Pulsar Detection

A Deep Learning Approach

Manideep Pendyala

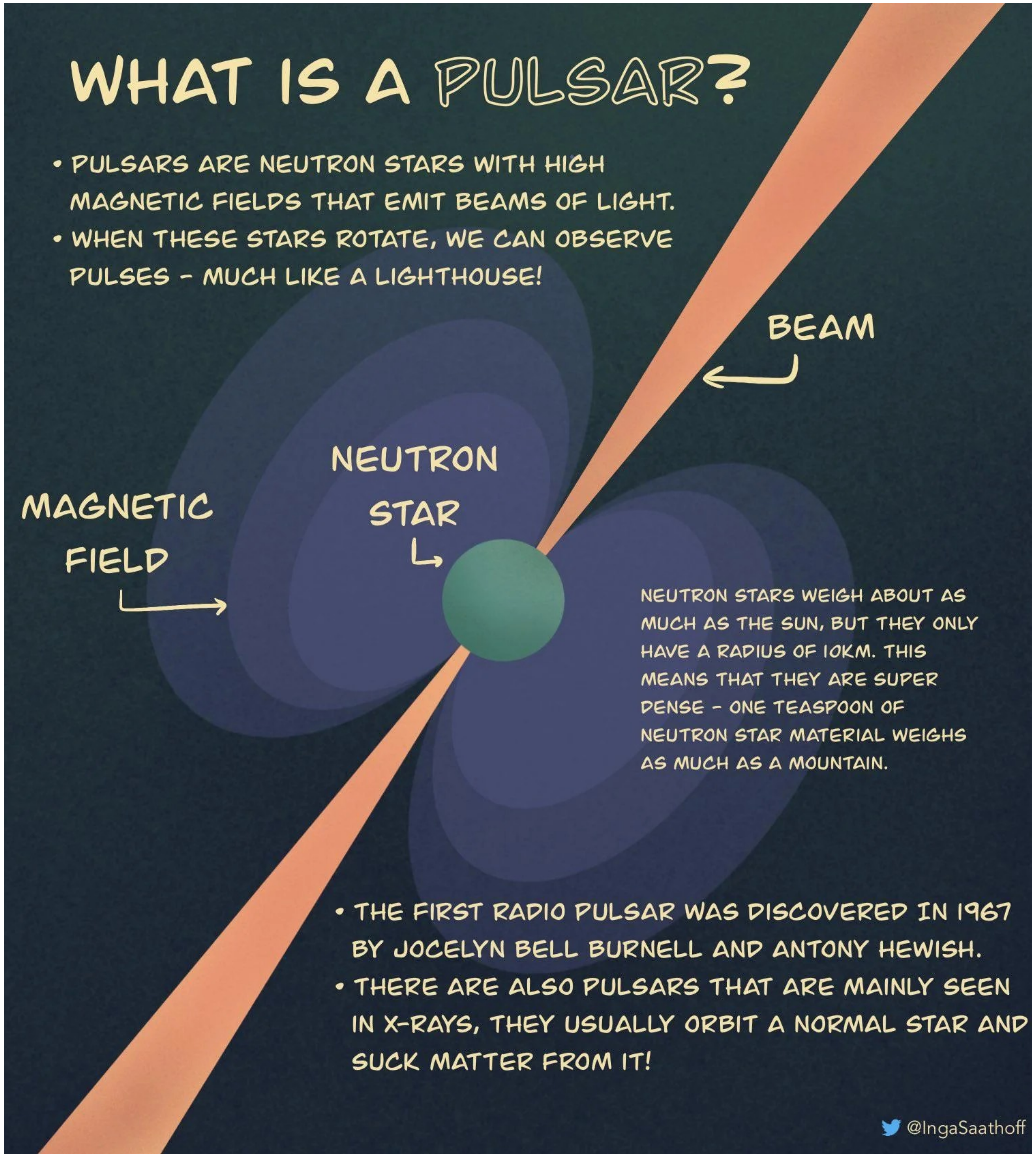
11 JAN 2023

Introduction to Pulsars

- Pulsars or Pulsating stars are the rotating neutron stars which can be observed in radio data as regular, dispersed pulses with pulsar specific periodicity.

- When looked at sky in wavelengths other than visible range astronomers discovered spots that flicker over-time, these points seem to emit extremely regular oscillations i.e they are pulsating hence called pulsars.

Look here for better understanding of pulsars



Why to automate the detection?

