Int. J. Advanced Networking and Applications
Volume: 17, Issue: 03, Pages : 6916-6921 (2025) ISSN: 0975-0290
https://doi.org/10.35444/IJANA.2025.17304

Ensemble Learning Based Pulsar Detection Along with Feature Engineering

Talha Mahamud
Physics Discipline, Khulna University, Khulna-9208
Email: iamtalhamahamud@gmail.com
Sadia Rahman
Physics Discipline, Khulna University, Khulna-9208
Email: sadianity1456@gmail.com
Abdullah-Al Nahid[1*]
Electronics and Communication Engineering Discipline, Khulna University, Khulna-9208
Email: nahid.eee.ku@gmail.com

*This Paper has one corresponding author: *Abdullah-Al Nahid.*

----------ABSTRACT----------

The study of pulsars is a contemporary research field with significant possibilities. In recent times as well as in the future, with the advancement of technology, the burgeoning number of pulsar-like objects may become the biggest constraint. To come up with a solution, many sophisticated methods have been taken into consideration over the fifty years, along with building machine learning models. In this paper we have built an efficient machine learning model using the HTRU2 dataset. The proficiency of a model does not depend on only one parameter, such as accuracy. Therefore, we have focused on building a model taking into account robustness, complexity, and run-time duration.

Keywords - **Pulsar Classification; Machine Learning; HTRU2 Dataset; Astronomy; LightGBM; Chi2.**

Date of Submission: August 17,2025 Date of Acceptance: Sep 29, 2025

1. INTRODUCTION

The study of pulsars is one of the most crucial research fields in physics and astrophysics. As extremely dense objects, pulsars have been used as a fundamental tool, from studying highly dense matter to experimental testing of gravitational theories [1]. This unique and exotic system allows us to probe the general theory of relativity, cosmological gravitational wave background, intergalactic medium, finding the superdense matter's equation-of-state, black holes, etc. This list goes on [2]. Some significant discoveries have already been made through investigating pulsars. For instance, the finding of the first extrasolar planet that revolves around a pulsar [3], the indirect confirmation of gravitational radiation by a binary pulsar system [4], the most recent discovery of evidence for gravitational-wave background [5], and many more. In that case, to advance our understanding and exploration of this complex universe through pulsar observation, identifying and confirming them among many other celestial objects is the first and foremost task.

2. LITERATURE REVIEW

Pulsars are highly magnetized neutron stars that rotate and continually radiate electromagnetic beams from their magnetic poles. Their period of pulse (the time interval between two successive pulses) is very regular, generally ranging from milliseconds to seconds, and based on this, they are categorized into two types of pulsars. The pulses are created due to the rotation of the star with high velocity, when the magnetic poles beaming with electromagnetic radiation (which is result of strong electric and magnetic field of the star) aligns with the earth, we see a pulse, and turns of when the beams rotate away from our sight; this can be compared with a 'light-house' [6]. This identical characteristic of pulsars helped to gain attention and accidentally led to the discovery of the first pulsar in 1968 [7], which was mainly an analysis of "paper-based pen chart recordings." However, this procedure became outdated a long time ago with the advanced computational technology and digital statistical applications that are able to identify astrophysical signals with higher accuracy [8]. The detection of the pulsars is done by observing the data collected from different telescopes around the world of pulsar-emitted radiations (radio emission, optical, x-rays, and γ-rays) and pulsating behaviors. But along with the advancement of technology, there showed up some serious difficulties in detecting pulsars. The volume of data from the telescopes is increasing rapidly, which contains less necessary data as well as noise, adding some friction to the process of detection, as now the scientists have to maintain and filter a large amount of data. Another crucial aspect is the presence of a proliferation of objects that show pulsar-like characteristics, which are not actually pulsars [9]. Finding such exquisite differences each time is a difficult task to do. In that case, machine learning model, an artificial neural network, can be a great help in detecting pulsars analyzing signals [10],[24]. As ML (Machine Learning) models can efficiently do any classification if they are

Int. J. Advanced Networking and Applications
Volume: 17, Issue: 03, Pages : 6916-6921 (2025) ISSN: 0975-0290
https://doi.org/10.35444/IJANA.2025.17304

