

# Comparative analysis of machine learning techniques for feature selection and classification of Fast Radio Bursts

Ailton J. B. Júnior,<sup>1,\*</sup> Jeferson A. S. Fortunato<sup>1b,2,†</sup> Leonardo J. Silvestre<sup>1b,3,‡</sup> Thonimar V. Alencar<sup>1b,4,§</sup> and Wiliam S. Hipólito-Ricaldi<sup>1b,4,5,¶</sup>

<sup>1</sup>*Departamento de Computação e Eletrônica, CEUNES, Universidade Federal do Espírito Santo (UFES), Rodovia BR 101 Norte, km. 60, CEP 29.940-540, São Mateus, ES, Brazil*

<sup>2</sup>*High Energy Physics, Cosmology & Astrophysics Theory (HEPCAT) Group, Department of Mathematics and Applied Mathematics, University of Cape Town, Cape Town 7700, South Africa*

<sup>3</sup>*Departamento de Computação e Eletrônica, CEUNES, Universidade Federal do Espírito Santo, Rodovia BR 101 Norte, km. 60, CEP 29.940-540, São Mateus, ES, Brazil*

<sup>4</sup>*Departamento de Ciências Naturais, CEUNES, Universidade Federal do Espírito Santo, Rodovia BR 101 Norte, km. 60, CEP 29.940-540, São Mateus, ES, Brazil*

<sup>5</sup>*Núcleo Cosmo-UFES, CCE, Universidade Federal do Espírito Santo, Av. Fernando Ferrari, 540, CEP 29.075-910, Vitória, ES, Brazil*

(Dated: June 24, 2025)

Fast Radio Bursts (FRBs) are millisecond-duration radio transients of extragalactic origin, exhibiting a wide range of physical and observational properties. Distinguishing between repeating and non-repeating FRBs remains a key challenge in understanding their nature. In this work, we apply unsupervised machine learning techniques to classify FRBs based on both primary observables from the CHIME catalog and physically motivated derived features. We evaluate three hybrid pipelines combining dimensionality reduction with clustering: PCA + k-means, t-SNE + HDBSCAN, and t-SNE + Spectral Clustering. To identify optimal hyperparameters, we implement a comprehensive grid search using a custom scoring function that prioritizes recall while penalizing excessive cluster fragmentation and noise. Feature relevance is assessed using principal component loadings, mutual information with the known repeater label, and permutation-based  $F_2$  score sensitivity. Our results demonstrate that the derived features including redshift, luminosity, and spectral properties, such as the spectral index and the spectral running, significantly enhance the classification performance. Finally, we identify a set of FRBs currently labeled as non-repeaters that consistently cluster with known repeaters across all methods, highlighting promising candidates for future follow-up observations and reinforcing the utility of unsupervised approaches in FRB population studies.

## I. INTRODUCTION

Fast Radio Bursts (FRBs) are among the most intriguing phenomena in modern astrophysics. These brief, highly energetic radio pulses, typically lasting only a few milliseconds, originate from extragalactic distances and exhibit high dispersion measures (DM). The DM provides information about the physical properties of the intervening medium and often exceeds the expected contribution from the Milky Way, offering strong evidence for their extragalactic origin. The first FRB was discovered in 2007 by Duncan Lorimer

and collaborators [1]. Despite significant observational efforts, the physical mechanisms underlying FRBs remain uncertain. A particularly compelling observational dichotomy has emerged between repeating and non-repeating FRBs [2], suggesting different progenitor scenarios or environmental conditions.

The short duration of the pulses, constrained by the light-crossing time of the emission region, implies that the source is extremely compact. The most plausible sources include highly magnetized neutron stars, black holes, and white dwarfs, each embedded in distinct astrophysical environments [3, 4]. Other possible progenitors involve binary interactions, collisions, tidal disruption events, and potentially even exotic physics [5–8]. So far, 862 FRBs have been published by multiple collaborations, 69 of which appear to repeat, and only 92 have been properly localized, with a redshift measurement [9]. The largest publicly available sample is provided by

\* ailton.brandao@edu.ufes.br

† jeferson.fortunato@edu.ufes.br

‡ leonardo.silvestre@ufes.br

§ thonimar.souza@ufes.br

¶ wiliam.ricaldi@ufes.br

the Canadian Hydrogen Intensity Mapping Experiment (CHIME) Collaboration [10].

One point that remains unclear is whether all FRBs repeat, as confirmation requires prolonged follow-up observations of the same sky region. While the occurrence rate of non-repeating FRBs was expected to align with that of cataclysmic events or compact-object births, recent studies report a significantly higher detection rate [11, 12]. This apparent inconsistency persists in the most recent CHIME/FRB catalog, which continues to classify the majority of sources as non-repeating. This suggests that many so-called non-repeaters may be repeaters with emissions below current sensitivity thresholds or recurrence intervals longer than typical monitoring durations. The definitive confirmation that an FRB is truly non-repeating will likely only come from the detection of a counterpart consistent with a cataclysmic event, such as a supernova or a gamma-ray burst [13].

Recently, several studies have explored the use of machine learning to identify hidden repeaters among apparently non-repeating FRBs. Unsupervised methods, such as Uniform Manifold Approximation and Projection (UMAP), have successfully clustered known repeaters and revealed new candidates directly from burst parameters [14]. Furthermore, topological data analysis has revealed structured groupings suggestive of distinct source populations [15]. Supervised classifiers have identified features such as brightness temperature and spectral width as strong discriminants [16], while hybrid approaches using t-SNE and UMAP with burst images or standardized data have further improved separation power [17]. Graph-based models, such as the Minimum Spanning Tree, have also been employed to isolate repeater-rich clusters [18]. These approaches demonstrate the potential of data-driven techniques to uncover hidden structure within the FRB population and guide targeted follow-up strategies. They also underscore the importance of developing robust classification frameworks and effective feature selection methods to reliably distinguish between repeating and apparently non-repeating sources.

Spectral properties, such as bandwidth, spectral index, and spectral running, which are sensitive to the underlying emission mechanism and propagation effects, including scattering and scintillation in the interstellar and intergalactic medium, have consistently emerged as important discriminants between repeating and apparently non-repeating FRBs. Several recent studies have incorporated these features into both supervised and unsupervised analyses, revealing that repeaters tend to exhibit narrower and more structured spectra, while non-repeaters often show broader, irreg-

ular profiles [13, 14, 19, 20]. However, despite these indications of heterogeneity, there is currently insufficient observational support to define firm subclasses within the repeater population. The available samples remain limited and affected by strong instrumental selection biases, which hinder the development of a robust taxonomy. Given these limitations, our analysis treats repeaters as a single class, avoiding premature sub-classification until future systematic observations can provide more definitive distinctions.

This paper builds upon recent advancements in FRBs classification by developing a systematic and interpretable machine learning framework designed to distinguish repeating from non-repeating sources. Our approach integrates three hybrid pipelines that combine dimensionality reduction with clustering – PCA + k-means, t-SNE + HDBSCAN, and t-SNE + Spectral Clustering – each evaluated under two feature configurations: one based solely on primary CHIME/FRB catalog observables, and another incorporating additional physically derived quantities. All pipelines are optimized through a grid search using a custom  $F_2$ -based scoring function to adjust the model parameters. This methodology complements previous approaches by integrating enriched feature representations with systematic clustering evaluation and optimization, contributing to robust classification performance.

In Section II, we describe the data used in this study, including catalog selection, preprocessing steps, identification of primary features, and the construction of derived features. In Section III, we describe our methodology in detail and implementation of our grid search. Section IV presents results and discussions comparing clustering performance, visualization diagnostics, and analysis of feature importance using PCA loadings, mutual information, and permutation-based  $F_2$  impact, as well as the identification of new repeating FRB candidates. Finally, conclusions are presented in Section V.

## II. DATA AND PREPROCESSING

### A. Data Acquisition and Preprocessing

The primary data set used in this work is the CHIME/FRB Catalog 1 [21]. This catalog contains more than 500 Fast Radio Bursts, including both repeating and apparently non-repeating sources, observed by between July 2018 and July 2019. Data include a variety of astrophysical observables such as flux densities, burst widths, and spectral properties derived from best-fit burst models [22]. To ensure the reliability of

the analysis, we first filter the dataset by removing the following sources that do not have flux measurements: FRB20190307A, FRB20190307B, FRB20190329B, FRB20190329C, FRB20190531A, and FRB20190531B. Additionally, we exclude duplicate entries of known repeating sources and remove rows with missing or misformed values.

### B. Primary Observables

Our analysis begins with a set of nine primary features directly extracted from the CHIME catalog, summarized in Table I.