- In modern pulsar surveys, we are finding so many candidate pulsars that it is no longer possible for humans to view all of these by eye and so we will have to use computer algorithms.

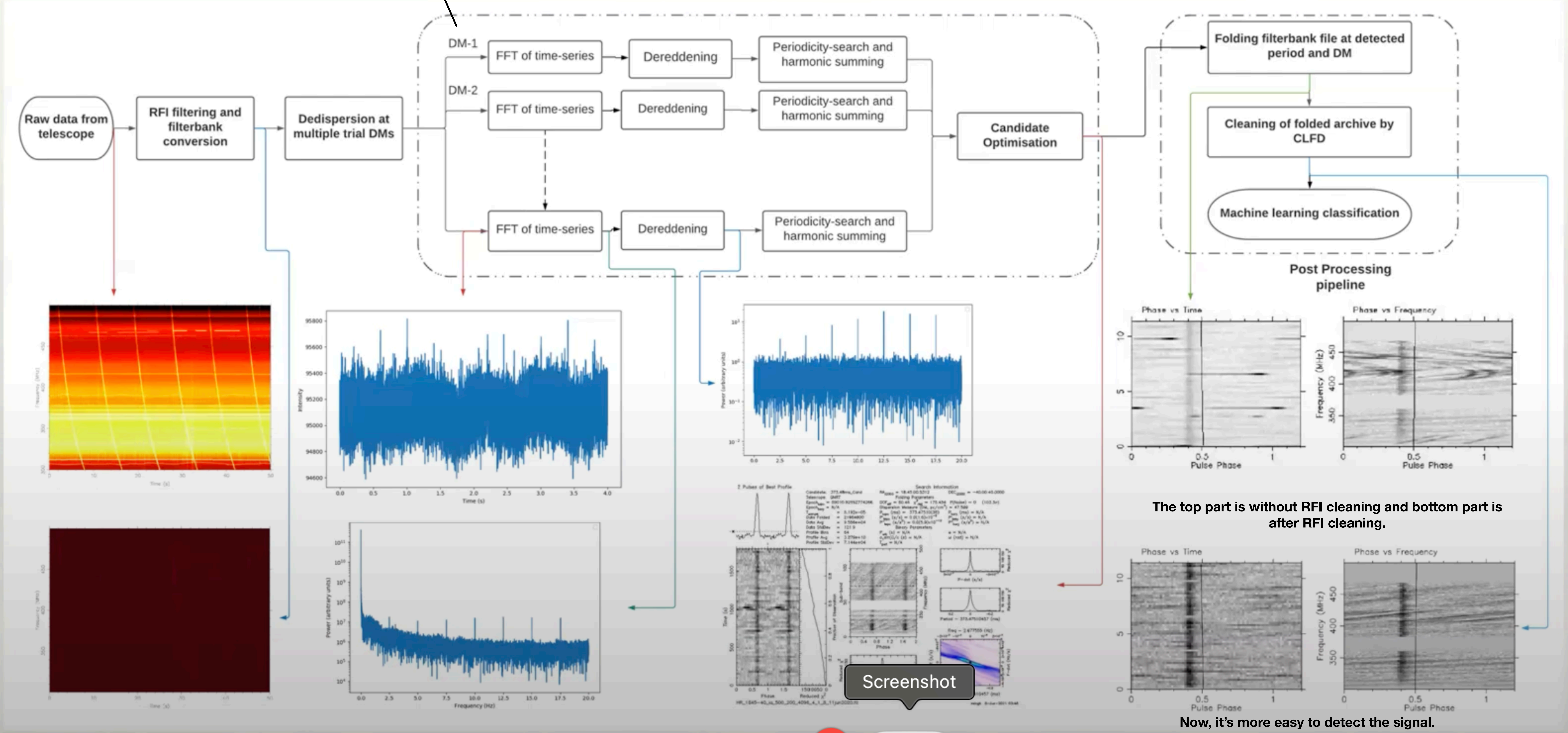
- Since the emergence of machine learning, its theory and technology have become increasingly mature, and has been successfully applied to astronomical research fields such as pulsar candidate screening.

- Searching for pulsar signals in radio data is a process of finding periodic pulsar signals with dispersion characteristics. Through the DM search, Fourier transform and periodic signal search, people get the candidate of real pulsar signal. The surveys are polluted by RFI that makes it hard to pick the pulsars from the candidates produced sing simple metrics such as SNR. We use periodicity here.

