The Astrophysical Journal, 982:46 (14pp), 2025 March 20 https://doi.org/10.3847/1538-4357/adb623

Rare Event Classification with Weighted Logistic Regression for Identifying Repeating Fast Radio Bursts

Antonio Herrera-Martin[1,2], Radu V. Craiu[2], Gwendolyn M. Eadie[1,2], David C. Stenning[3], Derek Bingham[3], B. M. Gaensler[1,4,5], Ziggy Pleunis[5,6,7], Paul Scholz[5,8], Ryan Mckinven[9,10], Bikash Kharel[11], and Kiyoshi W. Masui[12,13]

[1] David A. Dunlap Department of Astronomy & Astrophysics, University of Toronto, 50 St. George Street, Toronto, ON M5S 3H4, Canada
[2] Department of Statistical Science, University of Toronto, Ontario Power Building, 700 University Avenue, 9th Floor, Toronto, ON M5G 1Z5, Canada
[3] Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, Canada
[4] Department of Astronomy and Astrophysics, University of California Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, USA
[5] Dunlap Institute for Astronomy & Astrophysics, University of Toronto, 50 St. George Street, Toronto, ON M5S 3H4, Canada
[6] Anton Pannekoek Institute for Astronomy, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands
[7] ASTRON, Netherlands Institute for Radio Astronomy, Oude Hoogeveensedijk 4, 7991 PD Dwingeloo, The Netherlands
[8] Department of Physics and Astronomy, York University, 4700 Keele Street, Toronto, ON MJ3 1P3, Canada
[9] Department of Physics, McGill University, 3600 rue University, Montréal, QC H3A 2T8, Canada
[10] Trottier Space Institute, McGill University, 3550 rue University, Montréal, QC H3A 2A7, Canada
[11] Department of Physics and Astronomy, West Virginia University, P.O. Box 6315, Morgantown, WV 26506, USA
[12] MIT Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, USA
[13] Department of Physics, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, USA

Received 2024 October 21; revised 2025 January 27; accepted 2025 February 11; published 2025 March 18

Abstract

An important task in the study of fast radio bursts (FRBs) remains the automatic classification of repeating and nonrepeating sources based on their morphological properties. We propose a statistical model that considers a modified logistic regression to classify FRB sources. The classical logistic regression model is modified to accommodate the small proportion of repeaters in the data, a feature that is likely due to the sampling procedure and duration and is not a characteristic of the population of FRB sources. The weighted logistic regression hinges on the choice of a tuning parameter that represents the true proportion τ of repeating FRB sources in the entire population. The proposed method has a sound statistical foundation, direct interpretability, and operates with only five parameters, enabling quicker retraining with added data. Using the CHIME/FRB Collaboration sample of repeating and nonrepeating FRBs and numerical experiments, we achieve a classification accuracy for repeaters of nearly 75% or higher when τ is set in the range of 50%–60%. This implies a tentative high proportion of repeaters, which is surprising, but is also in agreement with recent estimates of τ that are obtained using other methods.

Unified Astronomy Thesaurus concepts: Radio transient sources (2008); Classification (1907); Astrostatistics techniques (1886); Sampling distribution (1899); Astrostatistics distributions (1884); Regression (1914); Nonlinear regression (1948); Linear regression (1945)

1. Introduction

Fast radio bursts (FRBs) represent an enigmatic phenomenon in astrophysics. FRBs are dispersed, isolated, millisecond-long radio pulses that are similar in appearance to single pulses from Galactic pulsars. The arrival of these radio pulses show a frequency-dependent delay (quantified by the dispersion measure (DM)) due to the electromagnetic wave's path through free electrons in the Universe. FRBs have the defining characteristic of a DM that exceeds the maximum DM expected from our Galaxy, suggesting that they are very luminous and of extragalactic origin (D. R. Lorimer et al. 2007; J. M. Yao et al. 2017).

The first FRB was identified in archival Parkes multibeam pulsar survey data by D. R. Lorimer et al. (2007), and was suggestive of an extragalactic origin. Since then, there has been rapid progress in the observation of these enigmatic events and their use as probes of the intergalactic medium (S. Chatterjee 2021; CHIME/FRB Collaboration et al. 2021; M. Bailes 2022). The discovery of FRBs has opened up new research avenues in astrophysics as they have the potential to help us better understand the distribution of matter in the Universe or the nature of dark matter (H.-N. Lin & Y. Sang 2021; Z.-W. Zhao et al. 2023).

The biggest mystery about FRBs is their origin or progenitor object(s). The FRB enigma is made more mysterious by the fact that some FRBs are observed to burst repeatedly (repeaters), while others have only been observed to burst once (nonrepeaters; L. G. Spitler et al. 2016). Models that explain the progenitors of FRBs are thus typically assigned to one of two broad categories. The first category considers noncataclysmic explanations, while the second category assumes FRBs are the result of a catastrophic event that destroys the astrophysical source. In the early years of FRB discoveries, the lack of repeating FRBs supported catastrophic models. The assumption of two distinct subpopulations of FRBs (repeaters and nonrepeaters) is now supported by arguments made by Z. Pleunis et al. (2021), which are based on the morphological differences between the repeating and nonrepeating FRBs. While the number of FRB repeater sources continues to grow, the total published sample is currently only 53.[14]

[14] https://www.wis-tns.org/ accessed 2024 May 30.

The number of known FRB repeaters is small compared to the total FRB population (roughly 2.6% are repeaters; CHIME/FRB Collaboration et al. 2023b), but observational selection effects strongly influence this number. The Canadian Hydrogen Intensity Mapping Experiment (CHIME) telescope, the world's leading detector of FRBs, relies on the rotation of the Earth to observe the whole northern sky over the course of each day. Due to its location in Penticton, BC at a latitude of 49°.32 (M. Amiri et al. 2018) CHIME observes some parts of the sky continuously, and observes other parts as little as 5 minutes day^{-1} (M. Amiri et al. 2018). This means that some areas of the sky are monitored as little as 0.3% of the time each day. Thus, many more FRB repeaters could exist, but have not been detected due to censoring. Recent works by C. W. James (2023) and S. Yamasaki et al. (2023), which take into account the observational selection effects of CHIME, have suggested that the true fraction of repeaters is closer to 50%. Surprisingly, we have reached a similar conclusion through an entirely independent approach that we present here.

In this paper, we propose a principled and interpretable statistical model to predict whether new FRBs are repeaters or nonrepeaters. Our method uses the morphological characteristics of FRBs (e.g., bandwidth and peak frequency) as input, and is based on weighted logistic regression for imbalanced data sets, the details of which are described in Section 3. In short, our classification algorithm can be used to identify potential FRB repeater candidates. Specifically, given a new FRB source for which only one burst has been recorded, the algorithm will provide the probability that this FRB source is actually a repeater. Having this kind of predictive tool could be useful for follow-up observations. For example, if an FRB source is given a high probability of being a repeater, but is in an area of the sky observed only 5 minutes day^{-1} by CHIME, then one could allocate time at other telescopes to observe it more regularly, or one could search through archival data from other telescopes to find previous bursts. While developing our prediction method and algorithm, using techniques entirely independent from C. W. James (2023) and S. Yamasaki et al. (2023), we arrive at a similar, albeit very tentative, conclusion about the true percentage of repeating FRBs.

Creating classification and prediction algorithms for FRBs is notably difficult because of the unique structure of the data. There are at least two major challenges to overcome.

1. *Challenge 1.* The observed number of repeaters is significantly outweighed by the observed number of nonrepeaters. This imbalance in the data leads to biased inference when it is not taken into account. Some studies do not account for the imbalance (e.g., B. H. Chen et al. 2021), while others have addressed the imbalance through resampling techniques (X. Yang et al. 2023). However, the resampling approach does not account for the intrinsic reduction of variability of features in the repeater subpopulation.
2. *Challenge 2.* For training data, the labels for FRB repeaters are almost entirely certain, but the labels for FRB nonrepeaters are not. Any nonrepeater FRB may actually be a repeater that we have not yet seen repeat. This corresponds to a mislabelling problem in the training data for any classification algorithm—i.e., there may be repeaters that are wrongly labeled as nonrepeaters. Statistical approaches to mitigate this usually rely on modeling the probability of an error in labeling (N. Nagelkerke & V. Fidler 2015; H. Hung et al. 2018) which, in the FRB classification case, is impossible.

Previous methods to predict and/or classify repeating FRBs have used black box machine learning approaches (e.g., B. H. Chen et al. 2021; J.-W. Luo et al. 2022; J.-M. Zhu-Ge et al. 2022; X. Yang et al. 2023, among others). However, these previous methods inadequately handle some of the challenges posed by the particular characteristics of FRB data. For instance, B. H. Chen et al. (2021), J.-W. Luo et al. (2022), and J.-M. Zhu-Ge et al. (2022) consider subbursts, which we will rename "substructures" as explained later, and repeater bursts as independent data points, and they do not differentiate between them when creating training and test data. Assuming independence makes it possible to split bursts from the same source and put them into the training and the test data. However, this approach will artificially (i) increase the similarity between the training and test sets, since FRBs from the same source will have some dependency and, consequently, (ii) enhance the model's classification performance because it is easier to identify a repeater source in the test data once one or more of its bursts have been used in the training data. Consequently, the use of subbursts from the same source in test and training data exaggerates the accuracy of the model's predictions.

The method that we propose accounts for the imbalance between the number of repeaters and nonrepeaters by weighting differently the information contained in each observation. This approach relies on a tuning parameter that represents the true proportion of repeater FRBs in the Universe. While a precise value is elusive, our analysis suggests that the model is robust to values of this tuning parameter between 50% and 60%. Our model is also able to identify which of the nonrepeaters to date are most likely to repeat. This information can be used for strategic and efficient monitoring of the sky.

Our paper is organized as follows. First, we introduce our data selection procedure, including which FRB features are used in our analysis (Section 2). Next, we present the proposed method of weighted logistic regression for imbalanced data sets, describe training, validation, and test data sets, and introduce the tuning parameter τ (Section 3). We then present our results (Section 4), and conclusions and directions for future work (Section 5).

2. Data

We construct a statistical prediction model that will identify repeaters based on their morphological features. Thus, we need a training set and a validation set with known labels (i.e., which FRBs are nonrepeaters and which are repeaters, albeit with the caveats described in the previous section) to determine the accuracy of our methodology. Once our statistical model is trained and validated, then we can apply it to a separate test set of data.

