Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **474,** 4571–4583 (2018)
Advance Access publication 2017 November 29

doi:10.1093/mnras/stx3047

Ensemble candidate classification for the LOTAAS pulsar survey

C. M. Tan,[1]★ R. J. Lyon,[1] B. W. Stappers,[1] S. Cooper,[1] J. W. T. Hessels,[2,3] V. I. Kondratiev,[2,4] D. Michilli[2,3] and S. Sanidas[1,3]

[1]*Jodrell Bank Centre for Astrophysics, University of Manchester, Oxford Road, Manchester M13 9PL, UK*
[2]*ASTRON, the Netherlands Institute for Radio Astronomy, Postbus 2, NL-7990 AA Dwingeloo, the Netherlands*
[3]*Anton Pannekoek Institute for Astronomy, University of Amsterdam, Science Park 904, NL-1098 XH Amsterdam, the Netherlands*
[4]*Astro Space Centre, Lebedev Physical Institute, Russian Academy for Science, Profsoyuznaya Str 84/32, 117997 Moscow, Russia*

Accepted 2017 November 22. Received 2017 November 20; in original form 2017 July 27

ABSTRACT

One of the biggest challenges arising from modern large-scale pulsar surveys is the number of candidates generated. Here, we implemented several improvements to the machine learning (ML) classifier previously used by the LOFAR Tied-Array All-Sky Survey (LOTAAS) to look for new pulsars via filtering the candidates obtained during periodicity searches. To assist the ML algorithm, we have introduced new features which capture the frequency and time evolution of the signal and improved the signal-to-noise calculation accounting for broad profiles. We enhanced the ML classifier by including a third class characterizing RFI instances, allowing candidates arising from RFI to be isolated, reducing the false positive return rate. We also introduced a new training data set used by the ML algorithm that includes a large sample of pulsars misclassified by the previous classifier. Lastly, we developed an ensemble classifier comprised of five different Decision Trees. Taken together these updates improve the pulsar recall rate by 2.5 per cent, while also improving the ability to identify pulsars with wide pulse profiles, often misclassified by the previous classifier. The new ensemble classifier is also able to reduce the percentage of false positive candidates identified from each LOTAAS pointing from 2.5 per cent (∼500 candidates) to 1.1 per cent (∼220 candidates).

Key words: pulsars: general – methods: data analysis – methods: statistical.

1 INTRODUCTION

Since the discovery of the first pulsar (Hewish et al. 1968), various large-scale surveys have been carried out to search for more of them, in order to characterize their Galactic population and to use them as physical probes (see Lyon et al. 2016 for a full list of pulsar surveys conducted). While there has been a resurgence in finding pulsars via detection of their single pulses, the majority of the pulsars discovered so far, were found through periodicity searches using a fast Fourier transform based method. One of the biggest issues faced by most modern pulsar surveys, is the number of periodicity search candidates produced by this approach. Early pulsar surveys by contrast produced relatively few candidates. For example the 2nd Molonglo survey produced only ∼2500 candidates in total (Manchester et al. 1978). However, as search techniques and telescope sensitivity improves, more and more candidates are generated, including a large number of false positive candidates generated by either noise or radio frequency interference (RFI). Modern surveys such as the Green Bank Northern Celestial Cap pulsar survey (Stovall et al. 2014) have produced more than 1.2 million candidates. It is not feasible to visually examine such a large number of candidates as it is a time consuming and error-prone process (Eatough et al. 2010). Hence, various techniques have been developed to address the issue of candidate numbers.

One of the earliest methods used to reduce the number of candidates to be examined, involved introducing a cut-off based on the signal-to-noise ratio (S/N) of a pulsar candidate in the spectral domain (Stokes et al. 1986). Those candidates with an S/N below the cut-off were considered likely to be noise, whilst those above the cut-off were highlighted for investigation. This method was used by the Arecibo Phase II survey, which reduced the number of candidates to a managable ∼5000. As candidate numbers grew, more sophisticated methods were developed to filter them. For instance, Faulkner et al. (2004) developed a graphical suite known as REAPER. The REAPER tool displayed a visual representation of candidate sets on a customizable 2D plot. This allowed users to rapidly summarize candidates according to key pulsar characteristics (e.g. S/N, DM etc.) and custom heuristics. By focusing attention on regions of the plot most likely to contain real pulsars, promising candidates could be quickly identified for follow-up analysis. An updated version known as JREAPER developed by Keith et al. (2009), assigned scores

★ E-mail: chiamin.tan@postgrad.manchester.ac.uk

based on the aforementioned heuristics, allowing candidates to be ranked based on the overall score achieved. However, as noted by Bates et al. (2012), the JREAPER method is biased against candidates with low S/N, or those having a period that is similar to known RFI sources. An alternate approach was developed by Lee et al. (2013), known as the PEACE algorithm, which linearly combines the scores obtained from six different heuristics in order to rank pulsar candidates.

Recently, several pulsar surveys have deployed machine learning (ML) tools to reduce the number of candidates to be inspected visually (Morello et al. 2014; Lyon 2016; Lyon et al. 2016). The most common branch of ML used for candidate classification is known as supervised learning (Mitchell 1997). In supervised learning, distinct subsets of candidates are labelled as pulsars and non-pulsars to form what is called a training set. A set of variables known as 'features' are then extracted from the candidates, and used by a ML algorithm to derive a mathematical model that can accurately separate candidates into their respective classes. The classification model produced by the algorithm is then used to search for pulsars by classifying unlabelled candidates collected during a survey.

The first application of ML to candidate filtering was accomplished by Eatough et al. (2010), for a reanalysis of the Parkes Multibeam Pulsar Survey data (Manchester et al. 2001). In that work, 12 numerical features inspired by JREAPER were extracted from a sample of candidates, and used to train a form of artificial neural network (ANN) known as the multilayer perceptron (Bishop 1995). This yielded a binary classifier able to assign 'pulsar' and 'non-pulsar' labels to candidates. Bates et al. (2012) later improved upon the classification process by extracting 10 extra features from the candidates collected during the High Time Resolution Universe Pulsar Survey (HTRU, Keith et al. 2010) to train a new classifier. The SPINN system developed by Morello et al. (2014) further built upon this work and developed an ANN system that uses six features extracted from candidates obtained from the HTRU-medlat survey. A different approach was followed by Zhu et al. (2014), who explored the selection problem as a visual learning task by developing the Pulsar Image-based Classification System (PICS). PICS takes the diagnostic plots generated for each candidate obtained from the PALFA survey (Cordes et al. 2006; Lazarus et al. 2015), and feeds them through a combination of several different ML algorithms to produce an image recognition classifier. More recently, Yao, Xin & Guo (2016) addressed the issue of class imbalance between pulsars and non-pulsars examples in training data, by building classifiers solely for the pulsar (positive) class. While Bethapudi & Desai (2017) explored the potential of several newer ML algorithms for pulsar classification. Ford (2017) undertook a similar study whilst also considering the class imbalance.

The LOFAR Tied-Array All-sky Survey (LOTAAS; 2014; Sanidas et al. in preparation) is an ongoing all-Northern-sky pulsar survey conducted using the LOw Frequency ARray (LOFAR; Stappers et al. 2011; van Haarlem et al. 2013). LOTAAS is a sensitive low-frequency (119–151 MHz) pulsar survey that utilizes the unique capabilities of phased-array telescopes in order to observe large areas of the sky for prolonged periods of time (1 h per pointing). The current periodicity search pipeline, which is run on by the Dutch national supercomputer Cartesius,[1] produces on average roughly 20 000 candidates per pointing. More than 1000 such pointings have so far been completed, producing more than 20 million candidates. Since visual inspection of all the survey's candidates is unfeasible on realistic time-scales without a large group of volunteers, Lyon et al. (2016) developed a ML classifier to significantly reduce the vast number of candidates. This is an algorithm which employs a tree-learning approach (Quinlan 1993), chosen specifically to overcome some of the distributional problems identified in pulsar data (Lyon et al. 2013, 2014). The classifier, from now on referred to as LOTAAS classifier 1 (LC1), reduces the number of candidates to be manually inspected from ∼20 000 to a more manageable ∼500 per pointing, filtering out ∼97.5 per cent of the candidates. Testing done on LC1 shows an estimated overall accuracy of 96.8 per cent.

Although LC1 is highly successful in identifying pulsars, shortcomings have been found with the data and features used to train the classifier. We hereby implemented several improvements to the heuristics and the ML classifier to overcome the issues. This resulted in the creation of a new classifier, LOTAAS classifier 2 (LC2). In Section 2, we describe the shortcomings in the classifier model of LC1. In Section 3, we describe the various improvements made to the classifier, primarily the introduction of new features and a separate class of RFI instances. In Section 4, we describe an ensemble approach of several distinct versions of the classifier used to improve classification. This approach improved the pulsar recall rate while reducing the false positive rate (FPR; see the definitions in Table 5). In Section 5, we discuss the performance of LC2 compared to LC1 on actual LOTAAS data. We conclude in Section 6 by describing the improvements achieved by LC2, potential future modifications and summarize how the new system has been implemented within the survey search pipeline.

