Chandra X-ray sources Classification using Machine Learning

Shivam Kumaran

Indian Institute of Space Science and Technology,

Thiruvananthapuram

To classify Globular cluster X-ray sources available in Chandra Source catalogue

Globular Cluster -
Gravitaionaly bound system of cluster of stars.
Large number of close encounters.

Classes -

- CV – Cataclysmic variable : Binary system with accretion onto a White Dwarf
- LMXRB – Accretion of a star onto BH or NS
- MSP – Milli second pulsar : Periodic rotation
- Other classes – AB , Stars

Image Credit Copyright: Jose Mtanous

Chandra Source Catalogue

- X-ray sources :
  - 317,167 unique sources -
  - **In 90 GC - 8275** sources
- Instruments :
  - ACIS - 5 energy bands
    - *broad band (b): 0.5-7.0 keV*
    - ultrasoft (u): 0.2-0.5 keV
    - *soft (s): 0.5-1.2 keV*
    - *medium (m): 1.2-2.0 keV*
    - *hard (h): 2.0-7.0 keV*
  - HRC
- Wide Band : 0.1-10 keV

Chandra Source Catalogue

Source Information

- RA-DEC
- Galactic-coordinates
- Exposure timings
- Flux significance

Source Fluxes

- Photon flux
- Energy flux

Spectral Properties

- Black Body
- Powerlaw
- Bremsstrahlung

Hardness Ratio

- **Source variability**
- Source Falgs

$$hard_x y = \frac{F(x) - F(y)}{F(x) + F(y)}$$

Chandra Source Catalogue

	Feature 1	Feature 2	Feature 3	...	Feature m	CLASS
Source 1						
Source 2	Nan	NAN				
...						
Source n	NAN					

How do we get labels

Using Other catalogues
Find RA-DEC

Cross-Match with CSC

Cross match using
HEASARC web tool
Radius - 3 arcsec

Choose Best cross-match

Select sources , assign class labels

Chandra Source Catalogue

We got training data with labels

	Feature 1	Feature 2	Feature 3	...	Feature m	CLASS
Source 1						
Source 2	Nan	NAN				
...						
Source n	NAN					

		Feature 1	Feature 2	Feature 3	...	Feature m	CLASS
Source 1	Obs 1			nan			
	Obs 2	Nan					
	Obs 3						
Source 2	Obs 1	Nan	NAN				
	...						

Using Other catalogues
Find RA-DEC

Cross-Match with CSC

Choose Best cross-match

Select sources , assign class labels

Cross match using
HEASARC web tool
Radius – 3 arcsec

Dataset

	Num Sources	Num obs
CV	66	516
NS	48	302
BH	248	160
PULSAR	1118	319

Number of Features – 56 , not using model-fit parameters

- Photon flux (b,h,m,s,u)
- Energy flux (b,h,m,s,u)
- Variability
- Hardness ratio

