Classification of Faint X-ray Sources Associated with Globular Cluster Using Machine Learning

December 2021

Shivam Kumaran

Indian Institute of Space Science and Technology

Problem Statement

We Need to classify X-ray sources associated with Globular cluster using properties available in Chandra Source Catalogue 2.0

Globular Cluster

- **System of stars gravitationally bound together**

- **GC dynamical Evolution**
  - simulation of dynamic evolution of GC[1]
  - Without XRB - mean collapse time scale < mean time scale of Galactic GC

[1]Carretta, et al .(2000). THE ASTROPHYSICAL JOURNAL,

Globular Cluster

- **Improved Simulation**
  - Included XRB
  - Mean time scale matches
- **Hypothesis -**
  - XRB helps in stability against gravitational collapse
- **Dynamical evolution process with XRB[1]**
  - Core contraction phase
  - binary burning phase - collapse halts
  - After max-possible binaries formed, collapse restarts
  - binary burning phase restarts

[1] .Pooley, D. (2009). Globular cluster x-ray sources. PNAS

Globular Cluster Evolution

- **Binary Burning phase and dynamical evolution governed by - Encounter Frequency**
- **Directly correlated with population distribution of XRB.**

Understanding XRB Population distribution

↓

Understanding Encounter Frequency

↓

Globular cluster Dynamics and Evolution

GC X-ray Binaries

Low Mass X-ray Binary

- Companion Neutron star or Black hole
- Donating star mass < 1.5 M_{solar}
- Identification : specteal studies , mostly during outburst

Cataclysmic Variable

- Binary system accretion onto White Dwarf
- Identification : Bright in UV , soft-X-ray

Millisecond Pulsar

- Rapidly rotating Neutron Star
- Formed from LMXB
- Identification : using radio timing

Example : 47-TUC , Heinke, et.al (2005)

- **About 47-TUC**
  - Mass : $10^6\ M_\odot$
  - Distance : 4.85 kpc
  - Size : core radius 24”

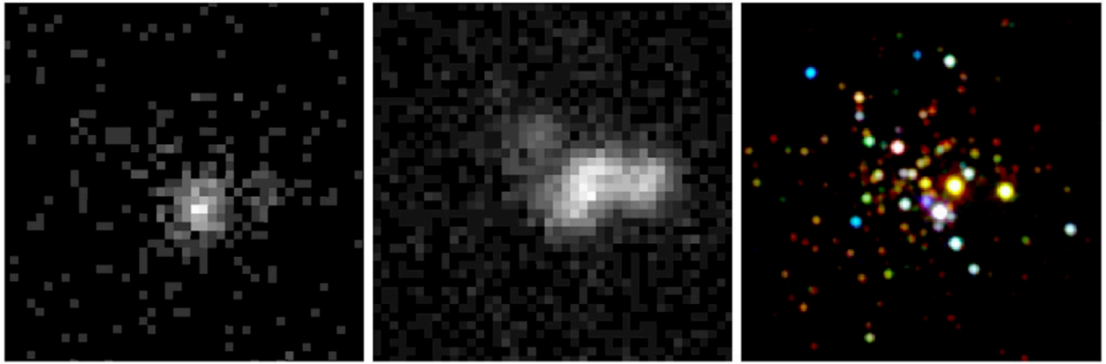
Class	**Identification Method**	**No of sources**	**No Expected**
CV	Optical identification	30	24-113
MSP	Radio cross match	27	~ 700
qLMXB	Spectral studies	5	~ 300

Manual Identification is not easy.

Need Better identification

Heinke, et.al (2005). A deep chandra survey of the globular cluster 47 tucanae: Catalog of point sources.

Why Chandra ?



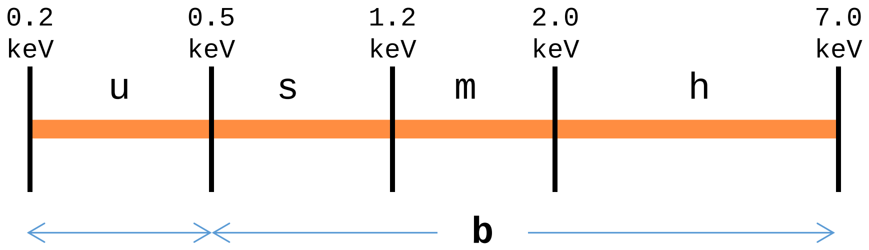
Images of the core of 47-Tuc made from 8 ks of Einstein data (Left), 77 ks of ROSAT data (Center), and 240 ks of Chandra data (Right) . Image and caption credits (Pooley, 2009)

- For any such identification we need telescope like Chandra
- Higher sensitivity
- Better resolution

Chandra : Instruments

- **High Resolution Camera (HRC)**
  - Energy band :
    - wide band (w) - 0.1-10keV

ACIS Energy bands



- **Advanced CCD Imageing Spectograph (ACIS)**
  - **Resolution** : 0.5 arcsec on-axis
  - **Sensitivity** : 4×10^{-15} ergs/cm^2/s - integration time 104 sec
  - **Energy Band** :
    - 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

Chandra : Chandra Source catalogue

Number of sources - 317,000
Number of sources associated with GC ~ 1700

Per-Obs Detection Table

contains detection properties based on observational data extracted independently from each individual observation

source	Obs	properties
s1	obs1	
	obs2	
	obs3	
s2	obs1	
s3	obs1	
	obs2	

Per-Stack Detection Table

The Stacked Observation Detections Table

Master Source Table

'best estimate' sources properties for each unique X-ray source in the catalog

source	properties
s1	
s2	
s3	
s4	
s5	
...	

Chandra Source catalogue : **features**

- What are these " **Properties** " ?

source	Obs	properties
s1	obs1	
	obs2	
	obs3	
s2	obs1	
s3	obs1	
	obs2	
	obs3	
	....	
...		

Chandra Source catalogue : **features**

Variability

- Inter observation Variability
- Intra Observation variability

Aparture Photometry

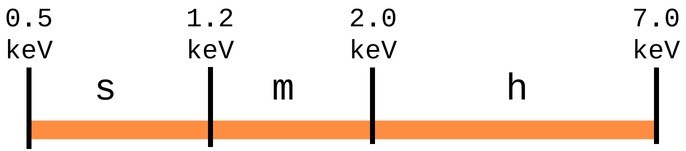
- Photon Flux
- Energy Flux

Spectral Properties

- Hardness Ratio
  - Hardness hm
  - Hardness ms
  - Hardness hs
- Model-Fit properties
  - Black Body model
  - Bremestralung model
  - Powerlaw model

Chandra Source catalogue : **features**

- **Hardness**
  - hard hm
  - hard ms
  - hard hs
- **Hardness calculation details**
  - Slope of the energy band vs flux curve



source	Obs	properties
s1	obs1	
	obs2	
	obs3	
s2	obs1	
s3	obs1	
	obs2	
	obs3	
	....	
...		

