Pulsar Candidate Classification with Multimodal Large Language Models

Fuyong Zhao[1] **Yuyang Li**[2,4] **Yanhao Wang**[3,*] **Hui Li**[1,*] **Mei Chen**[1]
Panfeng Chen[1] **Ningchen Sun**[2,4] **Cunshi Wang**[2,4] **Jifeng Liu**[2,4]
[1]Guizhou University
[2]University of Chinese Academy of Sciences
[3]East China Normal University
[4]National Astronomical Observatories, Chinese Academy of Sciences
gs.fyzhao22@gzu.edu.cn liyuyang22@mails.ucas.ac.cn
yhwang@dase.ecnu.edu.cn {cse.huili,mchen,pfchen}@gzu.edu.cn
sunnc@ucas.ac.cn {wangcunshi,jfliu}@nao.cas.cn

Abstract

Discovering pulsars is of great scientific value in the field of astronomy. Driven by the huge volume of data from radio telescope survey projects, machine learning (ML) methods have been proposed and widely adopted for this problem. However, existing ML methods rely on a single data modality, either visual or numerical, leading to suboptimal performance. In this paper, we first explore the usage of multimodal large language models (MLLMs) for pulsar candidate classification. Specifically, we propose a novel method called STARWHISPER-PULSAR that fine-tunes pre-trained MLLMs using labeled data in visual, textual, and numerical modalities to decide if a candidate is a real pulsar or a non-pulsar noise. We show that STARWHISPER-PULSAR outperforms state-of-the-art ML methods for pulsar candidate classification in a few training epochs. These results validate the potential of MLLMs in data-driven astronomical research, paving the way for their broader scientific applications.

1 Introduction

A pulsar is a highly magnetized, rotating neutron star that emits beams of electromagnetic radiation out of its magnetic pole, often observed as pulsed signals. It plays a crucial role in various astronomical and astrophysical problems, such as precise timing, gravitational wave detection, and the understanding of high-energy physical phenomena, to name a few. Due to their great scientific value, finding pulsars is an important task in astronomical research [7, 9]. The traditional method for pulsar discovery is to collect observation data of radiation signals using radio telescopes, to pre-process them into diagnostic plots, to ask human experts to manually check and determine whether these plots are from real pulsars or non-pulsar noise, and to verify the results with further observations.

In the last two decades, large-scale radio telescope surveys [5, 11, 12] have collected a huge volume of radiation signal observation data, from which millions of pulsar candidates have been extracted. As such, manual methods to screen the candidates become infeasible. To achieve higher efficiency, machine learning (ML) methods for automatic pulsar candidate classification have emerged in recent years. Generally, existing ML methods can be categorized into feature engineering-based and end-to-end methods. The first category of methods [2, 4, 18, 19] extracts the statistical characteristics

*Corresponding authors.

Foundation Models for Science Workshop, 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

of pulses, such as *signal-to-noise ratio* (SNR), *pulse width*, and *periodicity*, from the diagnostic plots to train classification models. But these methods rely heavily on carefully designed features, requiring in-depth astronomical knowledge, while having quite limited generalizability. To overcome such limitations, several end-to-end deep learning methods [15, 25, 27, 28] were proposed for pulsar candidate classification. Using plot images as input, these methods automatically extract hidden features and capture complex patterns, achieving better performance than numerical feature-based methods. However, they still suffer from two shortcomings. First, they focus only on visual features, while ignoring numerical features that are also crucial for classification. Second, since data published by different projects often contain various types of plots, an end-to-end method specific to one project might not be applied directly to other projects.

Our Contributions To address the above issues, in this paper, we explore how to utilize multimodal large language models (MLLMs), for pulsar discovery. Our work is inspired by the successful adoption of MLLMs in the astronomical domain [13, 20, 21]. However, to the best of our knowledge, none of these studies has considered the problem of pulsar candidate classification. We propose STARWHISPER-PULSAR, a novel method to decide if a candidate is a real pulsar or a non-pulsar noise based on visual, textual, and numerical inputs. Generally, our method is built on pre-trained MLLMs such as DeepSeek-VL [17] and InternVL [3], which are fine-tuned to enhance their understanding of pulsar data. To adapt flexibly to different types of plot images, it provides three modes, namely *single*, *multiple*, and *combined*, for visual input. It takes well-designed prompts that include task descriptions, instructions, pulsar statistical characteristics, and empirical rules as textual input. Then, we compose labeled pulsar candidate instances from radio telescope surveys, which comprise pre-processed diagnostic plot images, statistical characteristics, and ground-truth labels, as a visual question-answering (VQA) data set [1, 14] and adopt LoRA [8] to perform the fine-tuning process on the VQA data set.

Finally, we conduct comprehensive experiments on two real-world data sets, FAST [12] and HTRU [11], to evaluate the performance of STARWHISPER-PULSAR for pulsar candidate classification. The results demonstrate that STARWHISPER-PULSAR outperforms state-of-the-art ML methods in almost all evaluation metrics. It has accuracy rates of over 97% and 99% on the FAST and HTRU data sets within only six training epochs. These results validate the efficacy of MLLMs in data-driven astronomical research.