When different DM values are used for de-dispersion, the DM curve displays how the SNR of the pulse changes and we select the DM value with highest SNR.

Searching using the Fast Fourier Transform

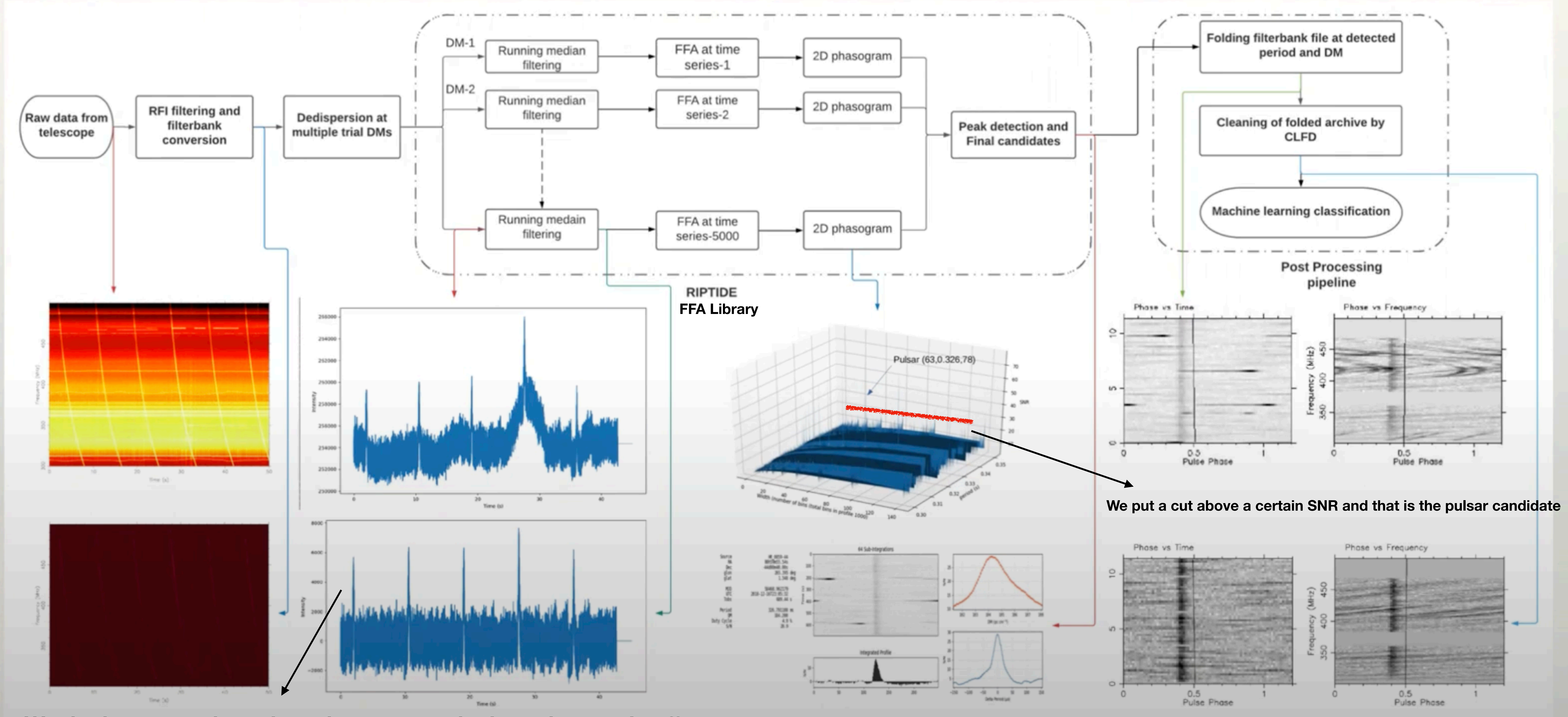
Dominant method by which we searched for pulsars for the last 30 years.



Advantage of FFA is it works well for pulsars with narrow width compared to their period and it keeps track of phase. In FFT if you don't sum the signal which is spread over many harmonics you won't recover the signal. FFA can't yet be used to look for fastest pulsars but it is very good at finding long period pulsars with narrow duty cycles.

Searching using the <u>Fast Folding Algorithm</u>

Computationally expensive that FFT



We don't remove the pulsars but remove the long time scale effects.