Introduction

Data Collection

Data Collection

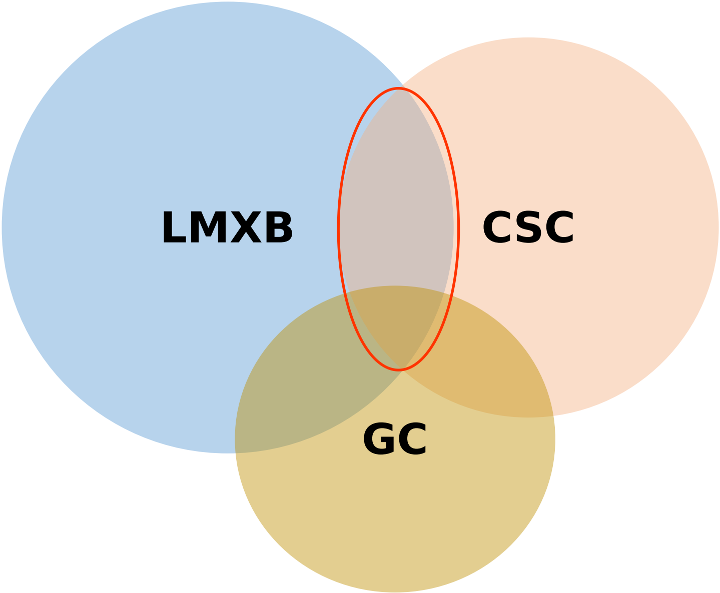
Problem

We do not have class labels in CSC

Solution

Look for other catalogue and in published literature

Data Collection



- CV - 314
- LMXB - 99
- Pulsar - 265

For a givne class - LMXB / CV / MPS

↓

Find RA/DEC for Known source

↓

Cross-Match with CSC

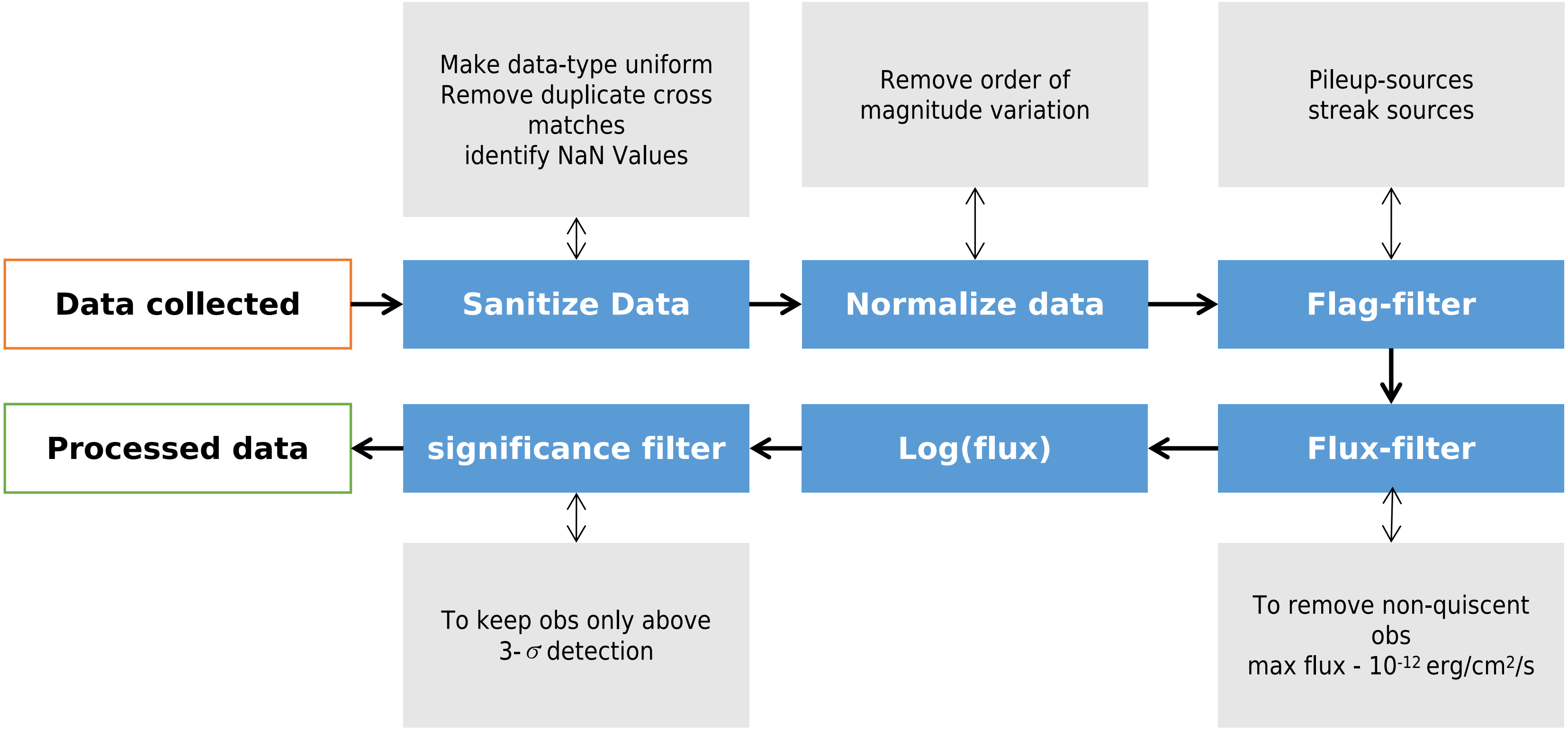
↓

Download data from CSC

Data Collection

Data Preprocessing

Data Preprocessing



Data Preprocessing

	CV		LMXB		Pulsar	
Sources Cross matched with Chandra	Source	Obs	Source	obs	Source	Obs
Unique id, ACIS obs available	314		99		265	
Source after flag filters	301	2101	99	735	245	1183
Sources after flux filter	282	2044	74	662	232	1104
Source above 3σ	184	1582	58	521	178	1000

Machine Learning

- Finally we have Data and labels as well.
- We need to learn feature-class label relation
- Typical Machine learning problem
- Is it that simple ?
- NO

source	Obs	properties	Class label
s1	obs1		
	obs2		
	obs3		
s2	obs1		
s3	obs1		
	obs2		
	obs3		
	....		
...			

We need to learn feature-class label relation

Data Preprocessing

- **Problems**
  - Very small dataset
    - CV - 184
    - MSP - 178
    - LMXB - 58
- **Missing data**
  - About 50% missing values
  - **NO** feature column with zero missing values
  - **only 7** Sources with zero missing values
- **Reason for missing values**
  - Source may be faint in some bands

We Need to fill in Missing values

Data Imputation

- **Statistical Imputation**
  - Impute with column mean
  - Impute with column median
  - Impute with zeros.
- **Correlation Imputation**
- **Regression Imputation**

Data Imputation

- **Regression Imputation**

src	properties			
	flux	variability	hardness	………
s1				
s2				
s3				
s4				
s5				
s6				
s7				
s8				
…				

Data Imputation

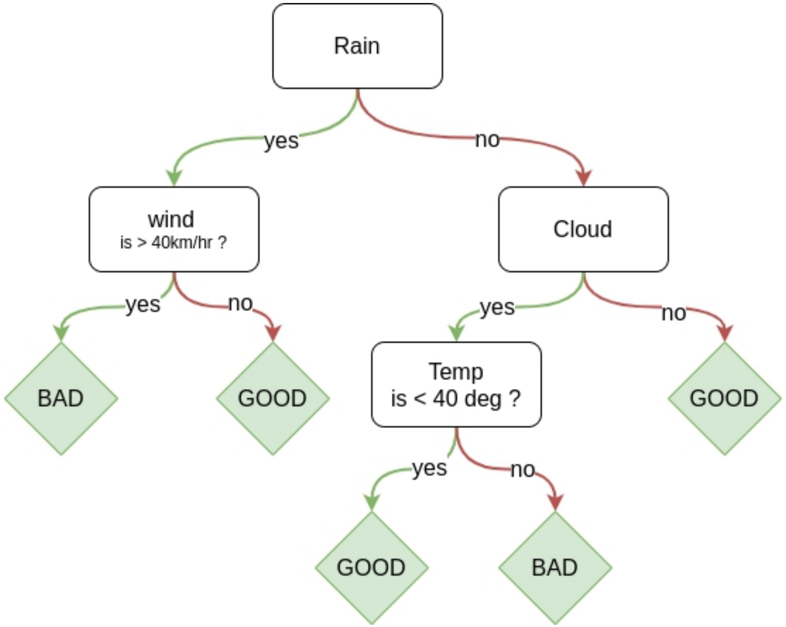
- How to compare which regression method works correctly for classification
- Need to do classification
- Need a classifier.