Data Processing

Data Scaling
- No-scaling
- Normalisation
- Standardisatsation

Data Imputation
- Zero
- Mean
- Median
- Correlation
- Random Forest

Classifier
- LR
- KNN
- FC
- CNN
- RF

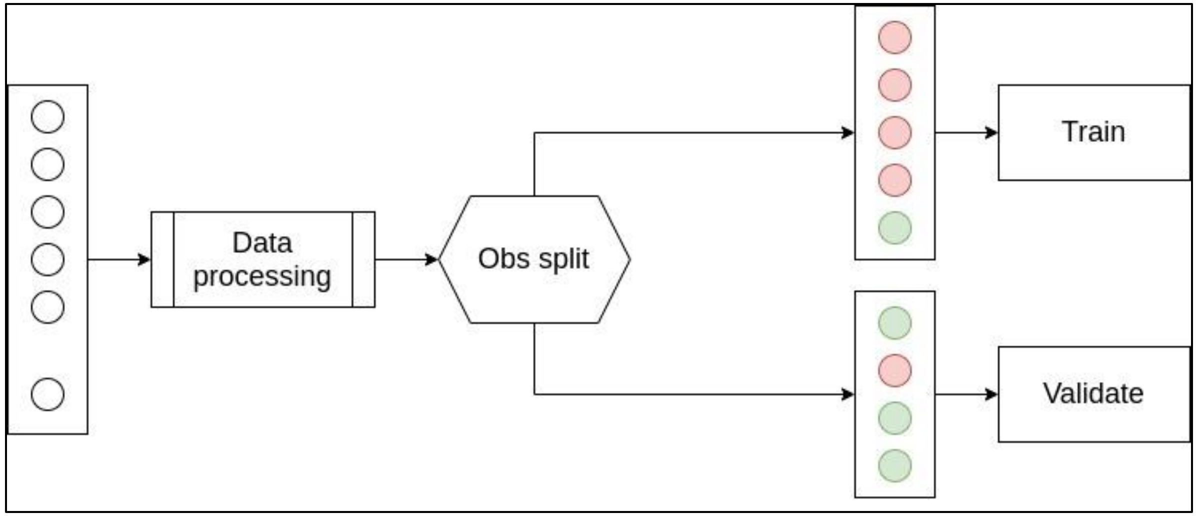
Data Processing

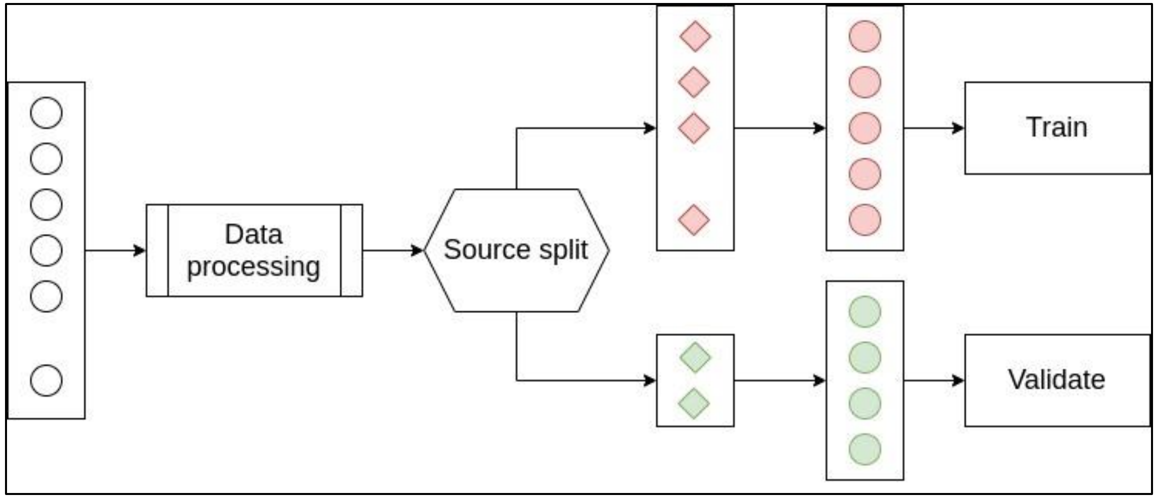
Data Scaling
- Standardisatsation

Data Imputation
- Random Forest

Classifier
- RF

Pipeline

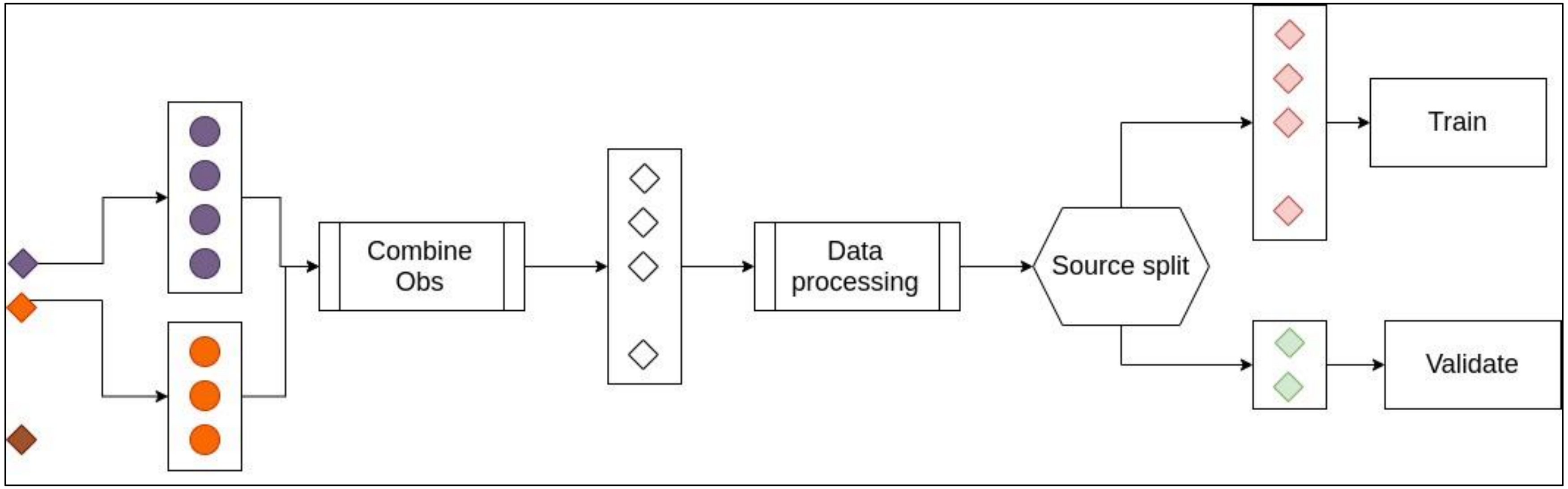




Validation accuracy
- Mean - 0.85
- Std - 0.02

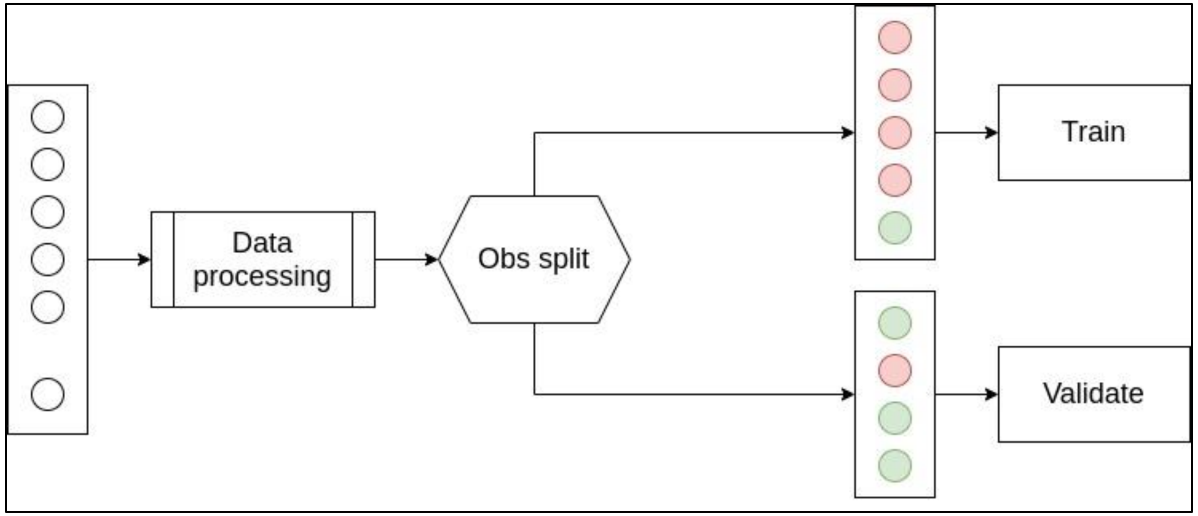
Validation accuracy –
- Mean – 0.45
- Std – 0.02