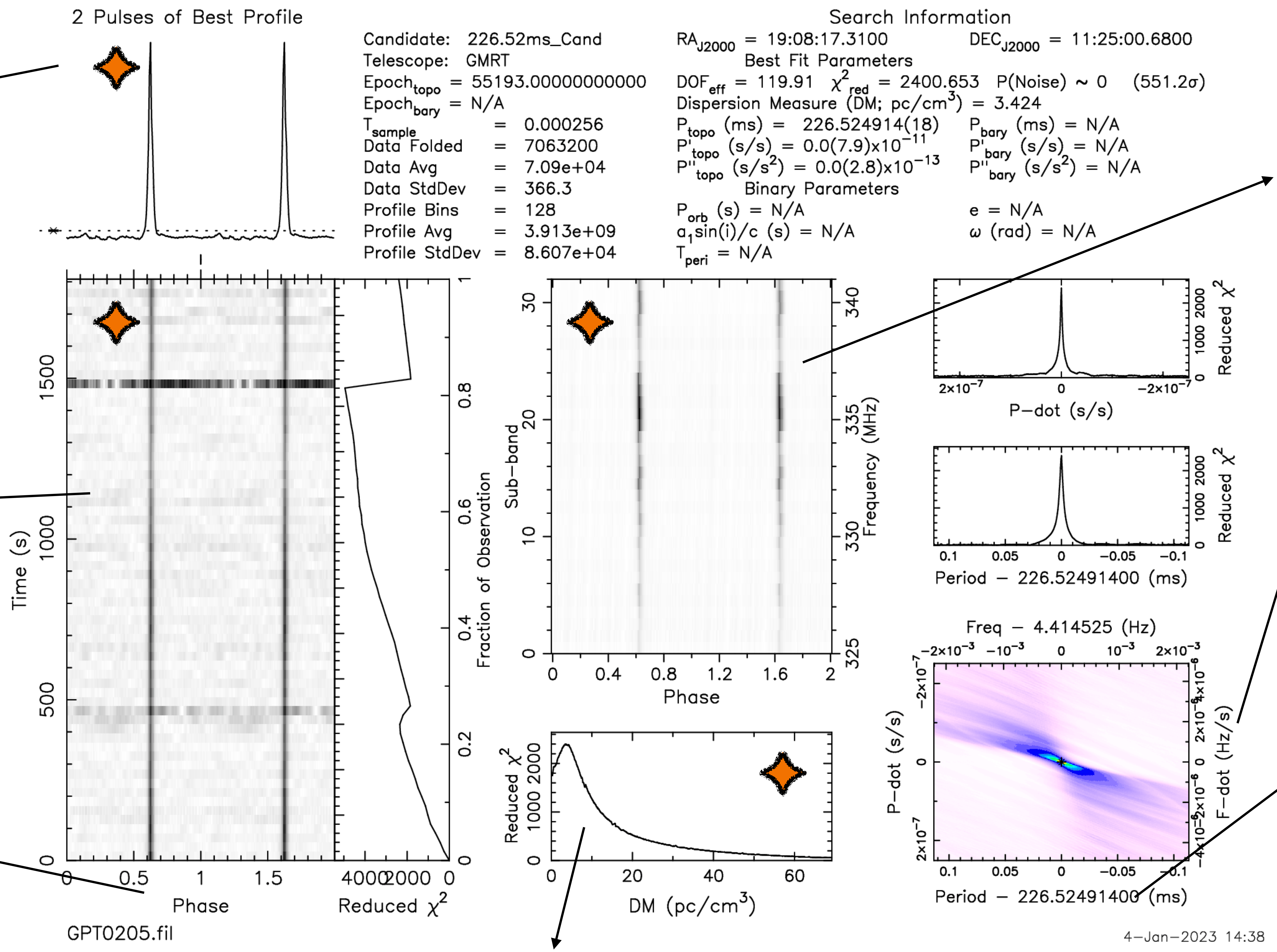
Diagnostic plot of a Pulsar candidate

Summed Pulse Profile: The total sum of the entire data folded at the best period.

The dark grey parts here correspond to the pulse. We've broken the data into small pieces and stacked them on top of each other.

We can see pulsar is constant approximately over all the time and you can also see other data in here that corresponds to RFI which has not yet been removed from the data.

Pulse phase means one rotation of pulsars going from 0 to 1. Here, it goes to 2 because we are showing two pulses.



Still pulse phase but as a function of frequency. We've broken the data into multiple frequency bands and we show if we use correct DM the pulsar signal is straight.

Best change of period

Best period. It may be that we don't get exactly the right period from our Fourier analysis. So, we search around that period to see if we can further improve the value or it may be that the period is slightly changing as a function of time especially if the system is in a binary.