Classifier

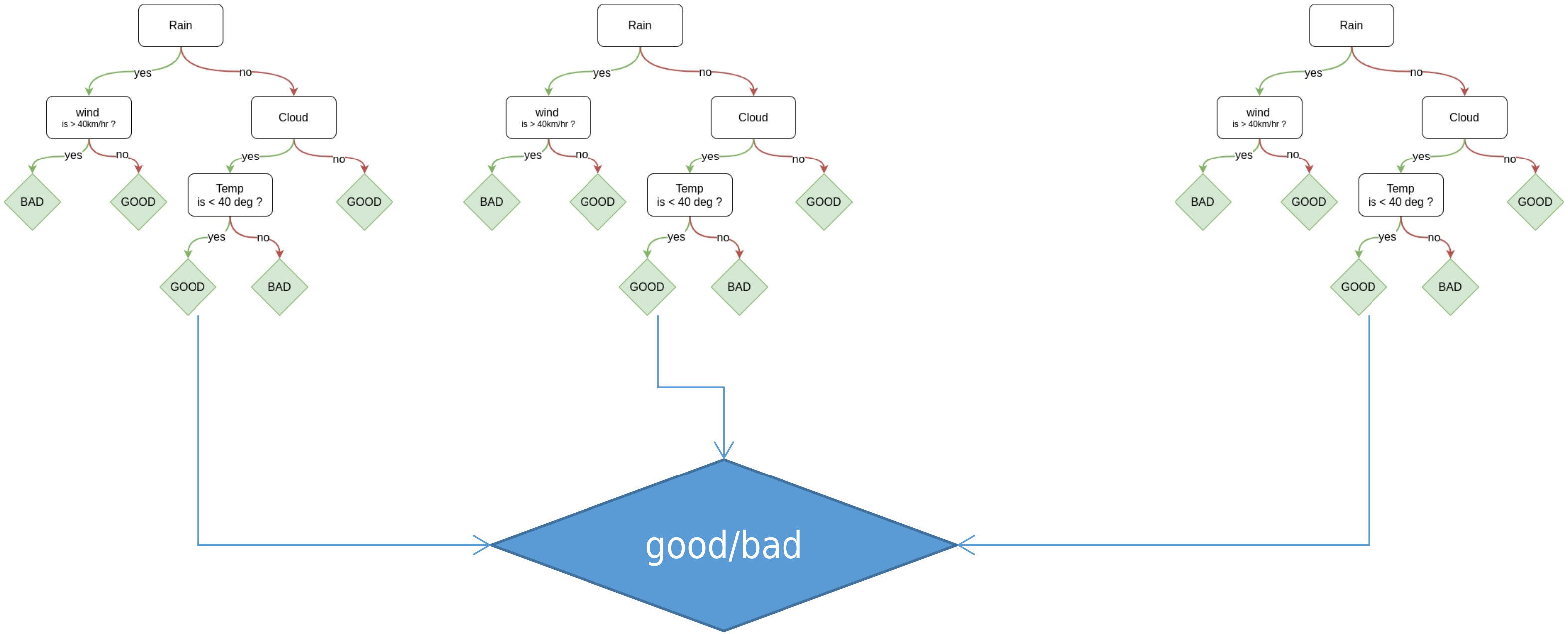
- Logistic Regression
- K-Nearest Neighbour
- Fully Connected Network
- Convolution Neural Network
- Random Forest Classifier

Classifier : Random Forest

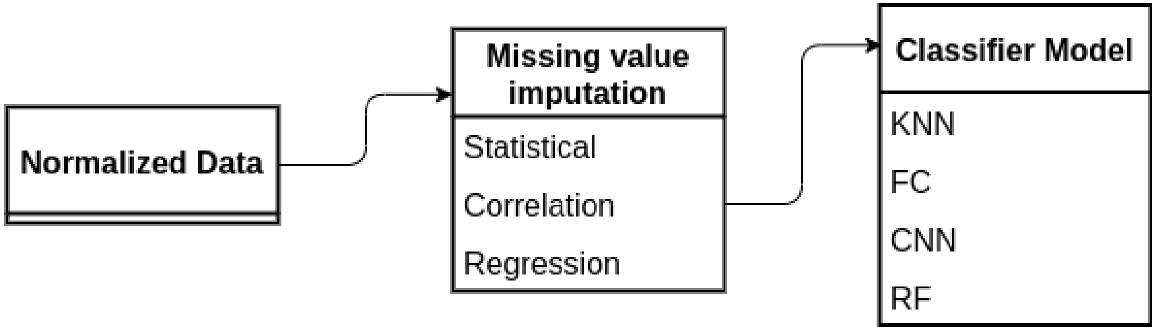
Day	Rain	Wind	cloud	Temp	How is the day
Day 0	yes	30	yes	10	**good**
Day 1	yes	55	yes	10	**Bad**
Day 2	no	10	no	55	**bad**
Day 4	no	30	yes	20	**Good**
…	…	….	….	….	**….**



Classifier : Random Forest



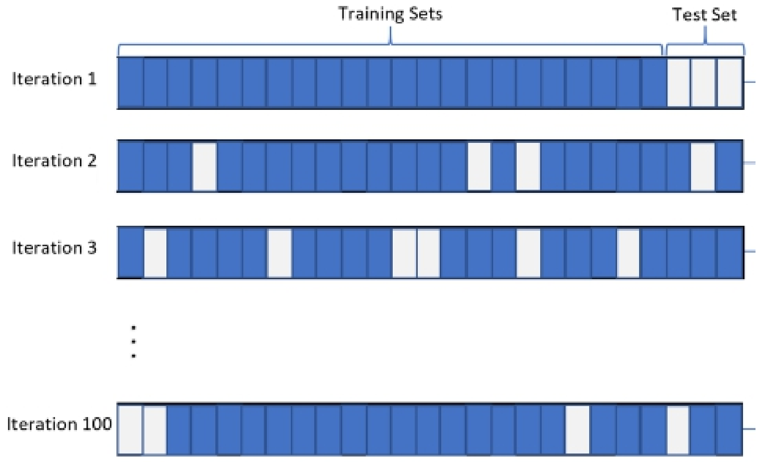
Classifier Pipeline



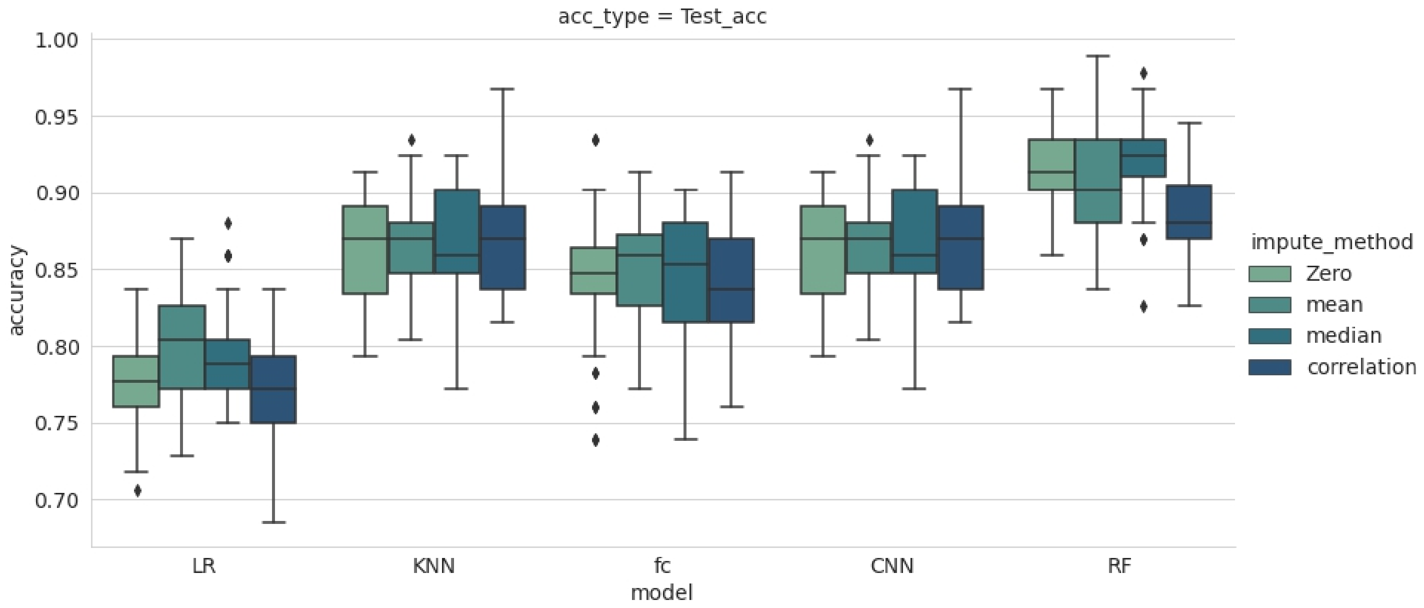
How to select which one works the best ?

