International Science and Technology Journal
المجلة الدولية للعلوم والتقنية
Volume 39 العدد
Part 2 المجلد
المجلة الدولية للعلوم والتقنية
International Science and Technology Journal
ISTJ

Optimizing Pulsar Star Prediction in Large-Scale Astronomical Surveys Using Hybrid Resampling and Automated Machine Learning

http://www.doi.org/10.62341/istj-vol39-2-fi14

Received	**2026/09/16**	**تم استلام الورقة العلمية في**
Accepted	**2026/10/11**	**تم قبول الورقة العلمية في**
Published	**2026/10/12**	**تم نشر الورقة العلمية في**

Optimizing Pulsar Star Prediction in Large-Scale Astronomical Surveys Using Hybrid Resampling and Automated Machine Learning

[1]Fatema. M Al-Asbli[1], Mohamed A. E. Abdalla[2], Tarek A M Nagem[3] Issa. M Al-magrhe[4]

[1,4] Higher Institute of Engineering Technologies, Al-Majouri Benghazi - Libya
[2,3] Faculty of Information Technology, University of Benghazi, Benghazi, Libya.
[1]Fatma-elasbli@mei.edu.ly, [2]mohamed.abdalla@uob.edu.ly
[3]Tarek.Nagem@uob.edu.ly [4]Issa.milad@mei.edu.ly

[2]https://orcid.org/0009-0002-1493-1095
[3]https://orcid.org/0000-0002-1347-4323
[4]https://orcid.org/0009-0009-9994-4231

Abstract

Automated pulsar candidate classification is challenging due to severe class imbalance, noise, and Radio Frequency Interference (RFI). This study presents a leakage-free comparative machine learning framework using the HTRU2 dataset of 17,898 instances, including 1,639 pulsars and 16,259 non-pulsars represented by eight statistical features. Data were split before preprocessing, with scaling and hybrid resampling restricted to the training set. The framework evaluates classical and ensemble models, XGBoost, MLP, hybrid undersampling–SMOTE, and H2O AutoML using comprehensive classification metrics, supported by 10-fold cross-validation, bootstrap confidence intervals, and McNemar's tests.

On the untouched test set, H2O AutoML achieved 98.12% accuracy, 89.22% F1-score, 84.96% recall, 99.45% specificity, 97.71% ROC-AUC, and MCC = 0.8833, while XGBoost achieved 98.06% accuracy and F1 = 0.8901. Hybrid resampling increased XGBoost recall from 85.57% to 91.87% and reduced FNR from 14.43% to

Optimizing Pulsar Star Prediction in Large-Scale Astronomical Surveys Using Hybrid Resampling and Automated Machine Learning
http://www.doi.org/10.62341/istj-vol39-2-fi14

8.13%, but decreased accuracy to 96.42%, demonstrating a sensitivity–precision trade-off rather than universal improvement. Overall, AutoML and hybrid resampling provide complementary strategies for imbalanced pulsar classification. Since no independent astronomical dataset was used, the study makes no claim of cross-survey generalization or deployment readiness.

Keywords: Pulsar candidate classification, HTRU2, class imbalance, SMOTE, hybrid resampling, XGBoost, ensemble learning, H2O AutoML, leakage-free evaluation, statistical evaluation.

تحسين التنبؤ بالنجوم النابضة في المسوحات الفلكية واسعة النطاق باستخدام إعادة العينات الهجينة والتعلم الآلي المؤتمت

فاطمة العسبلي[1]، محمد المنفي[2]، طارق نجم[3]، عيسى المقرحي[4]
[4-1]المعهد العالي للتقنيات الهندسية الماجوري – بنغازي - ليبيا
[3-2]كلية تقنية المعلومات، جامعة بنغازي، بنغازي - ليبيا.

Fatma-elasbli@mei.edu.ly, [2]mohamed.abdalla@uob.edu.ly[1]
[3]Tarek.Nagem@uob.edu.ly [4]Issa.milad@mei.edu.ly
[1]Fatma-elasbli@mei.edu.ly, [2]Issa.milad@mei.edu.ly

الملخص

تُعدّ عملية التصنيف الآلي للمرشحين للنجوم النابضة تحديًا بسبب الاختلال الشديد في توازن الفئات، والضوضاء، والتداخلات الراديوية (RFI). تقدم هذه الدراسة إطارًا مقارنًا للتعلم الآلي خاليًا من تسرب البيانات باستخدام مجموعة بيانات HTRU2 التي تضم 17,898 عينة، منها 1,639 نجمًا نابضًا و16,259 مرشحًا غير نابض، ممثلة بثماني خصائص إحصائية. قُسمت البيانات قبل إجراء المعالجة المسبقة، مع اقتصار التطبيع وإعادة موازنة الفئات الهجينة على بيانات التدريب فقط. ويقارن الإطار بين نماذج التعلم الآلي التقليدية والتجميعية، وXGBoost، وشبكاتMLP ، وإعادة الموازنة الهجينة باستخدام Under–samplingوSMOTE، وH2O AutoML، بالاعتماد على

Optimizing Pulsar Star Prediction in Large-Scale Astronomical Surveys Using Hybrid Resampling and Automated Machine Learning

http://www.doi.org/10.62341/istj-vol39-2-fi14

مجموعة شاملة من مقاييس التصنيف، مدعومة بالتحقق المتقاطع 10-fold، وفواصل الثقةBootstrap ، واختبار McNemar.

على مجموعة الاختبار غير المتأثرة، حقق H2O AutoMLدقة بلغت 98.12%، و F1-score قدره 89.22%، و Recall بلغ 84.96%، وSpecificity بلغت 99.45%، و ROC-AUCبلغ 97.71%، ومعامل MCC قدره 0.8833، بينما حقق XGBoost دقة 98.06%وF1-scoreقدره 0.8901. أدت إعادة الموازنة الهجينة إلى رفع معدل الاستدعاءRecall لـ XGBoost من 85.57% إلى 91.87 %وخفض معدل السلبية الكاذبة FNR من 14.43%إلى 8.13%، لكنها خفضت الدقة إلى 96.42%، مما يوضح وجود مفاضلة بين الحساسية والدقة بدلًا من تحقيق تحسن شامل. وبوجه عام، توفر AutoML وإعادة الموازنة الهجينة استراتيجيتين متكاملتين لمعالجة مشكلة عدم توازن الفئات في تصنيف المرشحين للنجوم النابضة. ونظرًا لعدم استخدام مجموعة بيانات فلكية مستقلة للتحقق، لا تدّعي الدراسة تحقيق تعميم عبر المسوحات الفلكية المختلفة أو الجاهزية للنشر.