2.1. CHIME/FRB Catalog

The majority of our FRB data comes from the first catalog of CHIME (CHIME/FRB Collaboration et al. 2021, 2023a), hereafter referred to as Catalog 1. CHIME is a transit radio telescope operating across the 400–800 MHz range and located on the grounds of the Dominion Radio Astrophysical Observatory in British Columbia, Canada. The original intention for CHIME was to map neutral hydrogen gas as a measure of dark energy, but the large collecting area, wide radio bandwidth, and powerful correlator make it an excellent instrument for the detection of FRBs as well.

The CHIME/FRB Collaboration's software pipeline searches for FRBs in real time (M. Amiri et al. 2018, 2022). The first FRB catalog recorded 536 FRBs between 2018 July 25 and 2019 July 1, and at the time increased the sample of FRB sources by more than a factor of 5 (E. Petroff et al. 2022). By using data from only one telescope, we lessen the potential for different observational biases. An FRB, for our purposes, is constrained to the start and end of a single FRB observation, as defined by the CHIME pipeline. CHIME/FRB's first catalog contains 474 apparent nonrepeating sources and 62 bursts from 18 repeating sources. For every FRB, the catalog contains 12 parameters that can be considered possible classification features.

2.2. Data Selection and Handling

Our exploratory variable selection analysis confirmed that we need only consider features that were identified as useful in previous work (see, e.g., Z. Pleunis et al. 2021), as some variables seemed to have no influence on the classification results. The morphological features we consider are taken from the CHIME/FRB catalog and consist of:

1. the boxcar width (a measure of the FRB burst duration in seconds),
2. the peak frequency (in units of MHz),
3. the intrinsic width of the FRB associated with the event in seconds, as modeled with the fitburst pipeline (i.e., without dispersion smearing and scatter broadening; E. Fonseca et al. 2024),
4. the number of substructures, referred in other works as subbursts,[15] and
5. the emitted bandwidth in units of MHz.

The emitted bandwidth is obtained by taking the difference between the high and low frequencies for detection at the full width at tenth maximum, which is the width of the FRB signal at the level corresponding to the difference between frequencies at which the signal reaches 10% of its peak intensity. Previous work has only looked at emitting bandwidth and not the peak frequency, but we consider the latter as a distinct feature. Detailed descriptions of the complete set of parameters are presented by CHIME/FRB Collaboration et al. (2021). In the case of FRBs with substructures, we define the corresponding FRB-specific features as the mean value across substructures for each feature.

3. Methods

The proposed model is based on logistic regression (P. McCullagh & J. Nelder 1989; C.-Y. J. Peng et al. 2002), which is reviewed in Section 3.1. In Section 3.2, we introduce from the statistics literature methods developed to account for imbalanced data sets in logistic regression, which not only help us address the low proportion of repeaters in the FRB sample, but also allows u to introduce a tuning parameter for the fraction of repeaters in the entire population.

3.1. Classification by Logistic Regression

Logistic regression is a widely used statistical model that captures the relationship between a binary dependent variable and a set of independent features, or covariates. Logistic regression falls under the broader class of generalized linear models (see for example C. E. McCulloch 1997), for which we assume that the information in the covariates about the dependent variable is conveyed through a linear combination of features,

$$\boldsymbol{\eta}(\boldsymbol{\beta}) = \boldsymbol{X}\boldsymbol{\beta}, \tag{1}$$

where $\boldsymbol{\eta}$ is called the linear predictor, $\boldsymbol{\beta}$ is a vector of model parameters, and $\boldsymbol{X}$ is the covariate or design matrix. That is, each row i in $\boldsymbol{X}$ is an FRB observation, with the first column representing the intercept in the regression model and the subsequent columns representing the covariates or features. The number of observations, or rows, in $\boldsymbol{X}$ is N, and the number of columns is $m+1$.

Let $\{(y_i,\ \boldsymbol{x}_i):\ 1 \leqslant i \leqslant N\}$ denote the sample of size n containing the observed FRBs, in which,

$$y_i = \begin{cases} 1 & \text{if the sample } i \text{ is from a repeater FRB source,} \\ 0 & \text{if the sample } i \text{ is from an apparently nonrepeating FRB source,} \end{cases} \tag{2}$$

and $\boldsymbol{x}_i \in \boldsymbol{R}^{(m+1)\times 1}$ is the column vector of observation i that contains m features. While the description is valid for all $m < N$, for the FRB model we selected five features ($m = 5$): the boxcar width, peak frequency, intrinsic width from fitburst, number of substructures, and emitted bandwidth, as described in Section 2.2.

The logistic regression model assumes that each y_i follows a Bernoulli distribution,

$$y_i \sim \text{Bern}(\pi_i(\boldsymbol{\beta})), \tag{3}$$

where,

$$\pi_i(\boldsymbol{\beta}) = \frac{e^{\eta_i(\boldsymbol{\beta})}}{1 + e^{\eta_i(\boldsymbol{\beta})}}, \tag{4}$$

and,

$$\eta_i(\boldsymbol{\beta}) = \boldsymbol{x}_i^T\boldsymbol{\beta} = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + x_{4i}\beta_4 + x_{5i}\beta_5, \tag{5}$$

where x_{ji} is the jth feature or covariate of observation i.

The inference for $\boldsymbol{\beta}$ is based on the likelihood function,

$$\mathcal{L}(\boldsymbol{\beta}|\text{Data}) = \prod_{i=1}^{N} [\pi_i(\boldsymbol{\beta})^{y_i}(1 - \pi_i(\boldsymbol{\beta}))^{(1-y_i)}]. \tag{6}$$

After replacing $\pi_i(\boldsymbol{\beta})$ using Equation (4) and taking the logarithm, the log likelihood becomes,

$$\log \mathcal{L}(\boldsymbol{\beta}|\text{Data}) = -\sum_{i=0}^{N} \log(1 + e^{(1-2y_i)\eta_i(\boldsymbol{\beta})}), \tag{7}$$

[15] Subbursts refer to substructure or multiple burst components under one burst envelope.

where Data $= \{(y_i,\ \boldsymbol{x}_i):\ 1 \leqslant i \leqslant n\}$. Estimates of $\boldsymbol{\beta}$ can be obtained through the maximum likelihood estimator, $\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}|\text{training data})$, and cross validation can also be used to further test and validate the classifier model.

With $\hat{\boldsymbol{\beta}}$ in hand, the logistic regression model (or classifier) can then be used to classify any new FRB with observed features $\boldsymbol{x}^* = (x_1^*,\ldots,x_5^*)$ as a repeater or nonrepeater. That is, one uses $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{x}^*$ to obtain $\eta_i^*(\hat{\boldsymbol{\beta}})$ and $\pi_i^*(\hat{\boldsymbol{\beta}})$ using Equations (5) and (4), respectively. $\pi_i^*(\hat{\boldsymbol{\beta}})$ is the probability P (ith FRB is a repeater). Although the model returns a probability value, in practice a threshold is used to enable metrics for the classification performance. The most common threshold is 0.5, and we use the criterion $\pi_i^*(\hat{\boldsymbol{\beta}}) > 0.5$ to predict that the ith FRB is a repeater.

One drawback of logistic classifiers is that they do not perform well with imbalanced data sets, i.e., situations in which the number of cases ($Y_i = 1$)[16] is vastly different from the number of controls ($Y_i = 0$) (H. He & E. A. Garcia 2009; A. Luque et al. 2019; M. Kim & K.-B. Hwang 2022). In the next section, we describe potential solutions to alleviate this challenge.

3.2. Imbalanced Data

In the statistics literature, imbalanced data refers to data for which—due to observational bias, poor sampling design, or population structure—the number of samples from different subpopulations are substantially different. The imbalance may be extrinsic, i.e., an unknown mechanism unrelated to the model describing the samples hides the real subpopulation ratio (M. Kim & K.-B. Hwang 2022). The imbalance may instead (or also) be due to exogenous sampling, i.e., the imbalance happens during the data sampling process (C. F. Manski & S. R. Lerman 1977). The imbalance may also be intrinsic in situations where only a small fraction of items belong to a subpopulation.

In this study, we have an imbalanced data set between repeaters and nonrepeaters (CHIME/FRB Collaboration et al. 2023b). The number of bursts from nonrepeaters vastly dominates the number of bursts from repeaters. Whether this imbalance is extrinsic, intrinsic, or due to exogenous sampling is still up for debate. Intuitively, it could be due to the fact that some of the putative nonrepeaters in the sample would yield another burst if they were monitored for a longer period of time. The idea that the actual percentage of repeaters in the population is much larger than the one computed directly from the samples collected to date is gaining momentum. For instance, C. W. James (2023) recently proposed a power-law model for the FRB population as a function of redshift, showing that at least 50% of bursts in CHIME's first catalog should come from intrinsic repeaters. Independently, S. Yamasaki et al. (2023) suggests the necessity of correcting the observed source count according to decl. to adequately model the evolution of repeaters and nonrepeaters, in particular the transition of apparent nonrepeaters when they are observed to repeat. A model population incorporating this correction results in an inferred repeater fraction that exceeds 50%.

Regardless of the cause of imbalanced data for FRBs, it is this imbalance that hinders the direct application of a logistic regression model to classify FRBs. Statistically, the data imbalance leads to biased estimates of the parameters and therefore biased predictive probabilities (P. McCullagh & J. Nelder 1989; G. King & L. Zeng 2001a; H. He & E. A. Garcia 2009; M. Maalouf & T. B. Trafalis 2011; R. van den Goorbergh et al. 2022).

There are two general approaches to alleviate the imbalance (H. He & E. A. Garcia 2009). The first is to make the data set more balanced by using a subsample of the larger population. In our case, this would mean using all the repeaters but only a subsample of nonrepeaters, and then proceeding as if that were the data in hand (see M. Kim & K.-B. Hwang 2022, for an empirical study of these methods). This approach only makes sense, however, if one knows the true proportion of repeaters in the population. The second approach to make the data set more balanced is to oversample the smaller class. However, such an approach introduces additional variability in the analysis from the subsampling procedure (see R. van den Goorbergh et al. 2022, for a discussion).