TABLE I: Primary features extracted from the CHIME/FRB catalog.

Feature	Description
snr_fitb	Signal-to-noise ratio of the burst based on the FitBurst model.
dm_exc_ymw16	Extragalactic DM component, computed as the excess above the YMW16 Galactic model [23].
flux	Estimated flux density in Jy.
fluence	Integrated fluence in Jy ms.
width_fitb	Temporal width of the burst (ms) from the FitBurst best fit.
scat_time	Scattering timescale estimated by the FitBurst model (ms).
sp_idx	Spectral index assuming a power-law spectrum.
sp_run	Spectral running, i.e., curvature of the spectrum.
low_freq	Lowest frequency of detection (MHz).

These features were selected based on both physical motivation and previous studies, highlighting their relevance for distinguishing between repeating and non-repeating FRBs. Fluence ( $F_\nu$ ), peak flux ( $S_\nu$ ), burst width ( $\Delta t$ ), scattering timescale ( $\Delta t_{ST}$ ), and spectral index parameters ( $\gamma$ ,  $r$ ) capture critical aspects of FRB emission and propagation, including pulse morphology, plasma scattering, and spectral complexity. In particular, the spectral index  $\gamma$  and spectral running  $r$  were found to exhibit a strong correlation and jointly reflect the underlying radiation mechanism and medium inhomogeneities [19]. Furthermore, the lowest detectable frequency is retained due to its physical interpretability and greater robustness against instrumental truncation effects within the CHIME telescope observing band (400 – 800 MHz). As noted in [19], higher frequency bounds

are often unreliable due to limitations at the upper end of the telescope frequency band. Therefore, the use of `low_freq` helps mitigate systematic errors when analyzing spectral coverage.

### C. Derived Quantities

To complement the catalog-provided observables, we compute six physically motivated derived quantities using standard cosmological and radiative relations, following the methodology described in [24]. These quantities serve as proxies for the intrinsic energetics, luminosity, and coherence scale of FRB emission.

- **Redshift ( $z$ ):** The observed dispersion measure,  $DM_{\text{obs}}$ , consists of contributions from both local and extragalactic (EG) environments. This is expressed as:

$$DM_{\text{obs}} = DM_{\text{local}} + DM_{\text{EG}}(z), \quad (1)$$

where the local component is given in terms of contributions from the Milky Way interstellar medium (ISM) and the halo surrounding our galaxy:

$$DM_{\text{local}} = DM_{\text{ISM}} + DM_{\text{halo}}, \quad (2)$$

while the extragalactic component includes contributions from the intergalactic medium (IGM) and the host galaxy of the FRB:

$$DM_{\text{EG}} = DM_{\text{IGM}} + \frac{DM_{\text{host}}}{(1+z)}. \quad (3)$$

The redshift is estimated from the excess dispersion measure using the Macquart relation [25], under the assumption of a fully ionized intergalactic medium. The intergalactic DM component is modeled as:

$$DM_{\text{IGM}}(z) = \frac{3cH_0\Omega_b f_{\text{IGM}}}{8\pi G m_p} \chi \int_0^z \frac{(1+z') dz'}{E(z')}, \quad (4)$$

where  $f_{\text{IGM}}$  and  $\chi$  represent the fraction and ionized fraction of baryons in the IGM. The cosmic baryon density, the mass of the proton, and the speed of light are denoted by  $\Omega_b$ ,  $m_p$ , and  $c$ , respectively.  $E(z)$  captures the cosmological dependence of this equation through the dimensionless Hubble parameter given by<sup>1</sup>

<sup>1</sup> We assume a flat  $\Lambda$ CDM cosmology throughout this work, specifically adopting the Planck 2018 best-fit parameters [26].

$E(z) = \sqrt{\Omega_m(1+z)^3 + \Omega_\Lambda}$ . We numerically invert this relation to estimate the redshift from `dm_exc_ymw16` for each burst. The values for the constants used in this work are listed in Table II. The values adopted for  $DM_{\text{halo}}$  and  $DM_{\text{host}}$  were  $30 \text{ pc cm}^{-3}$  and  $70 \text{ pc cm}^{-3}$ , respectively [27–31].

- **Frequency width ( $\Delta\nu$ ):** The effective bandwidth is approximated as the difference between the highest and lowest observed frequencies, scaled by redshift:

$$\Delta\nu = (\nu_{\text{high}} - \nu_{\text{low}})(1+z), \quad (5)$$

where  $\nu_{\text{high}}$  and  $\nu_{\text{low}}$  are provided in the CHIME catalog.

- **The time width in the reference frame ( $\Delta t_r$ ):** The intrinsic time width is computed from the observed time width and corrected for cosmic expansion:

$$\Delta t_r = \frac{\Delta t}{(1+z)}. \quad (6)$$

- **Isotropic-equivalent energy ( $E$ ):** The isotropic-equivalent energy emitted by the burst is estimated as:

$$E = 4\pi D_L^2(z) \cdot \mathcal{F}_\nu \cdot \Delta\nu \cdot (1+z)^{-1}, \quad (7)$$

where  $\mathcal{F}_\nu$  is the fluence in  $\text{Jy} \cdot \text{ms}$ ,  $\Delta\nu$  is the bandwidth in Hz and  $D_L$  is the luminosity distance derived from the redshift.

- **Luminosity ( $L$ ):** The burst luminosity is defined as the energy emitted per unit time:

$$L = 4\pi D_L^2(z) S_\nu \nu_c, \quad (8)$$

where  $S_\nu$  is the peak flux and  $\nu_c$  is the observed peak frequency.

- **Brightness temperature ( $T_B$ ):** The effective brightness temperature is computed as:

$$T_B = \frac{S_\nu \lambda^2}{2k_B} \left( \frac{D_L}{R} \right)^2, \quad (9)$$

where  $S_\nu$  is the peak flux density in Jy,  $\lambda = c/\nu$  is the observing wavelength, and  $R = c\Delta t/(1+z)$  is the transverse size of the emission region inferred from the burst duration.

TABLE II: Physical constants and cosmological parameters adopted in this work.

Constant	Value	Unit
$c$	$2.998 \times 10^8$	m/s
$G$	$6.674 \times 10^{-11}$	$\text{m}^3/\text{kg}/\text{s}^2$
$k_B$	$1.381 \times 10^{-23}$	J/K
$m_p$	$1.673 \times 10^{-27}$	kg
$H_0$	67.4	km/s/Mpc
$\Omega_m$	0.315	–
$\Omega_\Lambda$	0.685	–
$\Omega_b$	0.049	–
$f_{\text{IGM}}$	0.83	–
$\chi$	7/8	–

We organize our analysis around two complementary scenarios for feature selection:

- **Case I – Primary Only:** This configuration includes the nine features provided by the catalog discussed previously (signal-to-noise ratio, flux, fluence, width, scattering time, spectral index, spectral running, excess dispersion measure, and lowest detected frequency). This scenario isolates the predictive power of directly observed quantities, independent of any physical modeling assumptions.
- **Case II – Primary + Derived:** In this extended configuration, we supplement the nine primary features with six additional derived variables: redshift, frequency width, time width, isotropic-equivalent energy, luminosity, and brightness temperature. These quantities are computed using standard astrophysical relations and cosmological assumptions. Logarithmic transformations were applied to energy, luminosity, and brightness temperature to account for their broad dynamic range. While model-dependent, these derived quantities encode key physical constraints related to emission energetics and propagation physics, serving as proxies for intrinsic burst properties.

This dual approach allows us to evaluate the influence of incorporating physically derived parameters on the performance of unsupervised clustering algorithms. In both cases, the selected features are standardized using z-score normalization before being fed into dimensionality reduction and clustering pipelines. The full modeling process—including dimensionality reduction, grid search for clustering optimization, and evaluation of clustering performance—is described in the following sections.



### III. METHODOLOGY

Unsupervised learning provides a natural framework for exploring latent structures within FRB populations without imposing predefined labels. In this section, we detail the methodological pipeline used to uncover meaningful groupings and identify repeater candidates based on observational features.