الكلمات المفتاحية: تصنيف مرشحي النجوم النابضة،HTRU2 ، عدم توازن الفئات، SMOTE، إعادة الموازنة الهجينة، XGBoost ، التعلم التجميعي؛H2O AutoML ، التقييم الخالي من تسرب البيانات، التقييم الإحصائي.

1. Introduction

Radio pulsar surveys generate large numbers of candidate detections that must be distinguished between genuine pulsar signals and non-pulsar candidates caused by noise, radio-frequency interference (RFI), and other spurious effects. With the increasing sensitivity and processing capacity of modern radio telescopes, manual inspection of all candidates has become impractical, making automated candidate classification an important component of pulsar-search pipelines [1]–[5].

The High Time Resolution Universe 2 (HTRU2) dataset is a widely used benchmark for pulsar candidate classification. It contains 17,898 observations, including 1,639 confirmed pulsars and 16,259 non-pulsars, represented by eight statistical features derived from integrated pulse profiles and dispersion-measure signal-to-noise-

http://www.doi.org/10.62341/istj-vol39-2-fi14

ratio (DM-SNR) curves [4], [5]. The strong class imbalance makes accuracy alone insufficient for evaluating minority-class detection; therefore, complementary measures such as Recall, F1-score, Balanced Accuracy, Specificity, FNR, MCC, ROC-AUC, and PR-AUC are required [6], [7].

Resampling methods have been widely investigated to address this imbalance. SMOTE generates synthetic minority observations from neighboring samples [6], while hybrid over- and under-sampling strategies have also been studied for pulsar classification. In particular, Lee et al. [8] evaluated hybrid resampling on HTRU2 and reported approximately 99.3% accuracy using Extra Trees. Consequently, hybrid resampling and SMOTE are not claimed as novel algorithms in the present study. Instead, their effects are examined under a controlled leakage-free protocol, where the train/test partition is performed before preprocessing, scaling is fitted using training data only, and resampling is restricted to training data and cross-validation folds.

Pulsar candidate classification has also expanded toward deep learning, generative learning, and semi-supervised approaches. CNN- and DNN-based methods have explored diagnostic plots and image-like representations [9], [10], while representation-learning approaches have investigated residual convolutional architectures [11]. GAN-based methods have addressed class imbalance through synthetic candidates [12], and semi-supervised approaches have exploited unlabeled candidates [13], [14]. Ensemble methods, including Random Forest, XGBoost, and hybrid tree-based strategies, have likewise demonstrated competitive performance under imbalanced conditions [15], [16]. However, these approaches often employ different datasets, representations, and experimental protocols, limiting direct numerical comparison with HTRU2-based tabular models.

AutoML provides another approach by automating model exploration, hyperparameter configuration, and ensemble construction. H2O AutoML trains multiple supervised learning algorithms and can construct stacked ensembles, providing an automated model-selection strategy rather than a single manually specified classifier [17]. In this study, AutoML is therefore

http://www.doi.org/10.62341/istj-vol39-2-fi14

evaluated independently alongside conventional models and is not treated as either a resampling method or a replacement for supervised classifiers.

Despite extensive previous work, direct comparison among pulsar-classification studies remains difficult because of differences in datasets, feature representations, class distributions, preprocessing, resampling, and evaluation protocols. Accordingly, the methodological contribution of this study is not the introduction of a new classifier or resampling algorithm, but a controlled and leakage-free comparison of conventional machine learning, ensemble models, XGBoost, MLP, hybrid undersampling with SMOTE, and H2O AutoML using the same HTRU2 representation and independent test partition. The evaluation is further supported by 10-fold cross-validation, bootstrap confidence intervals, and McNemar's tests to assess stability and paired differences in classification behavior.

Accordingly, the study addresses four questions concerning (1) the effect of leakage-free hybrid resampling on minority-class detection, (2) model trade-offs across Recall, F1-score, specificity, and ranking metrics, (3) the contribution of H2O AutoML relative to manually specified classifiers, and (4) the extent of generalization beyond HTRU2. The fourth question remains an explicit limitation because no independent external astronomical dataset was used.

Overall, the study provides a controlled, leakage-free comparative benchmark for HTRU2, examining model family, class balancing, and automated model selection without claiming universal classifier superiority or cross-survey generalization.

2. RELATED WORK

2.1. Conventional Machine Learning and HTRU2

The application of machine learning to pulsar candidate selection predates modern deep-learning approaches. Eatough et al. [1] demonstrated the feasibility of artificial neural networks for automated candidate selection, followed by Bates et al. [2], who integrated neural-network classification into the HTRU survey pipeline to reduce manual inspection. Morello et al. [3] introduced SPINN for ranking HTRU candidates, while Lyon et al. [4]

http://www.doi.org/10.62341/istj-vol39-2-fi14

investigated historical candidate-selection methods and proposed a tree-based real-time classifier with engineered candidate features.
Feature representation remains important for computationally efficient classification. The HTRU2 benchmark provides eight statistical descriptors derived from the integrated pulse profile and DM-SNR curve, while Lin et al. [18] showed that effective classification can sometimes be achieved using a reduced subset of candidate features. These findings support the continued use of compact statistical representations in conventional machine-learning models.

2.2. Ensemble Learning and HTRU2 Classification

Tree-based ensembles are well suited to structured candidate features because they can model nonlinear relationships without requiring an explicit functional form. Random Forest provides an ensemble of randomized decision trees [19], while XGBoost uses gradient-boosted trees and has demonstrated strong performance on structured data [20]. In pulsar classification, Wang et al. [15] evaluated Random Forest, XGBoost, and hybrid ensemble strategies on HTRU datasets, while Rustam et al. [16] investigated tree-based boosting and voting methods under class imbalance. These studies demonstrate the effectiveness of ensemble learning while also highlighting the influence of the balancing and evaluation protocols used.

2.3. Class Imbalance and Resampling

Class imbalance is a central challenge in HTRU2 because pulsars represent only a small fraction of the observations. Imbalanced-learning research shows that conventional classifiers may become biased toward the majority class [7]. SMOTE addresses this problem by generating synthetic minority observations from neighboring samples [6].
For pulsar classification, Lee et al. [8] evaluated SMOTE, ADASYN, Cluster Centroids, and a concatenated hybrid resampling strategy on HTRU2, reporting approximately 99.3% accuracy with Extra Trees under their hybrid configuration. Other studies have combined resampling with ensemble methods or generative approaches [12], [13], demonstrating that class

http://www.doi.org/10.62341/istj-vol39-2-fi14

distribution can substantially influence classifier behavior. Importantly, hybrid resampling on HTRU2 is therefore an established approach rather than a novel algorithmic contribution of the present study.

