier, in addition to features used for model training, hyperparameters also affect single-pulse recognition results to some extent, with maximum tree depth and model learning rate being the primary parameters affecting performance.

[Figure 4: see original paper] shows the relationship between maximum tree depth and model F1-score. When maximum tree depth is less than 25, training time increases steadily and then remains essentially stable; the model' s F1-score shows a trend of first increasing, then decreasing, and then stabilizing. When maximum tree depth is set to 6, XGBoost achieves the highest F1-score on the test set with relatively short training time. This indicates that on the dataset used in this paper, a maximum tree depth of 6 provides a good balance between training time consumption and single-pulse classification performance. [Figure 5: see original paper] shows the impact of learning rate on XGBoost performance. The results indicate that XGBoost achieves optimal classification performance when the learning rate reaches 0.007.

Feature quantity also affects model performance for single-pulse recognition. On the dataset used in this paper, XGBoost combined with wrapper feature selection was used to analyze feature importance. For the wrapper feature selection algorithm, we obtained optimal feature subsets of different sizes by setting different thresholds and compared model performance based on these subsets. shows the F1-scores of XGBoost models trained on different-sized feature subsets for single-pulse signal recognition tasks.

The results demonstrate that feature quantity affects pulse signal classification performance. Although each feature impacts the model differently, the number and combination of input features are also key factors affecting model perfor-

mance. Different results are obtained with different numbers of input features. The highest F1-score is achieved when using 9 features: Duration, DM, S/N, DM_E, aDM, Time_E, N_{events}, KurtSigma, and KurtDuration.

## 5. Conclusion

In recent years, with the maturation of periodic pulse signal detection methods, single-pulse signal recognition has become an important field in pulsar research. Since relatively few features can be extracted from single-pulse signals, machine learning methods have become the primary approach. Designing key features and finding optimal machine learning algorithms are critical tasks in current pulsar signal recognition.

Building upon previous research, this paper combines the XGBoost classifier with wrapper feature selection, using the LOTAAS dataset for experimental comparison with AdaBoost, GBDT, Random Forest, and BP neural network models. The results show that XGBoost achieves lower misclassification rates and higher precision, recall, and F1-score in single-pulse recognition, making it an excellent method for single-pulse signal identification and extraction. In this study' s experimental design, DPGs from 47 known pulsars and RFI signals were randomly grouped 10 times separately, effectively avoiding data leakage impacts from dataset partitioning. If the 18,003 DPGs from 47 pulsars were directly partitioned for training and testing, the precision would reach as high as 99.79% and the F1-score would be 99.76%, demonstrating that the dataset partitioning method significantly affects recognition results.

Regarding feature selection, experimental results indicate that nine features—DM, S/N, Duration, DM_E, Time_E, N_{Events}, aDM, KurtDuration, and KurtSigma—were selected by most models, demonstrating strong discriminative power.

Annotating single-pulse signals to establish training datasets is labor-intensive and time-consuming work requiring long-term accumulation. Although this study uses the LOTAAS dataset, its results and methods can provide references for single-pulse signal research and applications in China, particularly those based on FAST-detected signals. Currently, mining and application of FAST data in China are being vigorously promoted, with successful detections of single-pulse fast radio bursts [**?**][**?**]. Additionally, as analysis and mining of single-pulse signal features continue, new research methods will be continuously proposed and improved.

## References

[1] Antony Hewish. Pulsars as Physics Laboratories[J], *Interdisciplinary Science Reviews*, 1994, 19:1, 70-74.
[2] Cordes J M; McLaughlin, M A Searches for Fast Radio Transients[J], *The Astrophysical Journal*, 2003, 596:2, 1142-1154.