You can think of Khi^2 is like signal to noise. It sharply peaks at best value for the DM and that's the DM we used for all the other analysis.

Early ML usage

- The machine learning method is a method in which a computer learns rules from training data, generates a model and then uses this model to predict new data.

- ANN, SVM, Decision tree, Ensemble Learning are the most used ML algorithms for pulsar candidate recognition.

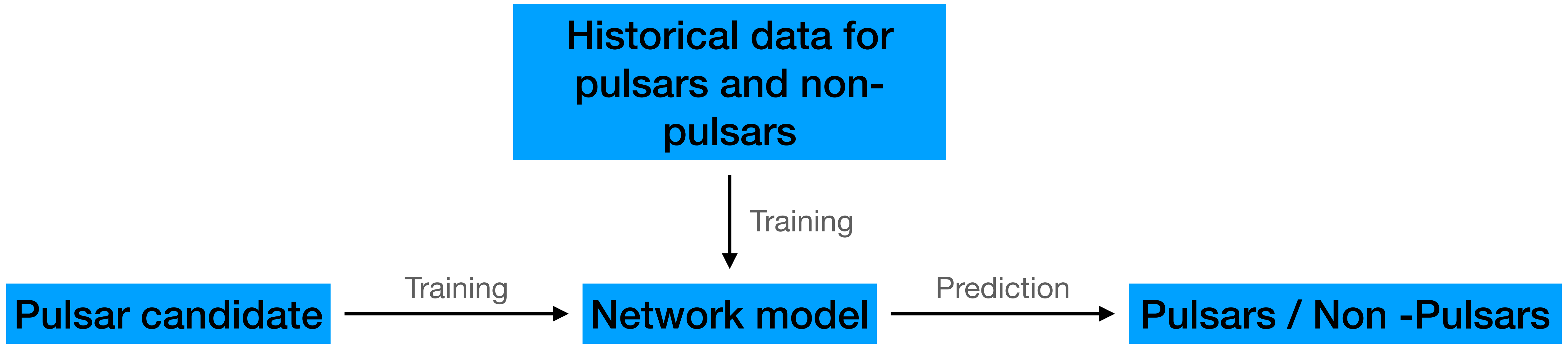


Image Classification (Early Work)

- Features : 1. Summed pulse profile 2. Time-versus phase plot 3. Frequency-versus-phase plot 4. DM curve.

- Problems - The integration time of survey observations may vary, and the number of phase bins with which the data are folded also changes depending on the period of the candidate. As a result, the number of data points in the subplots can vary from candidate to candidate.