2.4. Deep Learning and Representation Learning

Deep-learning methods have extended pulsar classification beyond compact statistical features by learning representations from diagnostic plots and image-like inputs. Wang et al. [9] developed an 11-layer CNN using sub-integration and sub-band plots, while Yin et al. [10] investigated deep neural networks for image-based candidate recognition. Yin et al. [11] further proposed a residual convolutional autoencoder combined with logistic regression using HTRU Medlat and PMPS-26k.

Generative and semi-supervised approaches have also been explored. Guo et al. [12] investigated GAN-based candidate classification, while Balakrishnan et al. [13] incorporated unlabeled candidates through a semi-supervised GAN. Liu et al. [14] later investigated self-tuning pseudo-labeling, and transformer-based approaches have extended representation learning using HTRU and CRAFTS data [21]. These studies demonstrate the growing use of learned and richer candidate representations; however, their results are not directly comparable with the present work because of differences in datasets, input modalities, preprocessing, class distributions, and evaluation protocols.

2.5. Semi-Supervised Learning and AutoML

Semi-supervised methods exploit unlabeled candidates to reduce dependence on manually labeled data. Balakrishnan et al. [13] and Liu et al. [14] demonstrated the potential of GAN-based and pseudo-labeling approaches, respectively. The present study remains fully supervised to maintain a controlled HTRU2 benchmark and therefore uses these methods as literature context rather than direct experimental competitors.

AutoML provides a complementary approach by automating model exploration, hyperparameter optimization, and ensemble construction. H2O AutoML trains multiple candidate models and can construct stacked ensembles [17]. In this study, AutoML is

treated as an independent model-selection strategy, not as a resampling method or inherently superior classifier; its contribution is evaluated empirically using the same HTRU2 training and independent test partitions.

2.6. Research Gap and Position of the Present Study

The literature demonstrates a progression from conventional feature-based classifiers to ensemble learning, resampling, deep learning, generative and semi-supervised methods, and automated model selection. However, direct comparisons remain difficult because studies differ in datasets, representations, preprocessing, balancing strategies, and evaluation protocols.

In particular, Lee et al. [8] established the effectiveness of hybrid resampling on HTRU2. Therefore, the present study does not claim novelty for SMOTE or hybrid resampling itself. Instead, it evaluates their effect under a leakage-free protocol, with training/test separation performed before learned preprocessing and resampling, and examines their impact across multiple metrics, particularly minority-class Recall and FNR.

The study further compares conventional and ensemble classifiers, XGBoost, an MLP baseline, hybrid undersampling with SMOTE, and H2O AutoML using the same eight-feature HTRU2 representation and independent test partition. Ten-fold cross-validation, bootstrap confidence intervals, and McNemar's tests are incorporated to assess stability and paired differences in prediction behavior [22]. Thus, the contribution is best characterized as a leakage-free and statistically supported comparative evaluation framework for HTRU2, rather than the introduction of a fundamentally new pulsar-classification algorithm.

3. PROPOSED METHODOLOGY

The proposed methodology fig (1) is designed as a leakage-free comparative framework for pulsar candidate classification under severe class imbalance. The overall workflow consists of dataset characterization, stratified hold-out partitioning, training-only preprocessing, conventional and ensemble classification, hybrid resampling, automated model selection, and statistical evaluation. A central design principle is that the independent test set is isolated

Optimizing Pulsar Star Prediction in Large-Scale Astronomical Surveys Using Hybrid Resampling and Automated Machine Learning
http://www.doi.org/10.62341/istj-vol39-2-fi14

before any learned preprocessing or resampling operation, ensuring that information from the final evaluation data cannot influence model development.

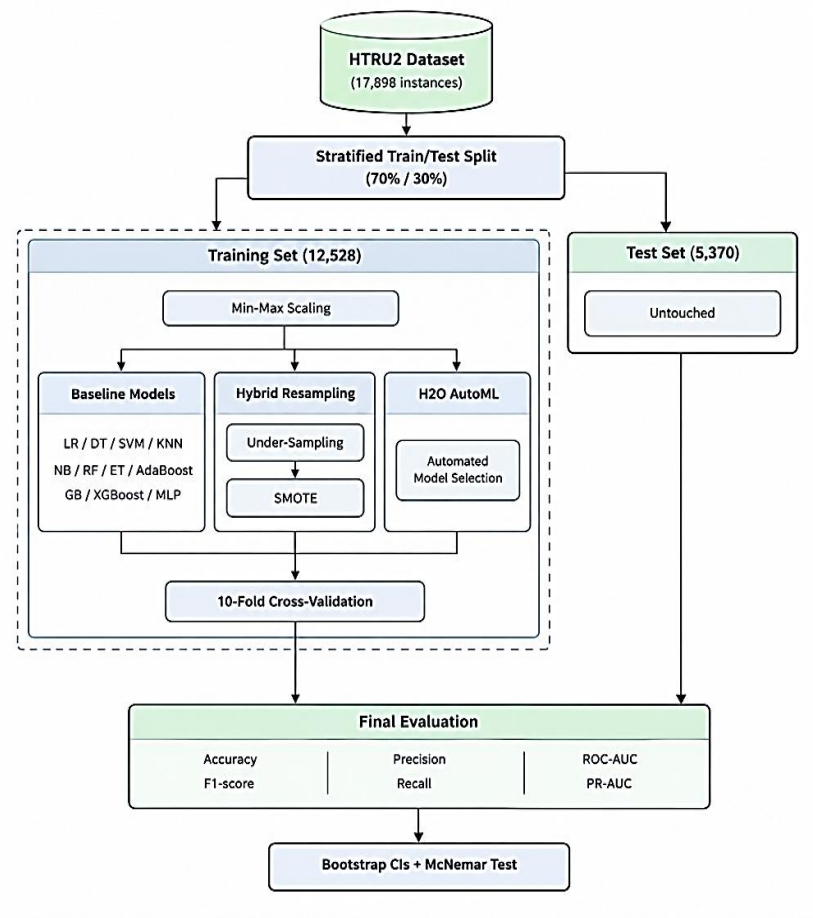
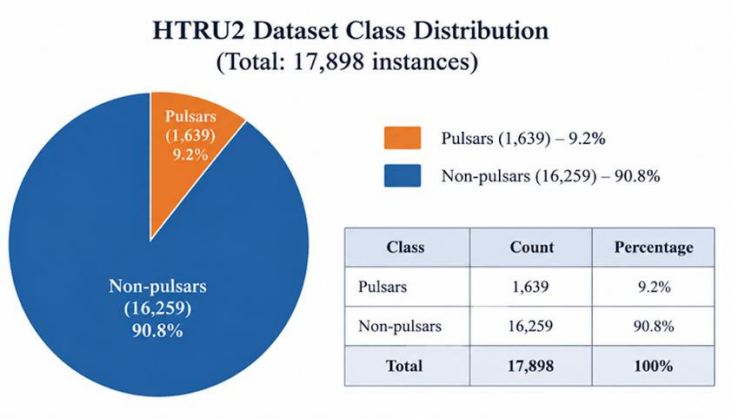


Fig. 1. :Overall leakage-free comparative framework.