2 ISSUES WITH LOTAAS CLASSIFIER 1

LC1 utilizes eight features obtained from the integrated pulse profile ($Prof_\mu$, $Prof_\sigma$, $Prof_S$, $Prof_k$) and the DM–$\chi^2_{\rm red}$ plot (DM_μ, DM_σ, DM_S, DM_k). These are plots A and D of Fig. 1. The mathematical description of the features used is shown in Table 1. The classifier model was built using a training set obtained by Cooper (2017) and a decision tree algorithm (Quinlan 1993). The tree algorithm examines the distribution of each feature variable. It uses an optimization function to find the feature variable that yields the 'best' class separation. Each time an optimal feature is found, a numerical split-point over the feature that maximizes class separation is identified. The training data are then split into two subsets based on the numerical split-point found. A single split-point over one feature is typically not very useful. However, many splits recursively combined partition the data into smaller and smaller groups, leading to more complex, and more accurate decision paths. Once constructed, the classifier is able to label each candidate as either a pulsar or a non-pulsar. An illustration of a decision tree is shown in Fig. 2.

Despite the success of LC1 in identifying many new pulsars (see Sanidas et al. in preparation and the LOTAAS website[2] for more details on the discoveries), it was suspected and subsequently verified through a series of tests, that some pulsars, particularly those with wide integrated pulse profiles, either intrinsically or due to large scattering tails, were regularly missed during classification (see Fig. 3 for the profiles of the pulsars that were misclassified). This was confirmed by folding the data from the survey pointings directed at positions of known pulsars with available rotational ephemerides. We then visually inspected the candidates from these pointings, and found that these pulsars were detected by the LOTAAS search

[1] https://userinfo.surfsara.nl/systems/cartesius

[2] http://www.astron.nl/lotaas

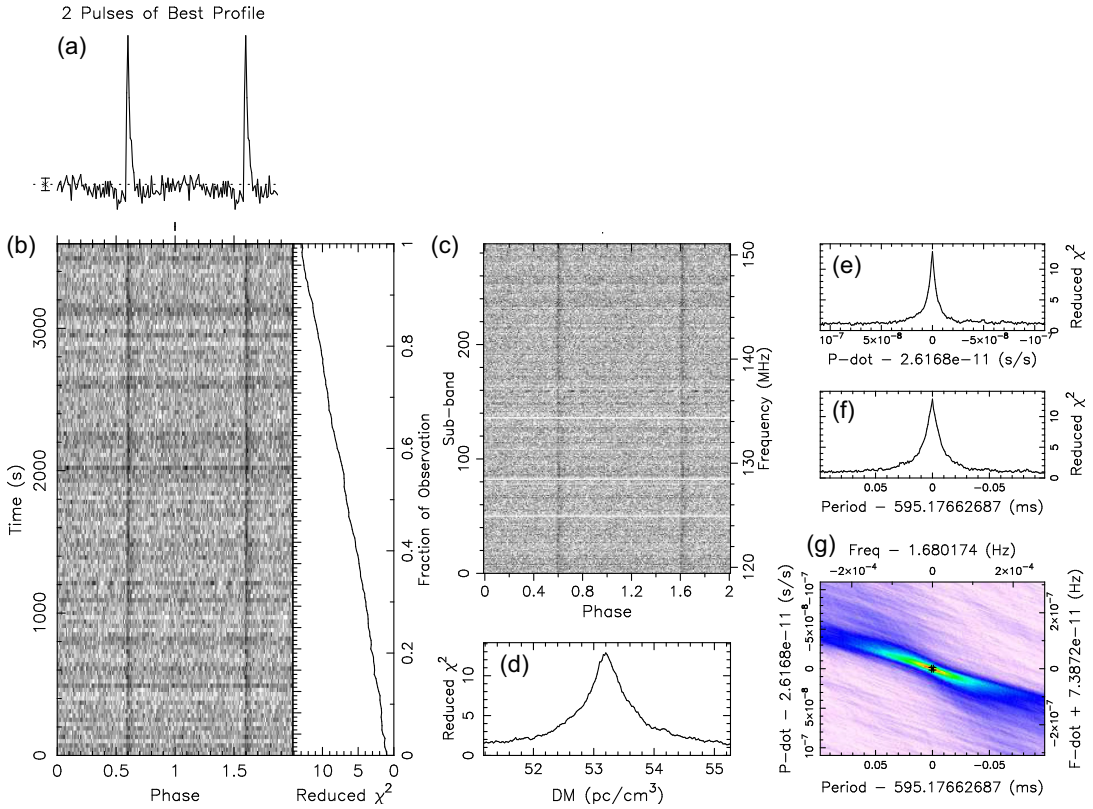

Figure 1. The diagnostic plot of a pulsar detected by the LOTAAS search pipeline obtained using the pulsar searching suite PRESTO (Ransom 2001; Ransom, Eikenberry & Middleditch 2002). The plot shows, A: the integrated pulse profile of the pulsar, B: pulse intensity (grey-scale) as a function of pulse phase and time, C: pulse intensity as a function of frequency sub-band and pulse phase, D: DM versus reduced χ^2, E: period versus reduced χ^2, F: period derivative versus reduced χ^2 plots and G: A heatmap of the reduced χ^2 of the pulsar over a range of period and period-derivative values. The calculation of reduced χ^2 is described in Section 3.1. A pulsar will exhibit a peak in the integrated pulse profile, and the peak will generally be consistent over time, and across the sub-bands, while showing a well-defined maximum value on the reduced χ^2 plots.

Table 1. The four features extracted from both the integrated pulse profile and the DM–$\chi^2_{\rm red}$ plot by Lyon et al. (2016). Here, n is either the total number of DM values or pulse phase bins plotted and y represents either the $\chi^2_{\rm red}$ or the pulse intensity values.

Feature	Definition
Mean, μ	$\frac{1}{n}\sum_{i=1}^{n} y_i$
Standard deviation, σ	$\sqrt{\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{n-1}}$
Skewness, S	$\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^3}{(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2})^3}$
Excess kurtosis, k	$\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^4}{(\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2)^2} - 3$

pipeline, but were incorrectly labelled by the classifier as non-pulsars.

While the presence of misclassified pulsars is concerning, they only represent roughly 3 per cent of all the different known pulsars in the survey. Although they appear to have wide profiles it was not clear from the current features why they were missed. However, the scatter plot in Fig. 4 suggests that it is due to class (i.e. pulsars and non-pulsars) inseparability. Fig. 4 shows the distribution of pulsars and non-pulsars in the space of the two most separable features: profile skewness ($Prof_S$) and kurtosis of the DM–$\chi^2_{\rm red}$ curve (DM_k). We found that $Prof_S$ is able to separate a large number of pulsars from non-pulsars. As most of the pulsars have a narrow integrated pulse profile, there is no emission at the majority of pulse phases. This gives rise to the high $Prof_S$ values as the distribution of the intensity values are highly skewed towards the positive side.

However, some pulsars with low $Prof_S$ are mixed deeply within the non-pulsar space. These are found to be the pulsars with wide integrated pulse profiles, either intrinsic to the pulsar or due to having large scattering tails. The scatter plot suggests that DM_k is unable to separate these pulsars from the non-pulsars. A study of the remaining features also revealed that they are unable to assist in separating these pulsars from non-pulsars. Hence, to derive an improved class separation in these dense regions, we needed to explore other possible features, as well as possible flaws in the preprocessing of the data. We also needed to improve the training data set, by including more pulsar examples that are misidentified by LC1, as these misclassified examples may assist with classification in the dense $Prof_S$–DM_k region.

3 IMPROVEMENTS TO THE CLASSIFIER IMPLEMENTATION

3.1 Data preprocessing

While we were investigating the cause of classifier error, we noticed that the script used to generate the features from the DM–$\chi^2_{\rm red}$ curve for LC1 contains an error which potentially affects the

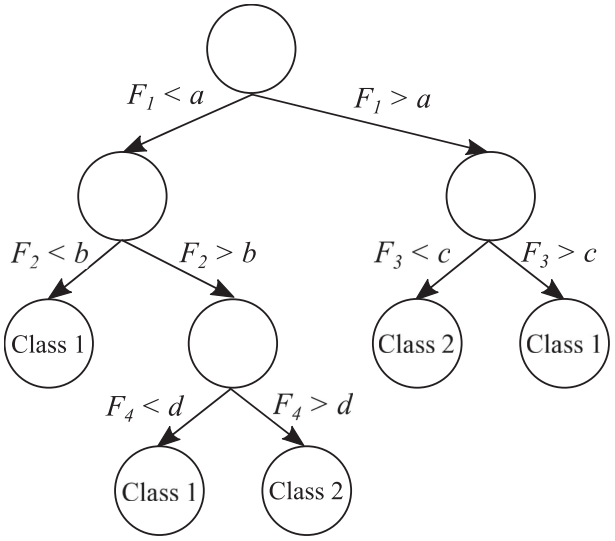

Figure 2. An illustration of a binary decision tree classifier. At each node of the decision tree the 'best' feature $F = \{F_1, F_2, \ldots, F_n\}$ is used to separate the data by identifying a numerical threshold. After each split-point, the separated data are re-evaluated to select a new 'best' feature to further split the data, until the tree reaches a decision at the leaf nodes where the data in each node are assigned a class.

classification process. The PRESTO toolkit, which is used in the LOTAAS search pipeline, calculates the significance of a signal by using a χ^2 measurement of the deviation of each bin in the integrated pulse profile from the average. When PRESTO makes its diagnostic plots it applies a correction to the $\chi^2_{\rm red}$ values that takes into account the effective number of bins used to fold the data for a particular candidate. However, this was not performed when the features were generated for LC1. This oversight was rectified when the features were generated for LC2. Without the correction the $\chi^2_{\rm red}$ values of candidates with short periods, especially millisecond pulsars, will have a much lower $\chi^2_{\rm red}$ value than expected.