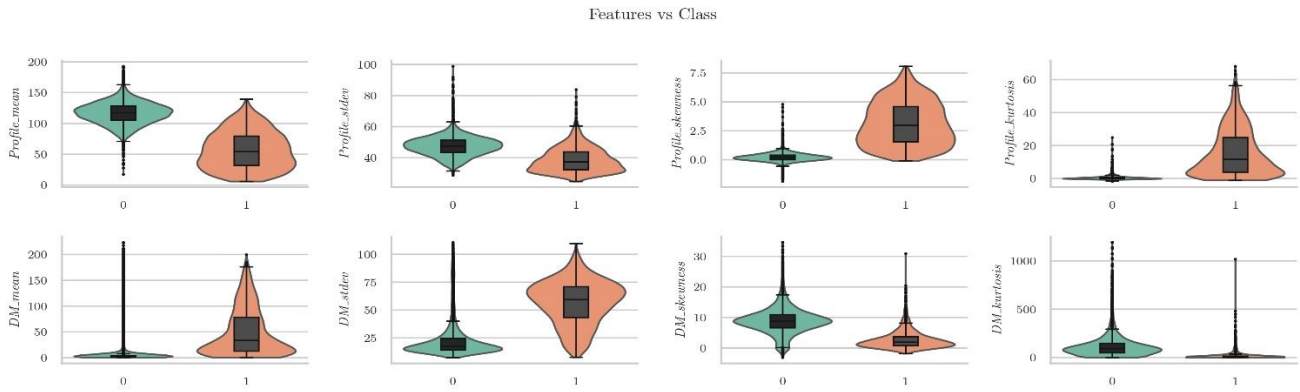

Fig 3: Distribution and Density of Features in the HTRU2 Dataset: Violin plots reveal the spread, symmetry, and concentration of values for each feature used in pulsar classification.

trained. Machine learning (ML), a subset of artificial intelligence (AI), focuses on enabling systems to acquire human-like learning capabilities. In supervised learning, classification is a common approach that leverages pattern recognition techniques to detect relationships among features. The human brain excels at complex classification tasks when adequately trained, primarily due to its inherent ability to learn through trial and error. Similarly, ML algorithms adopt this approach by iteratively improving their performance based on experience. For this, an ML model is first trained with a dataset; in that process, the model figures out the relation of the features with the target and predicts a possible outcome. And then the model is tested to see if it can classify properly when some specific features are given to it [8] [11], [23]. Hence, if we fed an ML model with the data of the signals (which is referred to as 'features' in the ML model) of already detected pulsars and non-pulsars, it will identify and learn the pattern and characteristics of the signals; therefore, it will be able to analyze a signal and detect if it is a pulsar or non-pulsar. The remainder of this paper is organized as follows: Section-2 outlines the methodology used in our study, including a brief overview of the datasets and the classifier employed. Section-3 presents the outcomes of our experiments, along with a summary of the analysis and key findings. Finally, Section-4 provides concluding remarks on our work.

3. METHODOLOGY

This paper focuses on solving the problem in an efficient way and gives a better model for the HTRU2 dataset and upcoming pulsar data. As the amount of data is increasing day by day at a large rate, our goal was to develop a machine learning that will make it easier to predict pulsers easily and precisely. A complete diagram of this section has been depicted in Fig. 2.

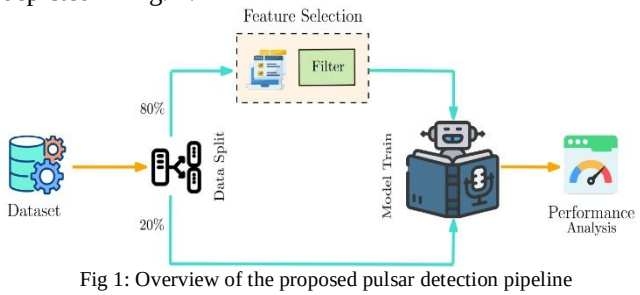

Fig 1: Overview of the proposed pulsar detection pipeline

3.1 Dataset

HTRU2 consists of 17,898 samples with 8 features of numerical values. This is a binary classification dataset.

$$S = \{C_1, C_2\}$$

Where C_1 = 1,639 positive samples (pulsars) and C_2 = 16, 259 negative samples (non-pulsars), and all have been checked by human annotators [12].

The ratio of the pulsar and non-pulsar is showing that the dataset is significantly class imbalance. The correlation matrix of features of this dataset has been shown in Figure 2.