Cross Validation

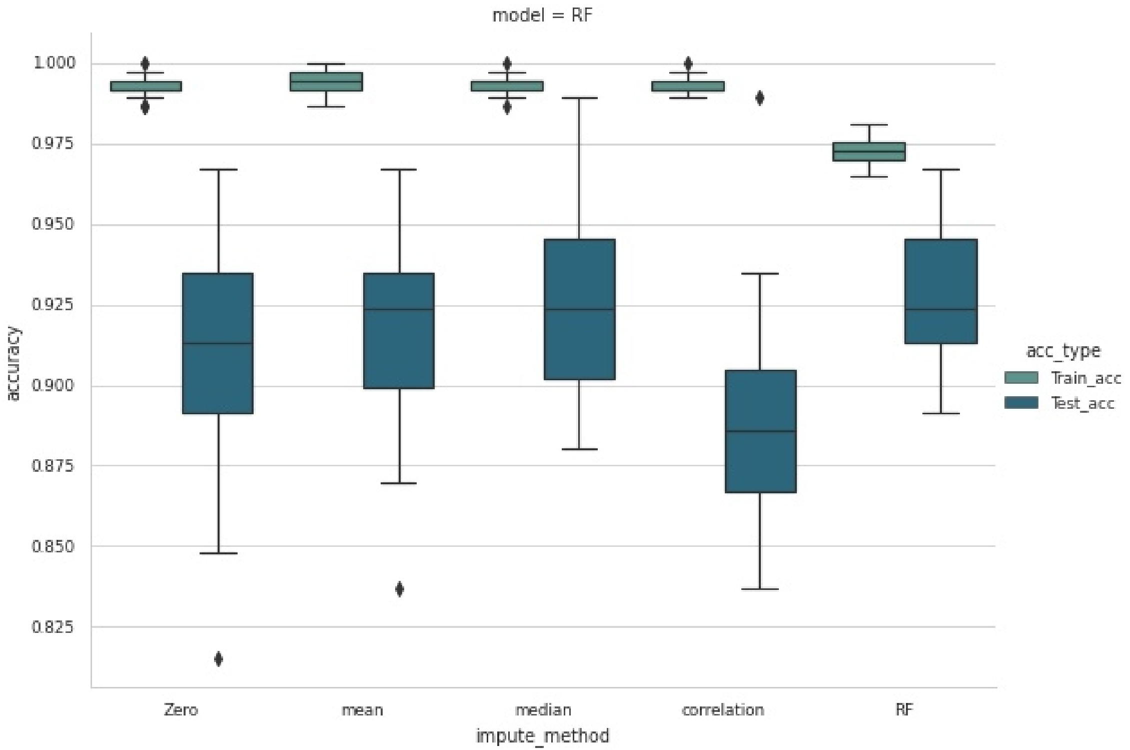
- **Algorithm**
  - Take randomly sampled examples
  - keep them aside
  - train on rest of the sample
  - check preformance on the kept-aside sample
- **A good model**
  - Higher mean accuracy
  - Least std in accuracy.



Monte-Carlo Evolution : Result



Imputer+ Classifier Result



Classifier selected

- **Validation Accuracy score**
  - Mean - 66.1
  - Std - 4.3
  - Min - 51.46
  - Max - 71.31

Preprocesssed data

↓

Fill missing values with random forest regressor

↓

Random Forest Classifier

Classifier selected

- **Combine Observation**
  - Improved statistics

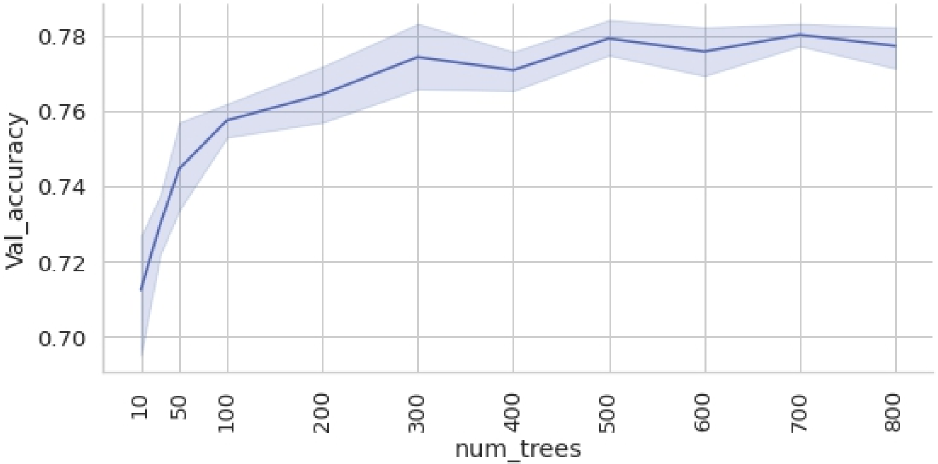
	Mean	Std	Min	Max
Observation-wise classification	66.1	4.3	51.46	71.31
Source-wise classification	76.37	1.83	71.4	79.36

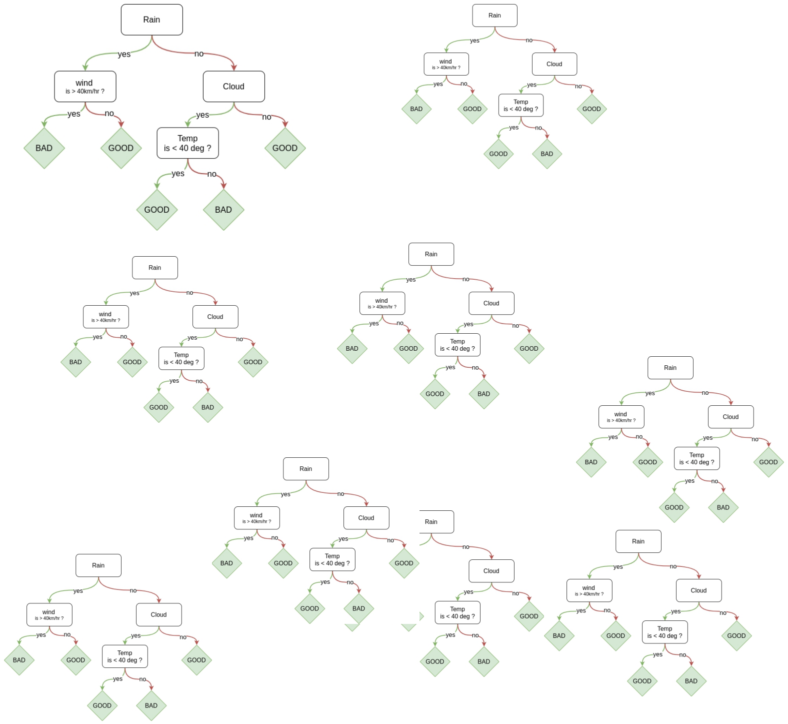
Classifier Designed..

Optimising Classifier

Optimizing RF

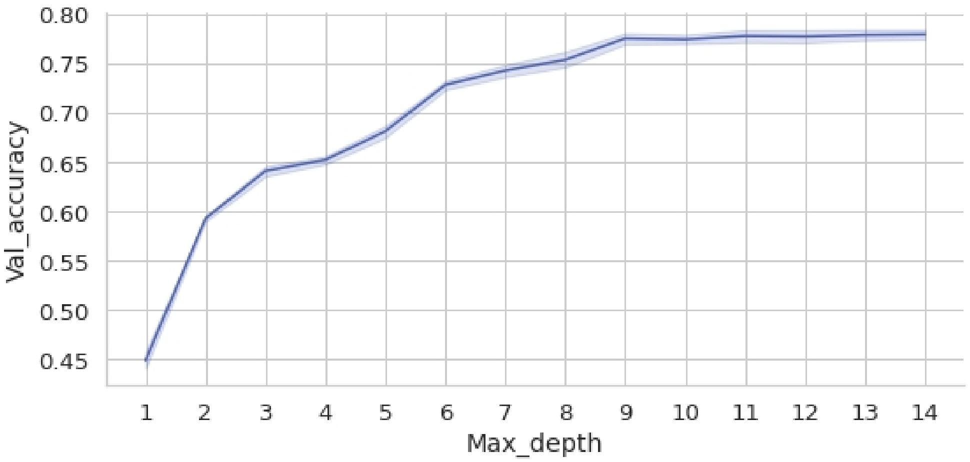
- **Hyperparameter tuning**
- **Parameters to tune**
  - Number of trees
  - Max-depth

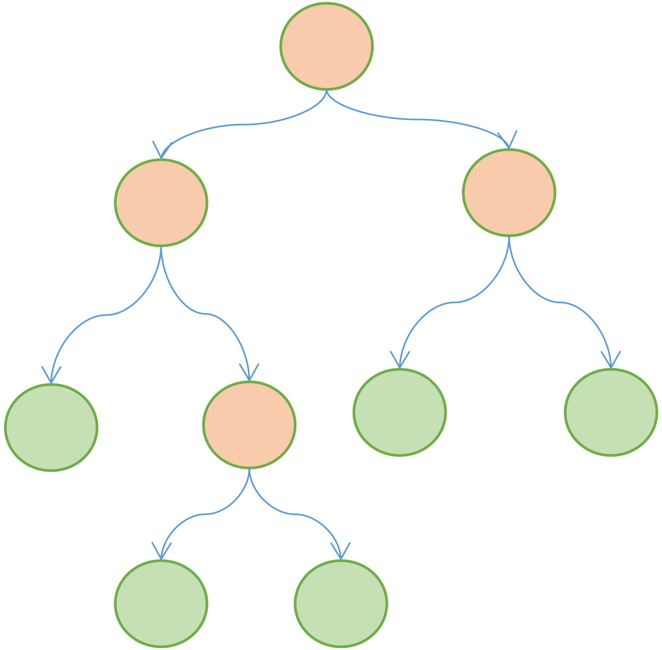




Optimizing RF

- **Hyperparameter tuning**
- **Parameters to tune**
  - Number of trees - 500
  - Max-depth - 10