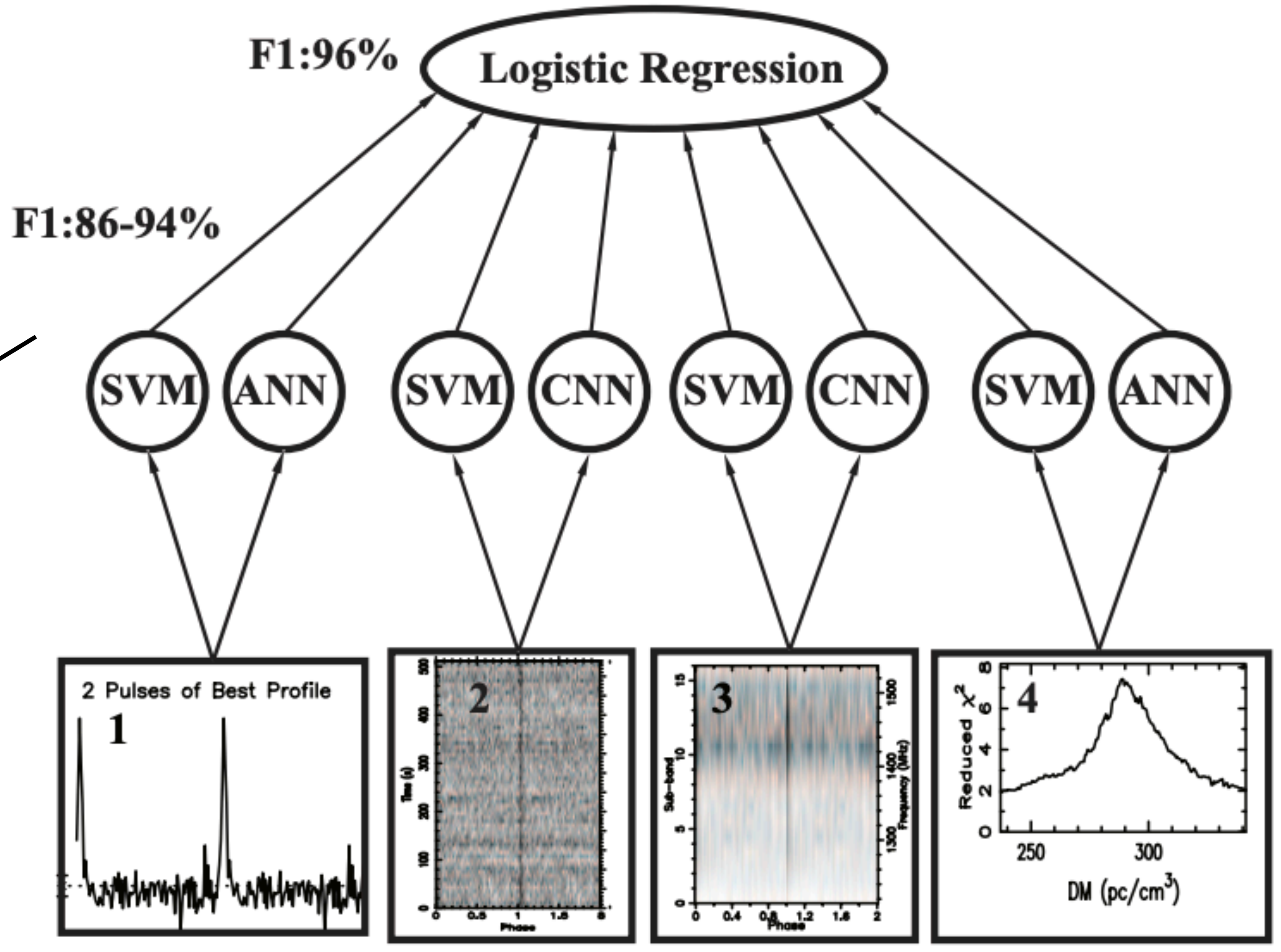
- Data - Key features are extracted from pfd files. These plots consist 1D data arrays (for Summed profile and DM curve) and 2D data. The sizes of the data array vary from candidate to candidate.

- We have to down-sample or interpolate the data to a uniform size.

- 1D and 2D data need different data preparation methods.

- The problem of AI developing phase-related bias was resolved by shifted candidate’s strongest peaks to phase 0.5 before feeding it to the AI.

Each first-layer classifier rates(we get 8 ratings in total, here) how pulsar-like a particular subplot of the candidate is with a number between 0 an 1, giving the prediction matrix that is the output of the first layer. These votes are fed into the second layer classifier which learns to properly weight these votes and forms final consensus on how pulsar-like a candidate is