Despite applying this correction to the $\chi^2_{\rm red}$ values, we suspect that DM–$\chi^2_{\rm red}$ curve is sub-optimal for pulsars with wide integrated pulse profiles, either intrinsically or due to scattering, and it may adversely affect the performance of our ML systems. To overcome this, we decided to use a box car convolution method to measure the S/N of the pulse (Edwards et al. 2001). First, the integrated pulse profile is convolved with a box car of various trial sizes to search for the baseline of the profile (minimum value obtained from the convolution). This value is then subtracted from the profile. A second box car convolution is computed to look for the best on-pulse region (which returns the maximum value of the convolution). The width of the box car ranges from 3 profile bins to 80 per cent of the

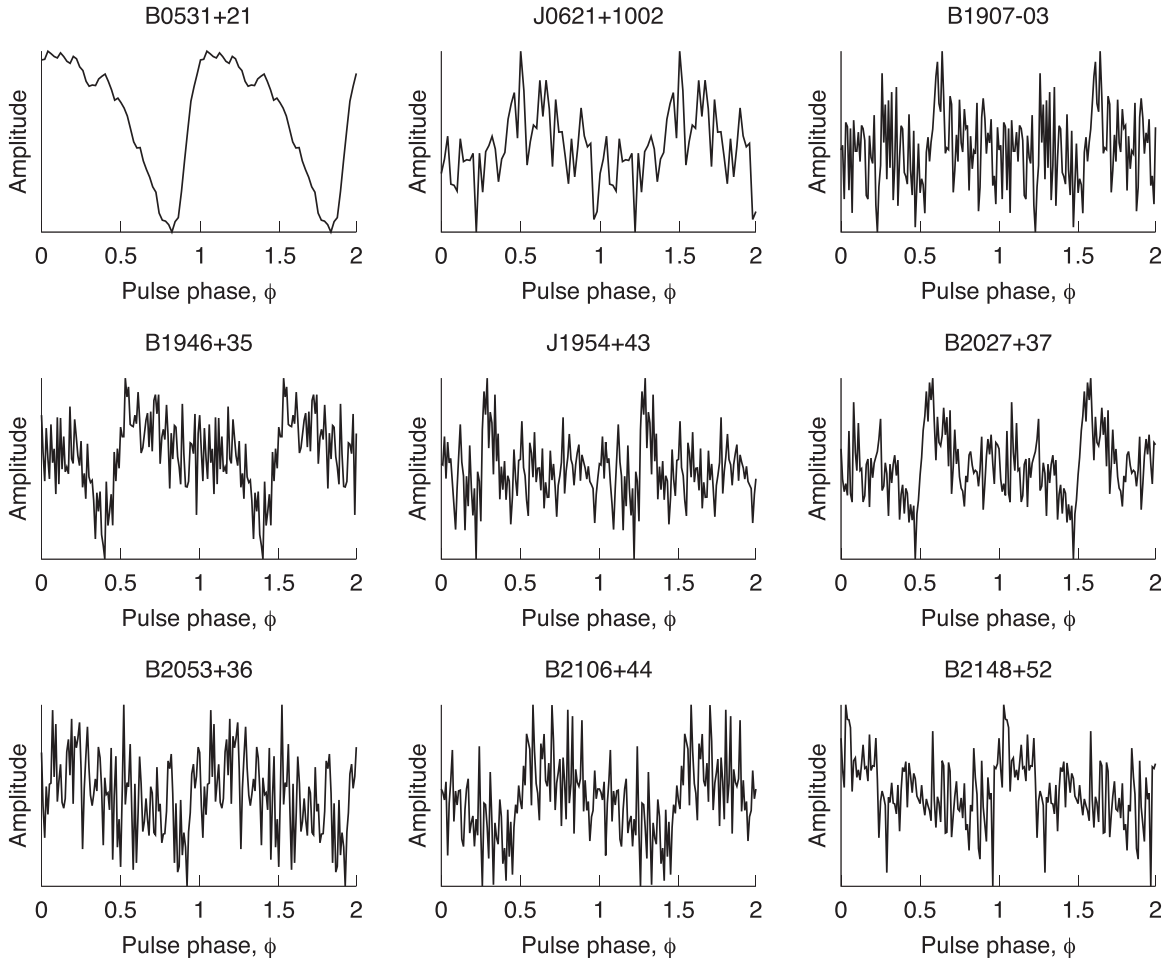

Figure 3. The integrated pulse profiles of the pulsars that were misclassified by the LC1. PSR J0621+1002 has a intrinsically wide integrated pulse profiles and the others are heavily scattered at LOFAR frequencies. Two periods of the integrated pulse profile are plotted to show the profile more clearly. The amplitudes of the integrated pulse profiles are in arbitrary units.

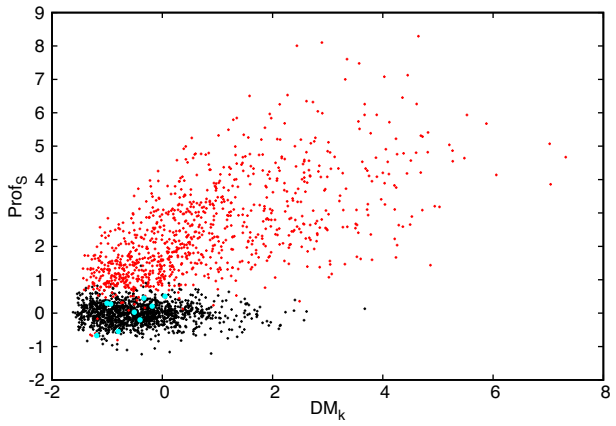

Figure 4. Scatter plot of the distribution of a sample of 986 pulsar detections (red) and 1267 non-pulsar detections (black) in the feature space of the profile skewness ($Prof_S$) and kurtosis of the DM–$\chi^2_{\rm red}$ curve (DM_k). The examples of pulsars misclassified by LC1 in Fig. 3 are shown in cyan. The plot shows that most of the known pulsars can be separated from the non-pulsars with just $Prof_S$, leaving a small number of pulsars embedded within the non-pulsars region.

pulse phase. This maximum obtained value is then used to calculate the S/N of a pulsar:

$$S/N = \frac{\max(convolution)}{\sqrt{\sigma^2}}\sqrt{\frac{n_{\rm on}n_{\rm off}}{n}}, \quad (1)$$

where σ^2 is the variance of the integrated pulse profile, $n_{\rm on}$ is the number of profile bins in the on-pulse region, $n_{\rm off}$ is the number of profile bins in the off-pulse region, and n is the total number of profile bins. As the convolution assumes the pulse to have a top-hat function, and the box car function used has a height of 1/width, a normalization factor, $\sqrt{\frac{n_{\rm on}n_{\rm off}}{n}}$ is required to correct for the shape of a real pulse. We then constructed the DM–S/N curve of each candidate using the method described above, using the same DM range as the DM–$\chi^2_{\rm red}$ curve from the diagnostic plots. We then compared the performance of two separate test classifiers using features generated from either the DM–S/N curve or DM–$\chi^2_{\rm red}$ curve. We found that we are able to recover more pulsars using the features generated from the DM–S/N curve. Hence, we decided to extract features from the DM–S/N curve, instead of the DM–$\chi^2_{\rm red}$ curve for LC2.

In the LOTAAS pipeline candidates can be folded with different numbers of profile bins, sub-integrations, and sub-bands depending on the period of the candidate. However, the different data ranges, that the candidates have, is problematic because of a basic assumption used in ML. For ML algorithms to work we need the input data to be independent and identically distributed (i.i.d.; Bishop 2006), which is violated in this case. We found that the distributions of the features generated from the pulse intensity as a function of time and pulse phase, and the pulse intensity as a function of frequency sub-band and pulse phase plots (plots B and C of Fig. 1, see Section 3.2 on the features generated) are different for candidates with different numbers of sub-integrations and/or sub-bands. To overcome this issue for LOTAAS, which like all pulsar surveys often produces candidates with varying data ranges, we ensured that the number of sub-bands and sub-integrations for all candidates are the same. However, for the number of profile bins, due to the large time resolution (492 μs), candidates with periods less than 50 ms have to be folded with 50 profile bins, as there are a limited number of independent time samples. For candidates with periods of more than 50 ms, 100 profile bins are used to preserve the resolution of the pulse profile, as long period pulsars often have small duty cycles that are not resolvable at 50 bins. Although it is advisable to have separate classifiers for the two types of candidates to preserve the i.i.d. assumption, we did not split up the data due to the lack of available pulsar examples in which the period is less than 50 ms.

Table 2. The new features extracted from the DM–S/N curve that was described in Section 3.1. Here, x and y are the DM and S/N values on the DM–S/N curve, respectively.

Feature	Definition
$DM_{\mu'}$	$\frac{\sum xy}{\sum y}$
$DM_{\sigma'}$	$\sqrt{\frac{\sum (x-DM_{\mu'})^2 y}{\sum y}}$
$DM_{\lvert S'\rvert}$	$\lvert\frac{1}{DM^3_{\sigma'}}\frac{\sum (x-DM_{\mu'})^3 y}{\sum y}\rvert$
$DM_{k'}$	$\frac{1}{DM^4_{\sigma'}}\frac{\sum (x-DM_{\mu'})^4 y}{\sum y} - 3$

3.2 New features

In order to improve the performance of the ML classifier, the incorporation of additional features to complement those by Lyon et al. (2016) was also investigated. While inspecting the diagnostic plots of candidates by eye, the two plots commonly used to identify pulsars, are the pulse intensity as a function of time and pulse phase, and the pulse intensity as a function of frequency sub-band and pulse phase plots. A pulsar will generally have a consistent signal that appears across both plots, aligned in phase with the peak seen in the pulse profile.[3] To look for the presence of such characteristics, we used a method first described by Bates et al. (2012). This involves calculating the correlation coefficient between each individual sub-band and the integrated pulse profile, as well as between each sub-integration and the integrated pulse profile. A pulsar should be characterized by larger correlation coefficients in both plots compared to a non-pulsar example. Two lists of the values of correlation coefficients are computed between all 36 sub-bands and the integrated pulse profile, as well as between all 40 sub-integrations and the integrated pulse profile. We then extract the features, specified in Table 1, from these lists. This gives us eight new features.