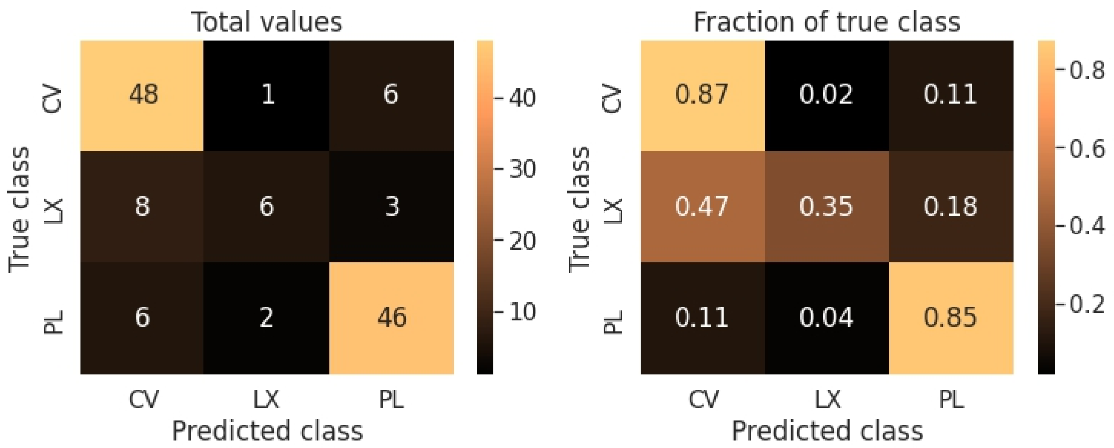
Tuned RF Result: Score

- **Best Random forest**
  - Number of trees - 500
  - Max-depth - 10
- **Result**
  - Accuracy :

	mean	Std	Min	Max
Baseline RF	75.27	1.78	72.22	78.57
Tuned RF	75.96	0.99	73.8	77.78

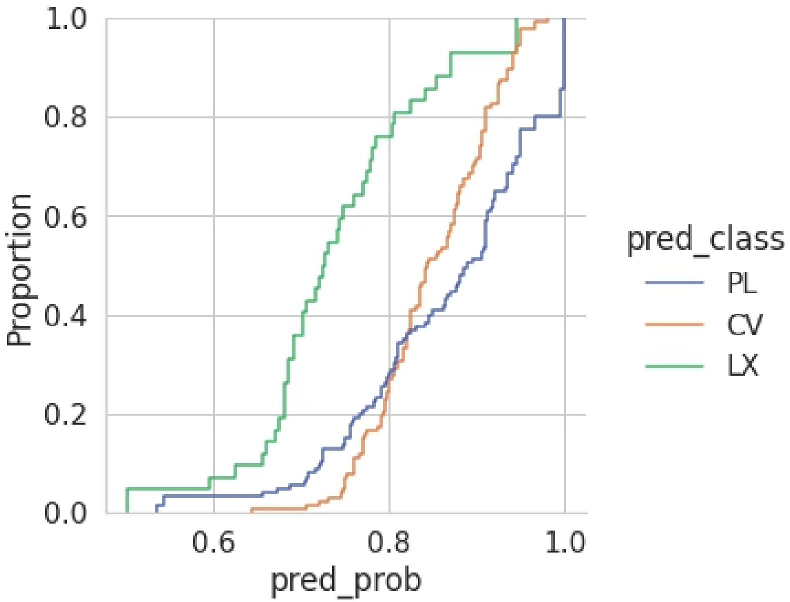
Tuned RF Result : Confusion Matrix

- **Confusion Matrix**
- **Probability quality**
- **Problem**
  - Class imbalance
  - not able to learn LMXB



Tuned RF Result : Probability quality

- **Probability quality**
- **Problem**
  - Class imbalance
  - not able to learn LMXB class.

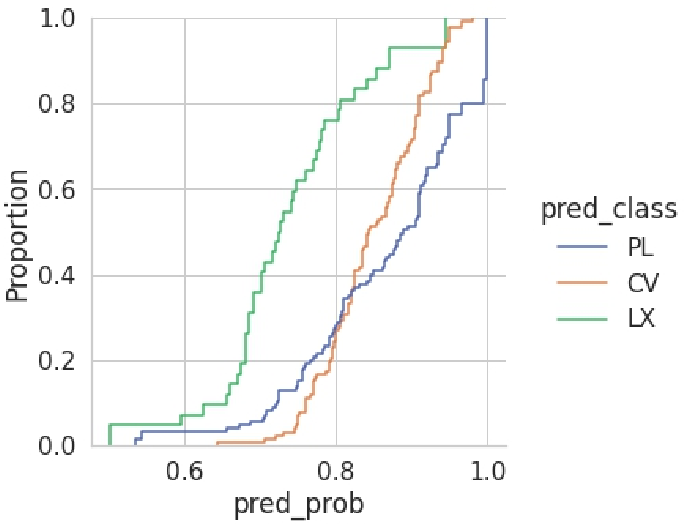


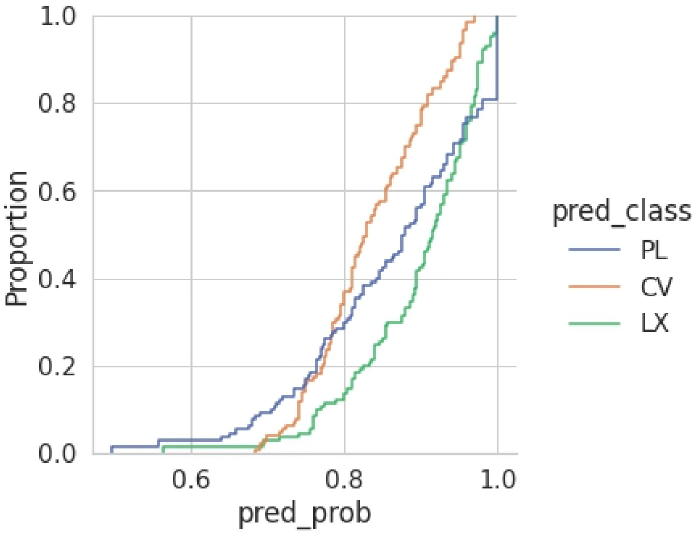
Balancing Class : SMOTE

- **SMOTE : Synthetic Minority Oversampling Technique**
- **Algorithm**
  - In higher dimension feature space
  - Each point represent one source
  - Linear interpolation between these points (source)
  - Sample points from nearest interpolations
  - Make each class equal.