We therefore favor a third approach that applies weights to rebalance the data. This is done using the method of weighted logistic regression for imbalanced data sets developed by G. King & L. Zeng (2001a, 2001b), M. Tomz et al. (2003), M. Maalouf & T. B. Trafalis (2011), and M. Maalouf & M. Siddiqi (2014). The weights are defined by introducing a parameter τ to denote the proportion of repeaters among the whole population of FRB sources. We assume τ is different than the observed fraction in the sample. For simplicity, we only consider the case of τ as a fixed fraction, independent of redshift, luminosity, etc.

In weighted logistic regression, a weight w_i is assigned to the ith sample (G. King & L. Zeng 2001a, 2001b; M. Tomz et al. 2003; M. Maalouf & T. B. Trafalis 2011; M. Maalouf & M. Siddiqi 2014). Following G. King & L. Zeng (2001b), the weight for the ith sample is,

$$w_i = \delta_r y_i + \delta_{nr}(1 - y_i), \tag{8}$$

where,

$$\delta_{nr} = \frac{1-\tau}{1-\bar{y}}, \qquad \delta_r = \frac{\tau}{\bar{y}}, \tag{9}$$

where r and nr stand for repeater and nonrepeater, respectively, and $\bar{y}$ is the number of repeaters in the data divided by the total number of observed sources. When $y = 1$ (a repeating FRB), $w = \delta_r = \frac{\tau}{\bar{y}}$. In this way, δ_r could be interpreted as the correction factor for the proportion of repeaters in the sample. Note that each weight w_i depends on the ratio τ. The latter is not known exactly, and so we first treat τ as a tuning parameter, trying various values of τ between zero and one. Ultimately, though, we rely on a combination of the evidence provided by S. Yamasaki et al. (2023), who state that $\tau \in (0.5, 1]$, and the results of our initial exploration of τ settle on the value $\tau = 0.55$. This will be discussed further in Section 4.1.

The resulting (weighted) logistic regression log likelihood is then,

$$\log \mathcal{L}(\boldsymbol{\beta}|\text{Data}) = -\sum_{i=0}^{n} w_i \log(1 + e^{(1-2y_i)\eta_i(\boldsymbol{\beta})}), \tag{10}$$

where η_i is defined as in Equation (5).

[16] We are using statistical notation here; a capital Y represents a random variable that follows a distribution, while y indicates a realization of that random variable, i.e., an observation with known value.

The maximum likelihood estimator for β derived from Equation (10) incurs an increase in bias and variance, given the number of samples, as discussed by G. King & L. Zeng (2001b) and M. Maalouf & M. Siddiqi (2014). The introduction of weights causes the estimators to become biased as proved by P. McCullagh & J. Nelder (1989) and in Appendix A of G. King & L. Zeng (2001b). However, one can correct for this: we use the maximization algorithm that corrects for bias and reduces variance using the methods proposed by M. Maalouf & M. Siddiqi (2014). Using their approach, we obtain estimates of β from our training data, and use cross validation to create our classifier.

We note that the mislabeling of a repeater as a nonrepeater simply because it was not observed for long enough can produce serious biases in any supervised learning-type analysis. This is a caveat that has not been addressed in previous work (J.-W. Luo et al. 2022; J.-M. Zhu-Ge et al. 2022; X. Yang et al. 2023). The introduction of the true proportion τ allows us to account for the uncertainty associated with the potential mislabeling of repeating sources as nonrepeaters.

3.3. Training and Test Data

Ideally, we would use multiple data sets to study the statistical performance of the proposed model. However, this is not possible because we have only one data set at our disposal. Thus, we instead use cross validation; we create multiple data sets by repeatedly splitting the available data set into a training set and a cross-validation set. For each split, we fit the model on the training data and then classify the cross-validation data as repeaters or nonrepeaters based on the fitted model. The classification is done by evaluating the the probability π_i in Equation (4), which returns the probability of belonging to one of the two possible classes.

To assess the performance of a classifier, it is customary to use a classification threshold of 50%. More precisely, if the estimated probability of an FRB being a repeater is greater than 50%, i.e., $\Pr(Y = 1) \geqslant 0.5$, then we classify the FRB as a repeater. Changing the threshold has an effect on the performance metrics, but the probability itself is not affected.

Instead of a single classifier, our approach yields an ensemble of classifiers, since every data split will yield a different set of parameter estimates. The procedure of data splitting and model fitting is repeated 500 times, which provides uncertainty quantification for the model's classification accuracy. While more replicates are possible, the results remain similar due to the small number of repeaters in the sample. When performing the data splitting, we carefully avoid situations in which repeated bursts from the same source appear in both the training and test data sets. In other words, repetitions from the same FRB source will all be included in the training set or will all be included in the cross-validation set at each data split.

The number of confirmed repeater sources in Catalog 1 is only 18, which is much smaller than the 474 nonrepeaters (together shown as white regions in Figure 1)—this limits the number of repeater sources in the test data. Since the release of Catalog 1, the CHIME/FRB collaboration has published an additional 25 newly discovered repeaters (the "Gold sample," shown as the yellow region in Figure 1 and in CHIME/FRB Collaboration et al. 2023b). For the purpose of our interpretation, we refer to these as "real repeaters."

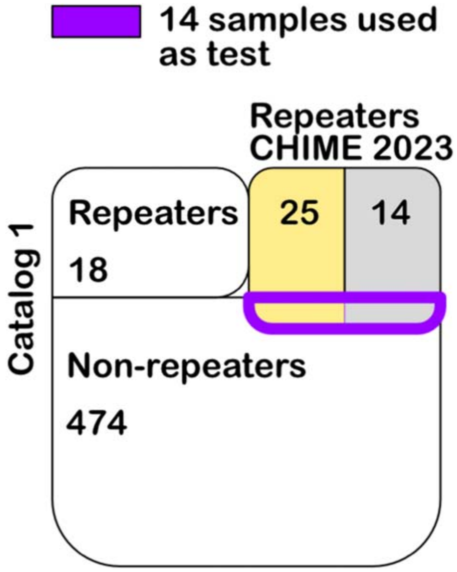

Figure 1. Venn diagram describing our training/validation set and "test" set. The training/validation set is comprised of the white, yellow, and gray regions, minus the region delimited by the purple outline. The latter is the "test" set of 14 FRBs and the performance of our model on these is shown in Figure 5.

CHIME/FRB Collaboration et al. (2023b) also include 14 newly discovered repeater candidates (the "Silver sample," shown as the gray region in Figure 1), which have a lower confidence of being repeaters than the Gold sample, but have a higher confidence than the nonrepeaters. We include the Silver sample as repeaters in our training and cross-validation splits.

Out of these 39 sources in the Gold and Silver samples, we exclude a subset of 14, which crossmatch in Catalog 1 and the repeaters catalog, from training and cross validation. These 14 FRB sources are used as our test set (delimited by the purple outline in Figure 1). The remaining 25 repeaters and nonrepeaters are added to the training/cross-validation set.

3.4. Selection of τ

The statistical model that we consider for classification depends on the parameter τ, which is not directly estimable from the data, or from theory.

In this paper, we treat τ as an unknown independent quantity and create a sequence of its possible values. For each value of τ, we train the proposed logistic regression classifier by splitting randomly 500 times the training and cross-validation sets, as described above. Together, the individual classifiers trained by the random splits form an ensemble of classifiers. For every value of τ, a different maximum likelihood estimator is obtained and a different classification for a newly observed FRB source is generated.

As mentioned previously, the logistic regression model gives a probability π_i of belonging to one of the classes (either repeater or nonrepeater), but this is not useful when measuring the performance of the model, as our response, y_i, is binary in nature. So it is necessary to apply a decision threshold—we use a standard value of 50%. The choice of this threshold is independent of τ or the imbalance of the data, and it only serves as a decision boundary to obtain a binary response. The standard value assumes that there is no preference for repeaters or nonrepeaters, and allows for better analysis of the performance of the logistic regression model.

When summarizing the performance of the model for each τ, we use the traditional metrics of accuracy, precision, recall, and

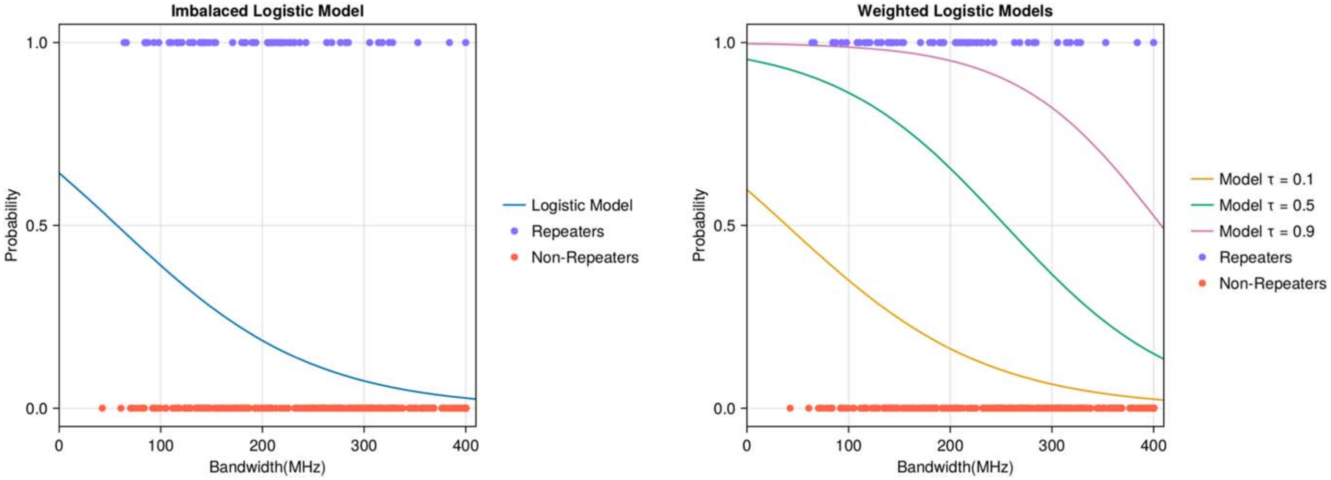

Figure 2. Example of a logistic classification model using the bandwidth of each FRB as a feature and the complete catalog as training data. The purple and orange dots represent the Catalog 1 samples, and the y-axis is the probability of belonging to one of the binary classes, where 1.0 is a 100% or the repeater class, and 0.0 is 0% or the nonrepeater class. As the data have a binary response, the samples are assigned to 0% or 100% depending of the assigned label of repeater or nonrepeater. Left: The classical logistic model ignores the imbalance in the data, and produces the fit shown as a solid blue line. Based on this model fit, and assuming a threshold probability of 50% to classify any future data as repeaters, only a few samples would be correctly classified. Right: the weighted logistic regression uses the rare events approach and an assumed value of τ (Section 3.2), resulting in the fit shown as solid lines ($\tau = 0.1$ in orange, $\tau = 0.5$ in green, and $\tau = 0.9$ in pink).