A violin plot shows the distribution of the dataset, visualizing where the data is spread out and where it is concentrated among the groups. The wider part shows where the data is more concentrated, as more data points represent a large area. Based on the HTRU2 dataset and our machine learning model, there are two groups: non-pulsar (represented as '0') and pulsar (represented as '1'). In Figure 3, for all the features from the HTRU2 dataset, the violin plot has been drawn, and it visualizes that Profile skewness, Profile kurtosis, DM skewness, and DM kurtosis—these four features—may not be useful for our pulsar detection model, as the pulsar violins (class 1) cover a larger range than the non-pulsar violins (class 0). But if we consider the more concentrated data in the pulsar violins, then these five features, Profile_mean, Profile_skewness,Profile_kurtosis,DM_mean,and DM_stdev seem to be good for our model. In this paper, we have divided the dataset by 80% and 20% for training and testing.

Int. J. Advanced Networking and Applications
Volume: 17, Issue: 03, Pages : 6916-6921 (2025) ISSN: 0975-0290
https://doi.org/10.35444/IJANA.2025.17304

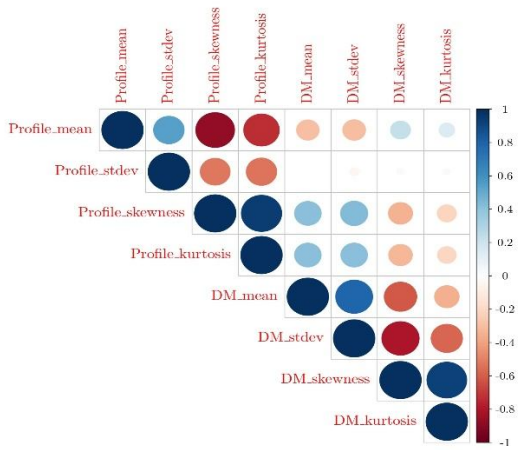

Fig 1: HTRU2 Correlation Matrix: Feature Interactions

3.2 Feature Selection

The machine learning model is significantly influenced by features, which can occasionally produce inaccurate or biased results. Three goals are pursued by variable or feature selection: enhancing predictor prediction performance, producing faster and more economical predictors, and enhancing comprehension of the underlying process that produces the data. [13]. In addition, this helps to build simpler and more comprehensible models, improve data mining performance, and prepare clean, understandable data [14]. We have used 'Mutual Information (MI),' a filter method that selects features based on how much information they share with the target variable, i.e., how much knowing a feature reduces uncertainty about the target. It performs well on both classification and regression tasks and captures non-linear relationships. Analysis of variance (ANOVA), a filter-based feature selection method that assesses each feature's relationship to the target variable using statistical tests, independent of any machine learning model, and Chi-Square (CS), a statistical method in discrete data testing that evaluates whether two variables are related or not, are two examples of how it captures non-linear relationships and functions for both classification and regression tasks. [15]. This algorithm performs to reduce the irrelevant features in the classification process [16]. The formula for Chi-square is

$$X_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where c is the degree of freedom, O is observed value and E is the expected value.

3.3 Model Description

In this study, we have used Ensemble Learning technique. With ensemble learning, multiple small models are used rather than just one. Even while each of these models might not be particularly good by itself, when we combine their findings, we get a more accurate and superior response. It's similar to asking a group of people for advice rather than just one; while each person may be slightly off, the group as a whole typically provides a better response.[16]

Light Gradient Boosting Machine (LightGBM) is one of the popular ensemble learning based machine learning model. We employed that as our predictive model. LightGBM is a gradient boosting framework that relies on tree-based learning algorithms and is optimized for speed and efficiency. Its key advantages include faster training times, reduced memory consumption, improved accuracy, and support for parallel, distributed, and GPU-based learning. Moreover, it is well-suited for handling large-scale