#### A. Dimensionality Reduction and Clustering Algorithms

To identify the structure within the FRB parameter space, we employ unsupervised machine learning techniques that combine dimensionality reduction with clustering. This approach allows us to visualize complex relationships and reveal natural groupings in the data. Beyond its utility for visualization, dimensionality reduction also improves the efficiency and effectiveness of learning algorithms by alleviating issues associated with high-dimensional feature spaces, such as increased sparsity, redundancy, and susceptibility to overfitting. Before clustering, the high-dimensional feature space is projected onto a two-dimensional subspace to simplify the structure while preserving key relationships. In this work, we consider two-dimensionality reduction methods:

- **Principal Component Analysis (PCA):** PCA is a linear transformation technique that projects the data onto orthogonal axes (principal components), capturing directions with the highest variance. It is computationally efficient and preserves global structure, but may struggle to capture non-linear patterns in the data [32].
- **t-distributed Stochastic Neighbor Embedding (t-SNE):** t-SNE is a non-linear dimensionality reduction method, particularly effective for visualizing local structures in high-dimensional data. Preserve neighborhood relationships by modeling pairwise similarities and projecting them into a lower-dimensional space. t-SNE is sensitive to hyperparameters such as perplexity and exaggeration, which control the balance between local and global structure preservation [33, 34].

Once the data are projected onto a two-dimensional plane, we apply clustering algorithms to group FRBs with similar features. The clustering algorithms used in this study are:

- **k-means clustering:** A classic centroid-based algorithm, first introduced in [35], which partitions

data into  $k$  clusters by minimizing the variance within each group. It assumes spherical clusters and equal cluster sizes.

- **HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise):** A density-based algorithm that identifies clusters of varying shapes and densities. It is robust to noise and does not require the number of clusters to be specified in advance. HDBSCAN labels points in low-density regions as noise, making it suitable for irregular astrophysical populations [36].
- **Spectral Clustering:** This graph-based algorithm constructs a similarity graph among data points and uses the eigenvectors of the Laplacian to group the data into clusters. It is particularly effective in detecting non-convex structures and complex cluster boundaries that k-means may fail to capture [37].

We then evaluate three combinations of dimensionality reduction and clustering, selected to balance interpretability, flexibility, and computational efficiency:

- **PCA + k-means:** A fast, fully linear baseline that projects features with PCA and applies k-means to the first two principal components.
- **t-SNE + HDBSCAN:** A flexible, noise-aware configuration that non-linearly projects the data and identifies arbitrarily shaped clusters while filtering out low-density outliers.
- **t-SNE + Spectral Clustering:** A hybrid model that uses the expressive t-SNE projection followed by a graph-based clustering algorithm to capture nuanced separations.

Each combination is independently optimized through grid search to identify the optimal configuration. The goal is to assess whether natural groupings – particularly between repeating and non-repeating FRBs – emerge under different algorithmic assumptions.

#### B. Grid Search and Custom Scoring

To determine the optimal configuration for each algorithmic pipeline, we perform a grid search over the relevant hyperparameter space. This process aims to identify parameter combinations that maximize clustering quality with respect to distinguishing repeating from non-repeating FRBs, while also avoiding overfitting and over-fragmentation.

Each combination of dimensionality reduction and clustering – namely, PCA + k-means, t-SNE + HDBSCAN, and t-SNE + Spectral Clustering – has its own set of hyperparameters. For each method, we evaluate all permutations of the following parameter values:

- **t-SNE parameters:**

- `perplexity`  $\in \{30, 50\}$  – controls the effective number of neighbors.
- `early_exaggeration`  $\in \{8, 12\}$  – influences the tightness of clusters in the early optimization stages.

- **HDBSCAN parameters:**

- `min_cluster_size`  $\in \{10, 15, 20\}$  – sets the minimum number of points required to form a cluster.
- `min_samples`  $\in \{1, 3, 5\}$  – defines the minimum density threshold for a point to be considered a core.

- **Spectral Clustering parameters:**

- Number of clusters  $k \in \{2, 3, 4\}$ .
- Assignment method  $\in \{\text{kmeans}, \text{discretize}\}$  – determines how clusters are extracted from the eigenvectors.

- **k-means parameters:**

- Number of clusters  $k \in \{2, 3, 4\}$ .

For each combination of parameters, the model generates cluster assignments for the input data. We then evaluate cluster quality using several standard classification metrics, computed by assigning a binary repeater label to each cluster based on the proportion of known repeaters it contains. This framework enables the calculation of precision, recall and the  $F_2$ -score for each configuration.

The *precision* is defined as

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (10)$$

where  $TP$  and  $FP$  represent the number of true and false positive predictions, respectively. It quantifies the proportion of bursts predicted as repeaters that are indeed known repeaters, penalizing models that incorrectly classify non-repeaters.

The *recall* is defined as

$$\text{Recall} = \frac{TP}{TP + FN}, \quad (11)$$

where  $FN$  is the number of false negatives. Measures the proportion of true repeaters successfully identified by the model, favoring configurations that minimize missed detections.

To balance these two metrics, we compute the  $F_2$ -score, a weighted harmonic mean that prioritizes recall over precision:

$$F_2 = \frac{5 \cdot \text{Precision} \cdot \text{Recall}}{4 \cdot \text{Precision} + \text{Recall}}. \quad (12)$$

This formulation is particularly appropriate for our astrophysical application, where failing to identify a potential repeater is more detrimental than overpredicting one.

For each combination of parameters, the model generates cluster assignments for the input data. We then evaluate the quality of the clustering using a custom objective function that balances three aspects: classification performance, cluster interpretability, and robustness to noise. The scoring function is defined as:

$$\text{Score} = F_2 - \frac{\alpha(n_c - 2)^2}{10} - \beta \cdot \frac{n_{\text{noise}}}{N}, \quad (13)$$

where  $F_2$  represents the  $F_2$ -score computed from binary predictions distinguishing repeaters from non-repeaters. The term  $n_c$  denotes the number of identified clusters, while  $n_{\text{noise}}$  corresponds to the number of data points labeled as noise – applicable only in the case of HDBSCAN.  $N$  is the total number of samples in the data set. The parameter  $\alpha = 1.0$  controls the penalty associated with deviations from a binary classification (with the ideal case being  $n_c = 2$ ), and  $\beta = 0.3$  penalizes the proportion of points left unclassified due to noise.

This metric was designed to promote solutions that (i) maximize the correct identification of repeaters, (ii) avoid overfragmentation of the dataset into too many clusters, and (iii) reduce sensitivity to outliers or sparsely populated regions. The weights  $\alpha$  and  $\beta$  were empirically chosen to balance the relative importance of interpretability and robustness without overpowering the  $F_2$  contribution. By incorporating these penalties directly into the evaluation, our grid search selects models that are both effective and physically interpretable in the astrophysical context of FRB classification.

All computational analyses conducted in this study were implemented using Python programming language. The core machine learning components, including dimensionality reduction, clustering algorithms, and performance evaluation metrics, were primarily built upon the `scikit-learn` library [38]. In addition, the `hdbscan` package [39] was employed for density-based clustering.

## IV. RESULTS AND DISCUSSIONS

### A. Clustering Performance

We evaluated three unsupervised clustering pipelines: PCA followed by k-means, t-SNE followed by Spectral Clustering, and t-SNE followed by HDBSCAN. Each method was applied to two feature sets: one with the nine primary catalog observables (Case I), and the other with all fifteen features, including derived quantities (Case II), as presented in Section II. Hyperparameters were optimized through grid search (see Section III).

Table III summarizes the clustering performance obtained for each combination of dimensionality reduction and clustering method, evaluated separately for Case I and Case II. The reported metrics include precision, recall, and the  $F_2$  score. Among all configurations, the best performance was achieved by the t-SNE + Spectral Clustering pipeline applied to the full feature set (Case II), with an  $F_2$  score of 0.76. This result outperformed both PCA + k-means ( $F_2 = 0.71$ ) and t-SNE + HDBSCAN ( $F_2 = 0.70$ ) using the same features. For Case I, the best  $F_2$  score was obtained with PCA + k-means (0.73), closely followed by t-SNE + Spectral Clustering (0.72). These results suggest that while Spectral Clustering is effective overall, PCA combined with k-means slightly outperforms it under the restricted feature configuration. In other words, although non-linear dimensionality reduction methods can capture more complex structures, PCA still provides a robust performance when only primary features are available.

Furthermore, these results demonstrate the benefit of incorporating additional physically motivated features that improve separability in feature space, as well as the advantage of using non-linear dimensionality reduction. The improvement in  $F_2$  score when including derived quantities in most pipelines indicates that physical properties such as redshift and luminosity, for example, carry a discriminative potential to distinguish repeater behavior. In particular, these features trace extragalactic distances and energetics, potentially associated with progenitor environments or emission mechanisms. Moreover, the high recall scores – especially in spectral clustering scenarios – are particularly relevant for identifying FRB repeaters, since missing a true repeater is more costly than allowing a few false positives. These results support the interpretation that repeaters and non-repeaters arise from distinct astrophysical subpopulations.