A second new set of features is obtained for the classifier when a new method for calculating the statistics of the DM–S/N curve is applied. The new set of statistics calculates the mean, standard deviation, absolute skewness, and excess kurtosis of the curve, which takes into account the values of the curve in the DM axis as well as the S/N axis as indicated in Table 2. When a pulsar candidate is found, the candidate might not be found in its optimal DM value that yields the highest S/N. Hence, a search around that DM value if done to find the best DM value for the candidate. For a real pulsar detection, the best DM value should be close to the DM value where the candidate is found, and hence, in the middle of the DM–S/N curve. For the new features to be able to identify the location of

[3] However, certain pulsars will show variation in signal over time and/or frequency. See Section 3.3 for a discussion on the variation in the frequency domain. The variation in time is usually caused by intermittency of the pulsar.

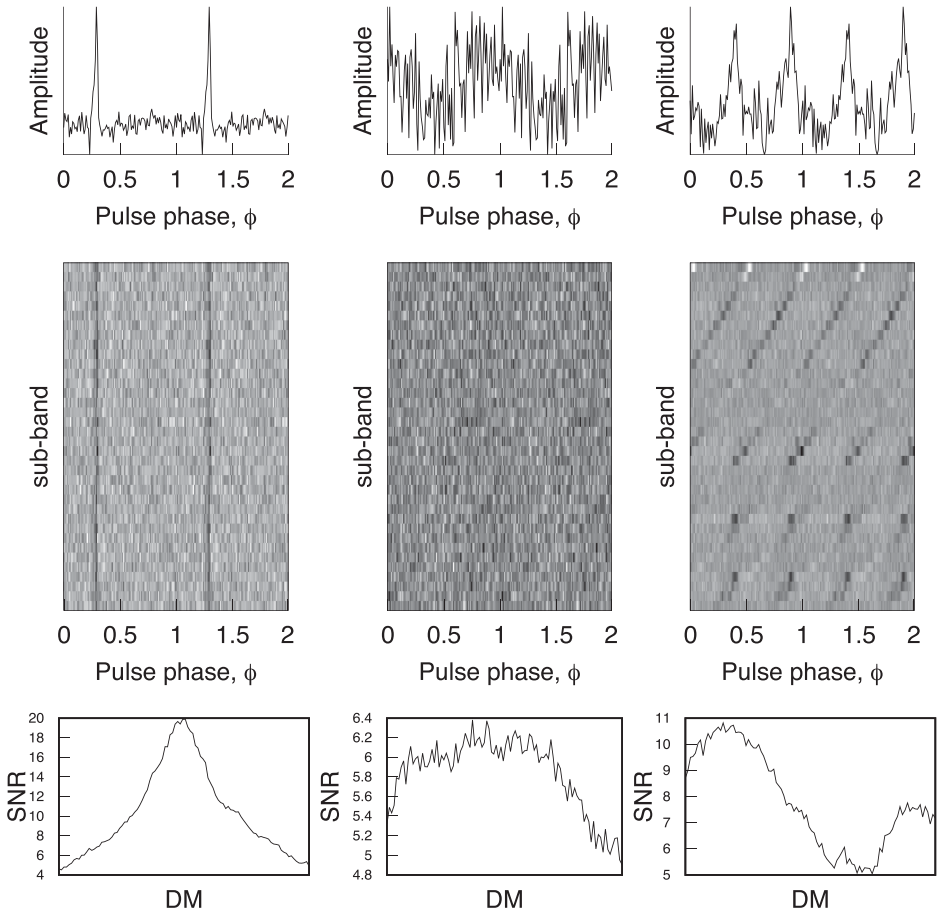

Figure 5. A comparison of the integrated pulse profile (top row), pulse intensity as a function of frequency sub-band and pulse phase plot (middle row), and DM–S/N curve (bottom row) of a typical pulsar (left), a noise-like detection (middle), and an RFI instance (right).

the peak, the values of the DM axis are changed from actual DM values of the trials to the integer number of the DM trials, so that the value of $DM_{\mu'}$ are similar for candidates where the peak is at the middle of the curve. These new features are in principle able to gauge the shape of the DM–S/N curve better than the statistics used by Lyon et al. (2016), which only takes into account the values of the curve on the S/N axis. The Lyon et al. (2016) features are able to detect the presence of a peak in the DM–S/N curve, without providing information on the location and symmetry of the peak. We note that the absolute value of the skewness is used instead of actual the skewness value, as we only want to know if the shape of the DM–S/N curve is skewed without knowing the direction towards which the curve is skewed. With the use of both sets of features for the DM–S/N curve, we are able to better gauge the shape of the curve.

3.3 Defining an RFI class

In addition to generating new features, we have added an RFI classification group alongside the pulsar and non-pulsar options included by Lyon et al. (2016). This new class has been added to represent specific types of RFI which strongly mimic pulsars. Many RFI instances were labelled as pulsars by the LC1. A close inspection of the diagnostic plots of these instances, revealed that they possess a pulsar-like pulse profile, and for certain types of RFI, the DM–S/N curves are similar too. However, these instances do exhibit a notable difference in the pulse intensity as a function of frequency sub-band and pulse phase plot. Fig. 5 shows the integrated pulse profile, pulse intensity as a function of frequency sub-band and pulse phase plot, and the DM–S/N curve of a typical pulsar, a noise-like non-pulsar, and an RFI instance, respectively.

The profile of the pulsar in Fig. 5 has a strong narrow peak, and the pulse intensity as a function of frequency sub-band and pulse phase plot has a consistent signal that correlates with the location of the peak of the pulse. On the other hand, the profile of the noise-like detection has no obvious pulse, and no clear signal is seen in the pulse intensity as a function of frequency sub-band and pulse phase plot. The RFI instance has two peaks in its profile that resembles a pulsar with an interpulse. However, the pulse intensity as a function of frequency sub-band and pulse phase plot shows that only a fraction of the sub-bands contain a signal that correlates with the pulse.

The DM–S/N curve for the three types of candidates are different from each other too. A typical pulsar will show a clear peak at a specific DM value of the curve, while the DM–S/N curve of a noise-like candidate will not show a clear peak. The RFI instance shows two maxima on the DM–S/N curve, which comes from the signal at different bands lining up at two different DM values. As the original features used did not describe the DM–S/N curve well on its own, the candidate is still likely to be misclassified as a pulsar by LC1.

To further illustrate the difference between the two types of non-pulsar candidates, we generated box and whisker plots showing the distribution of the standard deviation of the correlation coefficients,

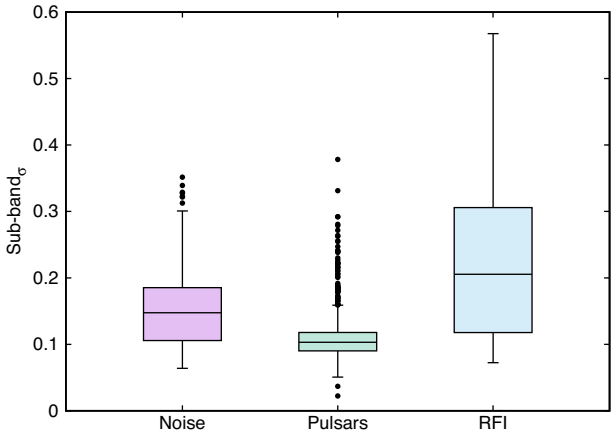

Figure 6. Box and whisker plots showing the distribution of the values of the standard deviation of the correlation coefficients between each sub-bands and the integrated pulse profile, *Sub-band*$_\sigma$, of the three different classes of candidates. The boxes show the median and interquartile ranges (IQRs, Q3−Q1, where Q1 and Q3 are the values at the end of the boxes) of the distribution of each class of candidates, with the middle 50 per cent of the distributions within the boxes. The whiskers show the range of the distribution, with the end of the whiskers being Q1−1.5IQR and Q3+1.5IQR. The dots show the outliers of the distributions.

between each sub-band and the integrated pulse profile for the three different types of candidate (see Fig. 6). It shows that noise-like candidates have higher median value of the feature compared to pulsars. RFI instances have the highest median among the three classes; this is because there are several sub-bands that are highly correlated with the pulse profile while others are not. Collectively this shows that the distributions of the three classes are different from each other. To further show that noise-like candidates and RFI instances have different distributions, we did an independent *t*-test on the distributions of the value of *Sub-band*$_\sigma$ of the two non-pulsar classes. The independent *t*-test gives the probability that the distributions of the two sets of values are the same. The test shows that the probability of the two classes having the same distribution is 7×10^{-12}. Hence, we decided to split these RFI instances into a separate class of non-pulsars to improve classification accuracy.