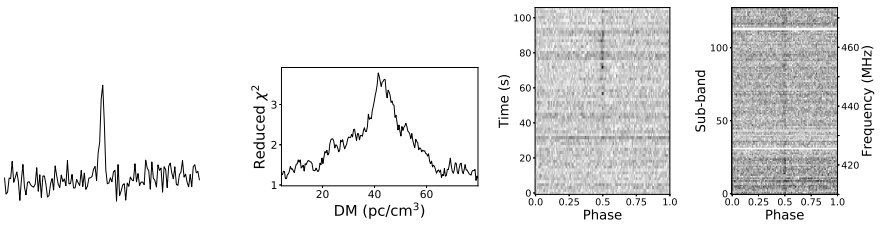
2 Related Work

The existing ML methods for pulsar candidate classification can generally be divided into two classes: feature engineering-based and end-to-end deep learning (DL) methods.

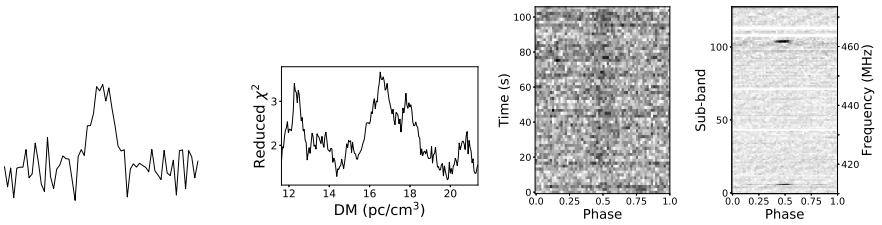
Eatough et al. [4] proposed a multi-layer perceptron model using 12 numerical features. Bates et al. [2] proposed an artificial neural network model trained on an expanded set of 22 features. Morello et al. [19] considered features with weak signal accommodation, robustness against noise, and reduced inter-feature correlation. Lyon et al. [18] introduced new features with better discriminative ability and reduced bias. They also proposed a decision tree-based algorithm to address class imbalances, i.e., positive samples (resp. real pulsars) are much fewer than negative ones. Tan et al. [24] improved upon the method in [18] by introducing time-phase and frequency-phase diagrams as new features and leveraging ensemble learning. The above feature engineering-based methods require in-depth astronomic knowledge and often show limited generalizability for different data sets.

Zhu et al. [28] utilized a convolutional neural network (CNN) to recognize image patterns in pulsar search. Wang et al. [26] adopted ResNet [6] to optimize the CNN model in [28]. Zeng et al. [27] proposed a concat CNN model to improve the performance of [26]. Tariq et al. [25] and Liu et al. [15] proposed new methods to address class imbalances in pulsar candidates. The above DL methods treat plots as images, totally ignoring numerical features. In addition, they cannot be easily adapted to different plot types.

Recent studies have revealed the excellent capabilities of (multimodal) LLMs in astronomy. Nguyen et al. [20] proposed AstroLLaMA, which is fine-tuned on LLaMA using more than 300,000 abstracts of astronomical papers, and verified that AstroLLaMA could generate scientifically more relevant texts than general-purpose LLMs. Parker et al. [21] proposed AstroCLIP, a cross-modal foundation model to embed galaxy images and spectra into a shared, physically meaningful latent space, which



(a) Pulsar diagnostic plots



(b) Non-pulsar diagnostic plots

Figure 1: Examples of diagnostic plots of a pulsar and a non-pulsar noise, where the *pulse profile*, *DM profile*, *sub-integrations*, and *sub-bands* plots are placed from left to right in each row.

could then be used for a variety of downstream tasks, such as photometric redshift estimation and morphology classification. Li et al. [13] considered the combination of deep learning and LLM for stellar light-curve classification. These attempts indicate that (multimodal) LLMs can serve as helpful research assistants, initiating a new paradigm in the astronomical domain. However, to the best of our knowledge, none of them considered using MLLMs for pulsar discovery.

3 The StarWhisper-Pulsar Model

Problem Formulation Each pulsar candidate is described by a set of four diagnostic plots $X_v = \{x_1, x_2, x_3, x_4\}$ in image format. These plots are (a) *pulse profile* x_1 that indicates the variation in the radiation signal over its period, (b) *dispersion measure (DM) profile* x_2 that measures the extent to which the signal deviates from white noise after dispersion correction, (c) *sub-integrations* x_3 that shows the persistence of the pulse signal over time, and (d) *sub-bands* x_4 that illustrates the distribution of the pulse signal across frequency channels. We provide two examples for the diagnostic plots of a pulsar and a non-pulsar noise in Figure 1. The data published by a ratio telescope survey project may not contain all four plots. However, we assume that each data set must contain at least one of these plots. Our method is versatile on any data set that contains one or more of the above four plots. Each pulsar candidate is also associated with a set $F = \{ft_1, ft_2, \ldots, ft_n\}$ of n numeric statistical features. As indicated in [2], a set of 22 features can be extracted from the diagnostic plots. However, not all of them are equally important for pulsar characterization. In particular, we pick the following four key features as input: (1) *best period* (i.e., the shortest interval between two subsequent pulse signals); (2) *best DM* (i.e., the optimal dispersion measure); (3) *best SNR* (i.e., the optimal ratio between the pulse signal and background noise); and (4) *pulse width* (i.e., the duration of the pulse signal). Finally, we use $y \in \{0, 1\}$ to denote the output of the model with $y = 0$ indicating a non-pulsar noise and $y = 1$ representing a pulsar. The goal is to learn a function that maps the input plot images X_v and numerical features F to the output y, i.e., $(X_v, F) \mapsto \{0, 1\}$.