In addition to quantitative metrics, we qualitatively examine the two-dimensional clustering results using

scatter plots of the t-SNE and PCA projections. Figures 1 and 2 present these visualizations for the primary and the primary + secondary feature sets, respectively. In the left panels of both figures, the true repeater labels – flagged by CHIME – are shown in red and blue, while the cluster assignments in the right panels are displayed in orange and blue.

For the primary-only case (Figure 1), the t-SNE + Spectral Clustering combination (middle panel) exhibits the most coherent separation between repeaters and non-repeaters, with minimal overlap and compact group structure. PCA + k-means (upper panel) also achieves a degree of separation, although its linear projection limits cluster boundaries. The t-SNE + HDBSCAN model (lower panel) captures partial structure but is affected by noise points and more diffuse groupings. These results suggest that non-linear embeddings better capture the intrinsic structure of FRB properties that distinguish repeaters from one-off bursts, while density-based clustering appears more sensitive to noise, likely reflecting the variability inherent in FRB observations.

When derived features are added (Figure 2), visual separability improves across all three models. The inclusion of physical quantities such as redshift and luminosity enhances manifold structure and highlights latent cluster shapes. Notably, the t-SNE + Spectral Clustering model (middle panel) maintains compact clusters and a clearer concentration of repeaters, consistent with its higher  $F_2$  score. PCA + k-means (upper panel) also benefits from the extended feature set, producing nearly disjoint clusters. Meanwhile, t-SNE + HDBSCAN (lower panel) reveals a more coherent structure with fewer noise-labeled points. These improvements reinforce the astrophysical expectation that incorporating distance and energy indicators is essential for uncovering meaningful clustering of FRBs sources.

To visualize the classification performance of each clustering pipeline, we present confusion matrices for all three models using both the primary-only and full feature sets. Figures 3 and 4 show the true vs. predicted repeater labels for optimal configurations selected via grid search.

For the primary-only case (Figure 3), the PCA + k-means configuration achieves a moderate balance between precision and recall, though some group confusion. The t-SNE + Spectral Clustering model improves both metrics, as seen in its denser diagonal entries. The t-SNE + HDBSCAN configuration yields slightly higher recall but introduces more noise-driven misclassifications. These results highlight trade-offs between precision and completeness across clustering approaches,

TABLE III: Clustering performance metrics for each combination of dimensionality reduction and clustering method, grouped by feature set. The custom score includes a penalty for noise and excessive cluster count.

Feature Set	Method	Precision	Recall	F <sub>2</sub> Score	F <sub>2</sub> Custom Score
Case I	PCA + k-means	0.61	0.77	0.73	0.73
	t-SNE + HDBSCAN	0.39	0.84	0.68	0.65
	t-SNE + Spectral Clustering	0.42	0.87	0.72	0.72
Case II	PCA + k-means	0.36	0.95	0.71	0.71
	t-SNE + HDBSCAN	0.67	0.71	0.70	0.69
	t-SNE + Spectral Clustering	0.43	0.95	0.76	0.76

a critical consideration for identifying repeaters, given their rarity and the inherent observational uncertainties.

When derived features are added (Figure 4), we observe clear gains in true positive detection. Notably, both t-SNE + Spectral and PCA + k-means achieve strong diagonal dominance, especially in correctly identifying repeaters. t-SNE + HDBSCAN performs well, but at the cost of a slightly higher false-positive rate. These visual insights align closely with the metric-based results from Table III, further validating the clustering assignments.

### B. Feature Importance

To identify which parameters contribute most to distinguishing repeaters from non-repeaters, we employed three complementary strategies: principal component analysis (PCA) loadings, mutual information (MI), and permutation importance based on the F<sub>2</sub> score. Each method offers a distinct perspective on feature relevance: PCA loadings reflect the contribution of each variable to the main axes of variance; mutual information quantifies the non-linear dependence between each feature and the repeater label; and permutation importance measures the drop in predictive performance when a feature is randomly shuffled.

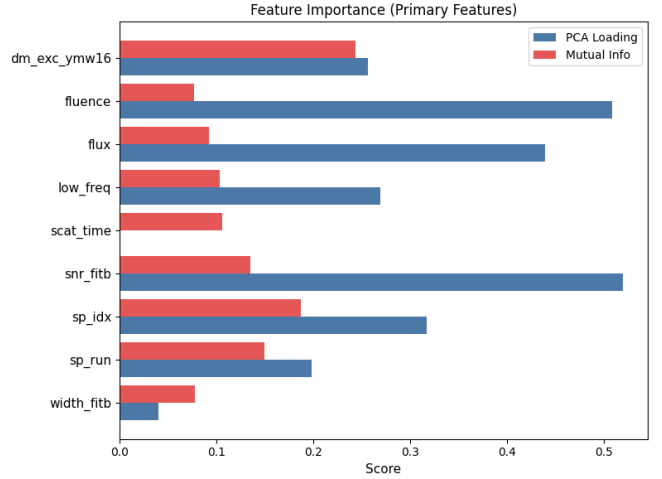
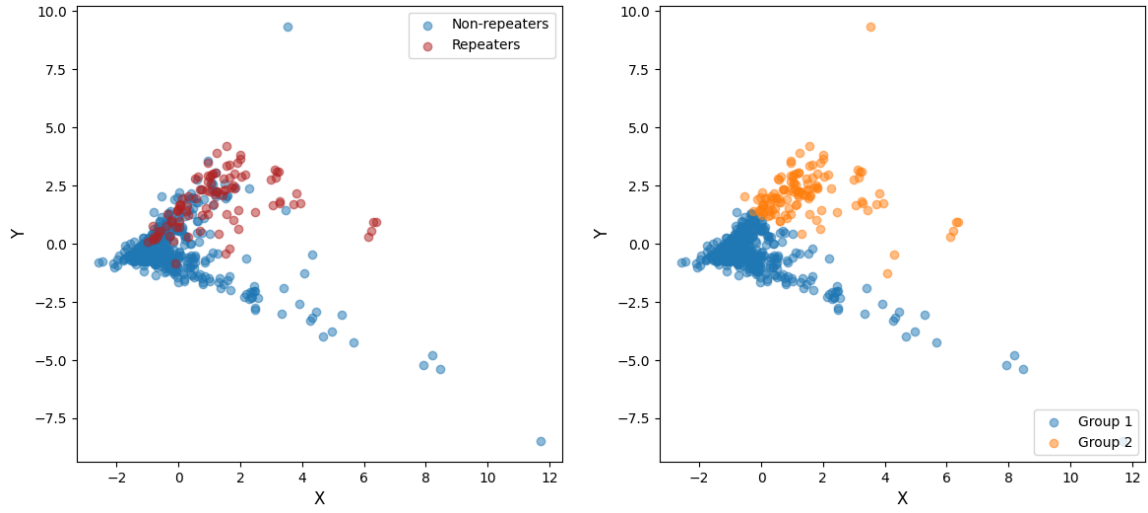


FIG. 5: PCA loadings and mutual information scores for the **primary-only** feature set. The most influential features are related to signal intensity, propagation effects, and spectral shape.

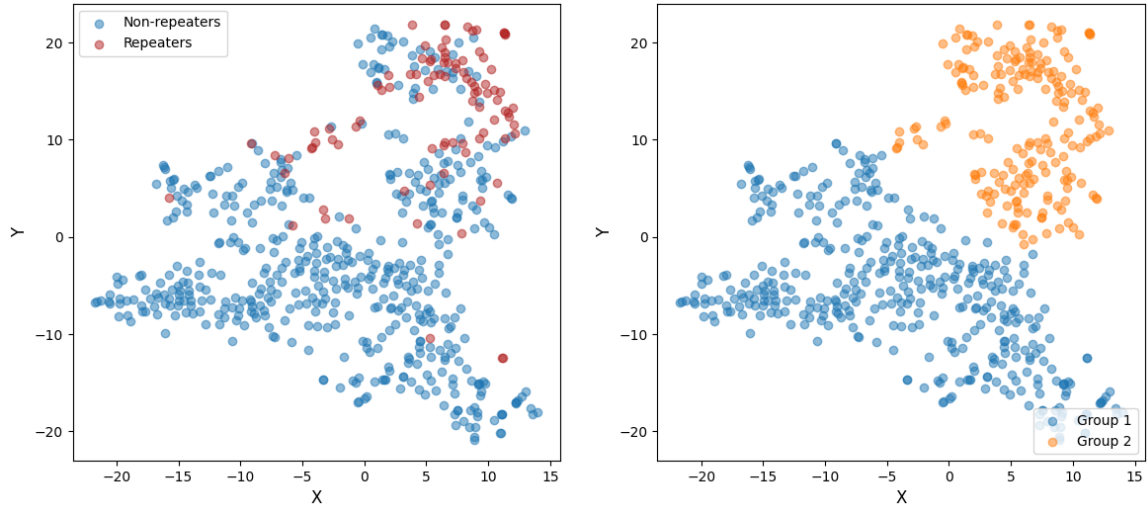
Figures 5 and 6 show the PCA loadings and mutual information scores for the primary-only and full feature sets, respectively. In the primary-only case, PCA loadings reveal that intrinsic properties like `snr_fitb`, `fluence`, and `flux` dominate the first principal component, indicating that signal strength and burst amplitude are the main sources of variation. However, when derived features are included, the variance shifts toward properties such as `log_luminosity`, `dm_exc_ymw16`, and `redshift`, emphasizing their importance in explaining the underlying FRB population structure.