F_1 score. These metrics are derived using empirical frequencies computed from the proposed model's performance on multiple splits of the training and cross-validation data. These metrics are defined as follows.

1. *Accuracy.* The proportion of correct predictions among all predictions, i.e., the proportion of correctly classified FRBs as repeaters and nonrepeaters, among all classifications,

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{11}$$

where TP, TN, FP, and FN are, respectively, the number of true positives, true negatives, false positives, and false negatives.

2. *Precision.* The proportion of positive predictions that are actually correct, i.e., the proportion of FRBs classified as repeaters that are real repeaters,

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{12}$$

3. *Recall.* The proportion of true positives that have been correctly predicted, i.e., the proportion of real repeaters that are correctly identified as repeaters,

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{13}$$

4. *F_1 score.* A measure that combines precision and recall. The F_1 score is the harmonic mean of the two, and can be interpreted as an optimal number of true positives without introducing too many false negatives or false positives. It can also be written as,

$$F_1\text{score} = 2\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{14}$$

The F_1 score is particularly useful in cases where the data set is imbalanced.

After training an ensemble of classifiers[17] for every τ, we analyze the above performance metrics (Section 4.2) and combine this information with literature estimates to settle on a single value of τ.

4. Results

4.1. Example Demonstrating Weighted Logistic Regression for Imbalanced Data Sets

In Figure 2, we compare classifications produced by traditional logistic regression that does not account for imbalance in the data (left) to that produced by weighted logistic regression (right). In this example, and only for ease of illustration, we use the bandwidth as a predictor and Catalog 1 as the training data. In the right-hand side of Figure 2, we also show how the logistic curve changes as τ takes the values 0.1, 0.5, and 0.9. For $\tau = 0.1$ the fitted logistic curve (orange line) is very similar to the fit obtained without any correction. When $\tau = 0.5$ the logistic fit changes shape, and predicts higher probability $\Pr(Y = 1)$ (i.e., predicts a repeater) for FRBs with lower bandwidths. When $\tau = 0.9$ we can see that most of the FRBs will be predicted to be repeaters, which is not surprising since the information provided by τ is that 90% of all FRB sources in the Universe are repeaters. Clearly, the choice of τ has a strong effect on the model. When judging the accuracy of the model, one must consider the rate of mislabeling a nonrepeater, which relates with $1 -$ precision, or mislabeling a repeater, which relates to $1 -$ recall. In general, we expect both type of errors to be nonzero and impossible to simultaneously optimize (see R. V. Craiu & L. Sun 2008, and references therein). One must therefore consider the effect of choosing a value of τ on both types of errors simultaneously.

[17] Example implementation is available at https://github.com/alfa33333/RE-FRB.

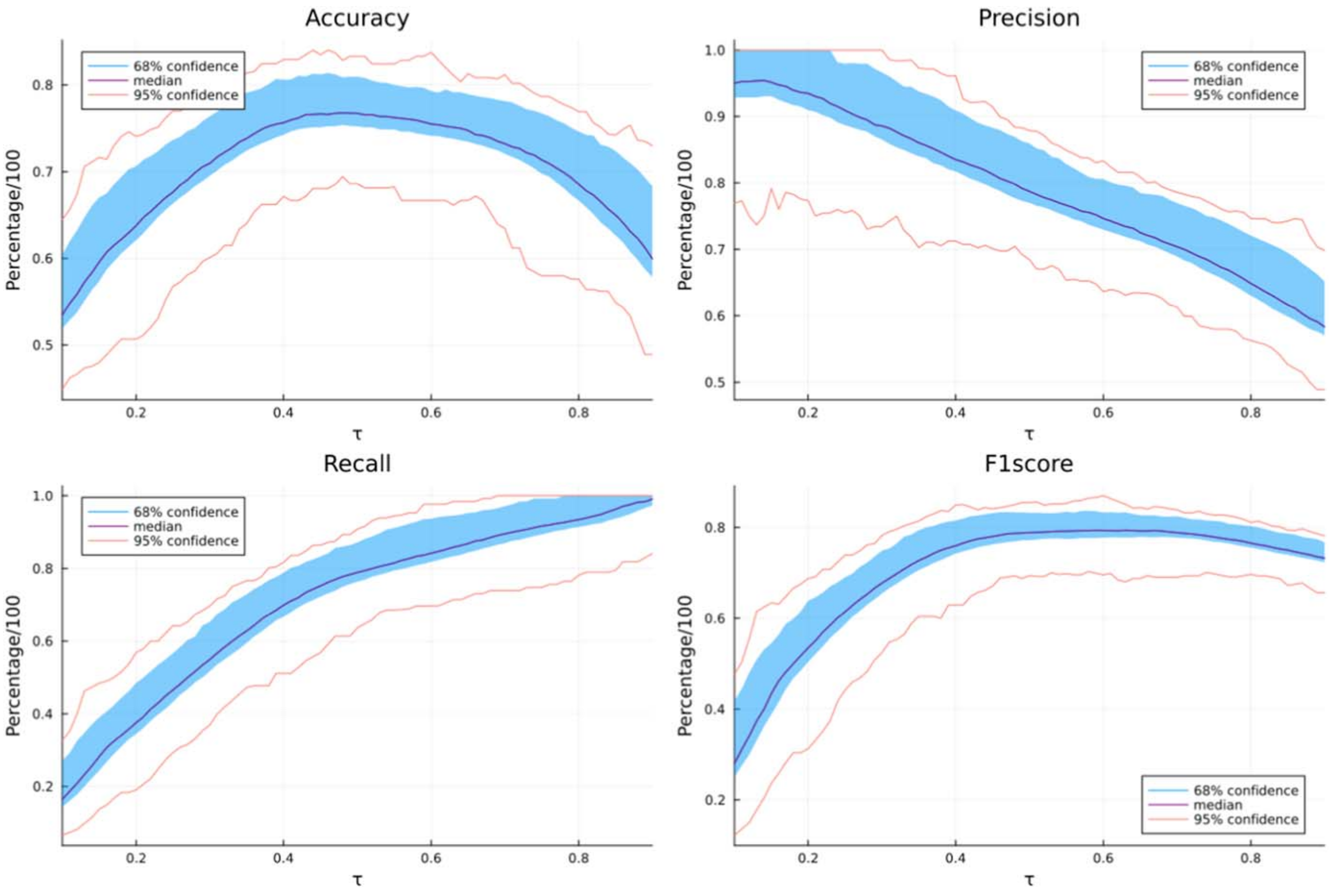

Figure 3. Performance metrics for the ensemble of classifiers as a function of τ. Values on the vertical axis represent the proportion for each metric (i.e., Equations (11)–(14)). Accuracy (the proportion of correctly classified FRBs as repeaters and nonrepeaters, among all classifications) is shown in the upper left, and is maximized for $0.4 \leqslant \tau \leqslant 0.6$. Precision (the proportion of FRBs correctly classified as repeaters out of all repeater predictions) is shown in the upper right, and decreases with increasing τ. Recall (the proportion correctly identified repeaters out of all real repeaters) is shown in the lower left, and increases with increasing τ. F_1 score (the average rate between precision and recall) is shown in the lower right, and appears to be maximized and stable in the approximate range $0.4 \leqslant \tau \leqslant 0.6$, similar to the accuracy. Precision and recall appear to be inversely related.

4.2. Ensemble of Classifiers: Finding a Soft Limit for τ

In Figure 3, we show the performance metrics and their estimated confidence intervals, as a function of τ, listed in Section 3 for our ensemble of classifiers. The bold purple line represents the median, the blue shaded regions represent the 68% confidence intervals, and the orange lines are the 95% confidence intervals. The most important feature of Figure 3 is that the range $0.4 \leqslant \tau \leqslant 0.6$ gives the most reliable model.

Recall from Equation (14) that the F_1 score provides an average of precision and recall and thus can be considered an overall measure of performance. Similarly, accuracy (Equation (11)) measures how often the model correctly predicts the output, regardless of the true value. As illustrated in the upper left and lower right panels of Figure 3, both accuracy and F_1 score show an optimal range of approximately $0.4 \leqslant \tau \leqslant 0.6$.

For values lower than $\tau = 0.4$, the accuracy, recall, and F_1 score return a lower percentage of correctly classified samples. Although the precision (upper right panel, Figure 3) is high for low τ, this is merely a consequence of the majority of the data being nonrepeaters; the precision (Equation (12)) only considers the performance as measured on positives (i.e., repeaters). Since the majority of the data consist of negatives (i.e., nonrepeaters), the method is more likely to misclassify a negative. We also note that precision is insensitive to the number of repeaters correctly classified. For example, we could achieve 100% precision by having just one repeater correctly classified and zero misclassified nonrepeaters.

Recall (lower left panel of Figure 3) shows the ability of our model to correctly identify repeaters. Ideally we want to achieve 100% recall; based on Equation (13), achieving 100% means that every repeater is correctly identified as a repeater. Our model indicates that a high value of τ is needed for such a result. Such a high value of τ would imply a belief that almost all FRBs in the Universe are repeaters.

For values higher than $\tau = 0.6$, the accuracy and F_1 score decline. Although it is possible to correctly classify all repeaters with $\tau = 0.8$, using $\tau = 0.6$ still correctly classifies all repeaters in the catalog for some of the test splits. While $\tau > 0.6$ is possible, we are not comfortable with the implication of having such a large proportion of repeaters, as we do not have other external evidence that justifies the decrease in accuracy and precision from the data.