Model Description The architecture of the STARWHISPER-PULSAR model is illustrated in Figure 2. Specifically, it is built on a pre-trained MLLM and is fine-tuned to learn a more comprehensive understanding of pulsar data. The visual module uses a set X_v of plot images as input. It uses a visual encoder $g(\cdot)$ to extract the features of X_v, i.e., $Z_v = g(X_v)$. Next, it aligns visual features with textual ones using a linear projection matrix W, i.e., $H_v = W \cdot Z_v$. Meanwhile, the textual module uses a prompt X_q consisting of task descriptions, instructions, pulsar characteristics, and empirical rules as input. It uses a word embedding $f(\cdot)$ for feature extraction, i.e., $H_q = f(X_q)$. Finally, an LLM $\Phi(\cdot)$ produces the output based on visual and textual features denoted as $\mathcal{T}(y) = \Phi(H_v, H_q)$.

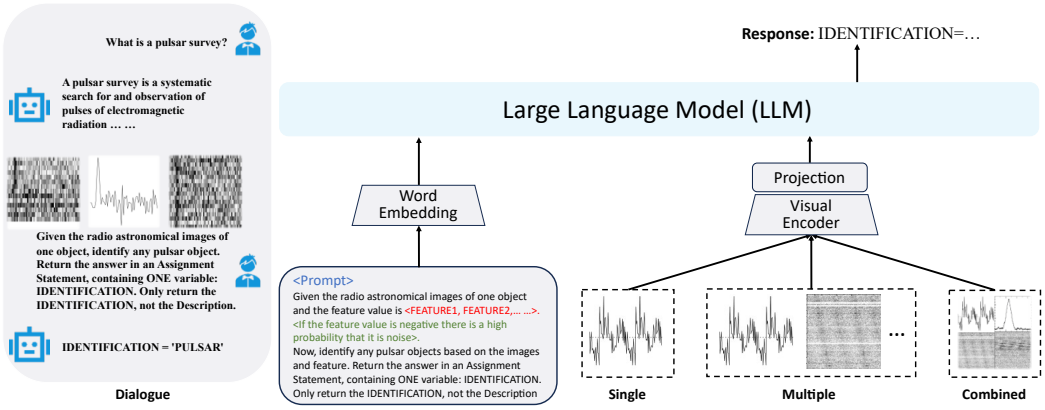


Figure 2: Illustration of the architecture of STARWHISPER-PULSAR.

In the implementation, we use DeepSeek-VL [17] or InternVL [3] as the base model, as they are open-source LLMs that can support both textual and visual modalities. As presented in the dialogue of Figure 2, the fine-tuned model can follow the prompt instructions with plot images and numeric attributes to distinguish pulsars from non-pulsar noise.

Existing DL methods for pulsar candidate classification are limited to processing one image at a time. However, there may exist multiple images (up to four) in X_v. Therefore, the visual module should be able to flexibly take one or more images as input. As such, we do not need to redesign it for different plot types provided by various projects. In particular, we consider three different input modes for X_v: *single*, *multiple*, and *combined*. In the single mode, we simply take a plot image as input. This mode is used only when there is only one plot available. In the multiple mode, we input all available images into the visual encoder one by one and concatenate their features for projection. In the combined mode, we combine all available images into a single image as input to the visual encoder. The visual encoder we use in STARWHISPER-PULSAR is the pre-trained ViT-L/14 model [22]. We then use a linear layer with a trainable projection matrix to align visual features with the word embedding space. After alignment, the visual features H_v have the same dimension as the textual features H_q in the embedding space.

To build the prompt X_q in STARWHISPER-PULSAR, we provide explicit instructions and qualifications for the tasks to be performed and to restrict the output according to the specifications. Furthermore, we embed numeric statistical features and relevant empirical rules in the prompt. In this way, the statistical features can signify the physical properties of pulsars, while the empirical rules are established based on expert experience to indicate what kinds of signals are not likely from real pulsars. With empirical rules, domain knowledge can be effectively leveraged in classification. In general, the procedure for building the prompt can be expressed as $X_q = \mathcal{T}(F, R)$, where $\mathcal{T}$ is the instruction template, F is the set of features, and R is the set of constraint rules. The instruction template in STARWHISPER-PULSAR is shown as follows:

> **[IMAGES X_v]** Given the radio astronomical images of one object and the feature values are $[ft_1, \ldots, ft_n]$. $[\text{RULE}_1, \ldots, \text{RULE}_m]$. Now, identify any pulsar objects based on the images and features. Return the answer in an Assignment Statement, containing ONE variable: **IDENTIFICATION**. Return only the IDENTIFICATION, not the Description.