The results of mutual information align with these findings: `dm_exc_ymw16` and `sp_idx` emerge as key predictors in both settings, while `redshift` and `log_temperature` gain importance when derived features were included. The consistent presence of `sp_idx` (spectral index) as a discriminative variable reinforces the idea that the spectral behavior of bursts plays a central role in repetition. These results support the interpre-

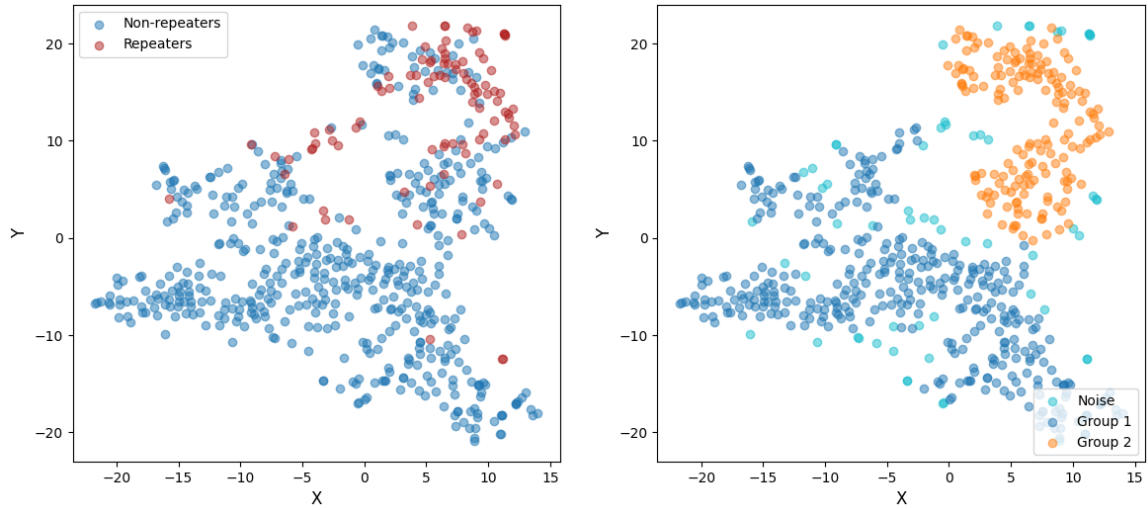




(a) PCA + k-means



(b) t-SNE + Spectral Clustering



(c) t-SNE + HDBSCAN

FIG. 1: Clustering visualizations using the **primary-only features**. Each panel shows a 2D projection of the FRBs colored by cluster assignment: (a) PCA + k-means, (b) t-SNE + Spectral Clustering, and (c) t-SNE + HDBSCAN.

tation that a combination of propagation and intrinsic properties influences repetition behavior.

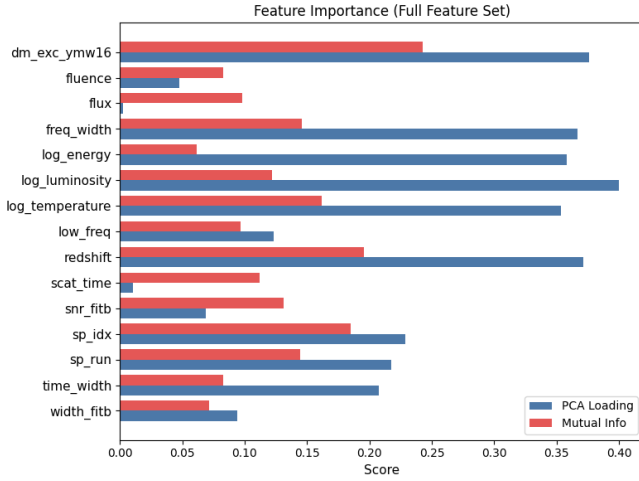


FIG. 6: PCA loadings and mutual information scores for the **full feature set** (primary + derived). The top contributors are *log\_luminosity*, *dm\_exc\_ymw16*, and *redshift*.

To evaluate the impact of each feature on classification performance, we computed permutation importance by measuring the drop in the  $F_2$  score when each feature was shuffled independently. This analysis was performed for each of the three clustering pipelines. Figures 7 and 8 summarize the  $F_2$  degradation for the primary-only and full feature sets, respectively.

In the primary-only case, *low\_freq*, *snr\_fitb*, and *sp\_idx* produced the greatest  $F_2$  drops, although the results were highly dependent on the clustering method. For instance, in the PCA + k-means pipeline, *fluence* and *low\_freq* have the strongest impact, while for t-SNE + HDBSCAN, *low\_freq* and *sp\_idx* were most impactful. With the full feature set, the most relevant variables varied across models: *width\_fitb*, *time\_width*, and *redshift* stood out in the HDBSCAN model, whereas *fluence*, *low\_freq*, and *dm\_exc\_ymw16* led in the spectral clustering configuration. These findings highlight the complex, model-dependent nature of feature relevance, but consistently reaffirm the central role of *dm\_exc\_ymw16*, *low\_freq*, and time-resolved or distance-based quantities.

Importantly, our feature importance results agree with previous studies such as [19], which used unsupervised machine learning techniques to assess the discriminative power among CHIME FRBs. In particular, the high ranking of *sp\_idx* (spectral index) and *sp\_run* (spectral running) in both mutual information and  $F_2$ -based permutation tests reinforce earlier findings that

frequency-dependent spectral behavior encodes key information about repetition. These spectral parameters likely reflect intrinsic emission processes and propagation effects that differ between repeaters and non-repeaters, suggesting that machine learning models are naturally sensitive to such signatures. Their persistent importance in our results could indicate that the pipelines capture genuine physical features that distinguish FRB types.

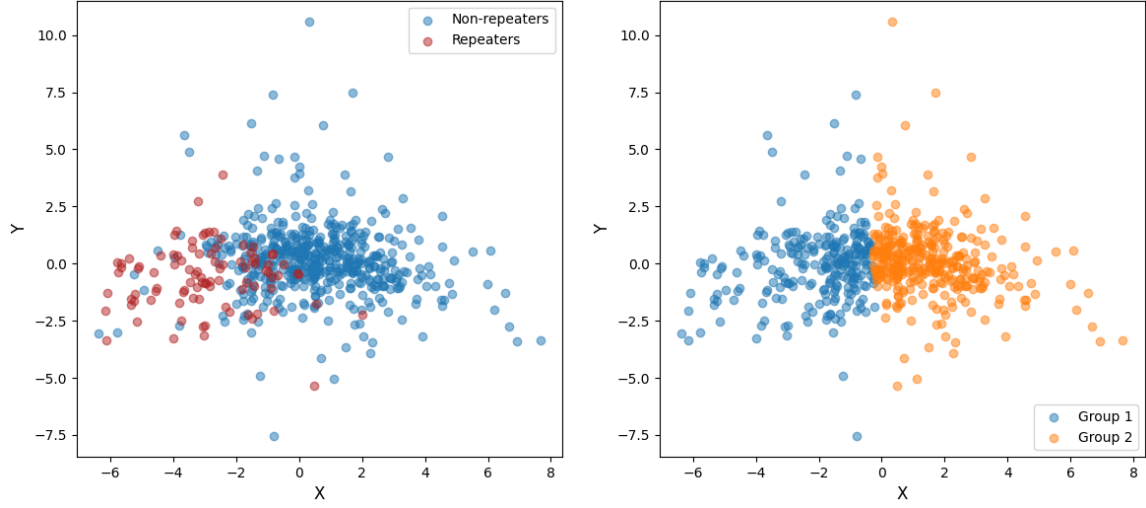
### C. Candidate Repeaters

To identify potential new repeating FRB sources, we implemented a voting scheme that combines the outputs of our three unsupervised clustering pipelines: PCA + k-means, t-SNE + Spectral Clustering, and t-SNE + HDBSCAN. For each method, FRBs were grouped based on their cluster assignments, and clusters where more than 15% of members were tagged as known repeaters were labeled as “repeater-dominant.” Any FRB in such a cluster received a tentative repeater classification for that method.