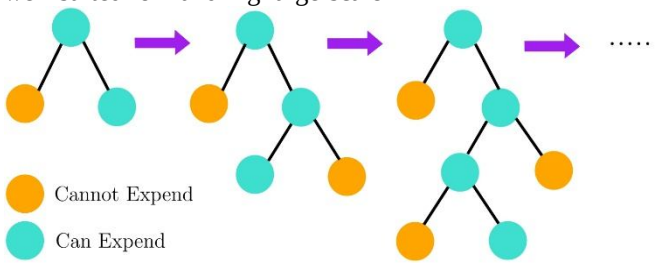

Fig 4: How LightGBM Learns: A Step-by-Step Breakdown

datasets [18]. Additionally, it provides information on feature importance, assisting in determining which traits are most important for prediction, which is helpful for interpretability in scientific study. The representation of how LightGBM predicts from different trees is expressed by the following formula,

$$Y = Base\_tree(X) - lr \sum_{i=1}^{N} Tree_i(X)$$

The first term represents the initial prediction generated by a simple decision tree, offering a preliminary estimate of the target variable. Each subsequent term corresponds to a correction made by additional trees (Tree1, Tree2, Tree3, etc.), where each tree is trained to minimize the errors of the preceding ones. The learning rate (lr) determines the extent to which these corrections influence the overall model. Ultimately, the final prediction is obtained by summing the initial prediction and all subsequent corrections, each weighted by the learning rate. This iterative and cumulative approach leads to more accurate and robust predictions compared to using a single decision tree [19]. The visual representation of working method of LightGBM model has been demonstrated in Figure 4.

3.4 Performance Analysis

In this section, we have presented the performance analysis of our model using 'Accuracy', 'Precision', 'Recall', and 'F1-score'. Accuracy gives an overall measure of correct predictions. It is calculated by the following formula.

$$Accuracy = \frac{No.of\ correcrt\ prediction}{Total\ Number\ of\ Unit\ Samples}$$

While precision tells us how many positive predictions made by the model are, actually correct

$$Precision = \frac{True\ Positive}{True\ Positive\ + False\ Positive}$$

and Recall measures how well the model captures all relevant positive cases.

$$Recall = \frac{True\ Positive}{True\ Positive\ +\ False\ Negative}$$

Given the nature of our dataset, F1-score is particularly significant because it combines the Precision and Recall scores of a model that shows us how precise (the number of correctly classified instances) and robust (does not miss any

Int. J. Advanced Networking and Applications
Volume: 17, Issue: 03, Pages : 6916-6921 (2025) ISSN: 0975-0290
https://doi.org/10.35444/IJANA.2025.17304

significant number of instances) our classifier is. Mathematically, it is expressed as,

$$F1\_Score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}$$

4. RESULT & DISCUSSION

As mentioned earlier, we have used three feature selection methods to get the best features for higher accuracy. From the Figure 5 given below, it is certain that, according to the all of the methods we have used, the highest accuracy is obtained when we take all the features of the HTRU2 dataset.

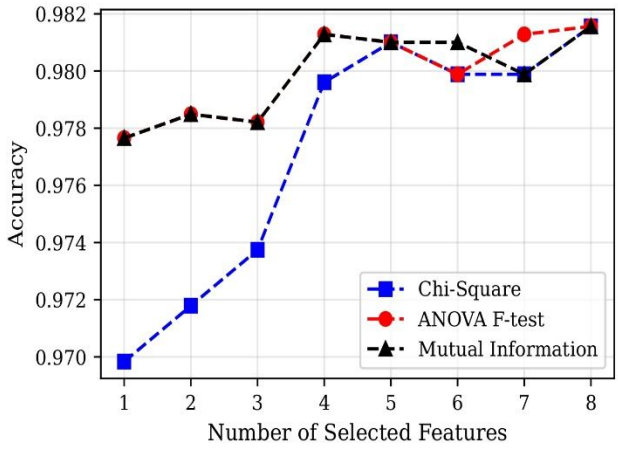

Fig 5: Accuracy of the model with varying number of selected features.

However, when building a machine learning model, the term "run-time" should be given significant consideration, as the model's efficiency depends on it. Run-time is the entire duration that a machine learning model requires to operate. The smaller the run-time, the more efficient the model is.

TABLE I: Comparison of Run-time, Accuracy, Recall, Precision and F1-score between models trained with 5 and 8 features.

No. of Feature	Run-time	Accuracy	Recall	Precision	F1-Score
Five	0.1389	0.981006	0.856698	0.925926	0.889968
Eight	0.1716	0.981564	0.856698	0.932203	0.892857

Following Figures 5 and 6, we get the best combination of accuracy and run-time for five features selected by Chi-square: accuracy of 0.981006 and run-time of 0.1389 s. Moreover, the difference between the highest accuracy and the accuracy for the five features is 0.000558, which is negligible.