Based on the overall performance metrics in Figure 3, it appears that our ensemble of classifiers performs best in the range $0.4 \leqslant \tau \leqslant 0.6$. Both C. W. James (2023) and S. Yamasaki et al. (2023) propose that the fraction of repeaters should be at least 50%, by modeling the DM with a low repeating population or correcting the source count evolution, respectively. Combining this lower limit of τ with our range, we are left with a proportion of repeaters somewhere between 50%

Table 1
Parameter Estimates, and Sample Means and Sample Standard Deviations Computed for Each Feature

Parameter	Estimates ($\hat{\beta}_j$)	$\bar{x}_j$	s_j
Intercept($\hat{\beta}_0$)	−0.20	⋯	⋯
bc_width($\hat{\beta}_1$)	0.11	9.759×10^{-3} s	14.376×10^{-3} s
Bandwidth($\hat{\beta}_2$)	−1.14	283.052 MHz	110.572 MHz
sub_struct($\hat{\beta}_3$)	0.27	1.371×10^{-1}	4.879×10^{-1}
width_fitb($\hat{\beta}_4$)	0.75	1.880×10^{-3} s	2.319×10^{-3} s
peak_freq($\hat{\beta}_5$)	0.61	512.520 MHz	110.531 MHz

Note. The second column shows the estimate of the median of each parameter's ensemble distribution. The last two columns show the sample mean and sample standard deviations of the data (i.e., the values used to standardize each feature).

and 60%. The analysis proposed in this paper is quite robust to the choice of τ in this range. For brevity, we present the results from a single value of $\tau = 0.55$ for the remaining analysis. The results corresponding to $\tau = 0.5$ and 0.6, the lower and upper limits of the range, are very similar.

4.2.1. Parameter Interpretation

On their original measurement scale, the covariates have different units and different sample means and sample standard deviations, as reported in the last two most column of Table 1. This complicates the interpretation of the regression parameters and of the importance of each feature. Intuitively, one can see that if covariates X_1 and X_2 are equally important in predicting the outcome, but X_1 has a much larger mean than X_2, then the regression coefficient of the former must be smaller than that of the latter. A similar phenomenon occurs when the values of feature X_1 are more spread out, i.e., have a larger standard deviation, than those of X_2. These challenges in interpreting the model are eliminated when the covariates are standardized, i.e., the values of each feature are realizations of a random variable with a mean of zero and a variance of one.

To standardize x_{ji}, one subtracts the sample mean $\bar{x}_j$ and divides by the sample standard deviation, s_j, of that feature, i.e.,

$$x'_{ji} = \frac{x_{ji} - \bar{x}_j}{s_j}. \tag{15}$$

This transformation turns the features into realizations of a dimensionless random variable with a mean of zero and a standard deviation of one. We note that while the classification accuracy is invariant to linear transformations of the covariates, standardizing the features allows direct comparison of the regression coefficients and gives a clear interpretation of $\hat{\beta}_0$. The latter becomes the logit transformation of the probability of a repeater when all features take their corresponding mean value (which is zero after standardization).

In Figure 4, we show the ensemble distributions for the different parameters $\hat{\beta}_j$ when $\tau = 0.55$. In Table 1, we provide the estimated median for each regression coefficient, computed from the ensemble distribution, and the sample means and standard deviations of the data used for standardization. We emphasize that the coefficient estimates are specific to the training data we used (Section 2).

From Figure 4 and Table 1, we observe that the biggest contributors to classification are the bandwidth and the fitburst width. This analysis of the coefficients agree with previous empirical observations in the literature, e.g., Z. Pleunis et al. (2021).

With the data standardized, each regression coefficient can be interpreted in terms of changes in the probability of a repeater when the corresponding feature value is one and the other feature values are fixed at zero. For instance, for bandwidth, the regression coefficient is $\beta_2 = -1.14$, implying that increasing the bandwidth from zero to one results in decreasing $P(Y = 1)$ from 0.45 to 0.21. Similar calculations can be done with the other features. When all features are equal to their mean zero value, the probability of the corresponding burst belonging to a repeater is $P(Y = 1) \approx 0.45$.

4.3. Testing Our Model with the Gold and Silver "Test" Data

We use the test data (i.e., the 14 FRB sources from the Gold and Silver samples described in Section 3.3) to assess the performance of the ensemble of classifiers for $\tau = 0.55$. The results are presented in Figure 5. The y-axis shows the names provided by the Transient Name Server for each FRB in the test set, and the x-axis shows the classification probability. If the classification probability for an FRB is higher than 0.5, then we classify the FRB as a repeater. Since each FRB is classified multiple times through the ensemble of classifiers, the results are summarized by violin plots of the predicted probabilities. The vertical black lines are the medians of all the prediction probabilities. We choose to present the median instead of the mean because the distribution of predicted probabilities is not symmetric and the mean may be misleading. The distributions are colored gold and silver, to indicate which FRBs belong to the Gold and Silver samples, respectively.

Of the 14 samples in our test set, six (four Silver FRBs and two Gold FRBs) are unambiguously identified as nonrepeaters because the entire distribution is below the threshold. Two FRBs remain ambiguous (FRB 20180909A and FRB 20190127B, both Silver FRBs) due to the distribution being broad and uninformative. We found that this result is consistent across three values for $\tau \in \{0.5, 0.55, 0.6\}$. Notably, five of the six samples identified as repeaters are from the Gold sample.

The remaining FRBs in Figure 5 are predicted to be repeaters, with only the tail of the distributions crossing the threshold to the nonrepeater region. Note that one could find several ways to make use of the distribution of classification probabilities. For instance, in the case of FRB 20190609C, all the ensemble models correctly predict it to be a repeater and the median of the classification probabilities value is significantly higher than 0.5, which strongly recommends it for future monitoring. In the case of FRB 20190127B (sixth from the top in Figure 5) the set of ensemble models that predicts it as a repeater is barely outnumbered by the complement set. An observer who is interested in finding new repeater signals might consider the ambiguity in the evidence and choose to continue monitoring this FRB.

At first glance, Figure 5 shows that six out of 14 FRB sources in the test set were correctly labeled as repeaters. A quick and naive interpretation of this result is that our model does no better than 50% chance at predicting repeaters. However, six out of 14 repeaters is merely a result of the 50% threshold for classification. Had a different threshold been chosen, then more or fewer FRBs would be classified as repeaters. Moreover, the classification is not random between the Gold and Silver samples. Almost all of the Gold sample are

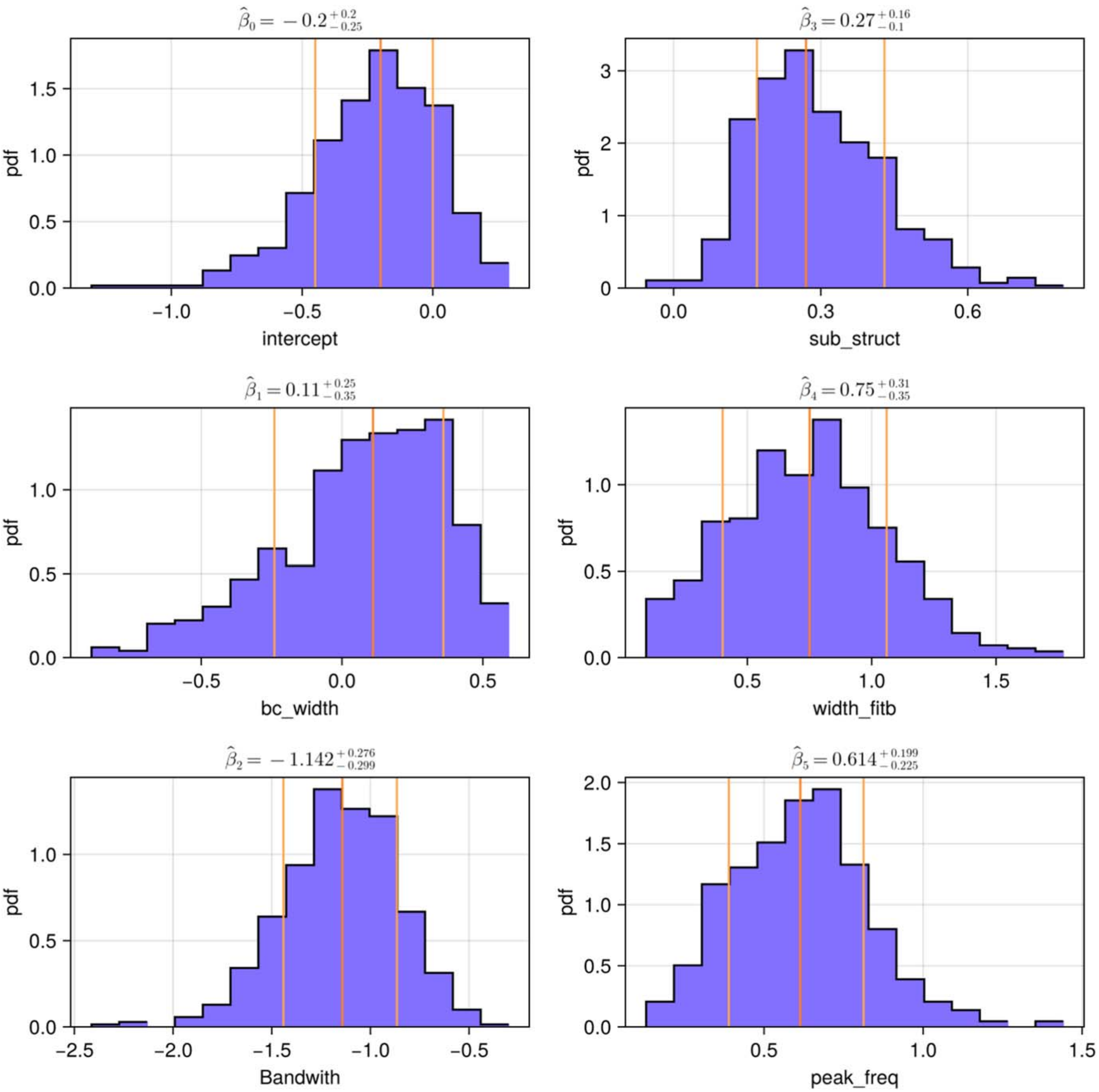

Figure 4. Ensemble distributions for the different β_i parameters for the selected fiducial value of $\tau = 0.55$. The orange lines correspond to the 16th percentile, median, and 84th percentile.

identified as repeaters, and most of the Silver sample are not, which implies that information from the covariates is informing our model. Furthermore, the training data are imbalanced to nonrepeaters, so if the classifiers give an FRB source in the test set a high probability of being a repeater, then that FRB must be quite different from the nonrepeater set. The precision in Figure 3 gives another measure of this idea—approximately 20% of nonrepeaters are misclassified at our selected τ. We include in the Appendix additional results for $\tau = 0.1$ (Figure 9) and $\tau = 0.9$ (Figure 10). These two plots illustrate the influence of imposing extreme values for the fraction of repeaters in the population. For instance, when τ is close to the lower bound, the vast majority of observed FRBs are assumed to be one-offs, and this influences the classification results.