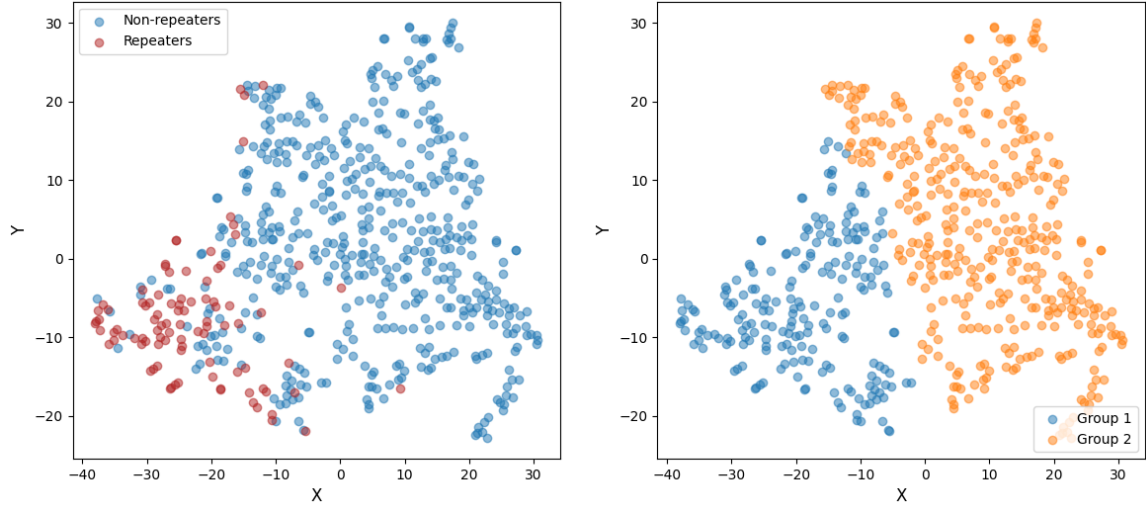
An FRB was considered a candidate repeater only if the three clustering methods labeled it as such. This criterion provides a strong consistency filter and reduces false positives. We applied this analysis to the two configurations considered in this study: (i) using only the primary observational features from the CHIME catalog, and (ii) using the full feature set, which includes both the primary and derived physical quantities described in Section II.

Tables A4 and A4 list the FRBs predicted to be repeaters in each configuration, with confirmed repeaters from CHIME/FRB (2023) Catalog highlighted in bold. From the primary-only configuration, we identified 37 candidate repeaters. The full feature set configuration yielded 41 candidates. We compared these predictions with the updated CHIME/FRB (2023) Catalog [40], which lists 25 new repeating sources, including six reclassifications of FRBs previously considered as non-repeaters.

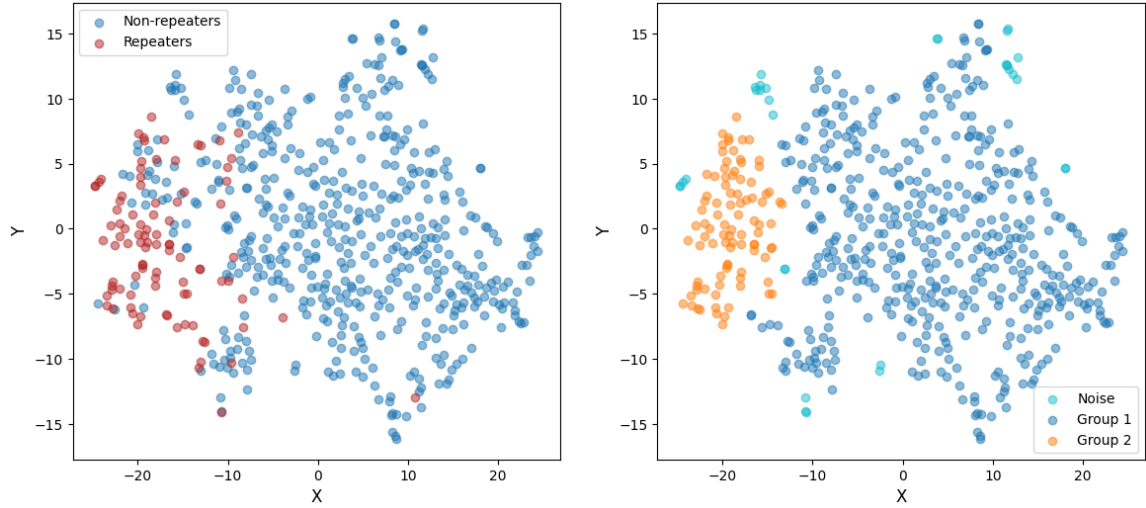
Upon cross-checking, we verified that two of the FRBs identified in our primary-only candidate set are now confirmed repeaters in CHIME/FRB (2023) Catalog: **FRB20190110C**, and **FRB20190430C**. For the full feature set, three matches were found: **FRB20190113A**, **FRB20190226B**, and **FRB20190430C**. Considering both configurations, we correctly predicted four of the six reclassified FRBs by the CHIME/FRB (2023) catalog, as they appeared in repeater-dominant clusters across all methods considered. This agreement supports the cred-



(a) PCA + k-means



(b) t-SNE + Spectral Clustering



(c) t-SNE + HDBSCAN

FIG. 2: Clustering visualizations using the **full feature set** (primary + derived). The panels show the same layout and clustering combinations as in Figure 1, now using the extended set of physical quantities.

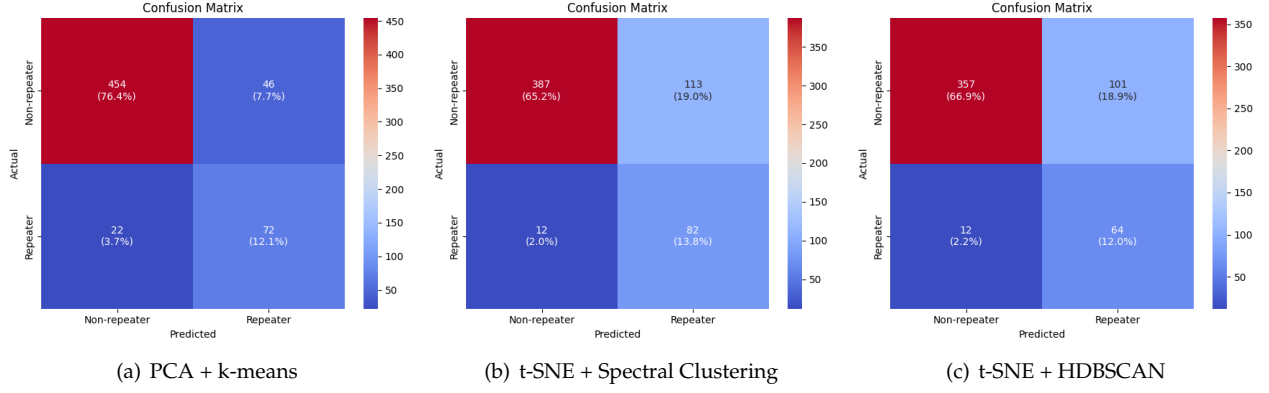


FIG. 3: Confusion matrices for the **primary-only feature set**. These plots show true vs. predicted repeater classifications for each clustering method. The t-SNE + Spectral Clustering configuration demonstrates clearer separation, while HDBSCAN provides good recall with moderate false positives.

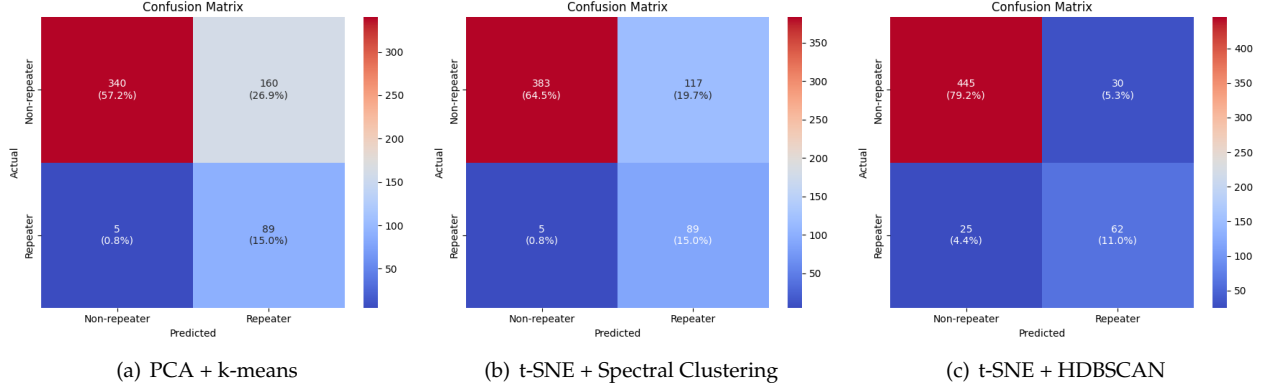


FIG. 4: Confusion matrices for the **full feature set** (primary + secondary). All three configurations show enhanced classification performance with derived quantities, particularly in the correct identification of repeaters.

ibility of our approach, given that these predictions were made independently of the updated classification and suggests that additional sources flagged by our models may also be repeaters awaiting confirmation.