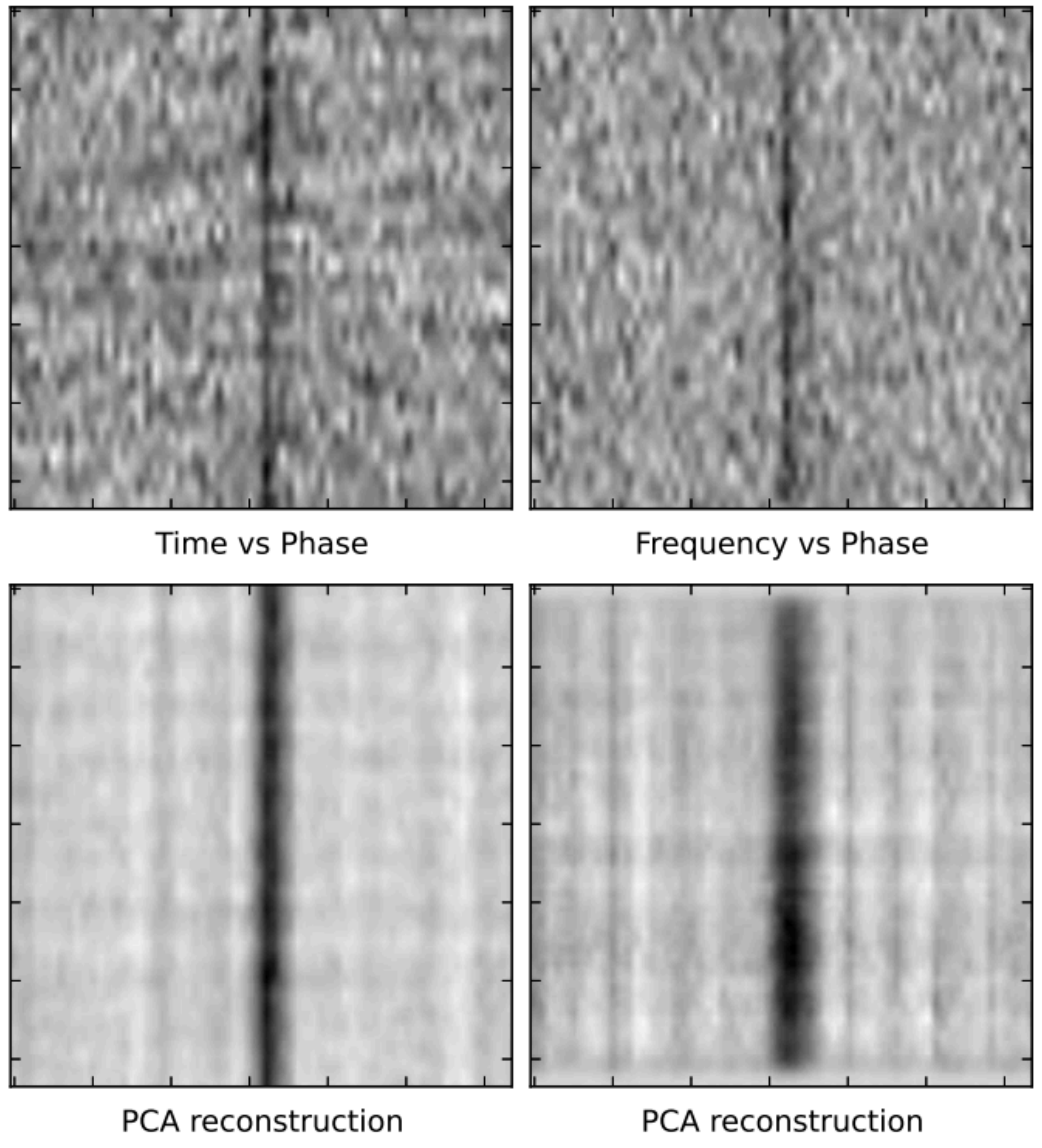


Preferred a classifier with higher completeness than precision. But, to emphasise on one metric would hinder our ability to improve the other. So, preferred a balanced model that maximised F1.

We used ANN for 1d subplots and CNN for 2D subplots. The CNN configuration is similar to LeNet-5 and is used for both the time-versus-phase plot and frequency-versus-phase plot .

The best CNN design, is determined through cross-validation tests. The labelled data are randomly shuffled and split into training set formed from 60% of the candidates and a validation set from the remaining 40%, and a large grid search is performed.