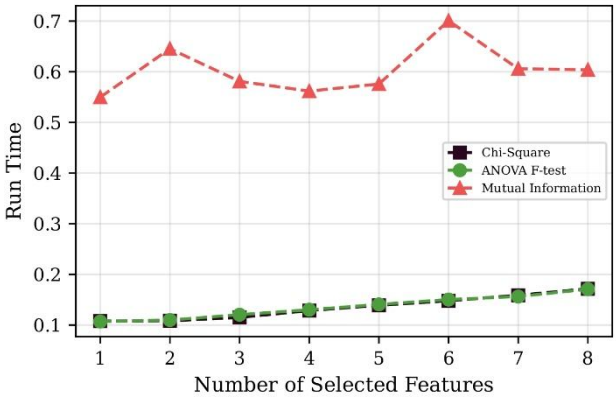

Fig 6: Effect of Feature Count on Model Run-time: Run-time increases as the number of selected features grows, highlighting the trade-off between feature selection and model efficiency.

Figure 7 shows the confusion matrix for 8 features and 5 selected features, respectively. Which tells us that if we choose to reduce the run-time and build our model with five selected features rather than just focusing on the accuracy, the difference in the number of false/correct predictions made by the model is not very large, and it is ignorable. We can get a clearer view by observing the ROC curves in both cases. The ROC curve explains how well the model distinguishes between the two classes. The closer the curve is to the top-left corner, the better the model is. If the curve is along the diagonal line, it depicts that the model is doing nothing but making predictions out of random guesses. Figure 8 contains the ROC curves of our model for the two distinct cases. The term AUC stands for 'Area Under the Curve,' indicating if the value of AUC is '1,' then the model is a perfect classifier, and if the value is '0,' it means the model only does random guesses. Figure 8(a) represents the model trained with all eight input features, gained an AUC of 0.9786, while Figure 8(b) represents the model trained with the five most significant features, yielding an AUC of 0.9760. Both models demonstrate high classification performance, but the model trained with all eight features shows a slightly superior discriminative ability. This improvement indicates that including all features contributes minimally to performance enhancement, which suggest that the reduced feature set retains most of the predictive information. Hence, working with only five features makes the model comparatively efficient and simpler; with less complexity, it becomes more robust. Therefore, we have chosen these five features, according to the Chi-square feature selection model, Profile_mean, Profile_skewness, Profile_kurtosis, DM_mean, and DM_stdev, for our model to effectively detect pulsars. These five features also support the findings of the violin plot explained earlier. Some previous works on the HTRU2 dataset show that they achieved accuracies around 97–98% [20] [21], whereas our model attains a slightly higher accuracy of 98.1%. Conversely, in another work, they obtained accuracy similar to ours (98.3%) [22], using all of the eight features. We used just five features, which resulted in significantly lower run-time. Although our model achieves a strong and resilient state, it has comparatively lower recall. Which indicates that, along with fewer false positives, it may miss some true pulsar candidates. Depending on the application scenario, this trade-off may be acceptable or further optimizable.

Int. J. Advanced Networking and Applications
Volume: 17, Issue: 03, Pages : 6916-6921 (2025) ISSN: 0975-0290
https://doi.org/10.35444/IJANA.2025.17304

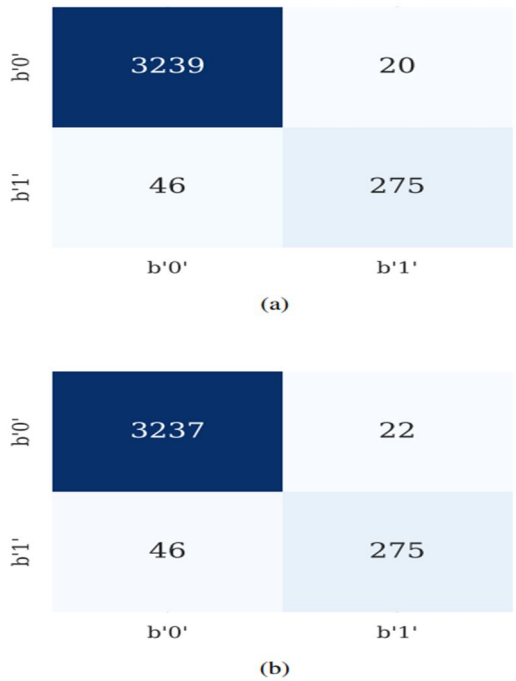

Fig 7: Comparison of model performance via confusion matrices: (a) shows the result using all 8 features, while (b) shows the result using 5 features selected through Chi-Square. Both highlight the trade-off between feature reduction and classification accuracy.