## V. CONCLUSIONS

We presented a comparative study of unsupervised machine learning techniques aimed at classifying FRBs and identifying new candidate repeaters. Our methodology combines dimensionality reduction techniques – PCA and t-SNE – with clustering algorithms including k-means, Spectral Clustering, and HDBSCAN. We evaluated their performance across two feature configurations: one using only primary observables from the CHIME/FRB Catalog 1, and another combining these with a set of astrophysically motivated derived quantities such as redshift, isotropic energy, and luminosity.

The performance of each method was assessed using standard metrics, complemented by a custom scoring criterion based on the  $F_2$  score that penalizes over-fragmentation and excessive noise. The best clustering performance was achieved by t-SNE + Spectral Clustering with the full feature set, supporting the view that physically informed features enhance the separability between repeaters and non-repeaters. More broadly, our results demonstrate that t-SNE-based approaches are particularly effective at capturing the complex, non-linear structure underlying the FRB parameter space.

To identify and understand the key factors behind successful classification, we evaluated feature importance using three complementary techniques: PCA loadings, mutual information with the repeater label, and permutation importance based on  $F_2$  score degradation. Features such as `dm.exc_ymw16`, `redshift`, and `sp_idx` consistently emerged as highly informative, re-



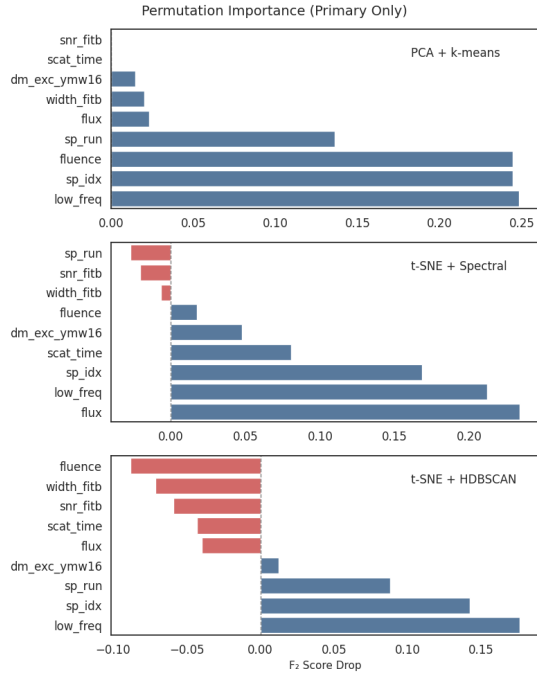


FIG. 7: Permutation importance ( $F_2$  drop) for the **primary-only feature set** using t-SNE + HDBSCAN, t-SNE + Spectral Clustering, and PCA + k-means.

enforcing previous findings and independently confirming the central role of spectral properties in distinguishing repeaters from non-repeaters [19].

In addition, we proposed new candidate repeaters through a voting scheme across all clustering pipelines. FRBs that consistently appeared in repeater-dominant clusters but had not been previously labeled as repeaters were flagged as potential repeaters. This yielded 37 candidates using only primary features, and 41 when including derived quantities. Upon cross-referencing with the CHIME/FRB (2023) Catalog, we found that some of our predictions aligned with sources recently reclassified as repeaters, suggesting that our methods successfully captured underlying patterns and demonstrated the efficacy of uncovering latent repeater behavior from observational features alone.

Taken together, these findings underscore the potential of unsupervised learning, especially when guided by astrophysically motivated features, to uncover latent structure in FRB populations and to support repeater

classification in a data-driven yet physically grounded manner. The framework developed here is readily adaptable to future FRB catalogs and extended feature configurations, and offers a foundation for more refined approaches, including probabilistic clustering, semi-supervised models, and time-resolved analyses of burst activity.

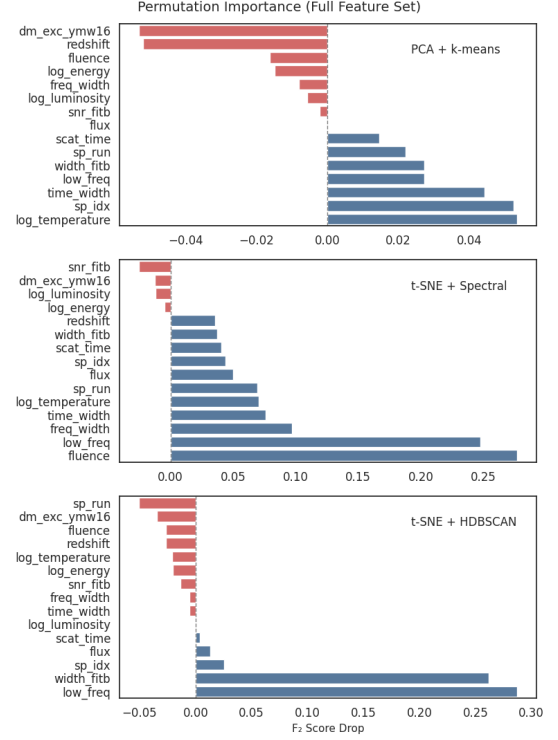


FIG. 8: Permutation importance ( $F_2$  drop) for the **full feature set** using t-SNE + HDBSCAN, t-SNE + Spectral Clustering, and PCA + k-means.

## ACKNOWLEDGEMENTS

JASF acknowledges support from the National Research Foundation of South Africa. WSHR acknowledges partial support from FAPES and CNPq.

## DATA AVAILABILITY

The data used in this study are available from the CHIME/FRB public catalog at <https://www.chime-frb.ca/catalog>.

[1] Duncan R Lorimer, Matthew Bailes, Maura Ann McLaughlin, David J Narkevic, and Fronev Crawford. A bright millisecond radio burst of extragalactic origin.

Science, 318(5851):777–780, 2007.

[2] Bing Zhang. The physics of fast radio bursts. *Rev. Mod. Phys.*, 95:035005, Sep 2023.