Combined observatios

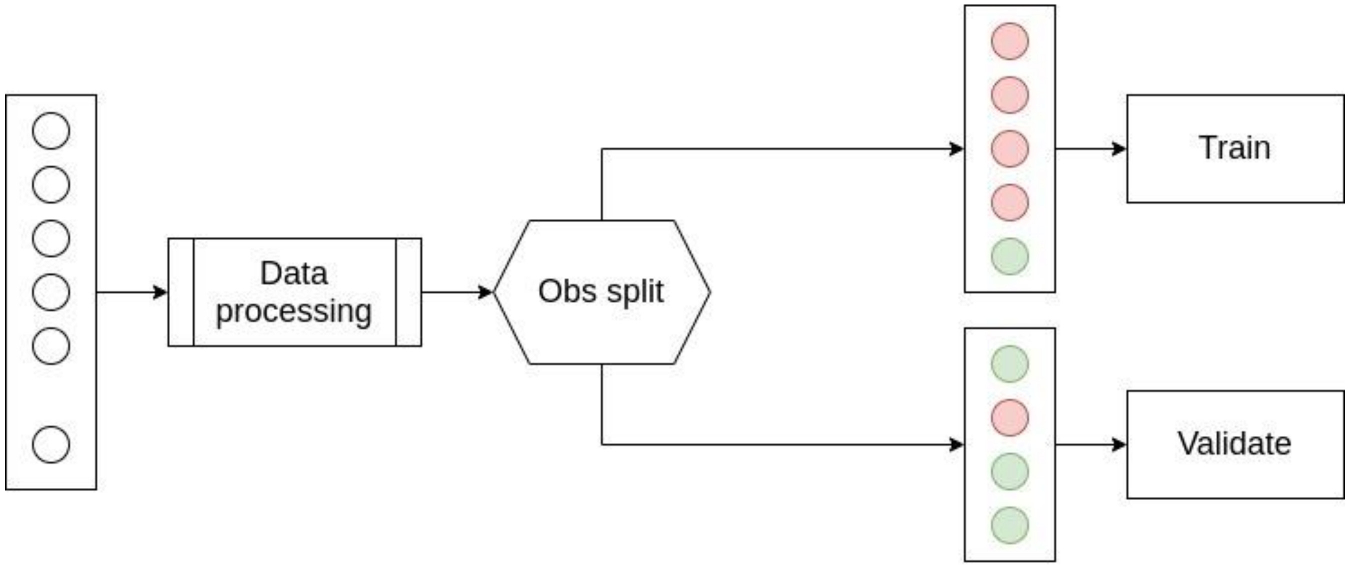


Validation accuracy –

- Mean – 0.73
- Std – 0.08



Validation accuracy – 0.85+/-0.02


Data
processing
Obs split
Train
Validate

Introduction

Data Preprocessing

Training

Result Analysis

Future Roadmap

Data Scaling

Data Imputation

Classifiers

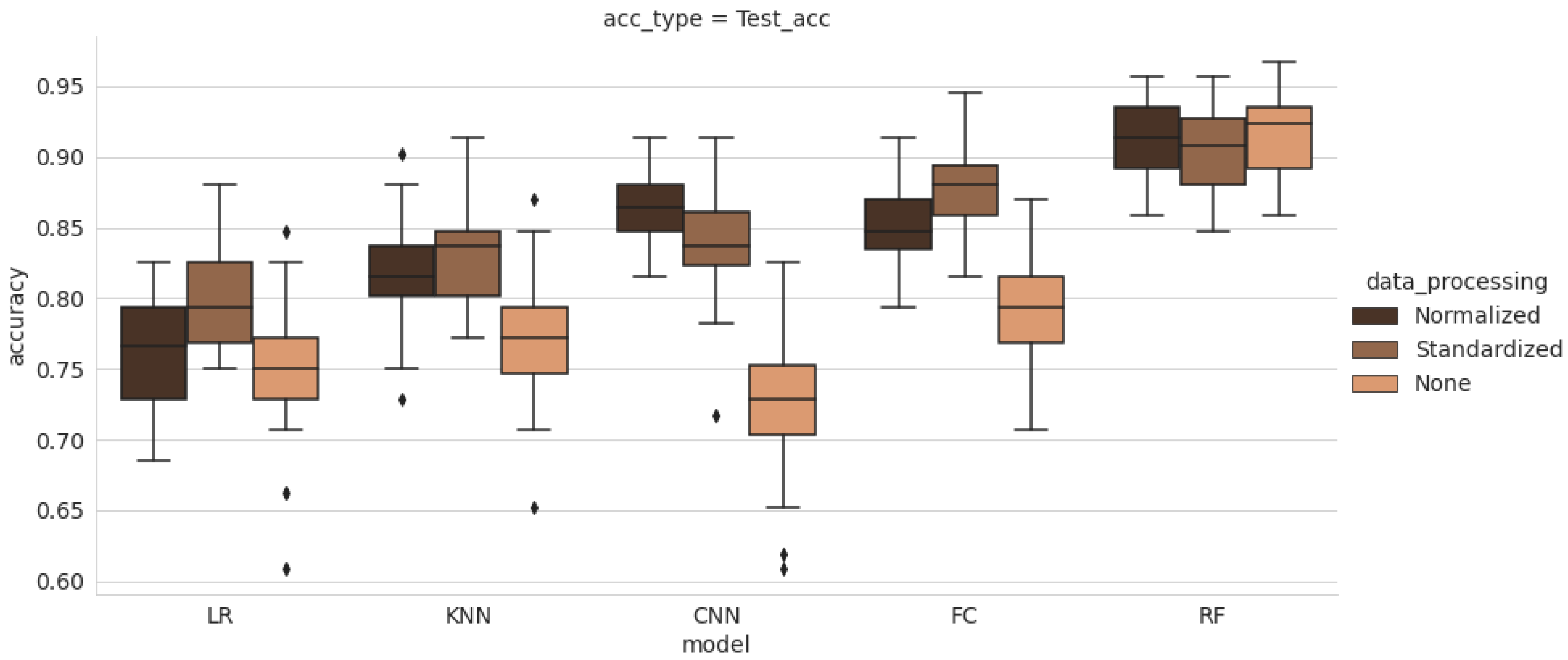
Best Schematic

Data Scaling
- No-scaling
- Normalisation
- Standardisatsation

Data Imputation
- Zero
- Mean
- Median
- Correlation
- Random Forest

Classifier
- LR
- KNN
- FC
- CNN
- RF

- Accuracy variations over data-scaling and classifier



Chandra Source Catalogue

We got training data with labels

	Feature 1	Feature 2	Feature 3	...	Feature m	CLASS
Source 1						
Source 2	Nan	NAN				
...						
Source n	NAN					

Quiescent / Outburst / Persistent source

- In quiescent state LMXRB x-ray luminosity : 10^{32} erg/s
- During outburst – luminosity : $10^{36} - 10^{38}$ erg/s
- But we have only flux information and not luminosity
- We are considering only galactic sources , distance is constrained

Dist | L_x	10^{36}	10^{38}
1kPc	8.4×10^{-9}	8.4×10^{-7}
8kPc	1.3×10^{-10}	8.4×10^{-8}
15kPc	3.7×10^{-11}	3.7×10^{-9}

- In quiescent state LMXRB x-ray luminosity : 10^{32} erg/s
- During outburst – luminosity : $10^{36} - 10^{38}$ erg/s
- We are considering only galactic sources
- 10^{-12} erg/s/cm^2

Introduction

Data Preprocessing

Training

Result Analysis

Future Roadmap

Problem

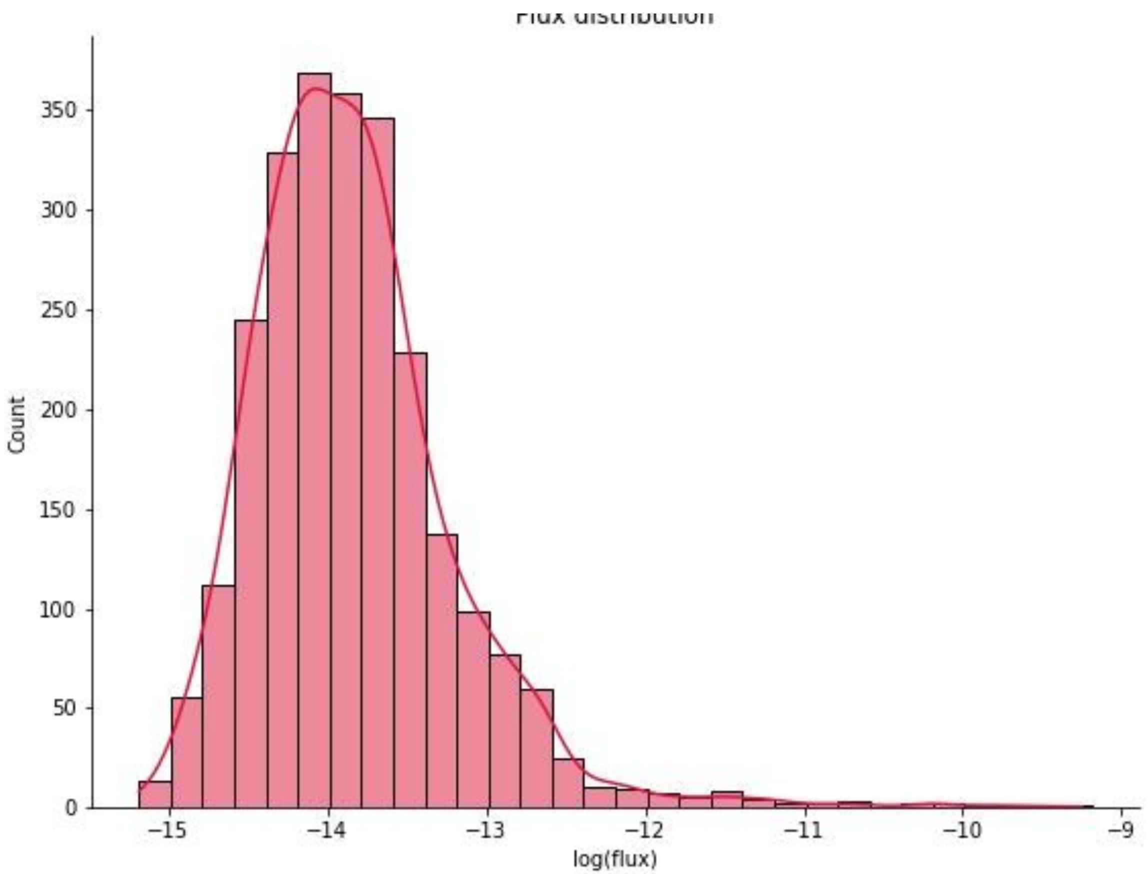
Chandra Source catalogue

Finding Labels

Source List

Quiescent / Outburst / Persistent source

- Globular cluster sources



Quiescent / Outburst / Persistent source

- Chandra source catalogue
  - Master Table
  - Per observation Table
- Identify individual observations.

		Feature 1	Feature 2	Feature 3	...	Feature m	CLASS
Source 1	Obs 1						
	Obs 2						
	Obs 3						
Source 2	Obs 1	Nan	NAN				
	...						
	Obs n						
...							
Source n		NAN					

Quiescent / Outburst / Persistent source

- Filtering
  - Flux filter
  - Pileup- flag
  - Streak source flag

Table 3: Total collected sources and observations , before obs filtering

	Num of sources	Num of Obs
NS lmxrb	84	493
BH lmxrb	33	227

Table 4: Number of sources and corresponding observations after all the filters applied

	Num of sources	Num of Obs
NS lmxrb	48	302
BH lmxrb	27	158

Order of Magnitude problem

		Feature 1	Feature 2	Feature 3	...	Feature m	CLASS
Source 1	Obs 1						
	Obs 2						
	Obs 3						
Source 2	Obs 1						
	...						