Although *Sub-band*$_\sigma$ showed promise in separating the three different classes, Fig. 6 suggests that a small number of pulsars exhibits large *Sub-band*$_\sigma$ values similar to typical RFI detections. A study of these pulsars revealed two main effects causing a high *Sub-band*$_\sigma$ value. First, some of these pulsar detections are affected by scintillation. This causes pulses to be brighter than average, for some sub-bands. Secondly, there are pulsars that are brighter in the lower part of the observing band. This can be caused by two different scenarios. The beam size at the lower part of the observing band of LOTAAS is larger than the beam size at the upper part of the observing band. Hence, this effect will show up if a pulsar is detected at the edge of a beam. There are also pulsars with steep spectra, in which the pulsars are intrinsically much brighter at lower observing frequencies. These resulted in variation in the correlation coefficient between the profile in each sub-band and the integrated pulse profile. However, inspections on pulsars affected by these two cases showed that the DM–S/N curve of the pulsars still shows a well-defined peak. Hence, these pulsars would still be identified by the features generated from the DM–S/N curve. There is also a possibility of profile evolution of the pulsar across the observing bandwidth. However, we expect the profile evolution is only visible at high S/N and the change across the band to be gradual and not show the distinct narrow-band signature of RFI. Hence, we do not expect pulsars that show profile evolution to produce a high *Sub-band*$_\sigma$ value.

3.4 New training set

We improve upon the training data used by Cooper (2017) to build the classification model for LC2. The newly acquired training data consists of 247 known pulsars redetected by LOTAAS, and 33 new pulsars discovered by the survey at the time of compilation of the training set. These include those that were not detected by LC1 (see Fig. 3). The new sample also included 17 pulsars observed as part of the LOFAR HBA pulsar census (Bilous et al. 2016), including 15 that were not detected by LOTAAS. These pulsars were either in the area not covered by LOTAAS yet, or were too faint to be detected with the LOTAAS setup which uses fewer LOFAR stations for the observations. The data from the HBA census have a larger bandwidth than a typical LOTAAS observation. Hence, in order to simulate a LOTAAS observation, the bandwidth is reduced to match the LOTAAS bandwidth before the data are folded using the ephemeris of the pulsar. In total, 986 different detections of 295 unique pulsars were used to build the classifier. How the sample of pulsars is used in the production of the ML classifier will be discussed below in Section 4.

The sample of noise-like candidates and RFI instances were obtained from 50 different LOTAAS pointings. The pointings from which the candidates are chosen, were randomly selected from all available survey pointings at the time of the analysis. These candidates were chosen at random from each pointing, manually labelled as either noise-like or RFI based on the correlation based criteria. A random sampling was used because the environment around the telescope varies over time. This in turn produces noise-like candidates, as well as RFI instances that differ over time. In total, 1267 noise-like candidates and 150 RFI instances were obtained for the two different non-pulsar classes.

3.5 Feature evaluation

Before any new features are used as inputs into the ML classifier, we need to test their viability, to ensure that they are able to improve the separability of the data. We used the method that Lyon et al. (2016) employed to study the performance of the features. This is known as Information Theoretic Analysis. In information theory, each feature is described in terms of entropy (Shannon & Weaver 1949). The entropy of a feature X^J is defined as

$$H(X^J) = -\sum_{x \in X^J} p(x) \log_2 p(x), \tag{2}$$

where x is a possible value that X^J can take, and $p(x)$ is the probability of the occurrence of x. Entropy provides information on the amount of uncertainty present in the distribution of X^J. The entropy is small if the values of X^J are biased towards a small range of x. On the other hand, if all values of x are equally likely to occur, then the entropy of X^J would be at its maximum. The entropy of a feature variable can be tied to another variable, in this case the class of the data. This is known as the conditional entropy. The conditional entropy of a feature X^J given the class Y is

$$H(X^J|Y) = -\sum_{y \in Y} p(y) \sum_{x \in X^J} p(x|y) \log_2 p(x|y), \tag{3}$$

where $p(x|y)$ is the probability of x given y. The conditional entropy quantifies the amount of uncertainty in X^J after knowing the outcome

of the class Y. If knowing Y reduces the uncertainty surrounding the value of X^J, then intuitively there must be some correlated information between the two. For example, if X^J represents the house price, and Y the number of bedrooms, one would expect the two to correlate, i.e. houses with more bedrooms are more expensive. If Y were house height however, the correlation would likely be weaker. By subtracting the conditional entropy (equation 3) from the entropy (equation 2), we can calculate the mutual information (MI; Brown et al. 2012) between X^J and Y. It is defined as

$$I(X^J; Y) = H(X^J) - H(X^J|Y), \tag{4}$$

where MI represents the amount of uncertainty in X^J that is removed after Y is known. If $I(X^J; Y) = 0$, then X^J and Y are independent of each other. On the other hand, if $I(X^J; Y) > 0$, then by knowing the class Y, we increase the understanding of the feature X^J. It is therefore desirable for classification features to have higher MI content. This is because we would need our pulsar features to be correlated with the class variable Y. By computing MI values for our new features we can not only determine their utility, but also compare them directly with the features used to build LC1. Furthermore, any new features developed in the future can be compared directly to those presented here using the same analysis.

Whilst MI provides details on how each feature performs against each other, it is possible for two or more features to possess redundant information (Guyon & Elisseeff 2003). Hence, if we only use a subset of the features with the highest MI value, we might end up with a sub-optimal subset of features for classification. A ranking system, known as the joint mutual information criterion (JMI; Yang & Moody 1999) was developed to rank the set of features after taking redundancy into account. JMI ranks the features by first choosing the feature with the most MI, X^1, and then uses a greedy iterative process, known as 'forward selection' (Kohavi & John 1997; Guyon & Elisseeff 2003), to look for the most complementary features to X^1, using the JMI score,

$$JMI(X^J) = \sum_{X^K \in F} I(X^J X^K; Y), \tag{5}$$

where $X^J X^K$ is the joint probability of the two features, and F is the set of features selected. The process carries on until a desirable set of features is chosen, or all the features are ranked accordingly.

The Information Theoretic Analysis outlined above requires the data to be discretized. We employed a Minimum Description Length approach (Fayyad & Irani 1993) to discretize the data. The entropy, MI, and JMI of the features are then calculated using the MITOOLBOX and FEAST suites, developed by Brown et al. (2012). Since the statistics above only consider features used for binary classification, we calculated the statistics between noise and pulsars, and pulsars and RFI, separately. The results are presented in Table 3.

The MI values obtained suggest that features extracted from the integrated pulse profile are the best at separating pulsars from both non-pulsar classes, with profile skewness being the highest ranked feature in both cases. When considering pulsars and noise-like candidates, the MI values for all the features extracted from the integrated pulse profile, are much higher than the rest. However, an analysis of the misclassified pulsars showed that features from the integrated pulse profile are poor at separating pulsars with wide profiles from non-pulsars. Hence, we looked into the JMI ranking for other features to assist us in classifying these pulsars.

The JMI ranking comparison of the features between pulsars and noise-like candidates, showed that *Sub-int*$_\sigma$ is the best feature after *Prof*$_S$ in separating pulsars from noise-like candidates. As we expect signals from most of the pulsars to be consistent in time, the *Sub-int*$_\sigma$ of pulsars will achieve low values to reflect that. On the other hand, noise-like candidates should have a random set of values for the correlation coefficients which should result in a larger value of *Sub-int*$_\sigma$. Besides that, other features extracted from the DM–S/N curve showed high JMI ranking, particularly the mean and standard deviation. This is likely due to the new method used to calculate the S/N of the pulse profile being more effective, producing a DM–S/N curve with a narrower peak and higher peak S/N. The combination of both gives a larger value for the mean and standard deviation of the DM–S/N curve compared to the noise-like candidates.

The JMI ranking of the features between pulsars and RFI instances, showed that *Sub-band*$_\sigma$ is the best feature after *Prof*$_S$ as expected. The next best feature is $DM_{\sigma'}$, as RFI instances usually have DM–S/N curves that have very different shapes compared to pulsars. Signals from RFI instances usually do not converge into a single DM value, and hence show multiple maxima in the DM–S/N curve. While the DM–S/N curve of a pulsar typically shows a well-defined maximum. These characteristics are better represented with the new features extracted from the DM–S/N curve.

The result from the Information Theoretic Analysis showed that the new features are helpful in separating pulsars from non-pulsars, with the JMI ranking showing the order of usefulness of each feature in separating the data. It also showed that those features giving the best separation between noise and pulsar examples, differ from those giving the best separation between pulsar and RFI examples. Hence, it is advisable to split the non-pulsar class into noise and RFI, in order to improve the ability of the classifier to identify pulsars.

4 ENSEMBLE APPROACH

A single LOTAAS pointing consists of three separate sub-array pointings (SAPs) formed by six Superterp stations, a 300 m diameter circular area where the LOFAR stations are most densely packed. The three SAPs are directed towards three different but nearby parts of the sky. An incoherent beam is formed for each of the three SAPs. The central region of each SAP is also tiled with 61 tied-array beams (TABs), in which the stations are combined coherently, arranged in a hexagonal grid. Also 12 additional TABs are formed in the direction of the known pulsars within the field of view (FoV) of a SAP but outside the tiled-up central part (see Sanidas et al. in preparation; Coenen et al. 2014; Cooper 2017, for details on the arrangement of the beams). Hence, when a bright pulsar is within the FoV of the pointing, it can be detected in multiple beams.[4] The pulsar can be detected either at the centre, the edge or the sidelobe of a beam. A bright pulsar can be detected in multiple adjacent pointings as well. The S/N of the detection will vary depending on the location of the pulsar within the beam, due to the sensitivity of the beam pattern on the sky (the structure of a LOFAR beam can be seen in fig. 27 of van Haarlem et al. 2013). This gives us a very large sample of pulsars of various S/N to train our ML classifier. However, most of the bright pulsars have the standard narrow integrated pulse profile which are easily detected by LC1. To properly represent the population of the pulsars visible to LOTAAS, and to avoid a training set biased towards certain types of pulsars, we decided to include only one instance of each of the 295 known pulsars in the initial training set.