3.1. Dataset Characterization

The HTRU2 dataset contains 17,898 candidate observations, including 1,639 confirmed pulsars and 16,259 non-pulsar candidates. This corresponds to a strongly imbalanced class distribution, with pulsars representing approximately 9.16% of the dataset and non-pulsars approximately 90.84%. The class distribution is illustrated in Fig.2.

Optimizing Pulsar Star Prediction in Large-Scale Astronomical Surveys Using Hybrid Resampling and Automated Machine Learning
http://www.doi.org/10.62341/istj-vol39-2-fi14



Class	Count	Percentage
Pulsars	1,639	9.2%
Non-pulsars	16,259	90.8%
Total	**17,898**	**100%**

Fig. 2. Class distribution of the HTRU2 dataset, showing the pronounced imbalance between pulsar and non-pulsar candidates.

3.2. Leakage-Free Data Partitioning

To prevent information leakage, the complete HTRU2 dataset is first divided into training and independent test partitions using stratified sampling. A fixed random seed of 42 is used to ensure reproducibility. Seventy percent of the observations are assigned to the training partition and the remaining 30% are retained as an untouched test set.

The resulting training partition contains 12,528 observations, including 11,381 non-pulsars and 1,147 pulsars, whereas the independent test partition contains 5,370 observations, including 4,878 non-pulsars and 492 pulsars. The test set is not used for feature scaling, resampling, model selection, hyperparameter optimization, or threshold adjustment. It is accessed only after the corresponding models have been trained to obtain the final performance estimates.

This ordering is essential because scaling parameters and synthetic samples are data-dependent. Applying these operations before the train-test split could allow information from observations reserved for final evaluation to influence model development and produce optimistic performance estimates.

Optimizing Pulsar Star Prediction in Large-Scale Astronomical Surveys Using Hybrid Resampling and Automated Machine Learning
http://www.doi.org/10.62341/istj-vol39-2-fi14

3.3. Training-Only Preprocessing

Feature scaling is performed using Min-Max normalization. Importantly, the scaler is fitted exclusively on the training data and subsequently applied to the training and independent test partitions using the parameters learned from the training partition.
For a feature x, the normalized value is calculated as

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \qquad (1)$$

where $x_{\min}$and $x_{\max}$are estimated from the training data only. No test observation contributes to these parameters.
The same principle is applied during cross-validation. Scaling is incorporated into the model pipeline so that each validation fold is transformed using parameters learned exclusively from its corresponding training folds. This prevents information from validation observations from entering the learning process.

3.4. Baseline Classification Models

A broad set of supervised learning algorithms is evaluated to establish a comprehensive baseline. The evaluated models include Logistic Regression, Decision Tree, Gaussian Naive Bayes, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, Extra Trees, AdaBoost, Gradient Boosting, XGBoost, and a Multilayer Perceptron (MLP).
These models represent different learning paradigms, including linear classification, probabilistic learning, distance-based classification, kernel-based learning, bagging ensembles, boosting ensembles, gradient-boosted trees, and neural networks. Evaluating multiple model families under the same data partition and preprocessing protocol allows differences in performance to be attributed more reliably to the learning strategy rather than to inconsistent experimental conditions.
XGBoost is included as a modern gradient-boosting baseline, while the MLP provides a same-feature neural-network comparison. The latter is particularly useful for addressing the limitation that several recent pulsar-classification studies employ deep neural networks with richer image-based diagnostic representations [8]–[10].

http://www.doi.org/10.62341/istj-vol39-2-fi14

Because the present study uses the eight-feature HTRU2 representation, image-based deep-learning results from the literature are treated as contextual evidence rather than direct head-to-head comparisons.

3.5. Hybrid Resampling Strategy

To investigate class-imbalance mitigation, a hybrid strategy combining Random Under-Sampling (RUS) and SMOTE is evaluated. RUS first reduces the majority class using a sampling strategy of 0.5, followed by SMOTE with a sampling strategy of 1.0 to generate synthetic minority observations through interpolation between minority-class neighbors [6].
The complete operation can be represented as

$$X_{\text{train}} \rightarrow \textbf{MinMaxScaler} \rightarrow \textbf{RandomUnderSampler} \rightarrow \textbf{SMOTE} \rightarrow \textbf{Classifier}. \quad (2)$$

All resampling operations are applied exclusively to the training data. During cross-validation, both RUS and SMOTE are embedded within the training folds, preventing synthetic observations from influencing the corresponding validation folds. The original test distribution remains untouched.
The hybrid strategy therefore enables evaluation of the trade-off between reducing majority-class dominance and increasing minority-class representation, while acknowledging the loss of some majority observations through undersamplingSince this approach has previously been investigated on HTRU2 by Lee et al. [8], it is treated here as an established technique rather than a novel algorithm.
The hybrid resampling procedure substantially changes the effective training distribution. The original training partition contains 12,528 observations, comprising 11,381 non-pulsar and 1,147 pulsar candidates. The majority class was first reduced through random undersampling to 1,638 observations, after which SMOTE was applied to increase the minority class from 1,147 to 1,638 synthetic-balanced observations. The resulting hybrid training set therefore contained 3,276 observations, with equal representation of the two classes. Consequently, 9,743 original non-pulsar observations were excluded by undersampling,

http://www.doi.org/10.62341/istj-vol39-2-fi14

corresponding to approximately 85.6% of the original majority-class training observations, while 491 synthetic minority observations were generated by SMOTE. Importantly, these operations were applied exclusively to the training data; the independent test set was left completely untouched. This design allows the effect of class balancing to be evaluated without altering the final test distribution.

3.6. Cross-Validation and Model Evaluation

Model stability is assessed using stratified 10-fold cross-validation. Stratification preserves the approximate class distribution across the folds, which is particularly important for the highly imbalanced HTRU2 dataset.

For each fold, preprocessing, resampling, and model fitting are performed using only the corresponding training portion. Predictions are then generated for the validation portion. This procedure is repeated across all ten folds, and the mean and standard deviation of the principal evaluation metrics are reported.