- Magnitude scale difference :
  - Flux features : 10^{-12}
  - Variance : 10^{1}
  - Hardness : -1 , 1
- Uneven weight for network based classifiers
- Incorrect feature importance
- Solution :
  - Data Normalization : xi = (xi – max)/(max – min)
  - Data Standardization : xi = (xi – mean)/var

Missing data problem

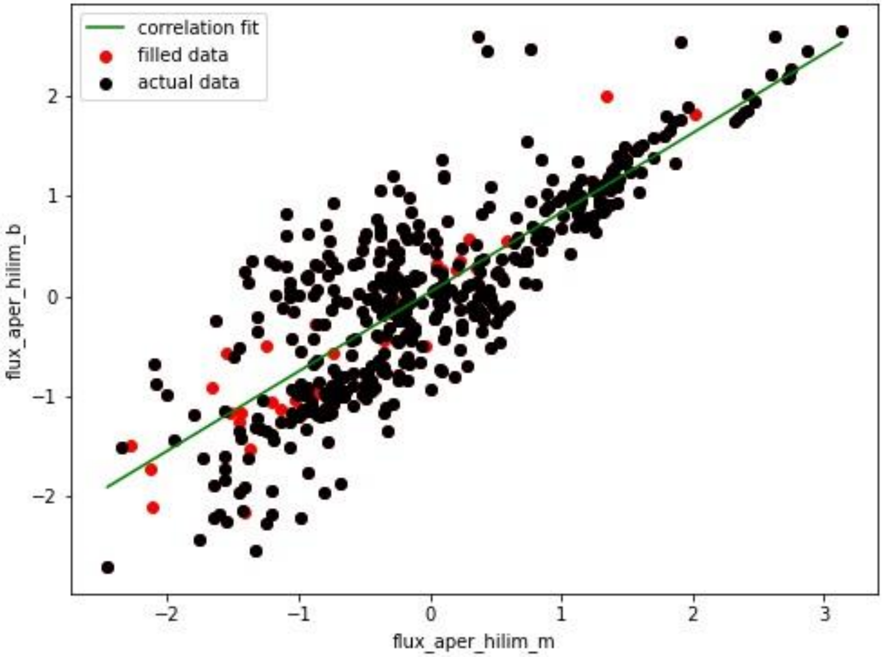
		Feature 1	Feature 2	Feature 3	...	Feature m	CLASS
Source 1	Obs 1			nan			
	Obs 2	Nan					
	Obs 3						
Source 2	Obs 1	Nan	NAN				
	...						

- Data Sparsity > 50%
- Why missing data
  - Not all obs are made in all bands
  - Model fit not done for observations made in <= 2 bands
- Solution
  - Impute with Zeros
  - Impute with feature mean
  - Impute with feature median
  - Imputation using feat correlation
  - Imputation using Random Forest

Missing data problem

		Feature 1	Feature 2	Feature 3	...	Feature m	CLASS
Source 1	Obs 1			nan			
	Obs 2	Nan					
	Obs 3						
Source 2	Obs 1	Nan	NAN				
	...						

- Imputation Using correlation
  - Find feature-feature correlation coefficient matrix
  - For each obs , fill in missing value using highest available correlated feature

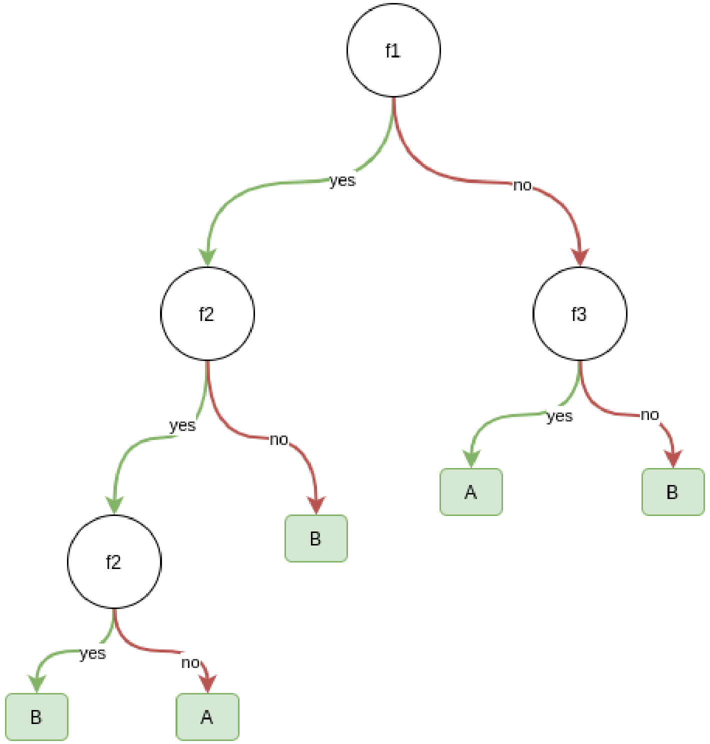


Missing data problem

		Feature 1	Feature 2	Feature 3	...	Feature m	CLASS
Source 1	Obs 1			nan			
	Obs 2	Nan					
	Obs 3						
Source 2	Obs 1	Nan	NAN				
	...						

- **Imputation Using Random Forest**
  - Fill in missing value with median
  - Calculate proximity matrix
  - Fill in missing value as weighted average of corresponding feature across all observation,
  - Weighing factor is proximity values
  - Recalculate proximity matrix
  - ..
  - ..

	x1	x2	...	xn
x1	1.0			
x2				
...				
xn				



Classifiers

- Logistic Regression
- K- Nearest Neighbour
- Fully connected network
- Convolution Neural Network
- Random Forest classifier