We construct a visual question-answering (VQA) data set [1, 14] to fine-tune the model. The data set is derived from training instances that include plot images, numerical features, empirical rules, and ground-truth labels. We follow the instruction template to compose a question-answer pair based on each training instance. We fine-tune the parameters in DeepSeek-VL, including those in the visual encoder, projection layer, and LLM, ensuring that they are well adjusted for the task of pulsar candidate classification. For InternVL2-40B, we adopt the LoRA (Low-Rank Adaptation) method [8] to perform the fine-tuning process, as full fine-tuning is unaffordable due to the large number of parameters.

Table 1: Parameter settings for the fine-tuning process.

Base Model	Method	# Epochs	Learning Rate	Weight Decay
DeepSeek-VL-7B	Full	6	1e-5	0.01
InternVL2-40B	LoRA	3	1e-4	0.1

4 Experiments

4.1 Experimental Setup

Data Sets We used two real observation data sets, FAST [26] and HTRU [19], in the experiments.

The FAST data set contains 1,163 known pulsars and 14,319 non-pulsar candidates. Each instance in the FAST data set is acquired by the PRESTO toolkit [23], which can effectively remove interference, eliminate dispersion, and store the processed 3D data along with detailed data description in PDF format. Then, it also generates time-phase and frequency-phase plots by summing the data along the frequency channel and time interval. In addition, by summing up the data along the frequency channel and time interval, a histogram of the pulse profile is also obtained. Finally, the DM profile is generated to graphically show the relationship between DM trials and the corresponding reduced (χ^2) values. After the above processing procedure, each instance is associated with four plot images with a resolution of 512×512 pixels but does not contain any numerical features. We randomly select 837 known pulsars and 1,593 non-pulsar candidates for training and the remaining instances for test.

The HTRU data set consists of 1,196 known pulsars and 89,996 non-pulsar candidates. The original instances in the HTRU data set are stored in NumPy format. We use the Matplotlib library to read each instance and draw the pulse profile, frequency-phase plot, and time-phase plot. Moreover, we leverage the feature extraction methods proposed in [18] to calculate the features used in [2, 4, 19], thereby obtaining a set of statistical features. Then, we only use the four key features in Section 3, as they are known to be the most important for classification. To keep consistency with FAST, we also randomly sampled 837 known pulsars and 1,593 non-pulsar candidates as the training set. We used the remaining 359 known pulsars and sampled 8,406 non-pulsar candidates as the test set. Each instance in HTRU is associated with three plot images except the DM profiles and four numerical features.

Baselines We compared STARWHISPER-PULSAR to several state-of-the-art deep learning methods in the experiments. In the FAST data set, we used the known results for PICS [28], PICS-RESNET [26], and CCNN [27] in their original papers. In the HTRU data set, we implemented PICS-RESNET [26] and H-CCNN [27] and ran them on the same training and test sets as STARWHISPER-PULSAR for the results.

Implementation We used DeepSeek-VL-7B and InternVL2-40B as base models in STARWHISPER-PULSAR and AdamW [16] as the default optimizer. Gradient clipping with a maximum norm of 1 was used for stable training. A cosine learning rate schedule was adopted to achieve smooth convergence, complemented by a warm-up ratio of 5%. Detailed parameter settings for the fine-tuning process are presented in Table 1. All experiments were carried out on a sever running Ubuntu 20.04 LTS with eight Nvidia RTX A6000 GPUs and eight Nvidia H100 GPUs. The code and data are publicly available at `https://github.com/ACMISLab/StarWhisper-Pulsar`.

4.2 Experimental Results

Overall Performance We compare the performance of our methods with baselines for pulsar candidate classification using five evaluation metrics, namely *accuracy*, *precision*, *recall*, F_1*-score*, and *#misses*, in Table 2. The first four metrics are standard for binary classification problems; the last metric, #misses (i.e., the number of real pulsars that are incorrectly identified as non-pulsar noise), is specific for pulsar discovery, as real pulsars are very rare among candidates, and thus missing any of them leads to more losses than checking suspected candidates with observations. We can see that STARWHISPER-PULSAR, using either DeepSeek-VL-7B or InternVL2-40B as the base model, performs significantly better than the state-of-the-art deep learning methods on almost all metrics on both data sets. In the FAST data set, STARWHISPER-PULSAR has significantly higher scores in

Table 2: Overall performance of different methods for pulsar candidate classification in the experiments. The best result for each measure on each data set is highlighted in bold font.