In Figures 6–8, we strengthen our case that our classifier is doing better than chance by showing the dynamic spectra, also called "waterfall plots," for each FRB in the test set. We present two plots for each FRB, at the left the intensity data and at the right the intensity data with an overlay of the fitburst model for visibility purposes. FRBs classified as repeaters are shown in Figure 6, nonrepeaters in Figure 7, and ambiguous FRBs in Figure 8. In these waterfall plots, we see the typical behavior of FRB morphology for repeaters and nonrepeaters: those with broad widths are classified as repeaters and those with a single short burst are classified as nonrepeaters.

Each of the repeating sources of FRBs within the Gold and Silver samples have a contamination rate of chance coincidence ($R_{\rm cc}$; CHIME/FRB Collaboration et al. 2023b). The latter is

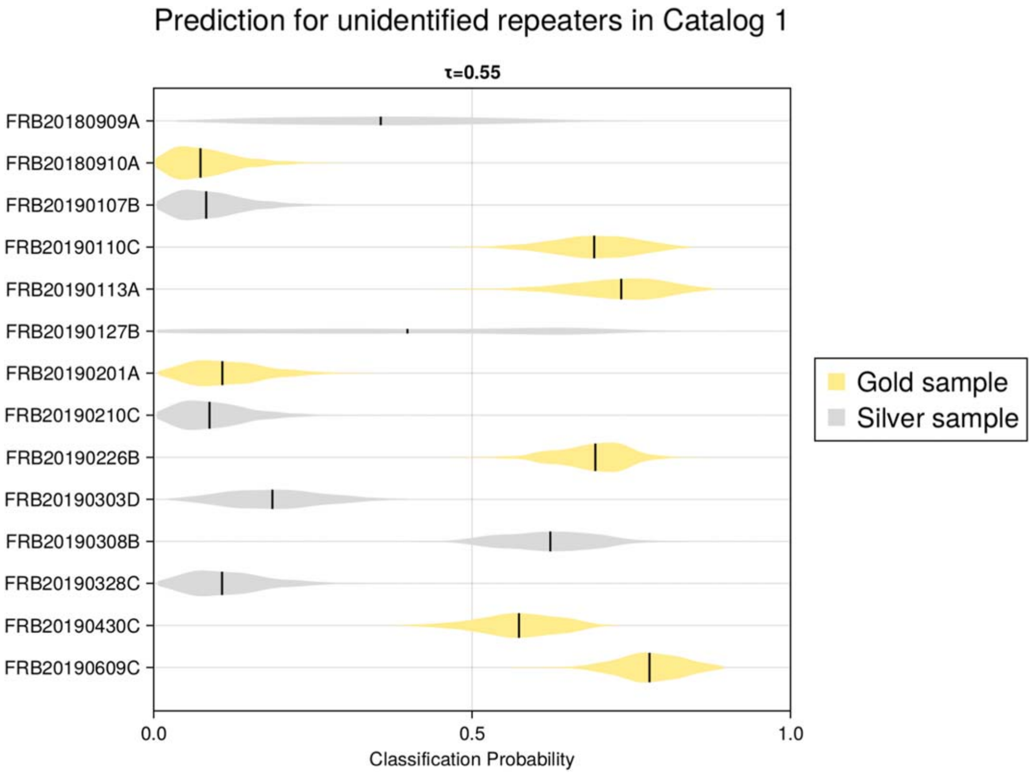

Figure 5. Violin plots of the classification probability for the ensemble of classifiers at $\tau = 0.55$ for the 14 unidentified repeater sources in Catalog 1. The nonrepeater to repeater threshold of 50% is shown as a gray vertical line; everything to the right of the gray line is classified as a repeater. The gold and silver colors correspond to the Gold and Silver samples, respectively. Each of the violin plots indicates where the majority of the distribution lies, and the dark lines inside the boxes are the medians of the distributions. The median values are enough to identify six out of 14 as repeaters across the different values of τ. The FRB 20180909A and FRB 20190127B classification distributions are stretched so thinly across the threshold that they become ambiguous to classify.

interpreted as a measure of uncertainty. A higher uncertainty is represented by a higher $R_{\rm cc}$ value, and is linked with the probability of being observed by chance in the same region as another source. The Gold sample includes sources with $R_{\rm cc} < 0.5$ while the Silver sample includes sources within the range $0.5 \leqslant R_{\rm cc} < 5$. This additional information provides a different perspective on our results. From the Gold sample, our method is able to correctly identify five out of seven sources as repeaters, yielding an accuracy on confirmed repeaters higher than 70%, which is more in line with the results generated via cross validation. The two FRBs from the Gold sample that were misclassified, FRB 20180910A and FRB 20190201A, can be considered atypical cases or outliers in their morphological features. While FRB 20180910A has the largest contamination rate from the Gold sample with $R_{\rm cc} = 0.43$, FRB 20190201A is similar to FRB 20200120E, which shows unusual larger bandwidths and narrower widths (M. Bhardwaj et al. 2021).

5. Conclusions and Future Work

The main aim of this paper has been to develop methodology that assists in the classification of newly observed FRBs as repeaters or nonrepeaters. Repeating FRBs are of scientific interest and any such classification has the potential to increase the number of confirmed repeating FRBs at a more rapid pace. For example, by identifying repeater candidates from the first observation with a certain repeater probability, one could rank potential repeating FRBs for follow-up. Alternatively, repeater candidates identified through our method may improve the efficiency of follow-up searches for repeat bursts in archival data.

The method proposed here is based on the widely used statistical model for studying the dependence of binary response variables on independent variables, known as logistic regression. Our logistic regression model is modified to account for the marked imbalance in the data (i.e., many fewer repeaters than nonrepeaters). The adjustment of the method relies on a tuning parameter that represents the proportion of repeaters, τ, in the whole population of FRBs.

Given the cross-disciplinary nature of the methodology, we have prepared takeaways according to the main interest of the audience.

Astronomy takeaways.

1. One of the first takeaways is efficiency. Compared with a deep-learning approach or other machine learning classification techniques, logistic regression requires a smaller volume of data.
2. The performance of the model is promising for finding potential repeaters. For the Gold sample, the model correctly identifies 70% of the test set FRBs, not previously used for training, as repeaters.
3. While logistic regression is well established in other fields, its application in astronomy, particularly for rare events, is relatively novel. We have introduced this approach to the astronomical literature and hope it can be successfully applied to other astronomy data sets that have binary responses. Moreover, we emphasize the introduction of a parameter that could help interpret the true population fraction in any study containing an imbalanced data set. We encourage the application of our method to other data sets where a similar pattern is present.

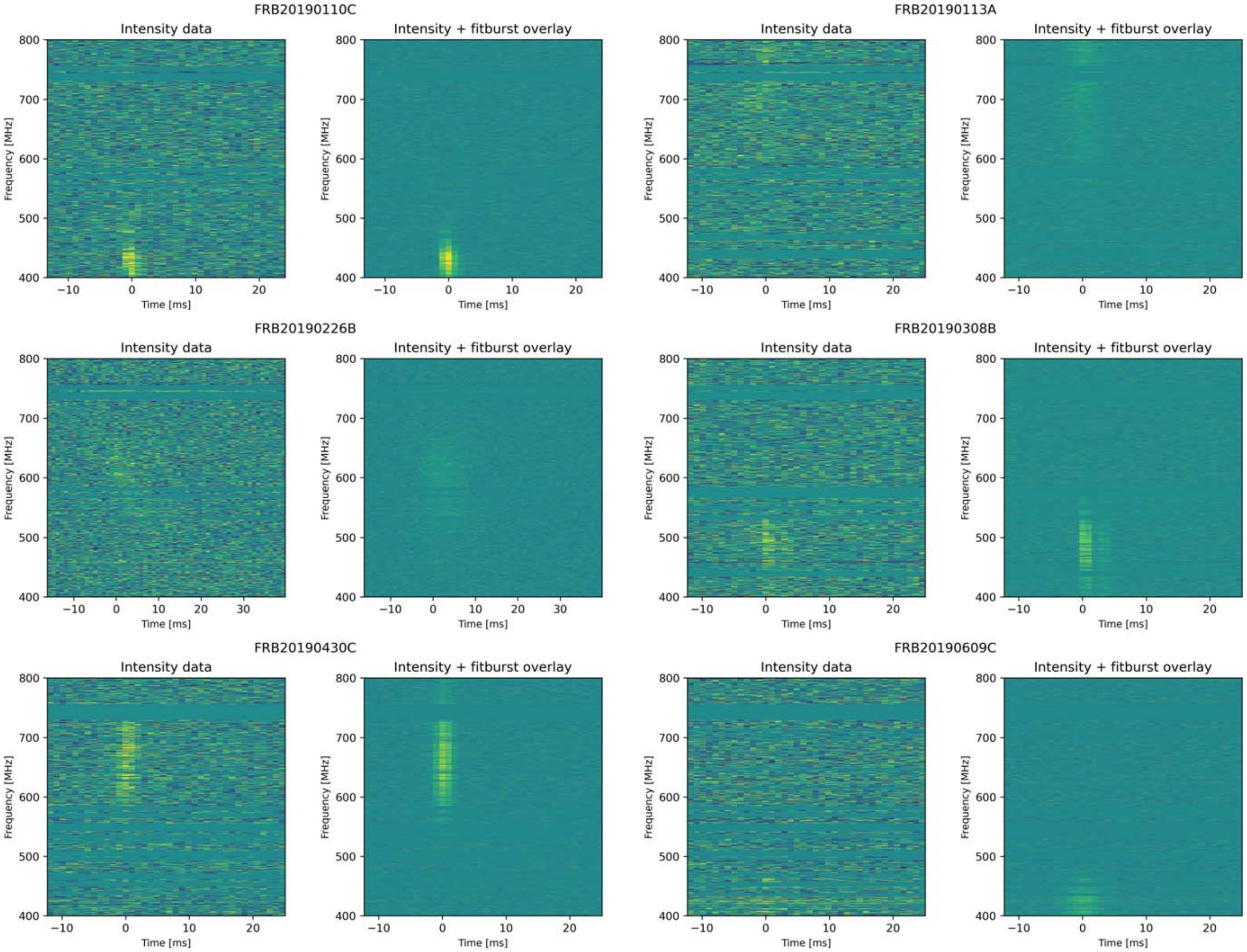

Figure 6. Dynamic spectra or waterfall plots for the six FRBs from the test set, from CHIME/FRB Collaboration et al. (2021), which are classified as repeaters by the ensemble classifier (i.e., FRB 20190110C, FRB 20190113A, FRB 20190226B, FRB 20190308B, FRB 20190430C, and FRB 20190609C from Figure 5). We present two plots for each sample: at the left is the intensity data and at the right is an overlay of the fitburst model to enhance the visual shape of the FRB. These bursts show evidence of being a repeater: wider widths, downward drifting, or narrow emitting bandwidths. This is in contrast with bursts from Figures 7 and 8, which do not show repeater characteristics.