Introduction

Data Preprocessing

Training

Result Analysis

Future Roadmap

Data Scaling

Data Imputation

Classifiers

Best Schematic

Data Scaling

- No-scaling
- Normalisation
- Standardisatsation

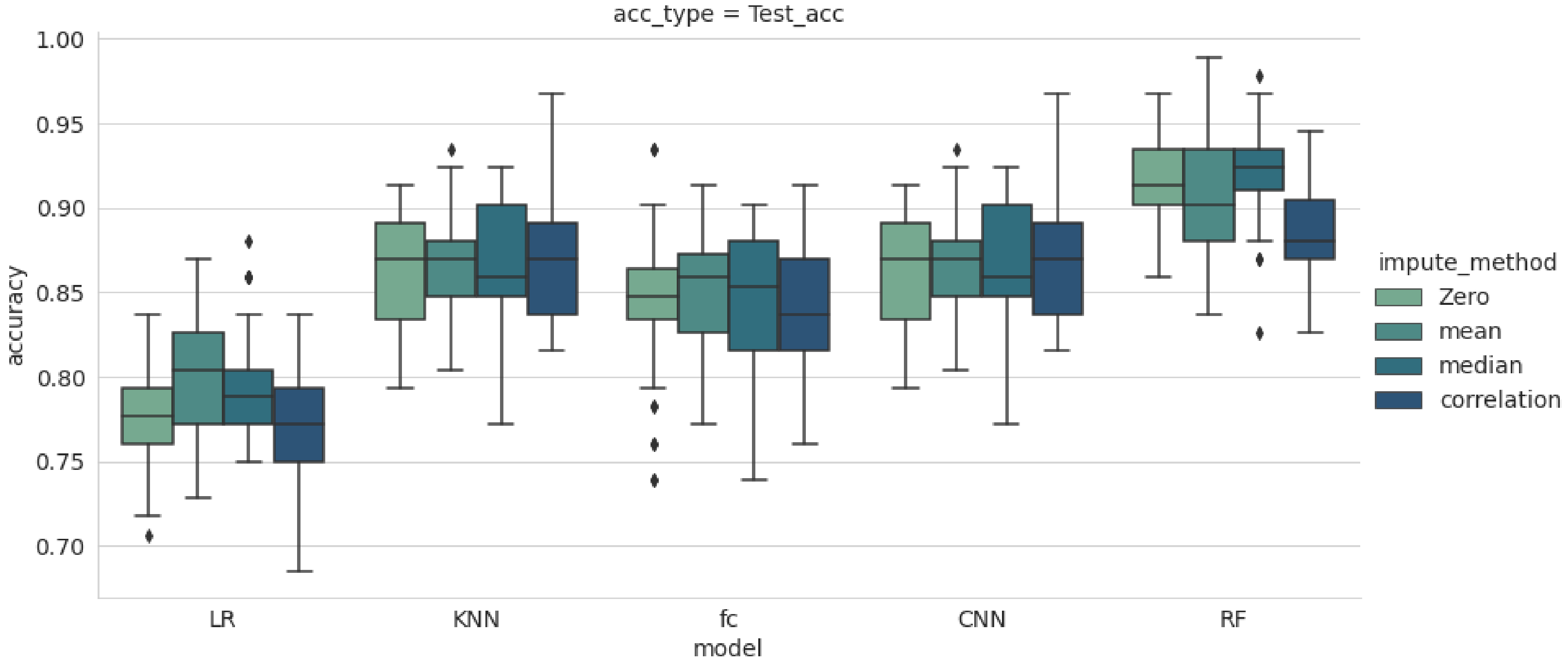
Data Imputation

- Zero
- Mean
- Median
- Correlation
- Random Forest

Classifier

- LR
- KNN
- FC
- CNN
- RF

- Accuracy variations over data-Imputation and classifier

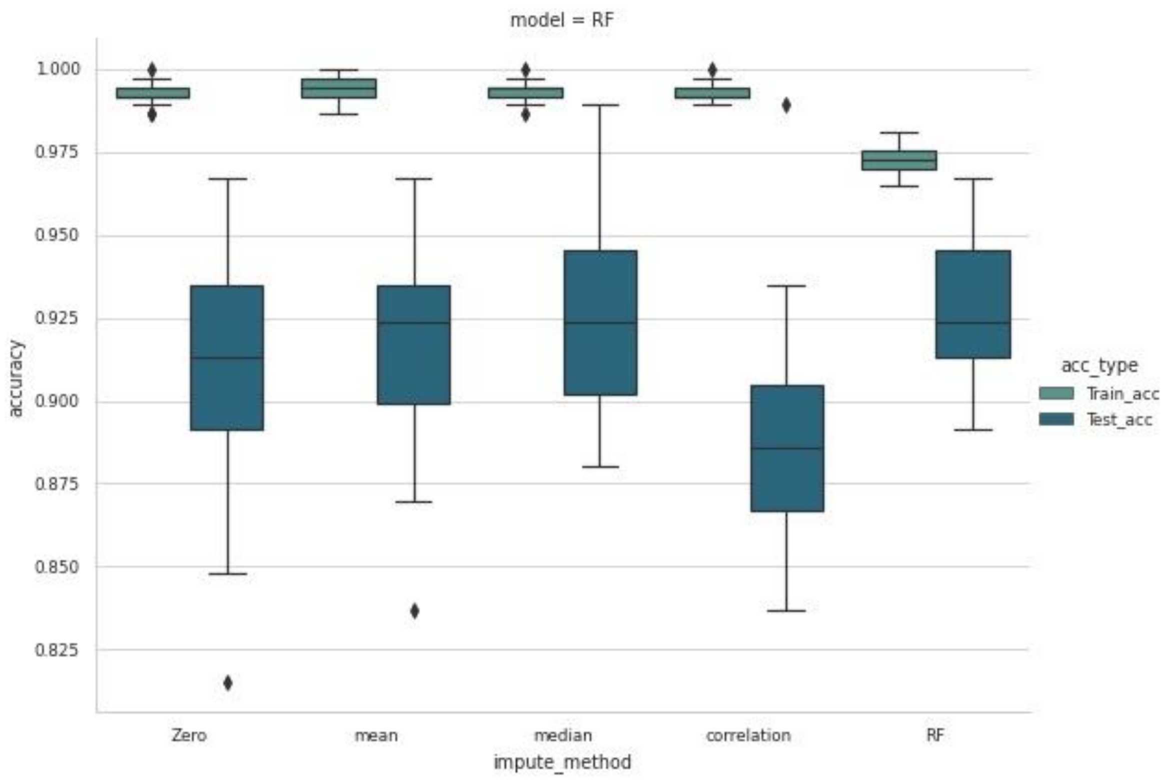


Data Scaling
- No-scaling
- Normalisation
- Standardisatsation

Data Imputation
- Zero
- Mean
- Median
- Correlation
- Random Forest

Classifier
- LR
- KNN
- FC
- CNN
- RF

- Accuracy variations over data-Imputation for Random forest classifier



• Best Classifier schematic

Data Scaling
- No-scaling
- Normalisation
- Standardisatsation

Data Imputation
- Zero
- Mean
- Median
- Correlation
- Random Forest

Classifier
- LR
- KNN
- FC
- CNN
- RF

Normalisation | **Random forest Imputation** | **Random Forest Classifier**

Classifier

Normalisation

Random forest Imputation

	fraction	Total obs	BH nos.	NS Nos.
Training sample	0.8	368	126	242
Test sample	0.2	92	32	60

- **Random Forest classifier**
- Hyper- Parameter tuning
- *Number of decision trees*
- *Maximum depth*
- Random search method
- Grid-search method