Data Set	Method	Accuracy ↑	F_1-score ↑	Recall ↑	Precision ↑	# Misses ↓
FAST	PICS	0.936	0.415	0.954	0.265	16
	PICS-ResNet	0.933	0.413	0.982	0.261	6
	H-CCNN	0.963	0.557	0.963	0.392	12
	V-CCNN	0.917	0.363	0.988	0.223	4
	H-CCNN+V-CCNN	0.948	0.363	0.982	0.311	4
	StarWhisper-Pulsar-D-7B	0.970	0.616	0.975	0.450	8
	StarWhisper-Pulsar-I-40B	**0.973**	**0.649**	**0.994**	**0.481**	**2**
HTRU	PICS-ResNet	0.989	0.875	0.967	**0.800**	12
	H-CCNN	0.969	0.710	0.933	0.573	24
	StarWhisper-Pulsar-D-7B	0.986	0.853	0.981	0.755	7
	StarWhisper-Pulsar-I-40B	**0.989**	**0.884**	**0.992**	0.798	**3**

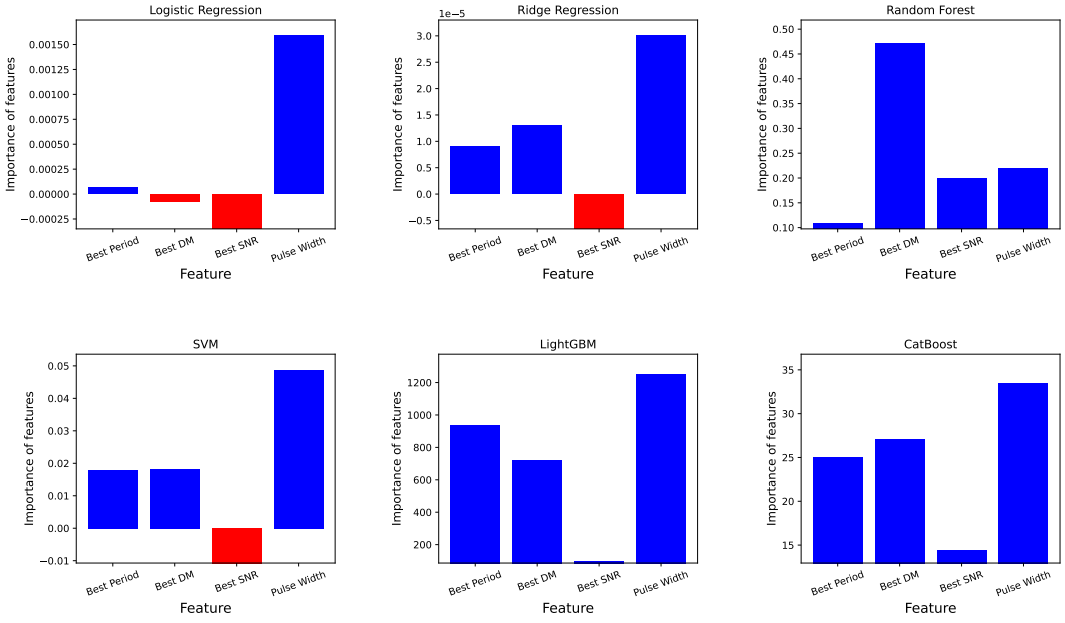


Figure 3: Results of numerical feature importance in pulsar candidate classification.

terms of *precision* and F_1-*score*, indicating much fewer false positives. In the HTRU data set, it has higher *recall* scores and greatly reduces the number of missed pulsars.

Furthermore, StarWhisper-Pulsar based on InternVL2-40B outperforms that based on DeepSeek-VL-7B. This validates the scaling law of LLMs [10]: Language models with more parameters (40B vs. 7B) show stronger generalizability than those with fewer parameters. The fine-tuned InternVL2-40B model achieves accuracy scores of nearly 0.97 and 0.99 on the FAST and HTRU data sets, respectively. Moreover, its recall scores are above 0.99 on both data sets, only missing 2 of 326 real pulsars in the FAST data set and 3 of 359 pulsars in the HTRU data set.

Finally, as already shown in Table 1, the fine-tuning process is performed only in six and three epochs on a training set of 2,430 samples, which takes around two hours. By analyzing the training loss over epochs, we observe that the model converges in a few epochs and remains stable when more fine-tuning epochs are performed. These results show the efficiency of StarWhisper-Pulsar.

Feature Selection As shown in Figure 3, we analyze the importance of the four numerical features to provide a basis for their usage in the subsequent binary classification task. To ensure the comprehensiveness and reliability of the selection results, multiple feature selection methods are employed. These methods encompass linear, tree-based, and gradient-boosting models as follows:

Table 3: Performance of STARWHISPER-PULSAR (DeepSeek-VL-7B) with different input formats on the HTRU data set. The best results on each metric are also highlighted in bold font.

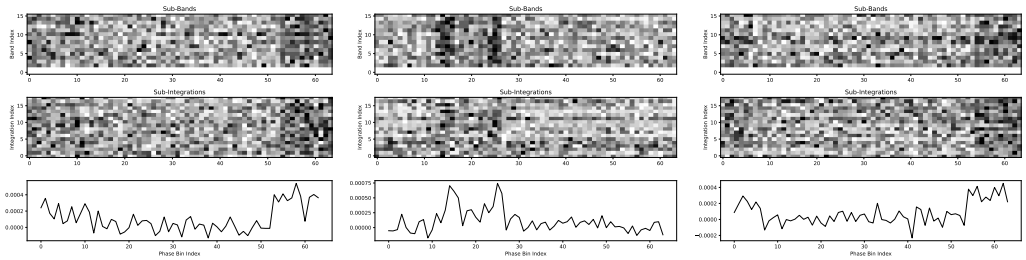
Input Format	Accuracy	F_1-Score	Recall	Precision	# Misses
S[1]	0.987	0.852	0.916	0.797	30
M[1, 2]	0.981	0.805	0.964	0.691	13
M[1–3]	0.987	0.860	0.969	0.773	11
M[1–3]+F[1]	**0.990**	**0.887**	0.969	**0.817**	11
M[1–3]+F[1–4]	0.987	0.864	0.969	0.779	11
M[1–3]+F[1]+R	0.986	0.853	0.981	0.755	7
C[1–3]	0.985	0.843	**0.983**	0.738	**6**