We then noticed that for different detections of the same pulsar, the features generated could vary. Table 4 shows the values of several features of two detections of PSR B0329+54 at different S/Ns. This showed that feature values generated from different detections of

[4] Beams refer to both the incoherent beams and TABs.

Table 3. The entropy, $H(X^J)$, MI, $I(X^J; Y)$, and JMI ranking, $JMI(X^J)$ of each of the features. Values are given for comparison between noise and pulsars, as well as between pulsars and RFI. *Prof* indicates features obtained from the pulse profile, *DM* indicates features obtained from the DM–S/N curve, *Sub-band* and *Sub-int* indicate the features obtained from the lists of correlation coefficients between each sub-band and the integrated pulse profile and between each sub-integration and the integrated pulse profile, respectively. The subscripts show the type of features as described in Tables 1 and 2.

Feature	Noise–pulsars			Pulsars–RFI		
	$H(X^J)$	$I(X^J; Y)$	$JMI(X^J)$	$H(X^J)$	$I(X^J; Y)$	$JMI(X^J)$
$Prof_\mu$	1.94	0.74	4	1.55	0.34	5
$Prof_\sigma$	2.49	0.60	6	2.02	0.25	7
$Prof_S$	1.57	0.83	1	1.27	0.37	1
$Prof_k$	1.91	0.79	3	1.45	0.37	4
DM_μ	2.14	0.19	7	2.34	0.05	12
DM_σ	1.96	0.30	5	1.65	0.13	10
DM_S	2.14	0.27	9	1.88	0.08	13
DM_k	1.36	0.09	14	1.01	0.02	17
$Sub\text{-}band_\mu$	1.37	0.19	13	1.28	0.21	6
$Sub\text{-}band_\sigma$	1.37	0.17	10	1.66	0.18	2
$Sub\text{-}band_S$	0.18	0.01	18	0.32	0.03	16
$Sub\text{-}band_k$	0	0	19	0	0	19
$Sub\text{-}int_\mu$	1.37	0.20	12	1.26	0.09	9
$Sub\text{-}int_\sigma$	1.40	0.20	2	1.01	0.06	11
$Sub\text{-}int_S$	0.78	0.02	17	0.91	0.01	18
$Sub\text{-}int_k$	0	0	20	0	0	20
$DM_{\mu'}$	1.25	0.03	16	1.11	0.04	14
$DM_{\sigma'}$	2.24	0.20	11	2.18	0.22	3
$DM_{\lvert S'\rvert}$	0.94	0.04	15	0.83	0.03	15
$DM_{k'}$	2.08	0.22	8	2.08	0.19	8

Table 4. The S/N of two different detections of PSR B0329+54 by LOTAAS, calculated using equation (1), and the corresponding values generated by several highly ranked features.

Feature	Detection 1	Detection 2
S/N	575	9.5
$Prof_S$	7.74	2.58
DM_μ	283	4.82
$Sub\text{-}band_\sigma$	0.06	0.12
$Sub\text{-}int_\sigma$	0.02	0.12
$DM_{k'}$	−0.56	−0.65

the same pulsar will be different. Hence, only using a single pulsar detection may not capture the various factors that characterize the data. This can result in a biased classifier if only a single training set is used. To overcome the issue we decided to employ an ensemble approach to produce LC2.

In our ensemble approach, five different classifier models are produced, each trained with a different subset of the training data available. In the case of the pulsar class, five unique detections of each pulsar are randomly assigned to one of the five subsets. However, 154 of the 295 training pulsars have been detected on fewer than five occasions. In this case, duplicate detections of the pulsar are used in several subsets of the training data. Hence, each training data subset still consists of 295 unique pulsars, for a total of 986 different examples. As for the sample of noise and RFI instances, five subsets of 600 noise-like candidates and 100 RFI instances, are randomly selected with replacement from the sample of 1267 noise-like candidates and 150 RFI instances. This approach is based on the bagging technique (Biau 2012) used in random forest classifiers (see Bishop 2006). By using multiple overlapping training data subsets for the non-pulsar classes in each different decision tree, we will be able to sample the feature space of the two classes without being biased towards one individual sample.

We used the decision tree algorithm (Quinlan 1993) from the WEKA data mining software (Frank, Hall & Witten 2016) to produce five different decision tree classifiers. The results of all five different classifiers were then combined together, with only candidates identified as pulsars by three or more different classifiers being kept for further visual inspection. This seemingly arbitrary choice to keep only those identified as pulsars by three or more classifiers is the standard majority vote rule that is popular in the ML literature (see section 3.4 of 'On combining classifiers'; Kittler et al. 1998). In the decision tree algorithm, a classifier will undergo pruning (Mitchell 1997) after building to reduce the bias of the classifier towards the training set. Pruning works by removing branches in the tree which are specialized (cover fewer examples) to prevent overfitting (Bishop 2006). However, He & Garcia (2009) showed that pruning in imbalanced data does not necessarily improve classification performance. Furthermore, in ensemble decision tree scenarios, overtraining each individual tree by reducing the amount of pruning done has been shown to yield improved performance over pruned decision trees (Sollich & Krogh 1995; Lyon 2016). Therefore, we decided to reduce the amount of pruning done by each individual decision tree classifier within our ensemble, by increasing the confidence factor of the decision tree from 0.25 to 0.5. A higher confidence factor results in less pruning for the decision tree.

We then tested the performance of LC2 on the full data set used to train the five different classifiers. The test data consists of all 1267 noise, 986 pulsar and 150 RFI instances used for training. Data used for training should never normally be used for testing. However due to the lack of labelled data describing the pulsar class (especially on pulsars with wide integrated pulse profiles), almost all the available samples are used for training, thus we were forced to use the same data for testing as well. Since we are only interested in

Table 5. The definition of the various standard metrics used to measure the performance of a classifier. True positive (TP) indicates pulsars that are classified correctly. True negative (TN) indicates non-pulsars (noise and RFI alike) that are classified correctly. False negative (FN) are pulsars that are misclassified as non-pulsars and false positive (FP) are non-pulsars that are misclassified as pulsars. All metrics have values ranging from 0 to 1.

Metric	Description	Definition
Accuracy	The overall accuracy of classification	$\frac{(TP+TN)}{(TP+FP+TN+FN)}$
FPR	The fraction of negative instances misclassified as positive	$\frac{FP}{(FP+TN)}$
Precision	The fraction of retrieved instances that are positive	$\frac{TP}{(TP+FP)}$
Recall	The fraction of correctly classified positive instances	$\frac{TP}{(TP+FN)}$
F-Score	The accuracy of classifier considering both precision and recall	$2 \times \frac{Precision \times Recall}{Precision + Recall}$
Specificity	The fraction of negative instances that are correctly classified	$\frac{TN}{FP+TN}$
G-Mean	Imbalanced data metric that describes the ratio between positive and negative accuracy	$\sqrt{Recall \times Specificity}$

Table 6. The performance of different ML classifiers when tested with the test data described in Section 4. LC1 is trained with the training set described in Section 2. Tree1–5 are the individual decision tree classifiers, each trained with a different training subset, used as part of the ensemble LC2.

Classifier	Accuracy	FPR	Precision	Recall	F-Score	Specificity	G-Mean
LC1	0.968	0.028	0.961	0.962	0.961	0.972	0.967
Tree1	0.983	0.018	0.975	0.984	0.979	0.982	0.983
Tree2	0.974	0.027	0.975	0.976	0.975	0.973	0.974
Tree3	0.983	0.020	0.971	0.987	0.979	0.980	0.983
Tree4	0.988	0.008	0.989	0.981	0.985	0.992	0.986
Tree5	0.979	0.011	0.984	0.965	0.974	0.989	0.977
LC2	0.992	0.005	0.993	0.987	0.990	0.995	0.991

the pulsar class, the performance of the classifiers is only measured in relation to the pulsar class. Table 5 shows the various metrics used to measure the performance of a classifier, and Table 6 shows the performance of each individual decision tree and the ensemble LC2, as well as LC1 deployed upon the new test data.

The performance data showed that the ensemble LC2 achieves better performance than any individual decision tree classifier across various metrics, with a lower FPR and a higher pulsar recall rate. LC2 also performed better than LC1 according to the G-Mean metric. While the original classifier is 3.3 per cent away from being perfect, the new classifier now achieved less than 1 per cent loss in performance. We also check how LC2 performs on the pulsars misclassified by LC1 (shown in Fig. 3). We found that seven out of the nine pulsars are now correctly identified by LC2 (except PSRs J0621+1002 and B2053+36), showing an improvement in the ability to discover pulsars that possess wide integrated pulse profiles.

5 APPLYING THE CLASSIFIER TO LOTAAS

Although LC2 exhibits better performance than LC1 when tested with the data set described earlier, it is more practical to apply the classifier to actual survey data to better gauge real-world performance. We would expect the classifier to produce fewer false positive candidates, given the results of Section 4, reducing the number of candidates to inspect. Besides that, we expect to see an improvement in the recall rate of the classifier. We applied LC2 to the candidates previously obtained by the search pipeline. We then compared the pulsars identified by LC2 with the pulsars identified by LC1. The pointings from which the candidates are obtained, contain some pulsar/noise/RFI examples also present in the training set, as well as examples from pointings obtained after the training set was compiled (independent samples).