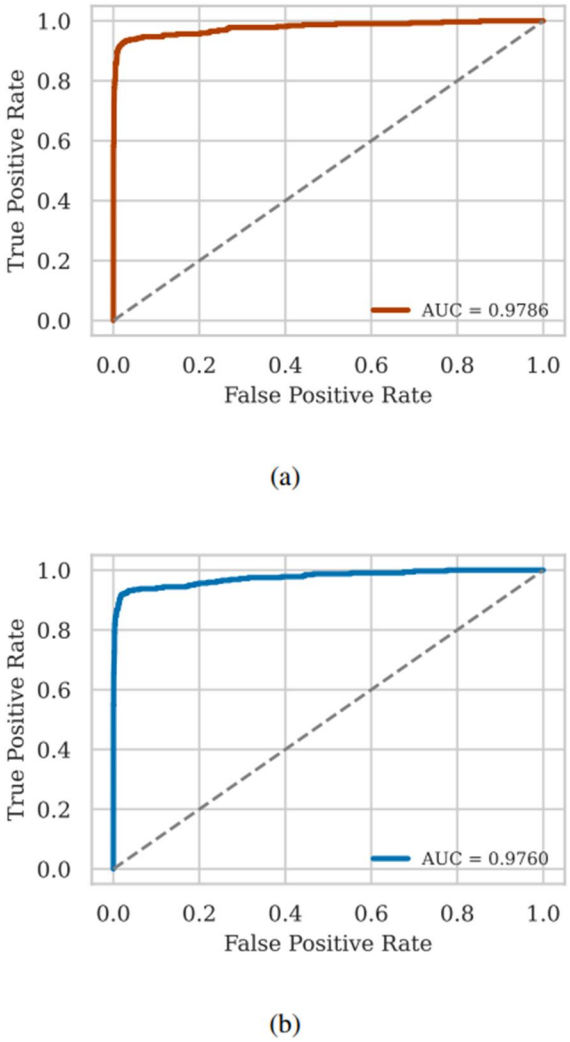

Fig 8: ROC curves for models trained on the HTRU2 dataset: (a) using all 8 features, and (b) using 5 selected features.

5. CONCLUSION

Throughout this study, an optimized Machine Learning model has been represented with a balanced run-time and accuracy with a view to classifying pulsar and non-pulsar candidates using HTRU2 dataset. Initially, we have studied the characteristics of the features of the dataset through the correlation matrix and violin plot. As for our research, we have selected the LightGBM model to train and test with the following dataset. Later on, we have used three feature selection methods and observed the model's performance based on these methods. Though the model gives higher accuracy for all of the features, we have considered the run-time and complexity to make the model more robust so that it can perform more efficiently with less data in the future. Also, the recall is comparatively lower than some previous works. Therefore, further study of this model can be based on gaining higher recall as well as F1-score. This can be done by re-adjusting the model and modifying the training process.

REFERENCES

[1] Kramer, M., "Pulsars as probes of gravity and fundamental physics," International Journal of Modern Physics D, vol. 25, no. 14, p. 1630029, Dec. 2016. [Online]. Available: http://dx.doi.org/10.1142/S0218271816300299

[2] Cordes, J., Kramer, M., Lazio, T., Stappers, B., Backer, D., and Johnston, S., "Pulsars as tools for fundamental physics astrophysics," New Astronomy Reviews, vol. 48, no. 11, pp. 1413–1438, 2004, science with the Square Kilometre Array. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1387647304001253

[3] Wolszczan, A. and Frail, D. A., "A planetary system around the millisecond pulsar PSR1257 + 12," , vol. 355, no. 6356, pp. 145–147, Jan. 1992.

[4] Weisberg, J. M., Taylor, J. H., and Fowler, L. A., "Gravitational waves from an orbiting pulsar," Scientific American, vol. 245, pp. 74–82, Oct. 1981.