- **Logistic Regression:** A linear model that evaluates the importance of features by estimating the feature coefficients. We used L_1-regularization in our experiments to avoid overfitting.
- **Ridge Regression:** A linear model that stabilizes estimates and prevents multicollinearity through L_2-regularization. The importance of the feature is also determined by the regression coefficients.
- **Random Forest:** An ensemble model based on decision trees that determines the importance of the features by averaging their importance in all trees. The importance of the feature is quantified by the reduction in the Gini impurity.
- **Support Vector Machine (SVM):** We assess the importance of the feature through the coefficients of the linear SVM model. The coefficients in the SVM model reflect each feature's contribution to the decision boundary.
- **LightGBM:** An implementation of gradient-boosting decision trees (GBDT). The important of the feature is calculated based on the number of times a feature is used for split in the model and the gain brought by those splits.
- **CatBoost:** Another GBDT-based model where the importance of the feature is evaluated using its feature importance score.

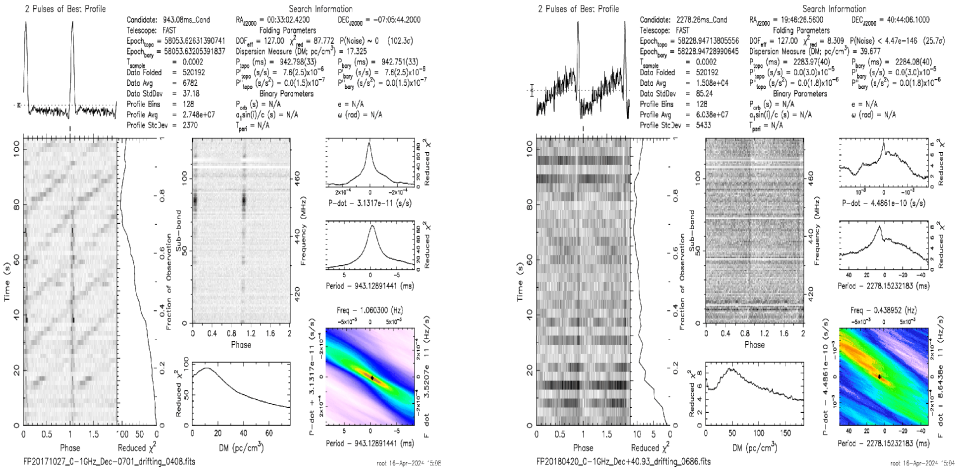
These methods cover a wide range of feature selection mechanisms, ensuring broad adaptability and robustness in evaluating feature importance. By integrating the evaluation results of multiple methods, we can more accurately identify key features, thus enhancing the predictive performance of the model. The experimental results show that, except for the Random Forest method that ranks *Pulse Width* as the second most important feature, all other methods evaluate *Pulse Width* as the most important feature. Therefore, *Pulse Width* is considered the key feature in pulsar binary classification in subsequent experiments.

Ablation Study We perform an ablation study for the effect of different input formats on the performance of STARWHISPER-PULSAR. The results are shown in Table 3. Using a single pulse profile image as input ("S[1]"), has achieved good accuracy but has a much lower recall score than any other method, resulting in 30 missed pulsars. Meanwhile, we can observe that the high performance achievable even with a single image is attributed to the extensive pre-training of MLLM and the rich features inherent in the image. Then, when the sub-integrations and sub-bands plots are added as input in the multiple mode ("M[1, 2]" and "M[1–3]"), the recall scores are obviously higher. This confirms that multiple plot images can provide more comprehensive visual features from different views, thus enhancing the capacity of the MLLM to understand pulsar data. Then, by further introducing one or four key statistical features ("+F[1]" and "+F[1–4]"), we observe further reductions in false negatives. Note that *pulse width* is the only key feature to use for "+F[1]", which shows better performance than when all the four features in F are used, i.e., "+F[1–4]". Moreover, we find that incorporating empirical rules on pulsar features ("+R") into the prompt can further improve the recall score. This result implies that domain knowledge can be effectively leveraged to improve the performance of pulsar candidate classification. Finally, we can see that combining three plots into a single image for input ("C[1–3]") achieves the highest recall score at the expense of a lower precision score. In the experiments, we opt for "M[1–3]+F[1]+R" by default, as it achieves a balance between precision and recall scores, with the emphasis of avoiding missing real pulsars.