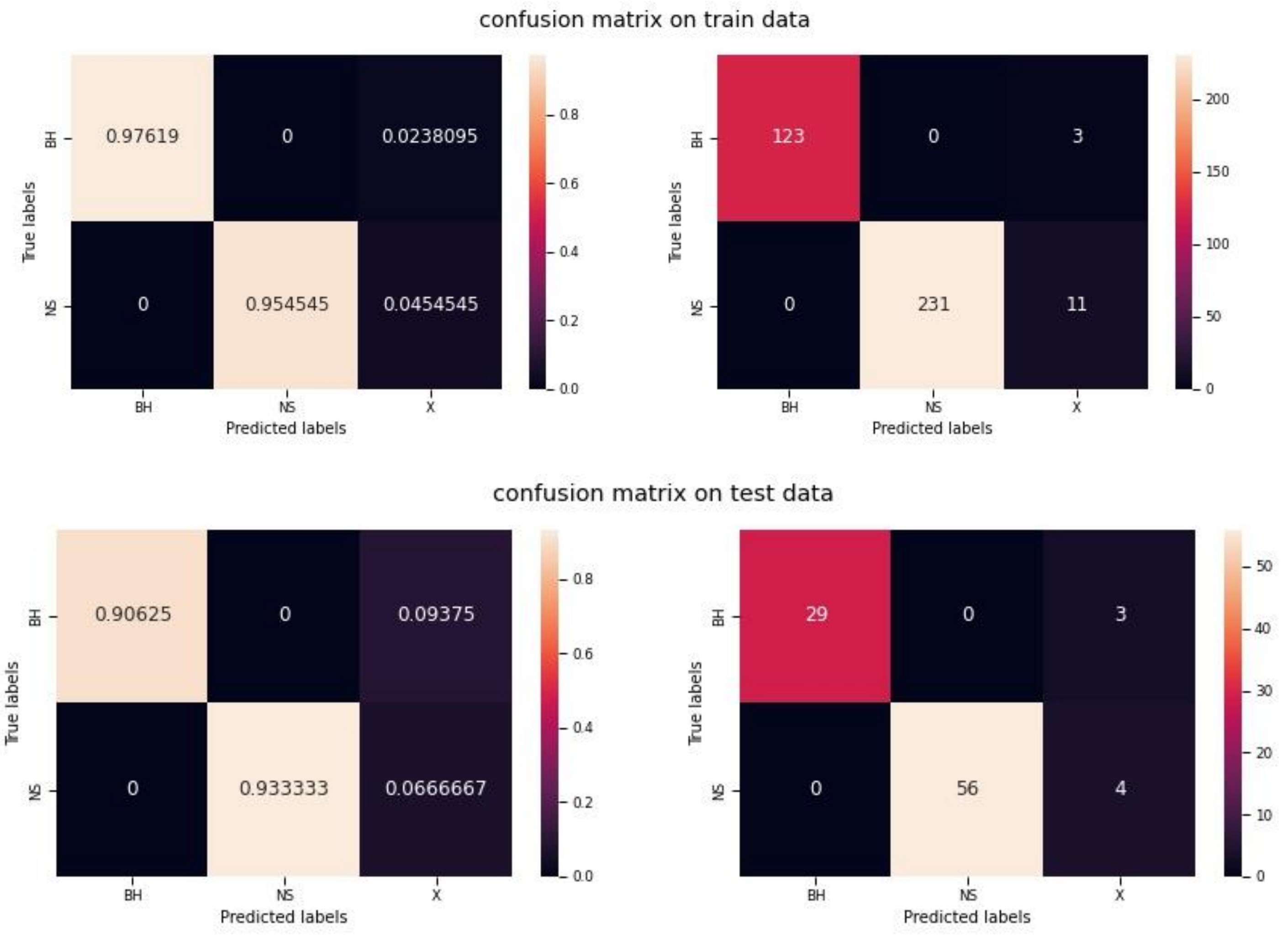
- **Random Forest classifier**
- Best Hyperparameters
  - Number of trees – 288
  - Max depth - 150

Prediction Scheme

- Probability threshold for reporting classification
  - Reduce chances of miss-classification
  - Set probability threshold for classification
  - Threshold is decided to keep false positive rate minimum
- Prediction classes :
  - NS
  - BH
  - Ambiguous
- Accuracy defined as

$$acc = \frac{(BH - BH) + (NS - NS)}{(BH - BH) + (NS - NS) + (BH - X) + (NS - X)}$$

Result


confusion matrix on train data
BH
NS
True labels
0.97619
0
0.0238095
0
0.954545
0.0454545
BH
NS
X
Predicted labels
0.0
0.2
0.4
0.6
0.8
123
0
3
0
231
11
0
50
100
150
200
confusion matrix on test data
0.90625
0
0.09375
0
0.933333
0.0666667
29
0
3
0
56
4
0
10
20
30
40
50

Result

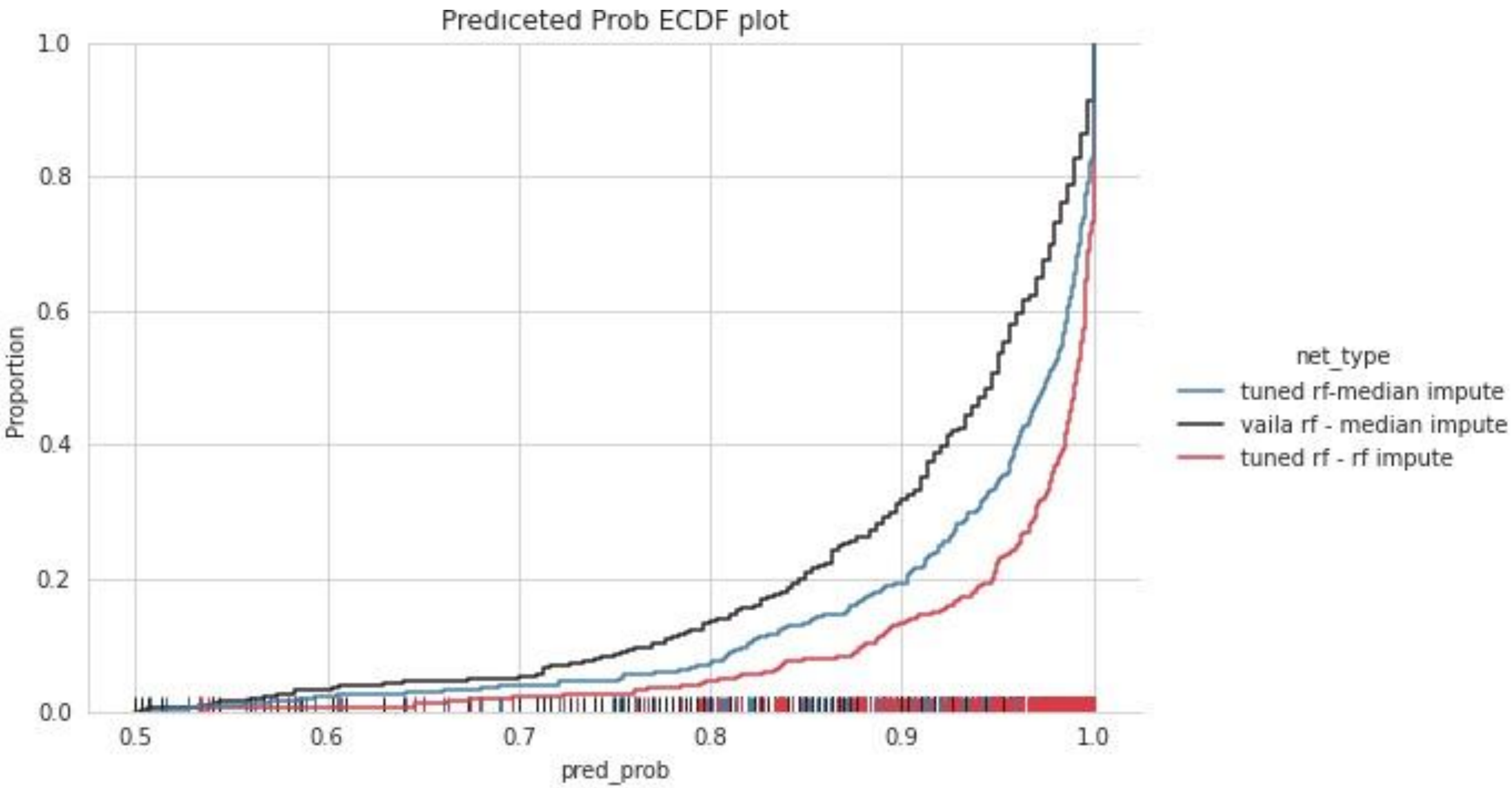
- With probability threshold for true positive set as 0.8 accuracy is :
- Training accuracy : 96.2 %
- Test accuracy : 92.1%

- **Training data**
- Total predictions – 368
- True prediction – 354
- Ambiguous predictions – 14
- Incorrect predictions - 0

- **Test data**
- Total predictions – 92
- True prediction – 85
- Ambiguous predictions – 7
- Incorrect predictions - 0