[5] Agazie, G. and et al., "The nanograv 15 yr data set: Evidence for a gravitational-wave background," The Astrophysical Journal Letters, vol. 951, no. 1, p. L8, Jun. 2023. [Online]. Available: https://dx.doi.org/10.3847/2041-8213/acdac6

[6] Goddard Space Flight Center, "Imagine the Universe!" 2017, [Online; accessed 011-May-2025]. [Online]. Available: https://imagine.gsfc.nasa.gov/science/objects/neutron stars1.html

[7] Hewish, A., Bell, S. J., Pilkington, J. D. H., Scott, P. F., and Collins, R. A., "Observation of a Rapidly Pulsating Radio Source," , vol. 217, no. 5130, pp. 709–713, Feb. 1968.

[8] Lyon, R. J., "Fifty years of candidate pulsar selection - what next?" Proceedings of the International Astronomical Union, vol. 13, no. S337, p. 25–28, Sep. 2017. [Online]. Available: http://dx.doi.org/10.1017/S1743921317007682

[9] "Why are pulsars hard to find," 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:125875662

[10] Eatough, R. P., Molkenthin, N., Kramer, M., Noutsos, A., Keith, M. J., Stappers, B. W., and Lyne, A. G., "Selection of radio pulsar candidates using artificial

neural networks," Monthly Notices of the Royal Astronomical Society, vol. 407, no. 4, pp. 2443–2450, 07 2010. [Online]. Available: https://doi.org/10.1111/j.1365-2966.2010.17082.x

[11] Bishop, C. M., Pattern Recognition and Machine Learning. New York, NY: Springer, 2016.

[12] Lyon, R., "HTRU2," UCI Machine Learning Repository, 2015, DOI: https://doi.org/10.24432/C5DK6R.

[13] Guyon, I. M. and Elisseeff, A., "An introduction to variable and feature selection," J. Mach. Learn. Res., vol. 3, pp. 1157–1182, 2003. [Online]. Available: https://api.semanticscholar.org/CorpusID:379259

[14] Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., and Liu, H., "Feature selection: A data perspective," ACM Comput. Surv., vol. 50, no. 6, Dec. 2017. [Online]. Available: https://doi.org/10.1145/3136625

[15] Statistics Solutions, "Using Chi-Square Statistic in Research," 2025, [Online; accessed 09-May-2025]. [Online]. Available: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/using-chi-square-statistic-in-research/

[16] GeeksforGeeks, "Ensemble Learning" 2025, [Online; accessed 15-Octobar-2025]. [Online]. Available: Ensemble Learning - GeeksforGeeks

[17] Setiyaningrum, Y. D., Herdajanti, A. F., Supriyanto, C., and Muljono, "Classification of twitter contents using chi-square and k-nearest neighbour algorithm," in 2019 International Seminar on Application for Technology of Information and Communication (iSemantic), 2019, pp. 1–4.

[18] LightGBM, "Welcome to LightGBM's documentation!" 2025, [Online; accessed 02-May-2025]. [Online]. Available: https://lightgbm.readthedocs.io/en/stable/

[19] Akash Pushkar, "What is LightGBM: The Game Changer in Gradient Boosting Algorithms," 2024, [Online; accessed 08-May2025]. [Online]. Available: https://intellipaat.com/blog/lightgbm/

[20] Lyon, R. J., Stappers, B. W., Cooper, S., Brooke, J. M., and Knowles, J. D., "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, 04 2016. [Online]. Available: https://doi.org/10.1093/mnras/stw656

[21] Abbas, S. A., Rajasekar, D., and Rahul, A., "A machine learning approach for the identification and estimation of pulsars upon statistical analysis," 09 2023, pp. 1–12.

[22] Rustam, F., Mehmood, A., Ullah, S., Ahmad, M., Muhammad Khan, D., Choi, G., and On, B.-W., "Predicting pulsar stars using a random tree boosting voting classifier (rtb-vc)," Astronomy and Computing, vol. 32, p. 100404, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2213133720300585

[23] Tiwari,V.K., Jain, A., Singh,R, and Singh, P, "Enhancing Outlier Detection and Dimensionality Reduction in Machine Learning for Extreme Value" Int. J. Advanced Networking and Applications, Volume: 15 Issue: 06 Pages: 6204 – 6210 (2024) ISSN: 0975-0290.

[24] Malaker, A.,Miad, A.H., Mim, K.F., Badhan, W.B.W., and Hossen, M.I., "An Approach to Detect Credit Card Fraud Utilizing Machine Learning" Int. J. Advanced Networking and Applications, Volume: 14 Issue: 05 Pages: 5619 - 5625 (2023) ISSN: 0975-0290.