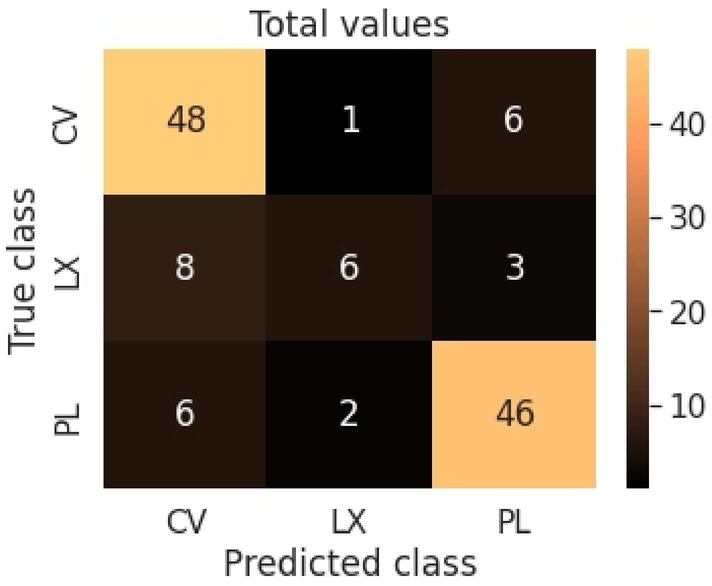
Balancing Class : SMOTE

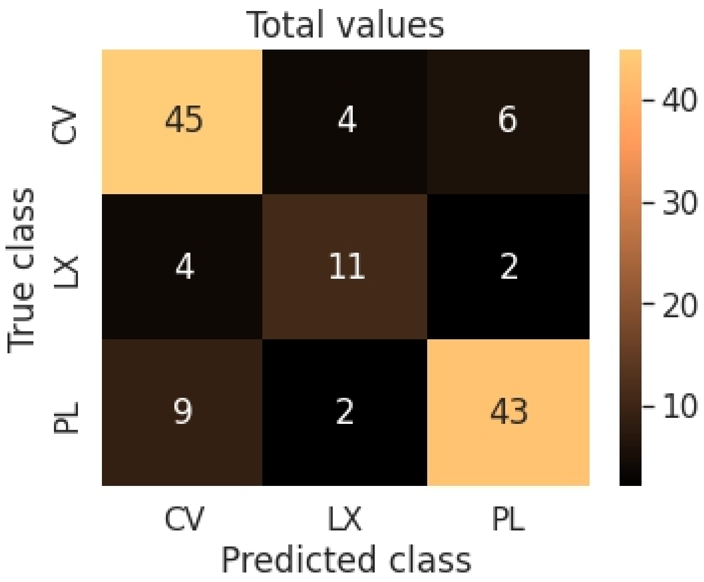
- **SMOTE result**



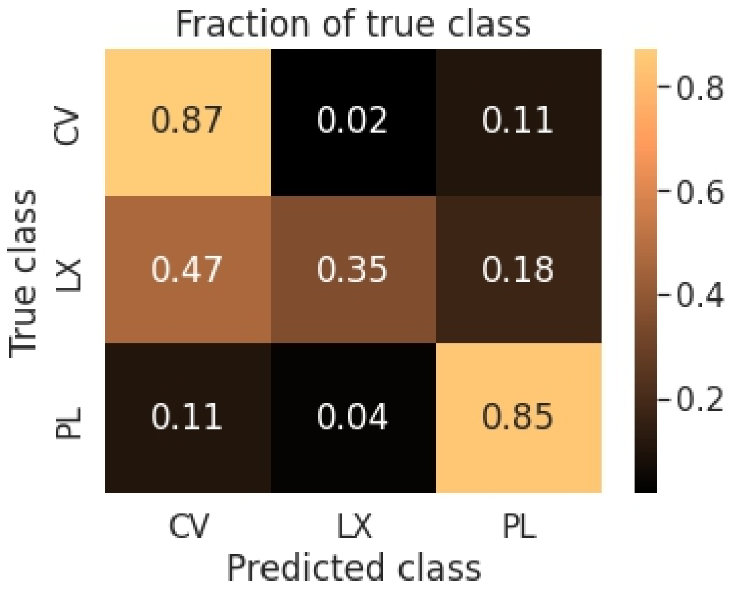


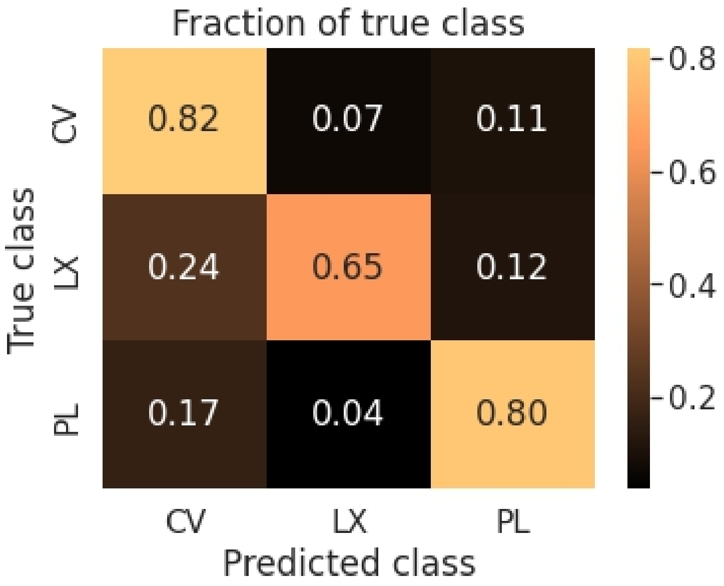
Balancing Class : SMOTE





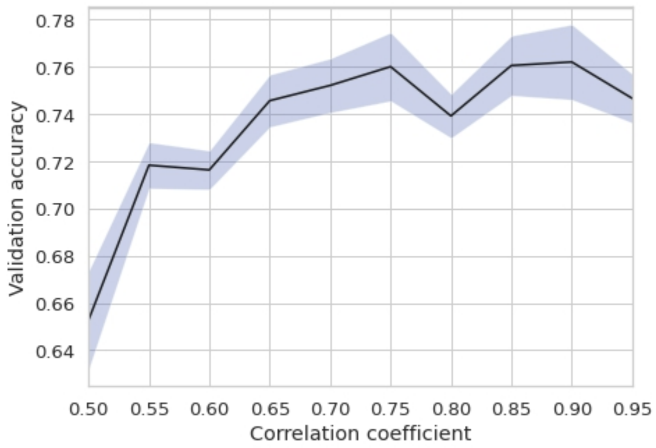
Balancing Class : SMOTE

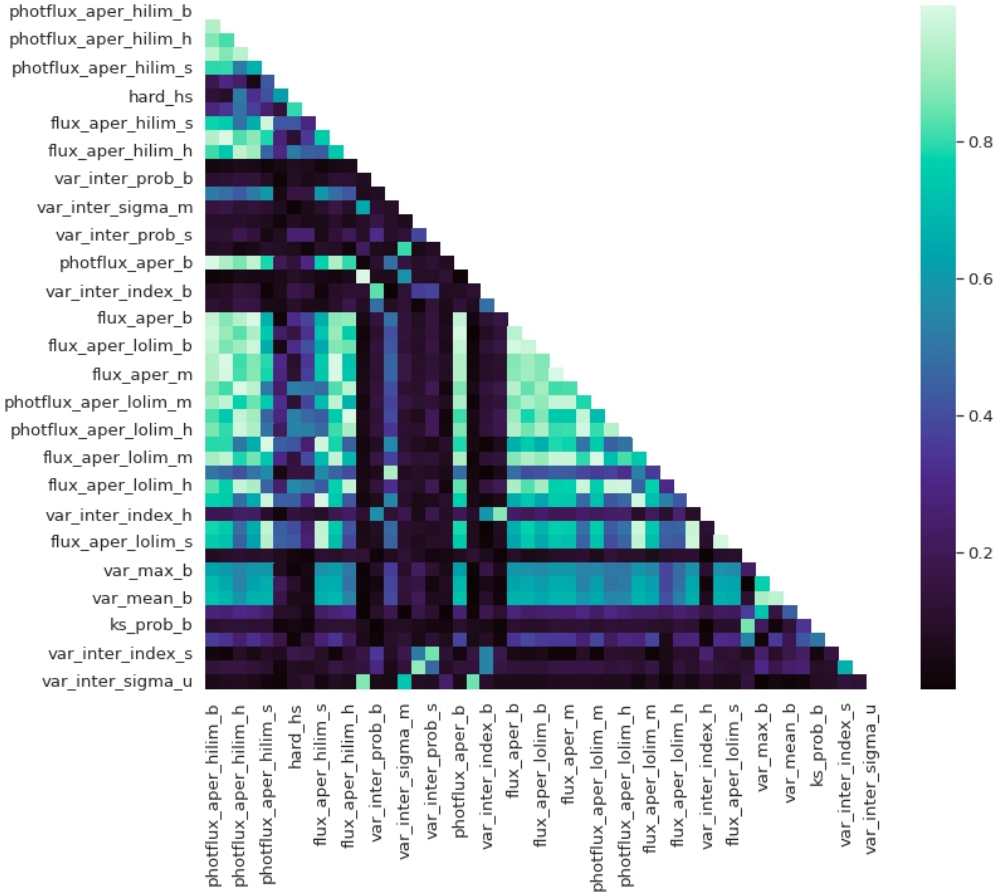




Feature optimization

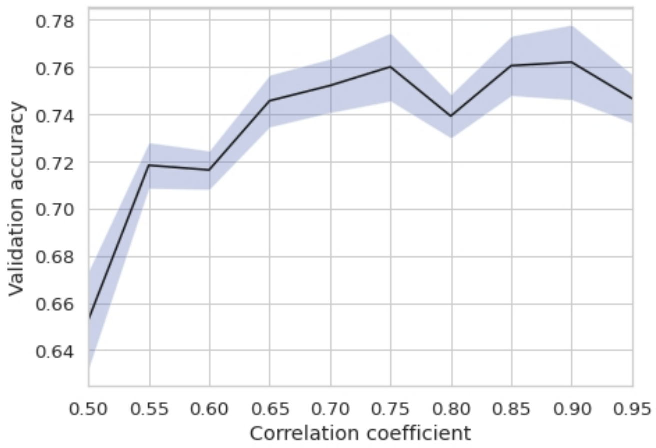
- Feature-feature correlation
- Need to remove correlated features.