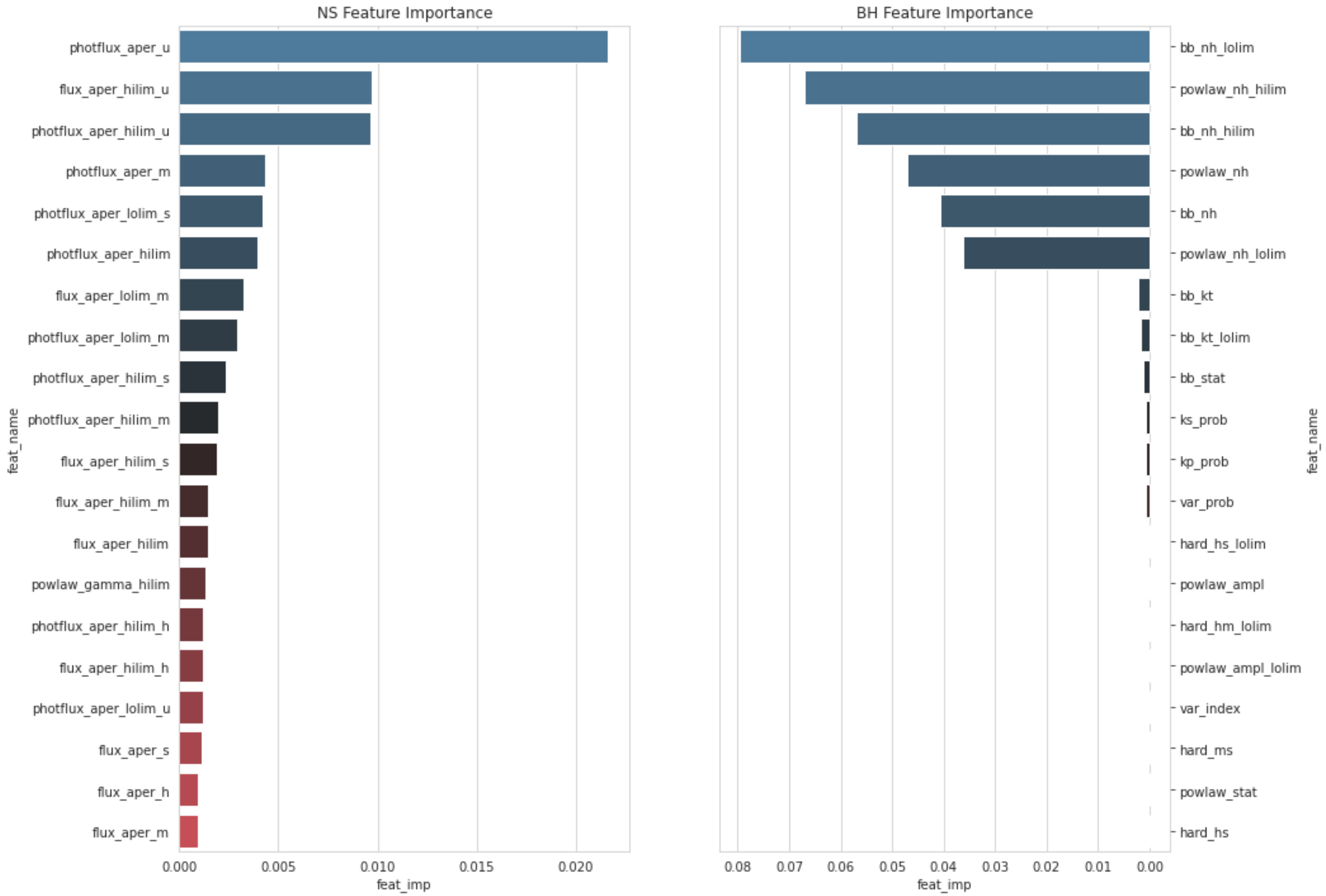
Predicted probability quality

- With probability threshold for true positive set as 0.8 accuracy is :
- Training accuracy : 96.2 %
- Test accuracy : 92.1%



Feature Importance

- RF gives feature importance to each feature
- Class-wise feature Importance :



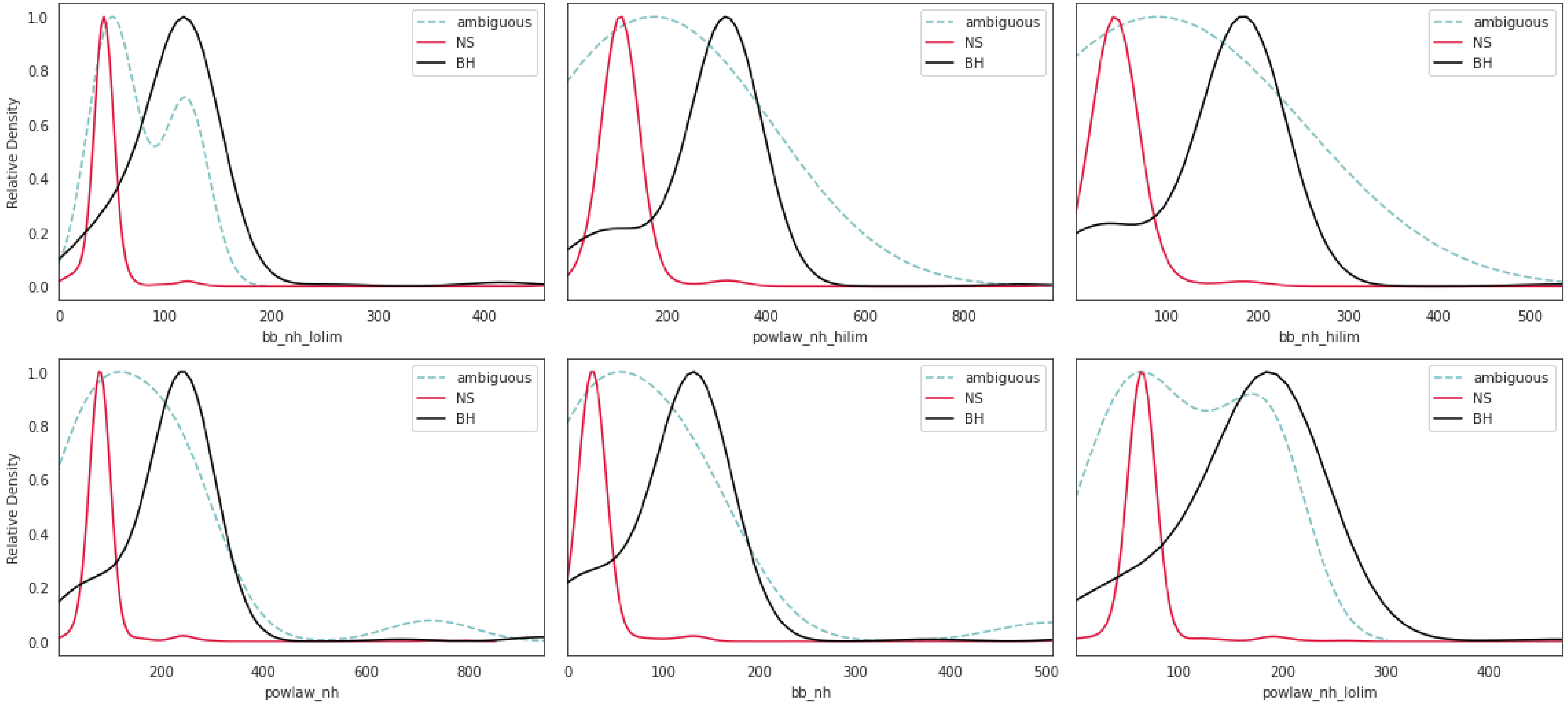
Feature Importance

- Based on Gini Impurity
- Class-wise feature Importance : $I_{fk,A} = I_k \times mean(f_k(X_i \in A))$

NS	BH
Photon flux -u band	Black body , column density lower limit
Energy flux - u band	Powerlaw column density upper limit
Photon flux -u band upper limit	Black body , column density upper limit
Photon flux - m band	powerlaw column density
Photon flux - s band Lower limit	black body column density
Band average photon flux upper limit	Powerlaw column density lower limit