The primary evaluation measures are Accuracy, Balanced Accuracy, Precision, Recall, F1-score, Specificity, False Positive Rate (FPR), False Negative Rate (FNR), Matthews Correlation Coefficient (MCC), ROC-AUC, and Precision-Recall AUC (PR-AUC). Mean Squared Error (MSE) is retained only as a supplementary measure and is not used as the principal criterion for ranking binary classifiers.

For binary classification, Recall and FNR are particularly important because they quantify the proportion of genuine pulsars detected and missed, respectively. Specificity and FPR characterize the ability to reject non-pulsar candidates, while F1-score provides a balance between Precision and Recall. MCC is additionally reported because it provides a correlation-based summary of binary classification performance that remains informative under class imbalance.

3.7. H2O AutoML Model Selection

H2O AutoML is incorporated as a separate automated model-selection branch rather than as part of the hybrid-resampling pipeline. This distinction is intentional: resampling modifies the

training class distribution, whereas AutoML automates model search and selection.

The AutoML experiment is trained on the original imbalanced training partition containing 12,528 observations. The independent test partition remains completely isolated. H2O AutoML is configured with a maximum of 30 models, a fixed random seed of 42, and five-fold internal cross-validation. Deep-learning models are excluded from the AutoML search to maintain a controlled comparison with the tabular machine-learning setting and to avoid conflating the separate MLP baseline with the AutoML search process.

The AutoML system explores multiple model families and ensemble configurations and produces a leaderboard based on the selected optimization criterion. The final selected model is then evaluated once on the untouched test set.

The role of AutoML in this study is therefore methodological rather than inherently performance-enhancing. A manually specified classifier may still obtain a higher value for an individual predictive metric. Accordingly, AutoML is evaluated according to its ability to automate model selection and configuration under a reproducible protocol rather than under an assumption that automation must produce the highest test accuracy.

3.8. Statistical Uncertainty and Paired Comparison

Point estimates alone may not adequately characterize the reliability of classification results. Therefore, bootstrap resampling is used to estimate 95% confidence intervals for selected performance measures, including F1-score, Recall, and ROC-AUC.

In addition, McNemar's test is applied when two classifiers generate paired predictions for the same independent test observations. The test focuses on the number of observations correctly classified by one model but incorrectly classified by the other. This provides a paired assessment of whether two classifiers exhibit statistically different prediction patterns on the same test set.

Importantly, statistical significance is interpreted as evidence of different prediction behavior on the evaluated test set rather than as proof that one algorithm is universally superior. This distinction is

http://www.doi.org/10.62341/istj-vol39-2-fi14

particularly important when comparing conventional models with hybrid-resampling configurations.

3.9. Independent External Validation

The methodology includes an external-validation stage to address the limitation of relying on a single benchmark dataset. An independent astronomical dataset would be required to evaluate cross-dataset generalization. HTRU Medlat is identified as a potential external benchmark; however, its diagnostic PHCX representation is not directly equivalent to the eight statistical HTRU2 features.

Consequently, external validation is not inferred from the HTRU2 test results. A valid external experiment would require documented feature extraction or a scientifically justified representation mapping, strict separation from the HTRU2 development data, and an independently defined test set. Because such an external dataset was not supplied for the current experiments, no claim of independent astronomical generalization is made.

4. EXPERIMENTAL RESULTS AND DISCUSSION

This section presents the experimental results obtained under the revised leakage-free evaluation protocol. The independent test set contains 5,370 observations, including 4,878 non-pulsars and 492 pulsars, and was kept completely isolated from preprocessing, resampling, and model selection. The analysis focuses on the principal classification results, the effect of hybrid resampling, the contribution of H2O AutoML, and the statistical reliability of the observed differences.

4.1. Overall Test-Set Performance

The performance of the selected baseline models and the H2O AutoML model was evaluated on the untouched 30% test partition. As shown in Table I, the models achieved high overall classification performance, although differences were observed in their ability to identify the minority pulsar class.

Optimizing Pulsar Star Prediction in Large-Scale Astronomical Surveys Using Hybrid Resampling and Automated Machine Learning
http://www.doi.org/10.62341/istj-vol39-2-fi14

Table I. Main Test-Set Performance of Selected Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score	ROC-AUC
XGBoost	98.06	92.73	85.57	0.8901	0.9759
Random Forest	97.99	92.48	84.96	0.8856	0.9713
Extra Trees	97.97	93.03	84.15	0.8837	0.9734
MLP	98.03	93.86	83.94	0.8863	0.9736
H2O AutoML	98.12	93.93	84.96	0.8922	0.9771

Among the evaluated baseline models, XGBoost achieved 98.06% accuracy, 92.73% precision, 85.57% recall, F1 = 0.8901, and ROC-AUC = 0.9759, while Random Forest and Extra Trees produced comparable performance with 97.99% and 97.97% accuracy and F1-scores of 0.8856 and 0.8837, respectively. The MLP also performed competitively, achieving 98.03% accuracy, 93.86% precision, 83.94% recall, and F1 = 0.8863, although its lower recall indicates that high overall accuracy does not necessarily imply better minority-class detection.

H2O AutoML achieved the highest test accuracy at 98.12%, with 93.93% precision, 84.96% recall, F1 = 0.8922, and ROC-AUC = 0.9771. Given the small differences from XGBoost, these results indicate comparable predictive performance rather than universal AutoML superiority. Overall, the results confirm that accuracy alone is insufficient under HTRU2 class imbalance; therefore, Recall, F1-score, and ROC-AUC are considered alongside accuracy in the subsequent hybrid-resampling analysis.

Fig 3 presents the Receiver Operating Characteristic (ROC) curves of the selected models on the independent HTRU2 test set. All curves remain substantially above the random-classifier reference, indicating strong discrimination between pulsar and non-pulsar candidates across a range of decision thresholds. The ROC-AUC values are closely grouped, ranging from 0.9713 for Random Forest to 0.9771 for H2O AutoML. This threshold-independent view complements the point-based performance measures reported in Table I by showing that the evaluated models maintain strong ranking capability across different operating thresholds.