First, we want to compare the FPR of LC2 on actual survey data with that of LC1. We did this by comparing the number of non-pulsars that are classified as pulsars by the two classifiers in 11 different LOTAAS pointings. LC1 identified 5406 different candidates as pulsars, giving an average of 491 pulsar predictions per pointing. LC2 identified 2400 different candidates as pulsars, giving an average of 218 pulsar predictions per pointing, i.e. roughly one candidate per beam. Given that each LOTAAS pointing produces roughly 20 000 candidates, the number of candidates flagged for inspection is reduced from about 2.5 per cent, to ∼1.1 per cent, approximately 56 per cent fewer candidates.

There is a difference between the FPRs obtained in survey data for LC1 and LC2, compared to test data. This result is expected, because in the real-world classification scenario the total number of non-pulsar examples in the training set is very small when compared to the total number of non-training data points to be classified. It is therefore difficult for such a small number of training examples to characterize all possible non-pulsar examples. This is especially true as our real-world data is full of variable forms of RFI/noise and interesting phenomena yet to be identified and therefore not included in the training sample, which are technically considered non-pulsars. The result is that FPRs are different on real-world data, compared to test data. That said, LC2 still reduces the FPR on real world data, and is even better than the FPR of LC1 on test data. Thus, the difference does not imply an issue with the new classifier.

Next, we compared the pulsar recall rate for both classifiers. The classifiers were applied to a total of 185 pointings known to contain pulsars. A total of 192 different pulsars are known to exist within these pointings, either through detection by LC1, or via plots made by the LOFAR Pulsar Pipeline (see Kondratiev et al. 2016), in which the raw data are folded with the ephemerides of known pulsars in the FoV. As some pulsars are detected in multiple different pointings, we have a total of 313 instances to recover. We compared the recall rate of LC2 on known pulsars, with LC1 on these pointings. LC1 misclassified 16 different instances, while LC2 misclassified only three, in which two of them were misclassified by both classifiers. This suggests an increase of pulsar recall rate from 94.9 per cent to 99.0 per cent. The integrated pulse profiles of the pulsars misclassified by LC1, but correctly identified by LC2, are shown in Fig. 7. The pulse profiles of the three pulsars misclassified by LC2 are shown in Fig. 8.

LC2 showed a marked improvement over LC1, as fewer pulsar instances were misclassified. LC2 was also able to identify pulsars with wide integrated pulse profiles that were otherwise not identified by LC1, specifically PSRs B0531+21, B1907-03, B1946+35, J2007+0809, B2027+37, B2106+44 and B2148+52, overcoming the main issue with LC1. LC2 was also able to classify some of the low S/N pulsars with narrow profiles and those comprised of two components. These pulsars, PSRs J1652+2651, B1907+03, B1935+25, and J2040+1657 were also misclassified by LC1 before.

However, a small number of pulsars have eluded the new classification system. We decided to inspect the features belonging to these pulsars, to understand what went wrong in the classifier. PSR B2324+60 has a pulse profile that is similar to PSRs B1935+25 and B2227+61, but was incorrectly classified. We checked the features of the detections and audited their decision paths within the classifier models produced by each of the five decision trees. We found that in four out of the five trees, the features $Sub\text{-}int_{\mu}$, $Sub\text{-}int_{\sigma}$, or $DM_{\sigma'}$ missed the positive prediction threshold by a small margin. This resulted in the pulsar being misclassified as either noise or RFI on those occasions. It showed that this observation of PSR B2324+60 lies on the edge of detectability of LC2. PSR B2053+36 has low S/N with a highly scattered integrated pulse profile. The pulsar is not immediately recognizable from its diagnostic plot, and only identifiable via its period and DM. Hence, it is not surprising that the classifier missed this instance.

PSR B0531+21, which has a pulse period of 33 ms is detected in three different LOTAAS pointings. However, one of the three detections is not identified by LC2 as a pulsar. The misclassified detection has a lower S/N than the detections from the other two pointings. The misclassified detection was made in a far side-lobe of an incoherent beam of the pointing. Hence, the pulsar is only being detected in the lower part of the observing bandwidth of LOTAAS, which resulted in a large value for $Sub\text{-}band_{\sigma}$ compared to a typical pulsar. The diagnostic plot of the detection also showed variation in signal intensity over time, due to the rotation of the beam pattern during the observation, resulting in the pulsar moving in and out of the side lobe. This gives a $Sub\text{-}int_{\sigma}$ value that is larger than for a typical pulsar. A combination of these two issues results in the pulsar being classified as noise by the classifier.

6 CONCLUSION

We have discovered shortcomings in LC1 that reduced its ability to identify pulsars with wide integrated pulse profiles. We therefore introduced a number of improvements. We first introduced a new method to construct the DM–S/N plot. We also introduced 12 new features for the ML classifier to learn from. Next, we defined a new class of non-pulsars that consists of notable RFI instances affecting LOTAAS observations. Besides that, we introduced a new training set consisting of a larger sample of candidates, with a larger number of pulsar examples that were misclassified by LC1. We then evaluated the usefulness of these new improvements via an Information Theoretic Analysis. We found that the new features are able to assist in the classification process, and that it is more advantageous to separate notable RFI instances from the rest of the non-pulsar class, forming a new class for our ML classifier to learn from.

Finally, we introduced an ensemble ML approach of five different decision tree classifiers, trained with five different training subsets, whereby a candidate is assigned the pulsar label only if three or more members of the ensemble agree. Compared to LC1 (96.7 per cent G-Mean), our improvements result in a new ensemble system that has a much better performance (99.1 per cent G-Mean) on the test data set. We then compared the performance of LC1 and LC2 on recent LOTAAS observations as the performance of the classifiers derived from actual survey data is more important. LC2 exhibited a big improvement compared to LC1, including a drop in false positives from ∼2.5 per cent to ∼1.1 per cent, which reduces the number of candidates per pointing from ∼500 to ∼220. A higher pulsar recall rate was also achieved, from 94.9 per cent to 99.0 per cent. More importantly, LC2 is able to identify pulsars with wide pulse profiles that LC1 could not.

The building of LC2 showed there are several important issues to keep in mind. First, we needed to make sure the features used to build a ML classifier are able to capture all the notable properties of the candidates. We found that only using features from the integrated pulse profile and the DM–$\chi^2_{\rm red}$ curve, yielded classification systems inadequate at separating pulsars from non-pulsars. Hence, we added features from the pulse intensity as a function of frequency sub-band and pulse phase plot, as well as the pulse intensity as a function of time and pulse phase plot, to assist in the classification process. A training set which properly represents the pulsar population is also crucial to the production of a successful classifier. We found that having separate classes for different types of non-pulsars is helpful to the classification process, as the noise and RFI instances have very different properties in terms of the features we used. Lastly, we found that for different detections of the same pulsar, the values of the features obtained are different. Therefore, we applied an ensemble approach in the production of the classifier to accommodate the variations we see in different detections of the same pulsar.

Although we have implemented several different 'modifications', there are still several potential improvements worth investigating. One potential change would involve separating the candidates to enable processing by different classification systems, according to the number of bins in the profiles. This can be done if we have a larger sample of short period pulsars. Investigation of the RFI class suggest that we can further separate the RFI class, depending on the properties of the RFI, i.e. narrow-band RFI or short duration burst. However, further subclassing the RFI instances would require a larger sample of training data for all the classes.

Now that the new ML classifier has proven to be successful, it is being implemented into the LOTAAS search pipeline. It is also being used to reclassify all older, archived LOTAAS candidates, in the expectation that previously missed new pulsars may yet be found.