Statistical takeaways.

1. The FRB data present interesting challenges for the statistician. The uncertain labeling of one-off FRBs is an open problem. One cannot be sure that a one-off source will never repeat in the future. This uncertainty in labeling is an unusual issue in statistical applications of logistic regression. This implies that the model we propose should not be interpreted as a classifier. Rather, an astronomer interested in repeating FRBs will want to use our model to select and prioritize monitoring of one-off sources that are more likely to repeat. In other applications, where both cases and controls are unambiguously labeled, one can use a model like ours for classification.
2. The imbalance of the sample can be accounted for, with caveats. The number of one-off FRBs vastly outnumbers the number of repeating FRBs in this application. This issue of imbalanced data is not unusual in statistical applications of logistic regression, and we have referenced works that tackle this problem. However, what is unusual in the application of FRBs is that we do not know the true, underlying proportion of repeating FRBs in the population. We have addressed this issue by allowing the proportion of repeaters in the population of FRBs, τ, to take on various values, and by assessing performance metrics to settle on a range of probable values. We found values of τ in the range 0.5–0.6 to perform well, and noted that the results do not change significantly within this range. Thus, we presented the analysis when $\tau = 0.55$. Our numerical experiments suggest that the model is most accurate when τ is around 60%. However, we hesitate to conclude that this qualifies as strong statistical evidence in favor of this being the *true* proportion, since the tuning parameter is not estimable from the data.
3. The dependence between observations produced by a repeating source stills needs to be taken into account. A one-off source generates a single burst while a repeater generates one to several subsequent bursts. In this paper, the information from multiple bursts is summarized by the mean. This strategy does not incorporate all the information contained in multiple bursts. Moreover, some repeating FRBs may have burst only twice, while others

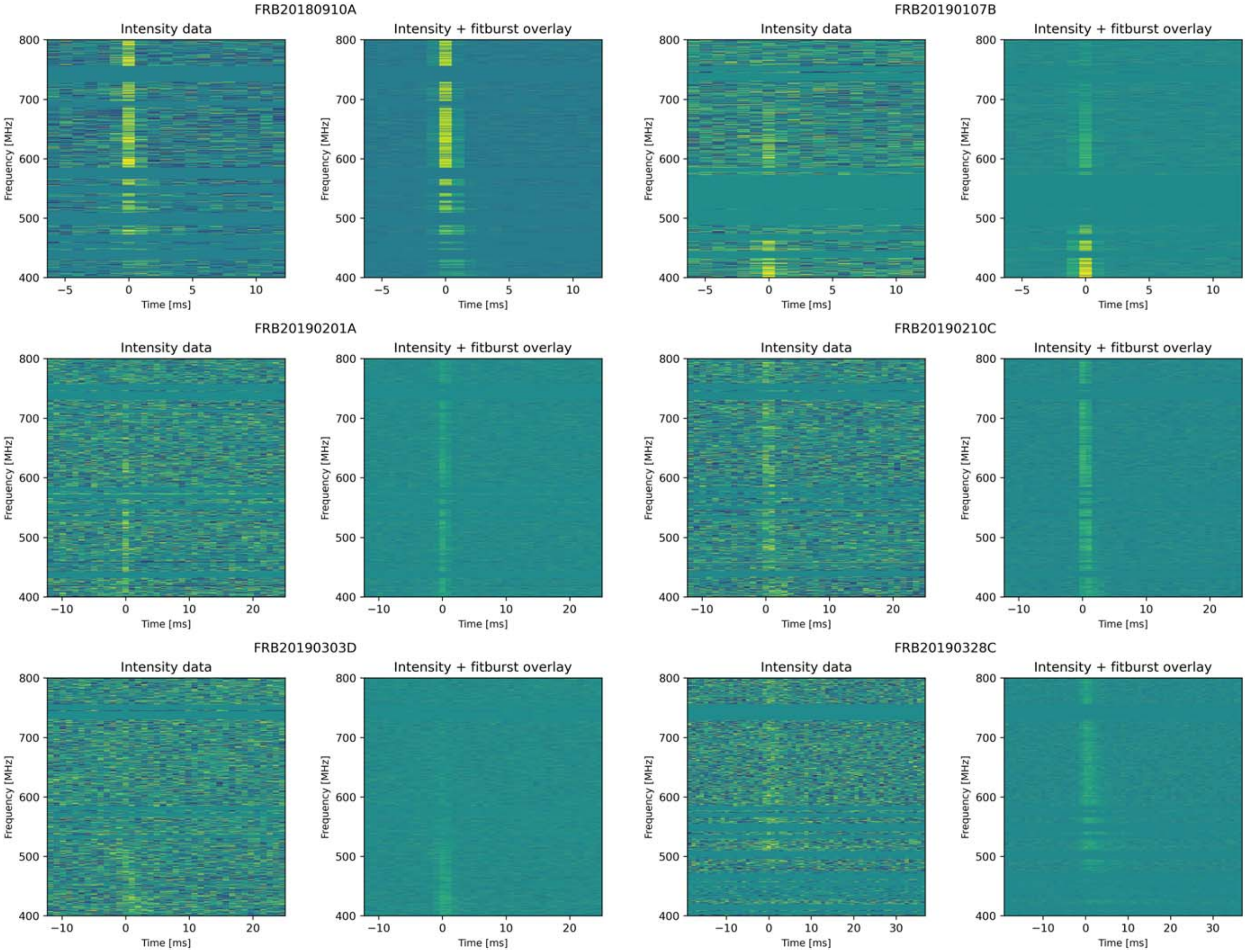

Figure 7. Dynamic spectra or waterfall plots for the six FRBs from the test classified as nonrepeaters by our method (i.e., FRB 20180910A, FRB 20190107B, FRB 20190201A, FRB 20190210C, FRB 20190303D, and FRB 20190328C from Figure 5). These FRBs' morphologies appear similar to nonrepeaters.

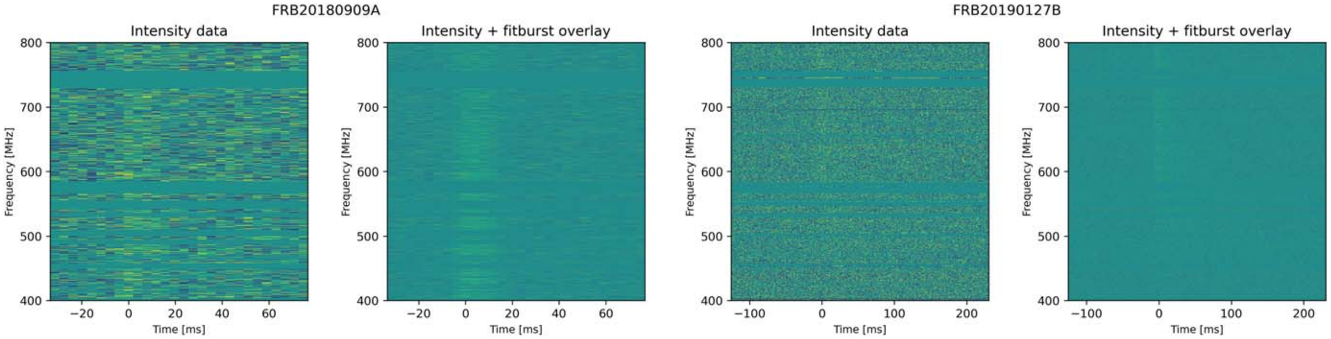

Figure 8. Dynamic spectra or waterfall plots of the two FRBs in the test set (FRB 20180909A and FRB 20190127B) that are ambiguously classified using our method.

may have burst tens or even hundreds of times. Thus, some repeaters have more data available, but this is not considered in the training. In other words, the amount of information provided by each repeater is not the same, a fact that the current model does not take into account. Moreover, subsequent bursts cannot be treated as independent observations since they have single origin. We are currently exploring the use of a mixed effects weighted logistic regression model that will automatically integrate the information from all bursts that have the same source. The results will be communicated in a future paper.

Acknowledgments

We thank an anonymous referee whose comments have helped us improve the paper. This work was supported by a Collaborative Research Team grant to G.M.E. from the Canadian Statistical Sciences Institute (CANSSI), which is supported by Natural Sciences and Engineering Research Council of Canada (NSERC). We thank Amanda Cook for providing specifics about the CHIME/FRB collaboration, helpful comments, and minor editing of the paper. The Dunlap Institute is funded through an endowment established by the David Dunlap family and the University of Toronto. B.M.G. acknowledges the support of the NSERC through grant RGPIN-2022-03163, and of the Canada Research Chairs program. Z.P. was a Dunlap Fellow and is supported by an NWO Veni fellowship (VI.Veni.222.295). G.M.E. also acknowledges the support of NSERC through Discovery Grant RGPIN-2020-04554. K.W.M. holds the Adam J. Burgasser Chair in Astrophysics.

Author Contributions

A.H.M. performed all of the analysis and wrote all of the code for this project. A.H.M. also made all figures and wrote the first and subsequent drafts of the paper. R.V.C., G.M.E., and D.C.S. cosupervised A.H.M. and helped write and edit significant portions of the paper.

Appendix

Figures 9 and 10 illustrate the case of edge or extreme values for the fraction of repeaters in the population.