Optimizing Pulsar Star Prediction in Large-Scale Astronomical Surveys Using Hybrid Resampling and Automated Machine Learning

http://www.doi.org/10.62341/istj-vol39-2-fi14

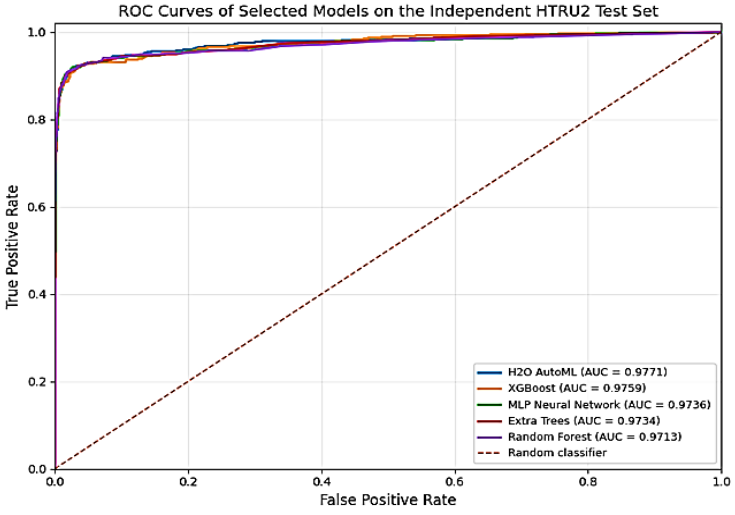


Fig. 3. ROC curves of selected models on the independent HTRU2 test set.

4.2. Effect of Hybrid Resampling

To examine the effect of class balancing on minority-class detection, a hybrid resampling strategy combining Random Under-Sampling and Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the training data. The independent test set was kept unchanged. The hybrid procedure was evaluated with XGBoost, Random Forest, and Extra Trees, and the corresponding results are presented in Table II.

Table II. Effect of Hybrid Resampling on Selected Ensemble Models

Model	Configuration	Accuracy (%)	Recall (%)	F1 score	FNR (%)	ROC-AUC
XGBoost	Baseline	98.06	85.57	0.8901	14.43	0.9759
XGBoost	Hybrid	96.42	91.87	0.8248	8.13	0.9741
Random Forest	Baseline	97.99	84.96	0.8856	15.04	0.9713
Random Forest	Hybrid	97.26	90.85	0.8588	9.15	0.9745
Extra Trees	Baseline	97.97	84.15	0.8837	15.85	0.9734
Extra Trees	Hybrid	97.36	90.45	0.8624	9.55	0.9752

Optimizing Pulsar Star Prediction in Large-Scale Astronomical Surveys Using Hybrid Resampling and Automated Machine Learning

http://www.doi.org/10.62341/istj-vol39-2-fi14

Hybrid resampling consistently improved minority-class detection. Recall increased from 85.57% to 91.87% for XGBoost, 84.96% to 90.85% for Random Forest, and 84.15% to 90.45% for Extra Trees, with corresponding reductions in FNR from 14.43%, 15.04%, and 15.85% to 8.13%, 9.15%, and 9.55%, respectively.

This improvement came at the cost of overall performance. XGBoost accuracy decreased from 98.06% to 96.42% and F1-score from 0.8901 to 0.8248, while Random Forest and Extra Trees also showed lower F1-scores. In contrast, ROC-AUC remained relatively stable, changing from 0.9759 to 0.9741 for XGBoost, 0.9713 to 0.9745 for Random Forest, and 0.9734 to 0.9752 for Extra Trees. This indicates that hybrid resampling primarily altered decision behavior at the applied threshold rather than substantially changing ranking ability.

Thus, hybrid resampling is best interpreted as a sensitivity-oriented configuration, reducing false negatives at the expense of accuracy and, in some cases, F1-score. Its suitability therefore depends on whether improved pulsar-candidate recovery is prioritized. These findings motivate the subsequent cross-validation and statistical analyses. Fig 4 summarizes the effect of hybrid resampling on minority-class detection. The consistent increase in Recall and corresponding decrease in FNR across the three ensemble models visually reinforces the sensitivity-oriented behavior observed in Table II.

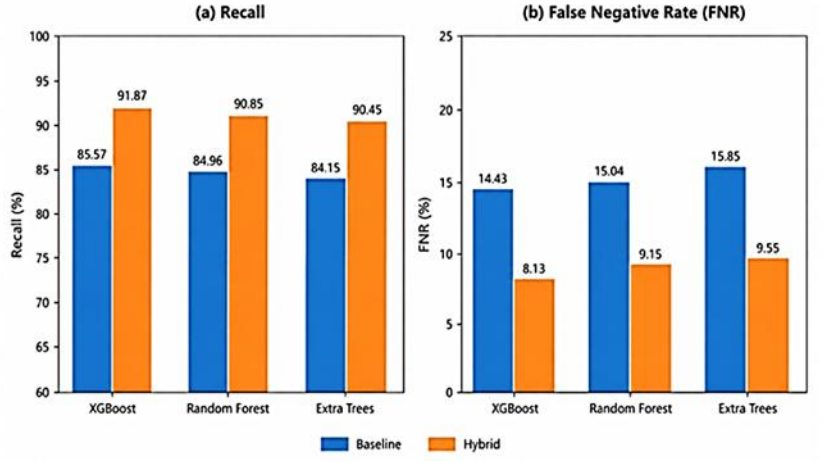


Fig. 4. Recall and False Negative Rate (FNR) before and after hybrid resampling for XGBoost, Random Forest, and Extra Trees.

http://www.doi.org/10.62341/istj-vol39-2-fi14

4.3. Cross-Validation Stability

To assess performance stability beyond the independent test split, stratified 10-fold cross-validation was performed using leakage-free pipelines. Preprocessing and hybrid resampling were fitted only within each training fold, while the corresponding validation fold remained independent.

Among baseline models, XGBoost achieved 98.00% ± 0.24% accuracy, F1 = 0.8868 ± 0.0124, and ROC-AUC = 0.9782 ± 0.0056. Random Forest and Extra Trees also showed stable performance, with F1-scores of 0.8842 ± 0.0146 and 0.8818 ± 0.0084, respectively, indicating limited variability across folds.

Hybrid configurations consistently emphasized minority-class detection. XGBoost-Hybrid achieved 91.02% ± 1.85% Recall, compared with 90.67% ± 2.15% for Random Forest-Hybrid and 90.06% ± 1.76% for Extra Trees-Hybrid. The results are consistent with the sensitivity-oriented behavior observed on the independent test set.

Overall, cross-validation supports the stability of the observed performance patterns: baseline ensembles maintained stable overall performance, whereas hybrid configurations consistently increased minority-class Recall. The independent test set remains the primary measure of performance on unseen data.

4.4. H2O AutoML Results

H2O AutoML was evaluated as an independent model-selection branch using the original imbalanced training set, with the test set excluded from model selection and hyperparameter optimization. Up to 30 models were trained using a fixed seed and five-fold internal cross-validation, with log-loss as the selection metric; deep-learning models were excluded.