[3] Thomas Ryan Devine, Katerina Goseva-Popstojanova, Maura McLaughlin, Detection of dispersed radio pulses: a machine learning approach to candidate identification and classification[J], *Monthly Notices of the Royal Astronomical Society*, 2016, 459:2, 1519–1532.
[4] Patel C, Agarwal D, Bhardwaj M, Boyce M M, Brazier A, & Chatterjee S, et al. Palfa single-pulse pipeline: new pulsars, rotating radio transients, and a candidate fast radio burst[J]. *Astrophysical Journal*, 2018, 869(2).
[5] McLaughlin M, Lyne A, Lorimer D, et al. Transient radio bursts from rotating neutron stars[J]. *Nature*, 2006, 439, 817–820.
[6] Lorimer D R, Bailes M, McLaughlin M A, Narkevic D J & Crawford F A. bright millisecond radio burst of extragalactic origin[J]. *Science*, 2007, 318, 777–780.
[7] Deneva J S, Cordes J M, Mclaughlin M A, et al. Arecibo Pulsar Survey Using ALFA: Probing Radio Pulsar Intermittency and Transients[J]. *Astrophysical Journal*, 2009, 703(2):2259-2274.
[8] Karako-Argaman C, et al. Discovery and Follow-up of Rotating Radio Transients with the Green Bank and LOFAR Telescopes[J]. *The Astrophysical Journal*, 2015, 809(1):67.
[9] Ryan D T, Katerina G P, Maura M L. Detection of Dispersed Radio Pulses: A machine learning approach to candidate identification and classification[J]. *Monthly Notices of the Royal Astronomical Society*, 2016(2):stw655.
[10] Wang Y C, Zheng J H, Pan Z C, Li M T. Review of pulsar candidate sample classification methods[J], *Journal of Deep Space Exploration*, 2018, 5:3, 203-211.
[11] McFadden R, Karastergiou A, Roberts S. Machine learning for pulsar detection[J]. *Proceedings of the International Astronomical Union*, 2017, 13(S337): 372-373.
[12] Eatough R P, Molkenthin N, Kramer M, et al. Selection of radio pulsar candidates using artificial neural networks[J]. *Monthly Notices of the Royal Astronomical Society*, 2010, 407(4).
[13] Bates S D, Bailes M, Barsdell B R, et al. The High Time Resolution Universe Pulsar Survey –VI. An artificial neural network and timing of 75 pulsars[J]. *Monthly Notices of the Royal Astronomical Society*, 2012, 427(2):1052-1065.
[14] Morello V, Barr E D, Bailes M, et al. SPINN: a straightforward machine learning solution to the pulsar candidate selection problem[J]. *Monthly Notices of the Royal Astronomical Society*, 2014, 443(2).
[15] Zhu W W, Berndsen A, Madsen E C, et al. Searching for Pulsars Using Image Pattern Recognition[J]. *The Astrophysical Journal*, 2014, 781(2):117.
[16] Michilli D, Hessels J, Lyon R J, et al. Single-pulse classifier for the LOFAR Tied-Array All-sky Survey[J]. *Monthly Notices of the Royal Astronomical Society*, 2018(3):3.
[17] Chen T, Guestrin C. XGBoost: A scalable tree boosting system[C]. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 16*. ACM, 2016:785-794.
[18] Coenen T, van Leeuwen J, Hessels J W T, et al. The LOFAR pilot surveys for pulsars and fast radio transients[J]. *Astronomy & Astrophysics*, 2014, 570.
[19] Sanidas S, Cooper S, Bassa C G, Hessels J W. T, Kondratiev V I, Michilli

D, et al. The LOFAR Tied-Array All-Sky Survey (LOTAAS): Survey overview and initial pulsar discoveries[J]. *Astronomy & Astrophysics*, 2019, 626.
[20] Lorimer D R, Kramer M. *Handbook of Pulsar Astronomy*[DB/OL], 2004.
[21] Li H. Statistical Learning Methods[M]. Tsinghua University Press, 2012.
[22] Suthaharan S. *Machine learning models and algorithms for big data classification*[M]. New York: Springer, 2016.
[23] Weiwei Zhu, Di Li, Rui Luo, et al. A Fast Radio Burst discovered in FAST drift scan survey[J]. *ApJL*, 2020, DOI: 10.3847/2041-8213/ab8e46
[24] Chen-Hui Niu, Di Li, Rui Luo, et al. CRAFTS for Fast Radio Bursts Extending the dispersion-fluence relation with new FRBs detected by FAST[J]. *ApJL*, 2021, DOI: 10.3847/2041-8213/abe7f0

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