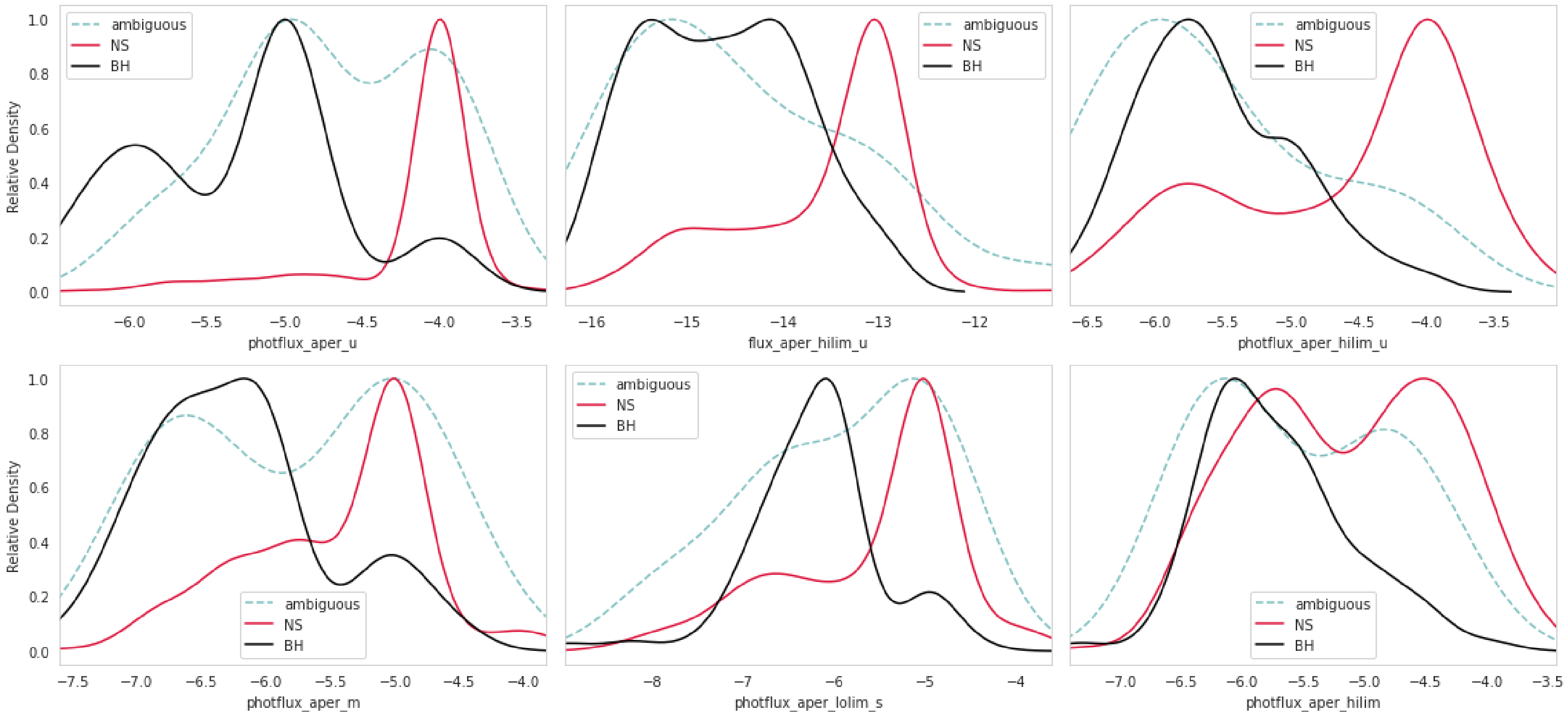
Feature Importance

- Black Hole lmxrb important features



Feature Importance

- Neutron Star important features

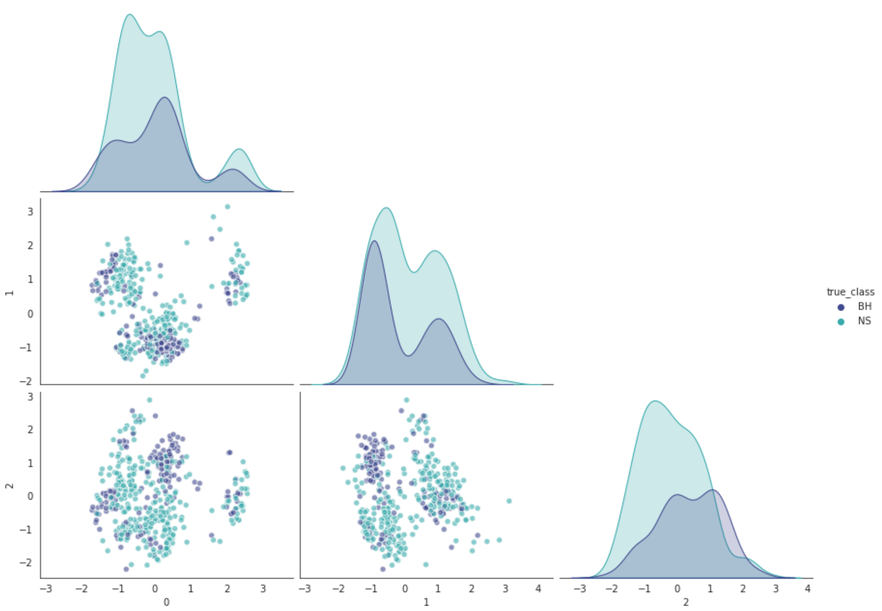


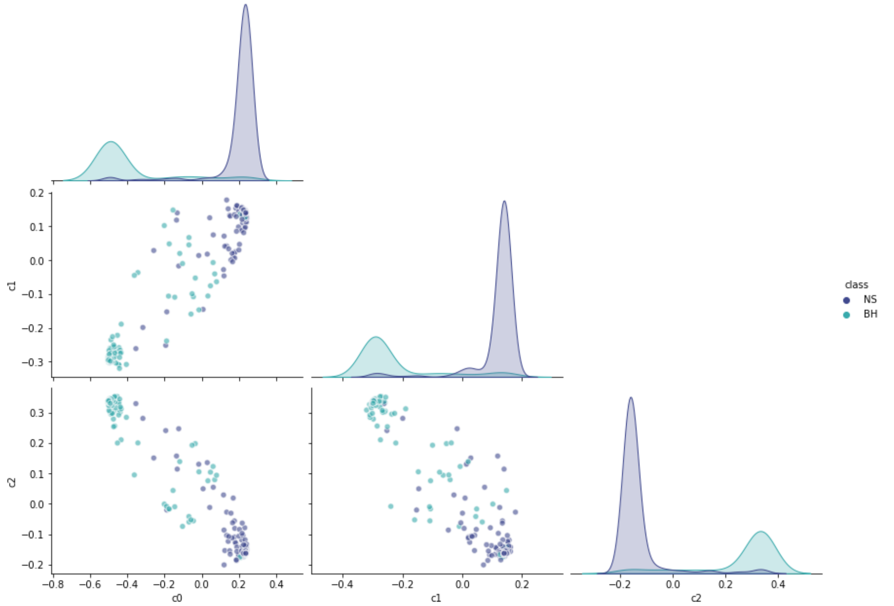
Conclusion and Future

- **Conclusion**
  - Identified best schematic for LMXRB classification into NS and BH
  - Achieved test accuracy – 92 %
- **Future Work Plan**
  - Study feature-feature correlation to drop not-so important features
  - Physical significane of the result
- Phase –02 :
  - Add CVs and Mili second puldars to classification
- Phase –03 :
  - Try Unsupervised learning with observations of all the GC sources in CSC
- Phase –04 :
  - Expand classification to non-gc and extraglactic sources also

Thank You

Why Probability improved





Introduction

Probability quality

Data Preprocessing

Feature Importace

Training

Result Analysis

Future Roadmap

Feature Importance

- RF gives feature importance to each feature
- Class-wise feature Importance : $I_{fk,A} = I_k \times mean(f_k(X_i \in A))$