The leading model was a Stacked Ensemble. On the untouched test set, it achieved 98.12% accuracy, 93.93% precision, 84.96% recall, F1 = 0.8922, ROC-AUC = 0.9771, and MCC = 0.8833. Its performance was comparable to the strongest manually specified models, indicating that AutoML provided an effective automated model-selection and ensemble-construction approach without demonstrating universal predictive superiority.

http://www.doi.org/10.62341/istj-vol39-2-fi14

Compared with XGBoost, which achieved 98.06% accuracy, 0.8901 F1-score, and 0.9759 ROC-AUC**,** H2O AutoML produced slightly higher values for these measures. However, the differences are small, and no statistical test was conducted specifically to establish that the AutoML model significantly outperforms XGBoost. Therefore, these results should be interpreted as competitive performance rather than evidence of statistically significant superiority**.**

The contribution of AutoML in this study is methodological rather than a claim of universal performance improvement. Unlike the manually specified models, the AutoML branch automates model exploration, configuration, and ensemble construction while maintaining the same independent test-set protocol.

Importantly, H2O AutoML was trained on the original imbalanced training distribution, not the hybrid-resampled data. Thus, the two approaches address different aspects of the problem: hybrid resampling modifies class distribution to enhance minority-class sensitivity, whereas AutoML automates model selection and optimization. This distinction is essential for interpreting their respective contributions within the proposed framework.

4.5. Statistical Reliability

To assess the reliability of the results, 95% bootstrap confidence intervals were calculated for the main performance metrics, while McNemar's test evaluated paired differences on the same independent test set. XGBoost achieved an F1-score of 0.8903 [0.8686, 0.9106], Recall of 0.8559 [0.8216, 0.8868], and ROC-AUC of 0.9760 [0.9666, 0.9838]. Extra Trees with hybrid resampling achieved an F1-score of 0.8628 [0.8386, 0.8845] and Recall of 0.9047 [0.8791, 0.9283], indicating relatively stable estimates.

McNemar's tests showed statistically distinguishable paired prediction patterns for RF vs. SVM-Hybrid ($\chi^2 = 17.284$, $p = 3.22 \times 10^{-5}$), XGBoost vs. XGBoost-Hybrid ($\chi^2 = 50.460$, $p = 1.22 \times 10^{-12}$), and Extra Trees vs. Extra Trees-Hybrid ($\chi^2 = 10.779$, $p = 1.03 \times 10^{-3}$). These tests indicate that hybrid resampling changes classification behavior, particularly for minority-class detection, but do not establish universal superiority of any classifier.

http://www.doi.org/10.62341/istj-vol39-2-fi14

Overall, bootstrap confidence intervals, cross-validation variability, and paired statistical testing provide stronger evidence for interpreting the results than single-point accuracy estimates alone.

4.6. Comparison with Deep Learning and Previous Studies

Recent studies have advanced pulsar candidate classification through transformer-based, image-based, and multimodal deep-learning approaches. Cao et al. [21] evaluated ViT and CvT architectures using CRAFTS and HTRU data, while Song et al. [28] investigated GGUS with RF, SVM, CNN, and ResNet50 on HTRU1. Luo et al. [29] proposed CFHAPNet based on multi-view candidate images and a Swin Transformer, reporting 97.5% accuracy, 98.4% recall, and 98.0% F1-score on FAST data. Tang et al. [30] introduced a multimodal multi-scale DenseNet combining diagnostic-plot images with structured features, reporting F1 = 0.904 and ROC-AUC = 0.978. These studies demonstrate the continued development of richer deep-learning representations.

In comparison, the present study uses an eight-feature tabular representation of HTRU2. Its MLP achieved 98.03% accuracy, F1 = 0.8863, and ROC-AUC = 0.9736, indicating competitive performance within this representation. However, these results are not directly comparable with image-based deep-learning studies because of differences in datasets, input modalities, sampling strategies, and evaluation protocols. Similarly, semi-supervised methods based on pseudo-labeling address a different setting because the present study uses a fully supervised HTRU2 protocol. Accordingly, the principal value of the comparison lies in methodological consistency: the evaluated models use the same HTRU2 representation and independent test partition, while AutoML follows a controlled automated-selection procedure. This enables the effects of model family and class-balancing strategy to be examined without conflating differences in datasets or input representations.

http://www.doi.org/10.62341/istj-vol39-2-fi14

4.7. Discussion

The combined experimental evidence provides several important observations regarding pulsar candidate classification under severe class imbalance.

First, the independent test results demonstrate that high overall accuracy can be obtained without necessarily maximizing minority-class detection. The baseline ensemble models achieved accuracies close to 98%, but their Recall values remained in the approximately 84–86% range. This confirms that accuracy alone does not adequately characterize performance when genuine pulsars constitute a small proportion of the observations.

Second, hybrid resampling produced a consistent increase in minority-class Recall across the evaluated ensemble models. This behavior was also observed during cross-validation, indicating that the effect was not restricted to a single train/test partition. However, the improvement in Recall was accompanied by additional false-positive predictions and, for several models, lower overall F1-score and accuracy. Hybrid resampling should therefore be viewed as a mechanism for changing the operating trade-off between missed pulsars and false detections rather than as an unconditional performance enhancement.

Third, the AutoML experiment demonstrates a different type of contribution. H2O AutoML achieved the highest accuracy and F1-score among the models included in the main comparison, but the numerical difference relative to XGBoost was small. Therefore, the principal advantage of AutoML in this study is its ability to automate model exploration, configuration, and ensemble construction rather than to guarantee a substantial increase in predictive performance.

Fourth, the statistical analyses reinforce the importance of examining model behavior rather than relying exclusively on headline accuracy. Bootstrap intervals demonstrate the uncertainty associated with the reported metrics, while McNemar testing confirms that selected baseline and hybrid configurations produce statistically distinguishable prediction patterns on the same test observations.

http://www.doi.org/10.62341/istj-vol39-2-fi14

Taken together, the experiments suggest that model selection and class balancing address different dimensions of the classification problem. Conventional and automated models primarily determine how the available feature space is exploited, whereas hybrid resampling modifies the learning distribution to increase sensitivity to the minority class. Treating these components as complementary rather than interchangeable provides a more accurate interpretation of the experimental results.

4.8. Limitations and Scope

The study is limited to the HTRU2 benchmark dataset, and no independent astronomical dataset was used; therefore, the results establish performance on the held-out HTRU2 test set but not cross-survey generalization across different instruments, observing conditions, or candidate-generation pipelines. The hybrid strategy also alters the training distribution through undersampling and SMOTE, improving minority-class sensitivity at the cost of increased false positives; its suitability therefore depends on the intended operating point. Comparisons with deep-learning and semi-supervised methods are literature-based and not direct experimental benchmarks because of differences in datasets and input representations.