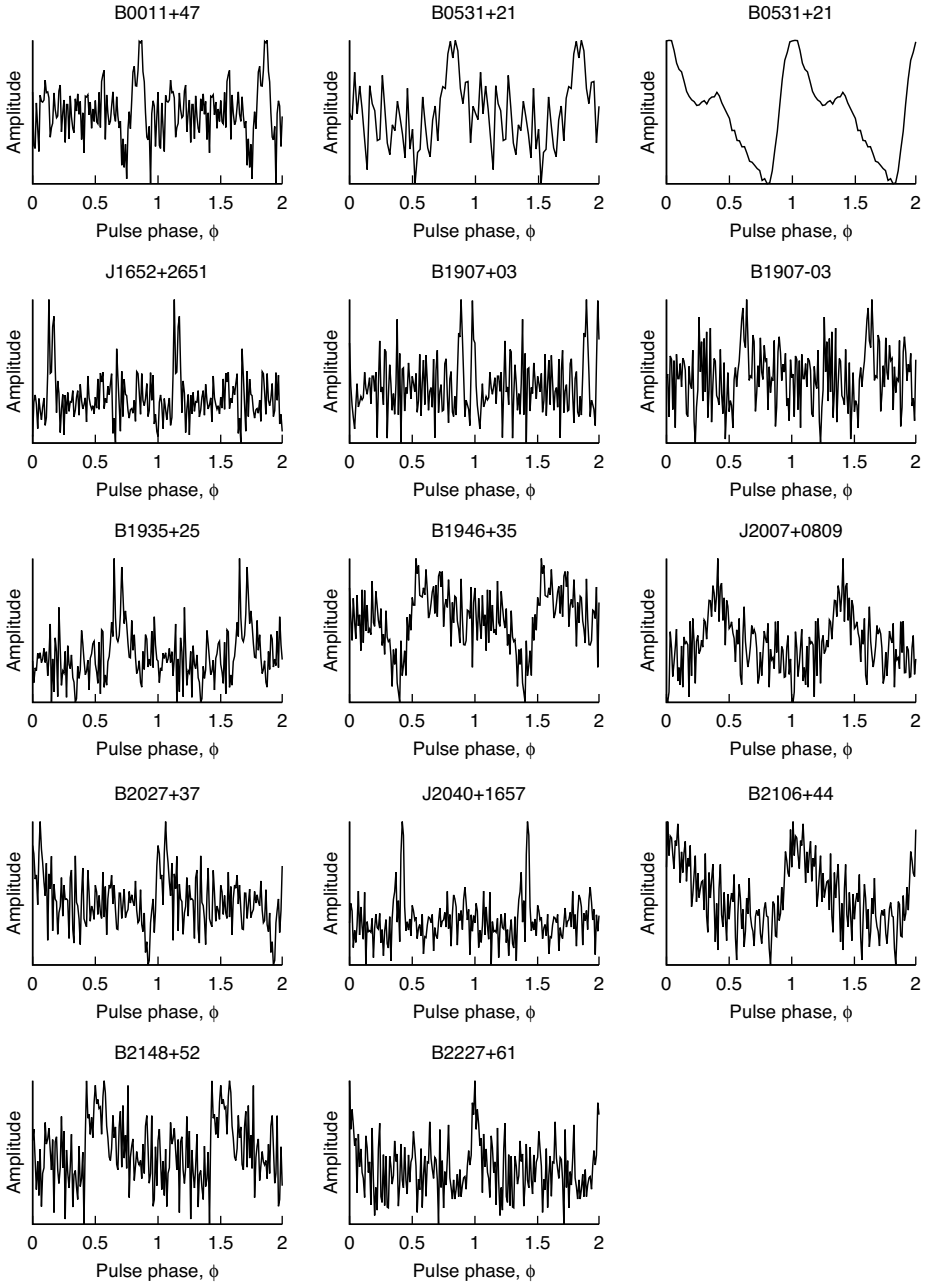

Figure 7. The integrated pulse profiles of the pulsars that are classified correctly by LC2, but not LC1. PSR B0531+21 was detected in two separate pointings with different S/Ns. The amplitudes are in arbitrary units. PSRs B0531+21, B1907−03, B1946+35, J2007+0809, B2027+37, B2106+44, and B2148+52 are pulsars with wide integrated pulse profiles, with several of them being heavily scattered. PSRs J1652+2651, B1907+03, B1935+25, and J2040+1657 are pulsars with low S/N and double peaked integrated pulse profiles.

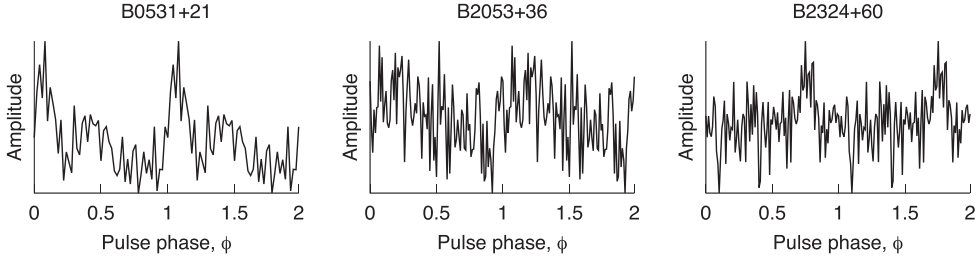

Figure 8. The pulse profiles of the pulsars that are classified incorrectly by LC2. This particular detection of PSR B0531+21, which was in a different pointing compared to the two detections in Fig. 7, was correctly identified by LC1.

ACKNOWLEDGEMENTS

This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative. Computing time was provided by The Netherlands Organisation of Scientific Research (NWO) Physical Sciences (project SH-242-15). The LOFAR facilities in the Netherlands and other countries, under different ownership, are operated through the International LOFAR Telescope foundation (ILT) as an international observatory open to the global astronomical community under a joint scientific policy. In the Netherlands, LOFAR is funded through the BSIK program for interdisciplinary research and improvement of the knowledge infrastructure. JWTH, VIK, DM, and S.S. acknowledge funding from an NWO Vidi fellowship and from the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013)/European Research Council Starting Grant agreement nr. 337062 (DRAGNET). We thank Joeri van Leeuwen for the useful discussions.

REFERENCES

Bates S. D. et al., 2012, MNRAS, 427, 1052
Bethapudi S., Desai S., 2017, preprint (arXiv:1704.04659)
Biau G., 2012, J. Mach. Learn. Res., 13, 1063
Bilous A. V. et al., 2016, A&A, 591, A134
Bishop C. M., 1995, Neural Networks for Pattern Recognition. Oxford Univ. Press, Oxford
Bishop C. M., 2006, Pattern Recognition and Machine Learning. Springer, Berlin
Brown G., Pocock A., Zhao Z., Luján M., 2012, J. Mach. Learn. Res., 13, 27
Coenen T. et al., 2014, A&A, 570, A60
Cooper S., 2017, PhD thesis, Univ. Manchester
Cordes J. M. et al., 2006, ApJ, 637, 446
Eatough R. P., Molkenthin N., Kramer M., Noutsos A., Keith M. J., Stappers B. W., Lyne A. G., 2010, MNRAS, 407, 2443
Edwards R. T., Bailes M., van Straten W., Britton M. C., 2001, MNRAS, 326, 358
Faulkner A. J. et al., 2004, MNRAS, 355, 147
Fayyad U. M., Irani K. B., 1993, in Bajcsy R., ed., Proc. Thirteenth International Joint Conference on Artificial Intelligence. Morgan Kaufmann, Burlington, MA, p. 1022
Ford J. M., 2017, PhD thesis, Nova Southeastern Univ.
Frank E., Hall M. A., Witten I. H., 2016, The WEKA Workbench. Online Appendix for 'Data Mining: Practical Machine Learning Tools and Techniques', 4th edn. Morgan Kaufmann, Burlington, MA
Guyon I., Elisseeff A., 2003, J. Mach. Learn. Res., 3, 1157
He H., Garcia E. A., 2009, IEEE Trans. Knowl. Data Eng., 21, 1263
Hewish A., Bell S. J., Pilkington J. D. H., Scott P. F., Collins R. A., 1968, Nature, 217, 709
Keith M. J., Eatough R. P., Lyne A. G., Kramer M., Possenti A., Camilo F., Manchester R. N., 2009, MNRAS, 395, 837
Keith M. J. et al., 2010, MNRAS, 409, 619
Kittler J., Hatef M., Duin R. P. W., Matas J., 1998, IEEE Trans. Pattern Anal. Mach. Intell., 20, 226
Kohavi R., John G. H., 1997, Artif. Intell., 97, 273
Kondratiev V. I. et al., 2016, A&A, 585, A128
Lazarus P. et al., 2015, ApJ, 812, 81
Lee K. J. et al., 2013, MNRAS, 433, 688
Lyon R. J., 2016, PhD thesis, Univ. Manchester
Lyon R. J., Knowles J. D., Brooke J. M., Stappers B. W., 2013, in Kellenberger P., ed., IEEE Int. Conf. Systems Man Cybern., Simple and Effective Machine Learning for Big Data, Special Session. IEEE Comput. Soc., Los Alamitos, CA, p. 1506
Lyon R. J., Knowles J. D., Brooke J. M., Stappers B. W., 2014, in O'Conner L., ed., 22nd IEEE Int. Conf. Pattern Recognit. IEEE Comput. Soc., Los Alamitos, CA, p. 1969
Lyon R. J., Stappers B. W., Cooper S., Brooke J. M., Knowles J. D., 2016, MNRAS, 459, 1104
Manchester R. N., Lyne A. G., Taylor J. H., Durdin J. M., Large M. I., Little A. G., 1978, MNRAS, 185, 409
Manchester R. N. et al., 2001, MNRAS, 328, 17
Mitchell T. M., 1997, Machine Learning. McGraw-Hill, New York
Morello V., Barr E. D., Bailes M., Flynn C. M., Keane E. F., van Straten W., 2014, MNRAS, 443, 1651
Quinlan J. R., 1993, C4.5: Programs for Machine Learning. Morgan Kaufmann, Burlington, MA
Ransom S. M., 2001, PhD thesis, Harvard Univ.
Ransom S. M., Eikenberry S. S., Middleditch J., 2002, AJ, 124, 1788
Shannon C. E., Weaver W., 1949, The Mathematical Theory of Communication. Univ. Illinois Press, Champaign, Illinois
Sollich P., Krogh A., 1995, in Touretzky D. S., Mozer M. C., Hasselmo M. E., eds, NIPS Conf., Advances in Neural Information Processing Systems 8. MIT Press, Cambridge, MA, p. 190
Stappers B. W. et al., 2011, A&A, 530, A80
Stokes G. H., Segelstein D. J., Taylor J. H., Dewey R. J., 1986, ApJ, 311, 694
Stovall K. et al., 2014, ApJ, 791, 67
van Haarlem M. P. et al., 2013, A&A, 556, A2
Yang H. H., Moody J. E., 1999, in Solla S. A., Leen T. K., Müller K., eds, NIPS Conf., Advances in Neural Information Processing Systems 12. MIT Press, Cambridge, MA, p. 687
Yao Y., Xin X., Guo P., 2016, in Bilof R., ed., IEEE 12th Int. Conf., Computational Intelligence and Security (CIS). IEEE Comput. Soc., Los Alamitos, CA, p. 120
Zhu W. W. et al., 2014, ApJ, 781, 117

This paper has been typeset from a TEX/LATEX file prepared by the author.