- [3] Emily Petroff, JWT Hessels, and DR Lorimer. Fast radio bursts. The Astronomy and Astrophysics Review, 27:1–75, 2019.
- [4] Paz Beniamini and Pawan Kumar. Hybrid pulsar-magnetar model for frb 20191221a. Monthly Notices of the Royal Astronomical Society, 519(4):5345–5351, 2023.
- [5] Sergei B Popov. Origin of sources of repeating fast radio bursts with periodicity in close binary systems. Research Notes of the AAS, 4(6):98, 2020.
- [6] Jing-Tong Xing and Tong Liu. Short-lived repeating fast radio bursts from tidal disruption of white dwarfs by intermediate-mass black holes. Monthly Notices of the Royal Astronomical Society: Letters, 528(1):L152–L156, 2024.
- [7] Dang Pham, Matthew J Hopkins, Chris Lintott, Michele T Bannister, and Hanno Rein. Fast radio bursts and interstellar objects. The Astrophysical Journal, 977(2):232, 2024.
- [8] Tanmay Vachaspati. Cosmic sparks from superconducting strings. Physical Review Letters, 101(14):141301, 2008.
- [9] Jiaying Xu, Yi Feng, Di Li, Pei Wang, Yongkun Zhang, Jintao Xie, Huaxi Chen, Han Wang, Zhixuan Kang, Jingjing Hu, et al. Blinkverse: A database of fast radio bursts. Universe, 9(7):330, 2023.
- [10] Mandana Amiri, Kevin Bandura, Anja Boskovic, Tianyue Chen, Jean-François Cliche, Meiling Deng, Nolan Denman, Matt Dobbs, Mateus Fandino, Simon Foreman, et al. An overview of chime, the canadian hydrogen intensity mapping experiment. The Astrophysical Journal Supplement Series, 261(2):29, 2022.
- [11] Vikram Ravi. The prevalence of repeating fast radio bursts. Nature Astronomy, 3(10):928–931, 2019.
- [12] Shotaro Yamasaki, Tomotsugu Goto, Chih-Teng Ling, and Tetsuya Hashimoto. The true fraction of repeating fast radio bursts revealed through chime source count evolution. Monthly Notices of the Royal Astronomical Society, 527(4):11158–11166, 2024.
- [13] Ziggy Pleunis, Deborah C Good, Victoria M Kaspi, Ryan Mckinven, Scott M Ransom, Paul Scholz, Kevin Bandura, Mohit Bhardwaj, PJ Boyle, Charanjot Brar, et al. Fast radio burst morphology in the first chime/frb catalog. The Astrophysical Journal, 923(1):1, 2021.
- [14] Bo Han Chen, Tetsuya Hashimoto, Tomotsugu Goto, Seong Jin Kim, Daryl Joe D Santos, Alvina YL On, Ting-Yi Lu, and Tiger YY Hsiao. Uncloaking hidden repeating fast radio bursts with unsupervised machine learning. Monthly Notices of the Royal Astronomical Society, 509(1):1227–1236, 2022.
- [15] Shruti Bhatporia, Anthony Walters, Jeff Murugan, and Amanda Weltman. A topological data analysis of the chime/frb catalogues. arXiv preprint arXiv:2311.03456, 2023.
- [16] Jia-Wei Luo, Jia-Ming Zhu-Ge, and Bing Zhang. Machine learning classification of chime fast radio bursts–i. supervised methods. Monthly Notices of the Royal Astronomical Society, 518(2):1629–1641, 2023.
- [17] X Yang, SB Zhang, JS Wang, and XF Wu. Classifying frb spectrograms using nonlinear dimensionality reduction techniques. Monthly Notices of the Royal Astronomical Society, 522(3):4342–4351, 2023.
- [18] CR García, Diego F Torres, Jia-Ming Zhu-Ge, and Bing Zhang. Separating repeating fast radio bursts using the minimum spanning tree as an unsupervised methodology. The Astrophysical Journal, 977(2):273, 2024.
- [19] Wan-Peng Sun, Ji-Guo Zhang, Yichao Li, Wan-Ting Hou, Fu-Wen Zhang, Jing-Fei Zhang, and Xin Zhang. Exploring the key features of repeating fast radio bursts with machine learning. The Astrophysical Journal, 980(2):185, 2025.
- [20] Da-Chun Qiang, Jie Zheng, Zhi-Qiang You, and Sheng Yang. Unsupervised machine learning for classifying chime fast radio bursts and investigating empirical relations. The Astrophysical Journal, 982(1):16, 2025.
- [21] Mandana Amiri, Bridget C Andersen, Kevin Bandura, Sabrina Berger, Mohit Bhardwaj, Michelle M Boyce, PJ Boyle, Charanjot Brar, Daniela Breitman, Tomas Casanelli, et al. The first chime/frb fast radio burst catalog. The Astrophysical Journal Supplement Series, 257(2):59, 2021.
- [22] E. Fonseca, Z. Pleunis, D. Breitman, K. R. Sand, B. Kharel, P. J. Boyle, C. Brar, U. Giri, V. M. Kaspi, K. W. Masui, B. W. Meyers, C. Patel, P. Scholz, and K. Smith. Modeling the morphology of fast radio bursts and radio pulsars with fitburst. The Astrophysical Journal Supplement Series, 271(2):49, mar 2024.
- [23] JM Yao, RN Manchester, and N Wang. A new electron-density model for estimation of pulsar and frb distances. The Astrophysical Journal, 835(1):29, 2017.
- [24] Jia-Ming Zhu-Ge, Jia-Wei Luo, and Bing Zhang. Machine learning classification of chime fast radio bursts–ii. unsupervised methods. Monthly Notices of the Royal Astronomical Society, 519(2):1823–1836, 2023.
- [25] Wei Deng and Bing Zhang. Cosmological implications of fast radio burst/gamma-ray burst associations. The Astrophysical Journal Letters, 783(2):L35, 2014.
- [26] N. Aghanim et al. Planck 2018 results. VI. Cosmological parameters. Astron. Astrophys., 641:A6, 2020. [Erratum: Astron. Astrophys. 652, C4 (2021)].
- [27] Jun Xu and J. L. Han. Extragalactic dispersion measures of fast radio bursts. Research in Astronomy and Astrophysics, 15(10):1629, oct 2015.
- [28] Shotaro Yamasaki and Tomonori Totani. The galactic halo contribution to the dispersion measure of extragalactic fast radio bursts. The Astrophysical Journal, 888(2):105, jan 2020.
- [29] K. Dolag, B. M. Gaensler, A. M. Beck, and M. C. Beck. Constraints on the distribution and energetics of fast radio bursts using cosmological hydrodynamic simulations. Monthly Notices of the Royal Astronomical Society, 451(4):4277–4289, 06 2015.
- [30] Tetsuya Hashimoto, Tomotsugu Goto, Alvina Y L On, Ting-Yi Lu, Daryl Joe D Santos, Simon C-C Ho, Seong Jin Kim, Ting-Wen Wang, and Tiger Y-Y Hsiao. No redshift

- evolution of non-repeating fast radio burst rates. Monthly Notices of the Royal Astronomical Society, 498(3):3927–3945, 08 2020.
- [31] Z Li, H Gao, J-J Wei, Y-P Yang, B Zhang, and Z-H Zhu. Cosmology-insensitive estimate of igm baryon mass fraction from five localized fast radio bursts. Monthly Notices of the Royal Astronomical Society: Letters, 496(1):L28–L32, 05 2020.
- [32] Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (pca). Computers & Geosciences, 19(3):303–342, 1993.
- [33] Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. Advances in neural information processing systems, 15, 2002.
- [34] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of machine learning research, 9(11), 2008.
- [35] Stuart Lloyd. Least squares quantization in pcm. IEEE transactions on information theory, 28(2):129–137, 1982.
- [36] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In Pacific-Asia conference on knowledge discovery and data mining, pages 160–172. Springer, 2013.
- [37] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. Advances in neural information processing systems, 14, 2001.
- [38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.
- [39] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. The Journal of Open Source Software, 2(11):205, 2017.
- [40] Bridget C Andersen, Kevin Bandura, Mohit Bhardwaj, PJ Boyle, Charanjot Brar, Tomas Cassanelli, S Chatterjee, Pragya Chawla, Amanda M Cook, Alice P Curtin, et al. Chime/frb discovery of 25 repeating fast radio burst sources. The Astrophysical Journal, 947(2):83, 2023.

## Appendix A: Appendix: Candidate Repeater Tables

TABLE A4: Predicted repeater candidates using **primary features only** (left) and **primary + secondary features** (right). FRBs already confirmed as repeaters in CHIME/FRB (2023) Catalog are indicated in **bold**.

FRB Source	FRB Source	FRB Source	FRB Source
FRB20180725A	FRB20190101B	FRB20180907E	FRB20190109A
FRB20180801A	<b>FRB20190110C</b>	FRB20180909A	FRB20190112A
FRB20180916C	FRB20190112A	FRB20180920A	<b>FRB20190113A</b>
FRB20181017B	FRB20190125A	FRB20180925A	FRB20190124E
FRB20181117C	FRB20190129A	FRB20181017B	FRB20190125A
FRB20181129B	FRB20190130B	FRB20181129B	FRB20190128C
FRB20181203B	FRB20190206A	FRB20181203B	FRB20190129A
FRB20181213B	FRB20190211A	FRB20181218C	FRB20190206B
FRB20181221A	FRB20190218B	FRB20181221A	FRB20190206A
FRB20181223B	FRB20190228A	FRB20181231B	FRB20190218B
FRB20181228B	FRB20190329A	FRB20190103B	FRB20190221B
FRB20181231B	FRB20190410A	FRB20190105A	<b>FRB20190226B</b>
FRB20190423B	FRB20190422A	FRB20190106A	FRB20190228A
FRB20190429B	FRB20190428A	FRB20190323D	FRB20190329A
FRB20190519J	<b>FRB20190430C</b>	FRB20190409B	FRB20190410A
FRB20190601C	FRB20190527A	FRB20190411C	FRB20190412B
FRB20190609A	FRB20190605D	FRB20190414B	FRB20190422A
FRB20190621C	FRB20190623B	FRB20190423B	FRB20190429B
FRB20190701C		FRB20190430A	<b>FRB20190430C</b>
		FRB20190609A	FRB20190617B
		FRB20190625A	