Overall, the leakage-free protocol, based on strict train/test separation and training-only preprocessing and resampling, supports a statistically grounded comparative evaluation on HTRU2. The findings should not be interpreted as evidence of universal astronomical generalization or deployment readiness.

5. CONCLUSION

This study presented a leakage-free comparative framework for pulsar candidate classification on HTRU2, evaluating ensemble learning, XGBoost, MLP, hybrid resampling, and H2O AutoML under a controlled independent test protocol. H2O AutoML and XGBoost achieved strong overall performance, while hybrid resampling substantially improved minority-class Recall and reduced FNR at the cost of increased false positives and lower overall accuracy. These results demonstrate that AutoML and hybrid resampling serve complementary purposes rather than

providing universal performance improvements**.** Cross-validation, bootstrap confidence intervals, and McNemar testing further supported the reliability of the observed performance patterns. However, because the study uses only HTRU2, cross-survey generalization and deployment readiness remain unestablished and require validation on independent astronomical datasets.

REFERENCES:

[1] Balakrishnan, V., Champion, D., Barr, E., et al. (2021). "Pulsar Candidate Identification Using Semi-Supervised Generative Adversarial Networks," Monthly Notices of the Royal Astronomical Society, vol. 505, no. 1, pp. 1180-1194.

[2] Bates, S. D., Bailes, M., Barsdell, B. R., et al. (2012). "The High Time Resolution Universe Pulsar Survey—VI. An Artificial Neural Network and Timing of 75 Pulsars," Monthly Notices of the Royal Astronomical Society, vol. 427, no. 2, pp. 1052-1065.

[3] Breiman, L. (2001). "Random Forests," Machine Learning, vol. 45, pp. 5-32.

[4] Cao, J., Xu, T., Deng, L., Zhou, X., Li, S., Liu, Y., and Zhou, W. (2024). "Pulsar Candidate Identification Using Advanced Transformer-Based Models," Chinese Journal of Physics, vol. 90, pp. 121-133.

[5] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). "SMOTE: Synthetic Minority Over-Sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357.

[6] Chen, T. and Guestrin, C. (2016). "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794.

[7] Cortes, C. and Vapnik, V. (1995). "Support-Vector Networks," Machine Learning, vol. 20, pp. 273-297.

[8] Eatough, R. P., Molkenthin, N., Kramer, M., et al. (2010). "Selection of Radio Pulsar Candidates Using Artificial Neural Networks," Monthly Notices of the Royal Astronomical Society, vol. 407, no. 4, pp. 2443-2450.

[9] Friedman, J. H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189-1232.
[10] Freund, Y. and Schapire, R. E. (1997). "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119-139.
[11] Guo, P., Duan, F., Wang, P., et al. (2019). "Pulsar Candidate Classification Using Generative Adversary Networks," Monthly Notices of the Royal Astronomical Society, vol. 490, no. 4, pp. 5424-5439.
[12] He, H. and Garcia, E. A. (2009). "Learning from Imbalanced Data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284.
[13] LeDell, E. and Poirier, S. (2020). "H2O AutoML: Scalable Automatic Machine Learning," in Proceedings of the 7th ICML Workshop on Automated Machine Learning (AutoML), July 2020.
[14] Lee, E., Rustam, F., Aljedaani, W., et al. (2021). "Predicting Pulsars from Imbalanced Dataset with Hybrid Resampling Approach," Advances in Astronomy, vol. 2021, Article ID 4916494.
[15] Lin, H., Li, X., and Luo, Z. (2020). "Pulsars Detection by Machine Learning with Very Few Features," Monthly Notices of the Royal Astronomical Society, vol. 493, no. 2, pp. 1842-1854.
[16] Liu, Y., Jin, J., Zhao, H., and Wang, Z. (2024). "Enhancing Pulsar Candidate Identification with Self-Tuning Pseudolabeling Semisupervised Learning," The Astrophysical Journal, vol. 967, no. 2, Article 155.
[17] Lyon, R. (2015). "HTRU2," UCI Machine Learning Repository, dataset.
[18] Lyon, R. J., Stappers, B. W., Cooper, S., Brooke, J. M., and Knowles, J. D. (2016). "Fifty Years of Pulsar Candidate Selection: From Simple Filters to a New Principled Real-Time Classification Approach," Monthly Notices of the Royal Astronomical Society, vol. 459, no. 1, pp. 1104-1123.

Optimizing Pulsar Star Prediction in Large-Scale Astronomical Surveys Using Hybrid Resampling and Automated Machine Learning
http://www.doi.org/10.62341/istj-vol39-2-fi14

[19] McNemar, Q. (1947). "Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages," Psychometrika, vol. 12, no. 2, pp. 153-157.
[20] Morello, V., Barr, E. D., Bailes, M., et al. (2014). "SPINN: A Straightforward Machine Learning Solution to the Pulsar Candidate Selection Problem," Monthly Notices of the Royal Astronomical Society, vol. 443, no. 2, pp. 1651-1662.
[21] Rustam, F., Mehmood, A., Ullah, S., et al. (2020). "Predicting Pulsar Stars Using a Random Tree Boosting Voting Classifier (RTB-VC)," Astronomy and Computing, vol. 32, Article 100404.
[22] Song, Y.-N., Chen, M.-Z., and Liu, Z.-Y. (2025). "Improving Pulsar Candidate Identification with Grid Group Uniform Sampling," Research in Astronomy and Astrophysics, vol. 25, no. 5, Article 055007.
[23] Tang, J., Xie, X., and Xiong, X. (2025). "Efficient Pulsar Candidate Recognition Algorithm Under a Multi-Scale DenseNet Framework," Applied Sciences, vol. 15, no. 24, Article 13097.
[24] Wang, Y., Pan, Z., Zheng, J., Qian, L., and Li, M. (2019). "A Hybrid Ensemble Method for Pulsar Candidate Classification," Astrophysics and Space Science, vol. 364, no. 8, Article 139.
[25] Wang, Y.-C., Li, M.-T., Pan, Z.-C., and Zheng, J.-H. (2019). "Pulsar Candidate Classification with Deep Convolutional Neural Networks," Research in Astronomy and Astrophysics, vol. 19, no. 9, Article 133.
[26] Yin, Q., Li, J., Zheng, X., Li, Y., Cui, H., and Bao, Z. (2022a). "Pulsar Candidate Selection with Residual Convolutional Autoencoder," Monthly Notices of the Royal Astronomical Society, vol. 516, no. 2, pp. 3082-3091.
[27] Yin, Q., Wang, Y., Zheng, X., and Zhang, J. (2022b). "Pulsar Candidate Recognition Using Deep Neural Network Model," Electronics, vol. 11, no. 14, Article 2216.