Missed Pulsars Analysis Despite the impressive performance of STARWHISPER-PULSAR, it still misses several true pulsars in the test sets. Specifically, it misidentifies three pulsars from the HTRU



(a) HTRU



(b) FAST

Figure 4: Diagnostic plots of missed pulsars from the HTRU and FAST test sets.

test set and two from the FAST test set as non-pulsar noise. The diagnostic plots presented in Figure 4 provide a thorough insight into the characteristics of these missed pulsars. In fact, all these diagnostic plots for missed pulsars still exhibit clear "pulsar-like" patterns, whereas containing various types of misleading information:

- **Signal Intensity Fluctuation:** The intensity of the signal varies over time, and in certain subplots it seems to disappear intermittently (e.g., the time-phase plot in Figure 4b). This behavior might be related to the rotation of the beam pattern during observations.
- **Presence of RFI:** In the case of missing pulsars, RFI is most evident in the time- and frequency-phase plots. Intense and continuous RFI, like the time-phase plots, nearly masks the pulsar signal. Periodic interference creates diagonal lines in the time-phase plot, and these diagonal lines are indicative of RFI with zero DM. Such misleading patterns can hinder the ability of STARWHISPER-PULSAR to detect pulsar signals.

5 Conclusion

In this paper, we pioneer the application of MLLMs in the pulsed signal processing domain. We propose STARWHISPER-PULSAR, which fine-tunes pre-trained MLLMs using labeled pulsed signal data in visual, textual, and numerical modalities for pulsar candidate classification. Through experiments on two real data sets, we demonstrate that STARWHISPER-PULSAR significantly outperforms several state-of-the-art deep learning methods for this problem. It can also achieve good performance in a few training epochs, exhibiting fast convergence, high efficiency, and a good few-shot learning capability. In future work, we will explore how to extend our method to handle other signal processing tasks in astronomy, including fast radio burst identification and gravitational wave detection. Our ultimate goal is to build a general-purpose MLLM-based AI agent for data-driven astronomical research.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (Grant Numbers 62162010, 11988101, 11973054, and 11933004).

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015.

[2] S. D. Bates, M. Bailes, B. R. Barsdell, N. D. R. Bhat, M. Burgay, S. Burke-Spolaor, D. J. Champion, P. Coster, N. D'Amico, A. Jameson, S. Johnston, M. J. Keith, M. Kramer, L. Levin, A. Lyne, S. Milia, C. Ng, C. Nietner, A. Possenti, B. Stappers, D. Thornton, and W. van Straten. The high time resolution universe pulsar survey – VI. an artificial neural network and timing of 75 pulsars. *Monthly Notices of the Royal Astronomical Society*, 427(2):1052–1065, 2012.

[3] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198, 2024.

[4] R. P. Eatough, N. Molkenthin, M. Kramer, A. Noutsos, M. J. Keith, B. W. Stappers, and A. G. Lyne. Selection of radio pulsar candidates using artificial neural networks. *Monthly Notices of the Royal Astronomical Society*, 407(4):2443–2450, 2010.

[5] A. J. Faulkner, I. H. Stairs, M. Kramer, A. G. Lyne, G. Hobbs, A. Possenti, D. R. Lorimer, R. N. Manchester, M. A. McLaughlin, N. D'Amico, F. Camilo, and M. Burgay. The Parkes multibeam pulsar survey – V. finding binary and millisecond pulsars. *Monthly Notices of the Royal Astronomical Society*, 355(1):147–158, 2004.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[7] A. Hewish, S. J. Bell, J. D. H. Pilkington, P. F. Scott, and R. A. Collins. Observation of a rapidly pulsating radio source. *Nature*, 217:709–713, 1968.

[8] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations (ICLR)*, 2022.

[9] R. A. Hulse and J. H. Taylor. Discovery of a pulsar in a binary system. *The Astrophysical Journal*, 195:L51–L53, 1975.

[10] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv:2001.08361*, 2020.

[11] M. J. Keith, A. Jameson, W. van Straten, M. Bailes, S. Johnston, M. Kramer, A. Possenti, S. D. Bates, N. D. R. Bhat, M. Burgay, S. Burke-Spolaor, N. D'Amico, L. Levin, Peter L. McMahon, S. Milia, and B. W. Stappers. The high time resolution universe pulsar survey – I. system configuration and initial discoveries. *Monthly Notices of the Royal Astronomical Society*, 409 (2):619–627, 2010.

[12] Yichao Li, Yougang Wang, Furen Deng, Wenxiu Yang, Wenkai Hu, Diyang Liu, Xinyang Zhao, Shifan Zuo, Shuanghao Shu, Jixia Li, Peter Timbie, Reza Ansari, Olivier Perdereau, Albert Stebbins, Laura Wolz, Fengquan Wu, Xin Zhang, and Xuelei Chen. FAST drift scan survey for hi intensity mapping: I. preliminary data analysis. *The Astrophysical Journal*, 954(2):139, 2023.

[13] Yu-Yang Li, Yu Bai, Cunshi Wang, Mengwei Qu, Ziteng Lu, Roberto Soria, and Jifeng Liu. Deep learning and LLM-based methods applied to stellar lightcurve classification. *arXiv:2404.10757*, 2024.