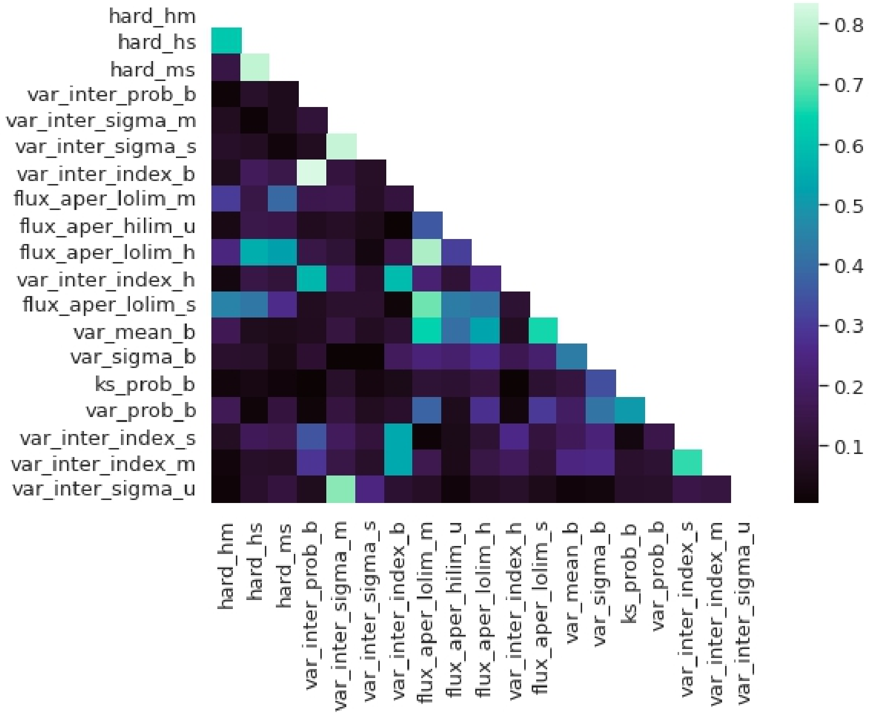




Feature optimization

- Feature-feature correlation
- Need to remove correlated features.





Feature optimization

- Feature -feature correlation
- why we need to remove correlated features
- Comparison of result -

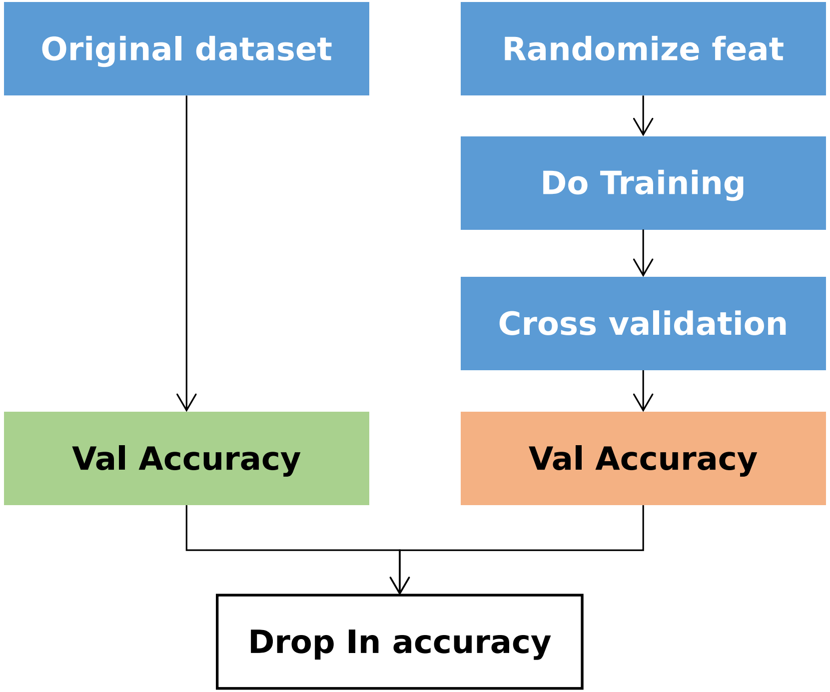
	Number of features	Mean accuracy	Std accuracy
Before removing correlated features	49	76.2	1.2
After removing correlated features	19	77.3	1.1

Feature Importance

- Contribution of each feature for classification
- Understanding physical significance
- Learn what machine has learnt.

Feature Importance

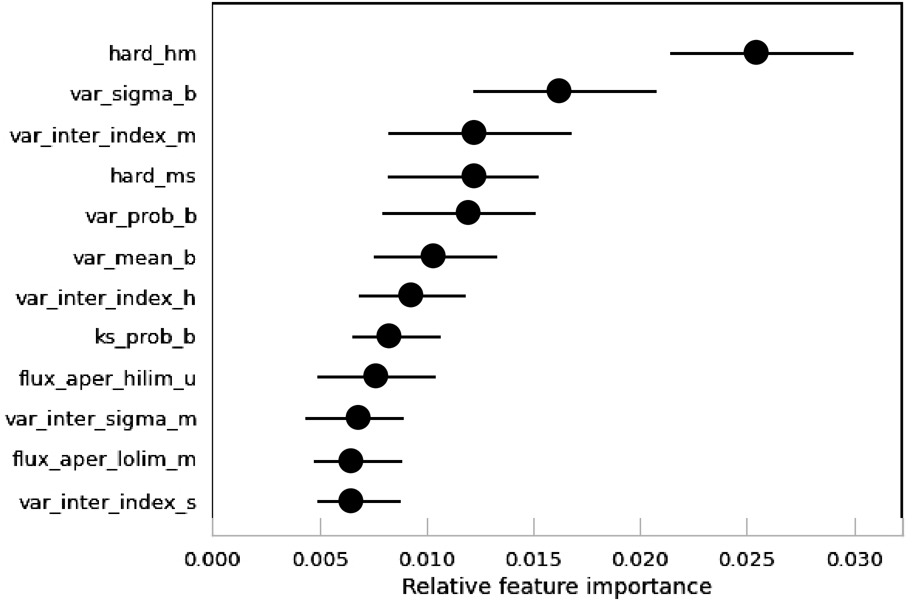
- **Algorithm**



src	properties				Class label
	flux	variability	hardness	....	
s1					
s2					
s3					
s4					
s5					
s6					
s7					
s8					
...					

Feature Importance

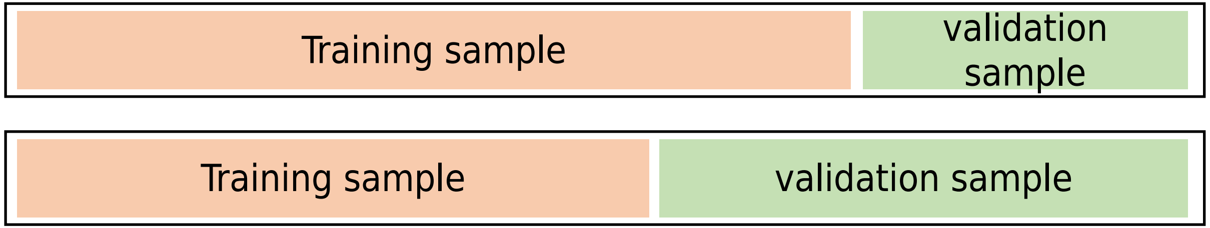
- **Result**
- **Discussion**
- **Important feature**
  - Hardness in hm band
  - Short term variability
  - Long term variability



Result significance

- **Problem with validation result**
  - small sample available for validation
    - LMXB - 17
    - CV - 55
    - MSP - 54
- **Validation result - training quality tradeoff**

- **Other method**
  - Permutation Significance



Permutation - Test

- **Null Hypothesis -**
  - No relation between features and the class label
- **p-Score**
  - Probability that accuracy on label-permuted data will be more than or equal to accuracy on original data
  - Null-hypothesis p-score ~ 0.5

S1	features s1	LMXB
S2	features s2	CV
S3	features s3	CV
S4	features s4	LMXB
S5	features s5	MSP
S6	features s6	CV
S7	features s7	MSP
S8	features s8	CV
S9	features s9	MSP
S10	features s10	LMXB