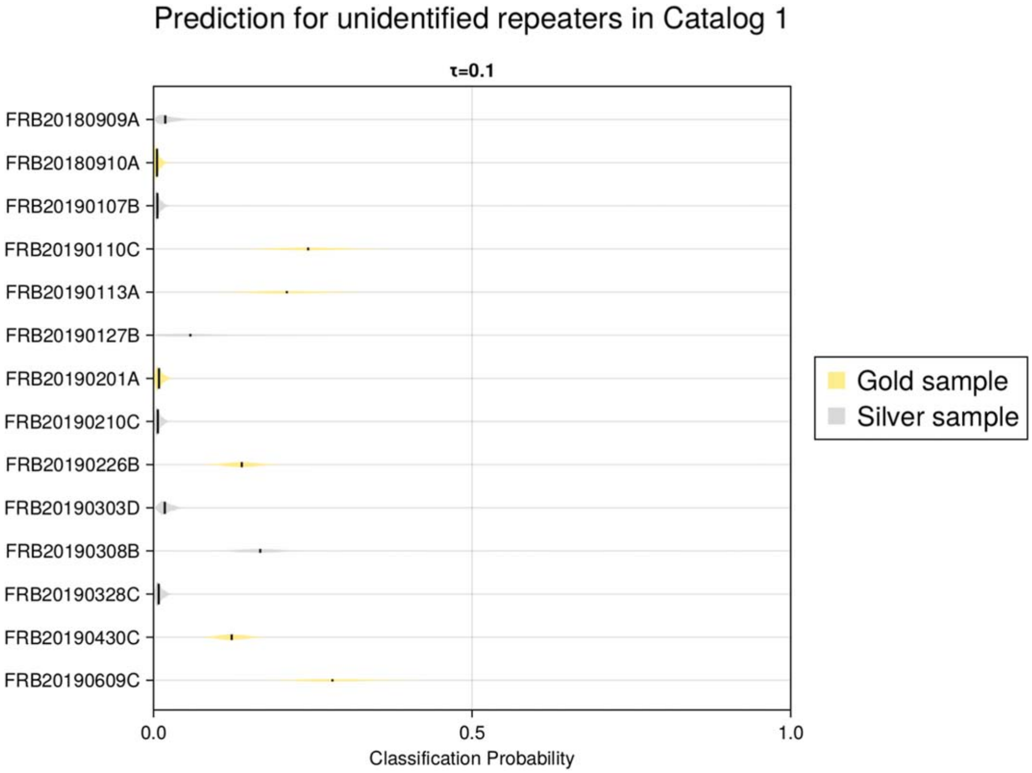

Figure 9. Violin plots of the classification probability for the ensemble of classifiers at $\tau = 0.1$ for the 14 unidentified repeater sources in Catalog 1, similar to Figure 5. At this value of τ everything is classified as a nonrepeater.

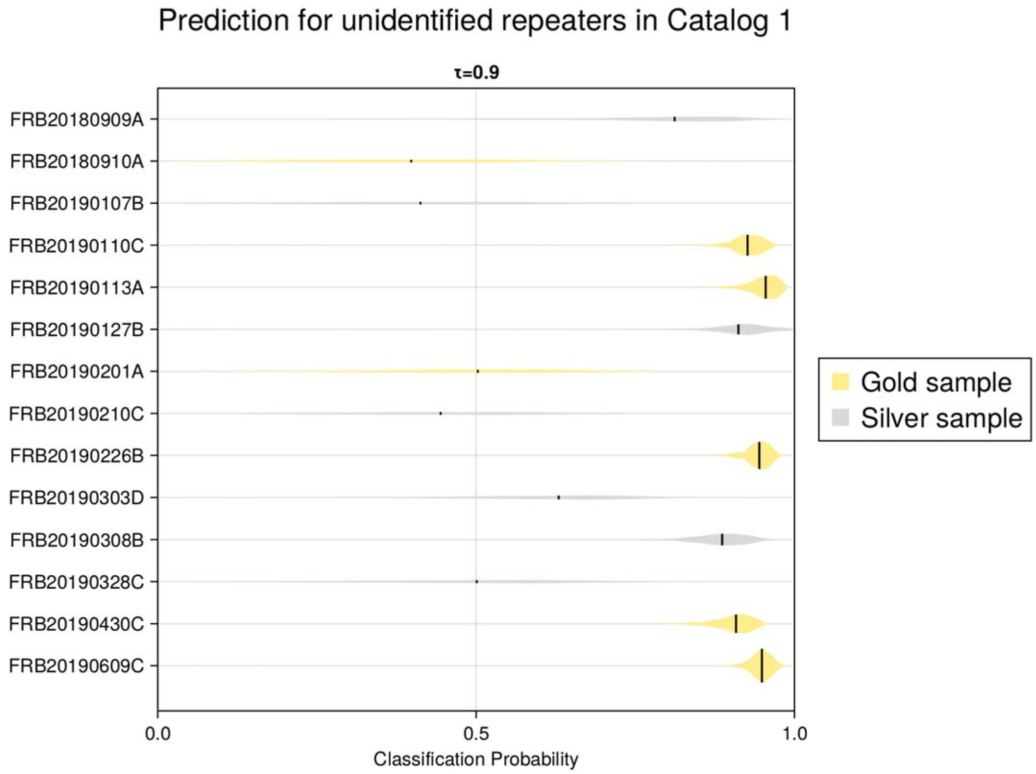

Figure 10. Violin plots of the classification probability for the ensemble of classifiers at $\tau = 0.9$ for the 14 unidentified repeater sources in Catalog 1, similar to Figure 5. At this value of τ there more security of repeater classification, and no event is completely classified as nonrepeater.

ORCID iDs

Antonio Herrera-Martin https://orcid.org/0000-0002-3654-4662
Radu V. Craiu https://orcid.org/0000-0002-1348-8063
Gwendolyn M. Eadie https://orcid.org/0000-0003-3734-8177
David C. Stenning https://orcid.org/0000-0002-9761-4353
Derek Bingham https://orcid.org/0000-0001-5628-7256
B. M. Gaensler https://orcid.org/0000-0002-3382-9558
Ziggy Pleunis https://orcid.org/0000-0002-4795-697X
Paul Scholz https://orcid.org/0000-0002-7374-7119
Ryan Mckinven https://orcid.org/0000-0001-7348-6900
Bikash Kharel https://orcid.org/0009-0008-6166-1095
Kiyoshi W. Masui https://orcid.org/0000-0002-4279-6946

References

Amiri, M., Bandura, K., Berger, P., et al. 2018, ApJ, 863, 48

Amiri, M., Bandura, K., Boskovic, A., et al. 2022, ApJS, 261, 29

Bailes, M. 2022, Sci, 378, eabj3043

Bhardwaj, M., Gaensler, B. M., Kaspi, V. M., et al. 2021, ApJL, 910, L18

Chatterjee, S. 2021, A&G, 62, 1.29

Chen, B. H., Hashimoto, T., Goto, T., et al. 2021, MNRAS, 509, 1227

CHIME/FRB Collaboration, Amiri, M., Andersen, B. C., et al. 2021, ApJS, 257, 59

CHIME/FRB Collaboration, Amiri, M., Andersen, B. C., et al. 2023a, ApJS, 264, 53

CHIME/FRB Collaboration, Andersen, B. C., Bandura, K., et al. 2023b, ApJ, 947, 83

Craiu, R. V., & Sun, L. 2008, Statistica Sinica, 18, 861, https://www.jstor.org/stable/24308520

Fonseca, E., Pleunis, Z., Breitman, D., et al. 2024, ApJS, 271, 49

He, H., & Garcia, E. A. 2009, IEEE Transactions on Knowledge and Data Engineering, 21, 1263

Hung, H., Jou, Z.-Y., & Huang, S.-Y. 2018, Biome, 74, 145

James, C. W. 2023, PASA, 40, e057

Kim, M., & Hwang, K.-B. 2022, PLoSO, 17, e0271260

King, G., & Zeng, L. 2001a, International Organization, 55, 693

King, G., & Zeng, L. 2001b, Political Analysis, 9, 137

Lin, H.-N., & Sang, Y. 2021, ChPhC, 45, 125101

Lorimer, D. R., Bailes, M., McLaughlin, M. A., Narkevic, D. J., & Crawford, F. 2007, Sci, 318, 777

Luo, J.-W., Zhu-Ge, J.-M., & Zhang, B. 2022, MNRAS, 518, 1629

Luque, A., Carrasco, A., Martín, A., & de las Heras, A. 2019, PatRe, 91, 216

Maalouf, M., & Siddiqi, M. 2014, Knowledge-Based Systems, 59, 142

Maalouf, M., & Trafalis, T. B. 2011, Computational Statistics & Data Analysis, 55, 168

Manski, C. F., & Lerman, S. R. 1977, Econometrica, 45, 1977

McCullagh, P., & Nelder, J. 1989, Generalized Linear Models (London: Chapman and Hall)

McCulloch, C. E. 1997, JASA, 92, 162

Nagelkerke, N., & Fidler, V. 2015, PLoSO, 10, e0140718

Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. 2002, The Journal of Educational Research, 96, 3

Petroff, E., Hessels, J. W. T., & Lorimer, D. R. 2022, A&ARv, 30, 2

Pleunis, Z., Good, D. C., Kaspi, V. M., et al. 2021, ApJ, 923, 1

Spitler, L. G., Scholz, P., Hessels, J. W. T., et al. 2016, Natur, 531, 202

Tomz, M., King, G., & Zeng, L. 2003, Journal of Statistical Software, 8, 1

van den Goorbergh, R., van Smeden, M., Timmerman, D., & Van Calster, B. 2022, Journal of the American Medical Informatics Association, 29, 1525

Yamasaki, S., Goto, T., Ling, C.-T., & Hashimoto, T. 2023, MNRAS, 527, 11158

Yang, X., Zhang, S.-B., Wang, J.-S., & Wu, X.-F. 2023, MNRAS, 522, 4342

Yao, J. M., Manchester, R. N., & Wang, N. 2017, ApJ, 835, 29

Zhao, Z.-W., Wang, L.-F., Zhang, J.-G., Zhang, J.-F., & Zhang, X. 2023, JCAP, 2023, 022

Zhu-Ge, J.-M., Luo, J.-W., & Zhang, B. 2022, MNRAS, 519, 1823