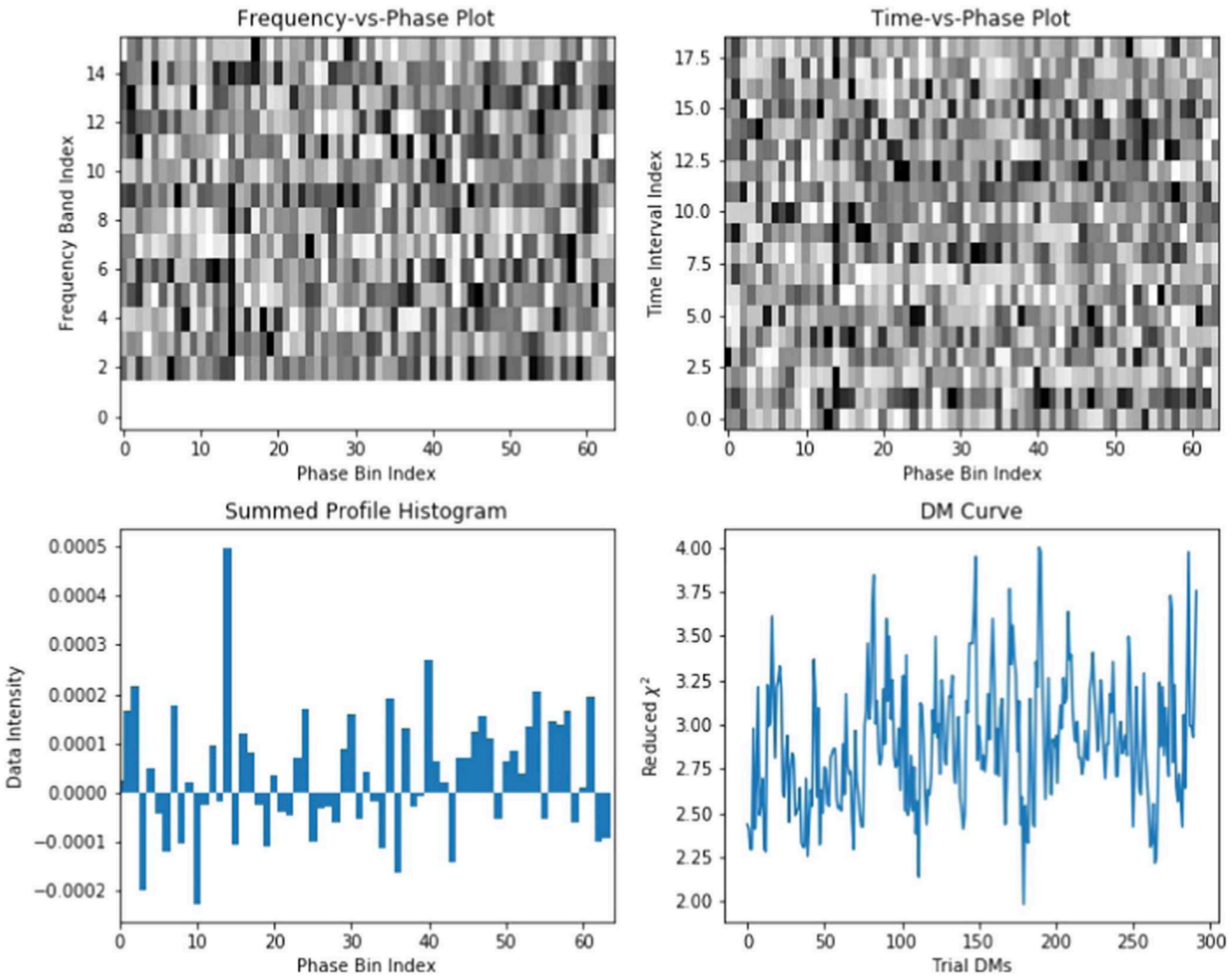
This CNN performs well on frequency-versus-phase plot because a large fraction of RFI is narrow-band emission. For continuous emission this RFI shows up as a horizontal line while burst-like or periodic RFI appears as a small dot. This is opposed to the broadband pulsed signal of a pulsar, which shows up as a vertical stripe. Since the CNN excels at detecting small-scale features, it can easily detect this form of RFI. A second reason this classifier is the best discriminant of pulsars is simply because there is more information in the 2D plots than the 1D plots. For similar reasons, the algorithm can recognise and reject burst-like RFI or signals that drift around in phase using the time-versus-phase plot.



To speed up the computation, the image features are characterised by PCA. The PCA-reconstructed plots from the top 24 PCA components. By filtering out the weaker PCA components we can also reduce the noise level in the reconstructed image. It is not needed for CNN but needed for SVM.

Using GAN and ResNext (2023 JAN)

- We first run two training sessions on the identical DCGAN structure to generate FPP and TPP. The original data set and the generated images are combined to create a balanced data set, which is then used to train ResNeXt. For both FPP and TPP, we separately train the identification model ResNeXt.
- The input is a grey-scale pulsar-candidate image, and the output is the predicted probability that the candidate is a positive or negative sample.



Possible direction to proceed

1. For de-dispersing (either using neural networks or some ML approach)

2. To classify the pulsar candidate after FFT or FFA (a combination of ML and image recognition DL)

3. Image classification on the sub plots. (Pure DL)

4. We can also do combinations of the whole process, we can apply techniques on de-dispersing, de-reddening and then use that to generate pulsar candidate and maybe apply some cleaning methods on the plot and the use the plots to train image classifier to do the binary classification.

References

Pulsar Search 1
Pulsar Search 2
Pulsar Search 3
https://arxiv.org/pdf/1309.0776.pdf
Machine Learning Workshop : Hunting for Pulsars
https://iopscience.iop.org/article/10.3847/1538-4365/ac9e54/pdf
https://academic.oup.com/mnras/article/516/2/3082/6692891
https://arxiv.org/pdf/2106.04407.pdf
https://www.sciencedirect.com/science/article/pii/S2314728817300776
https://arxiv.org/pdf/1704.04659.pdf