[14] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36*, pages 34892–34916, 2023.

[15] Yi Liu, Jing Jin, and Hongyang Zhao. Deep learning-based pulsar candidate identification model using a variational autoencoder. *New Astronomy*, 106:102125, 2024.

[16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *The Seventh International Conference on Learning Representations (ICLR)*, 2019.

[17] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. DeepSeek-VL: Towards real-world vision-language understanding. *arXiv:2403.05525*, 2024.

[18] Robert J. Lyon, B. W. Stappers, Sally Cooper, John Martin Brooke, and Joshua D. Knowles. Fifty years of pulsar candidate selection: from simple filters to a new principled real-time classification approach. *Monthly Notices of the Royal Astronomical Society*, 459(1):1104–1123, 2016.

[19] V. Morello, E. D. Barr, M. Bailes, C. M. Flynn, E. F. Keane, and W. van Straten. SPINN: a straightforward machine learning solution to the pulsar candidate selection problem. *Monthly Notices of the Royal Astronomical Society*, 443(2):1651–1662, 2014.

[20] Tuan Dung Nguyen, Yuan-Sen Ting, Ioana Ciuca, Charles O'Neill, Ze-Chang Sun, Maja Jabłońska, Sandor Kruk, Ernest Perkowski, Jack Miller, Jason Jingshi Li, Josh Peek, Kartheik Iyer, Tomasz Rozanski, Pranav Khetarpal, Sharaf Zaman, David Brodrick, Sergio J. Rodriguez Mendez, Thang Bui, Alyssa Goodman, Alberto Accomazzi, Jill Naiman, Jesse Cranney, Kevin Schawinski, and Roberta Raileanu. AstroLLaMA: Towards specialized foundation models in astronomy. In *Proceedings of the Second Workshop on Information Extraction from Scientific Publications*, pages 49–55, 2023.

[21] Liam Parker, Francois Lanusse, Siavash Golkar, Leopoldo Sarra, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Geraud Krawezik, Michael McCabe, Rudy Morel, Ruben Ohana, Mariel Pettee, Bruno Régaldo-Saint Blancard, Kyunghyun Cho, Shirley Ho, and The Polymathic AI Collaboration. AstroCLIP: A cross-modal foundation model for galaxies. *Monthly Notices of the Royal Astronomical Society*, 531(4):4990–5011, 2024.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763, 2021.

[23] Scott Ransom. PRESTO: PulsaR Exploration and Search TOolkit. `https://ascl.net/1107.017`, 2011. Accessed: 2024-09-08.

[24] C. M. Tan, R. J. Lyon, B. W. Stappers, S. Cooper, J. W. T. Hessels, V. I. Kondratiev, D. Michilli, and S. Sanidas. Ensemble candidate classification for the LOTAAS pulsar survey. *Monthly Notices of the Royal Astronomical Society*, 474(4):4571–4583, 2018.

[25] Irfan Tariq, Qiao Meng, Shunyu Yao, Wei Liu, Chenye Zhou, Adnan Ahmed, and Apostolos Spanakis-Misirlis. Adaboost-DSNN: an adaptive boosting algorithm based on deep self normalized neural network for pulsar identification. *Monthly Notices of the Royal Astronomical Society*, 511(1):683–690, 2022.

[26] Hongfeng Wang, Weiwei Zhu, Ping Guo, Di Li, Sibo Feng, Qian Yin, Chenchen Miao, Zhenzhao Tao, Zhichen Pan, Pei Wang, Xin Zheng, Xiaodan Deng Zhijie Liu, Xiaoyao Xie, Xuhong Yu, Shanping You, and Hui Zhang. Pulsar candidate selection using ensemble networks for FAST drift-scan survey. *Science China Physics, Mechanics & Astronomy*, 62:959507, 2019.

[27] Q. Zeng, X. Li, and H. Lin. Concat convolutional neural network for pulsar candidate selection. *Monthly Notices of the Royal Astronomical Society*, 494(3):3110–3119, 2020.

[28] W. W. Zhu, A. Berndsen, E. C. Madsen, M. Tan, I. H. Stairs, A. Brazier, P. Lazarus, R. Lynch, P. Scholz, K. Stovall, S. M. Ransom, S. Banaszak, C. M. Biwer, S. Cohen, L. P. Dartez, J. Flanigan, G. Lunsford, J. G. Martinez, A. Mata, M. Rohr, A. Walker, B. Allen, N. D. R. Bhat, S. Bogdanov, F. Camilo, S. Chatterjee, J. M. Cordes, F. Crawford, J. S. Deneva, G. Desvignes, R. D. Ferdman, P. C. C. Freire, J. W. T. Hessels, F. A. Jenet, D. L. Kaplan, V. M. Kaspi, B. Knispel, K. J. Lee, J. van Leeuwen, A. G. Lyne, M. A. McLaughlin, X. Siemens, L. G. Spitler, and A. Venkataraman. Searching for pulsars using image pattern recognition. *The Astrophysical Journal*, 781(2):117, 2014.