Permutation - Test : Algorithm

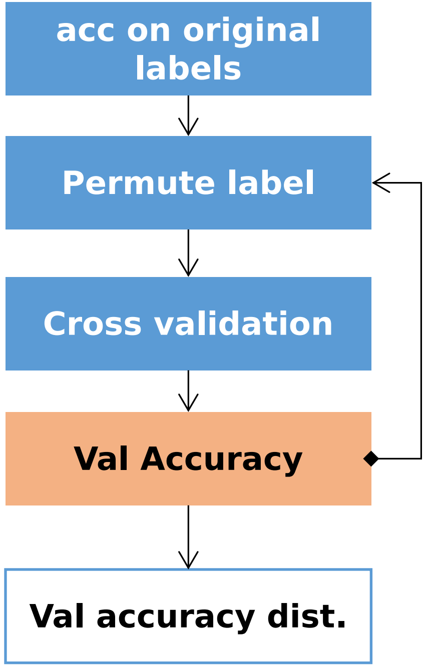
acc on original labels

↓

Permute label

S1	features s1	LMXB
S2	features s2	CV
S3	features s3	CV
S4	features s4	LMXB
S5	features s5	MSP
S6	features s6	CV
S7	features s7	MSP
S8	features s8	CV
S9	features s9	MSP
S10	features s10	LMXB

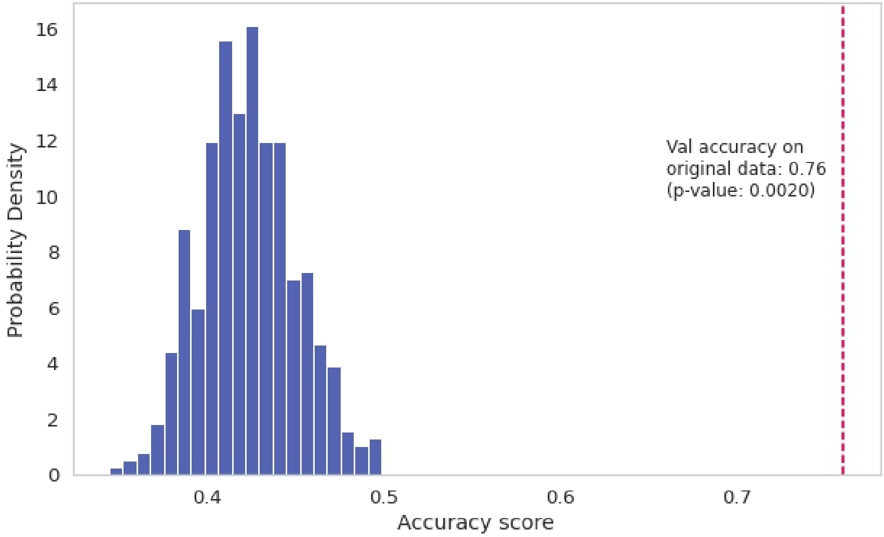
Permutation - Test



S1	features s1	MSP
S2	features s2	CV
S3	features s3	LMXB
S4	features s4	MSP
S5	features s5	CV
S6	features s6	CV
S7	features s7	LMXB
S8	features s8	CV
S9	features s9	LMXB
S10	features s10	MSP

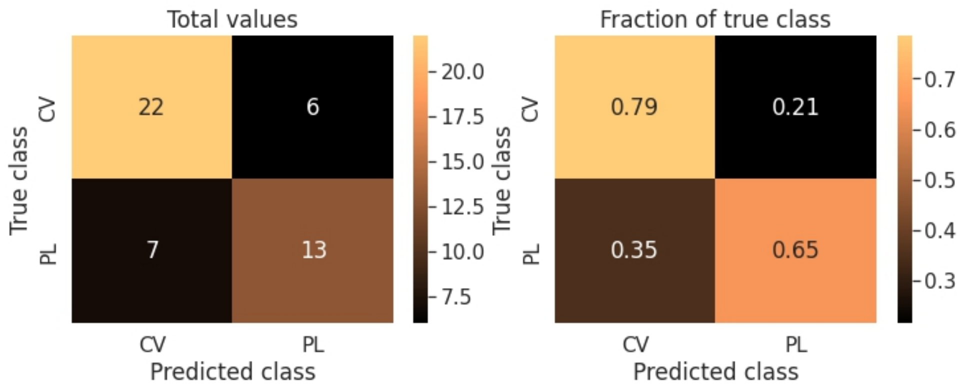
Permutation - Test : Result

- **Permutation test result**
- **p-score: 0.002**



Application on 47-TUC

	Confirmed	Candidate	In CSC After Filtering
MSP	16	11	19
CV	22	8	28
LMXB	2	3	0



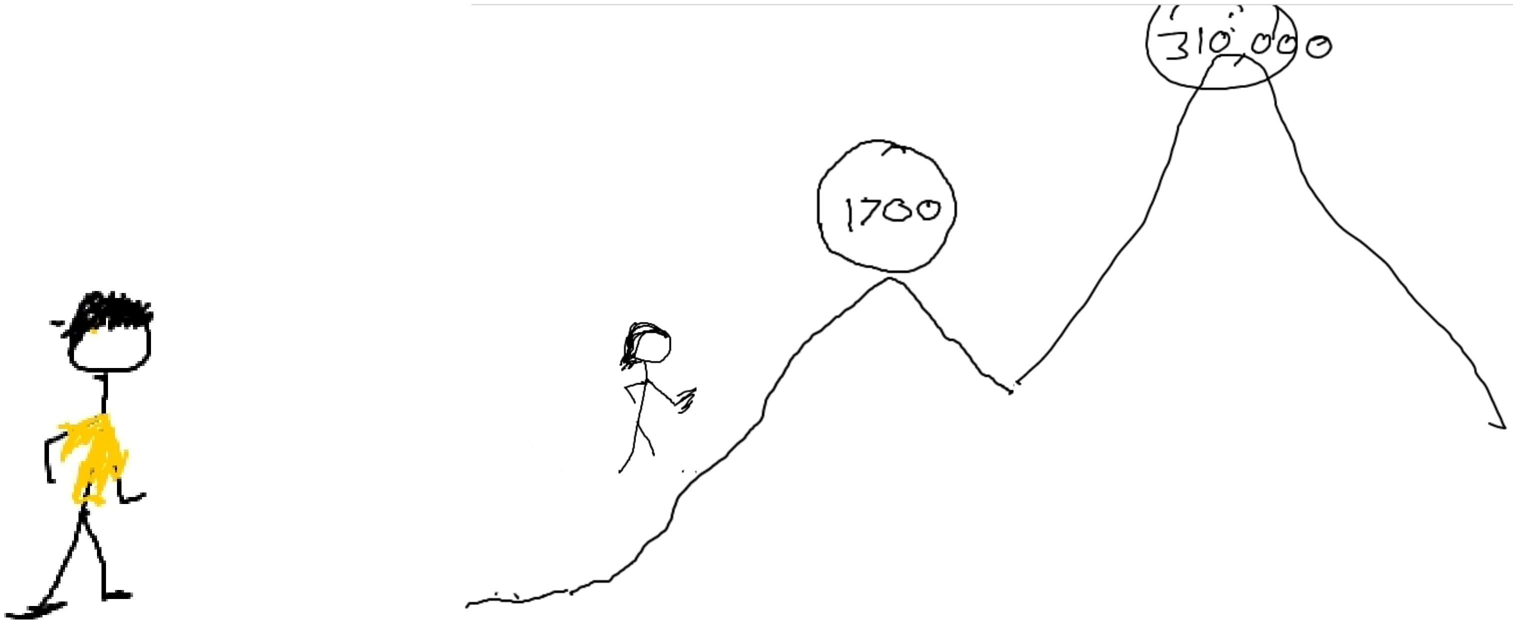
Conclusion

- Using Published catalogues , literature survey a **subset catalog of CSC** was created.
- Explored various methods of **filling in missing values**
- Imputation with RF works best
- Explored various classifier models. > RF chosen
- Classification result
  - Validation accuracy ~ 75%
  - Permutation test **p-score 0.002**
- Applied to **47-TUC**

- **Road Blocks**
  - Result not accurate enough
  - Predicted probabilities are low
  - Data sample is small

Future Plan

- **Immediate Future**
  - Gaussian Resampling of minority class
  - Upsampling of all classes
  - Deep Learning - Auto Encoder for missing value prediction
  - Cross match with NED
- **Ahead**
  - Application on other Globular clusters
  - Understanding GC dynamics
  - Adding more classes
- **Classification on entire Chandra Source Catalogue**


1700
310,000

Thank You

until next time....

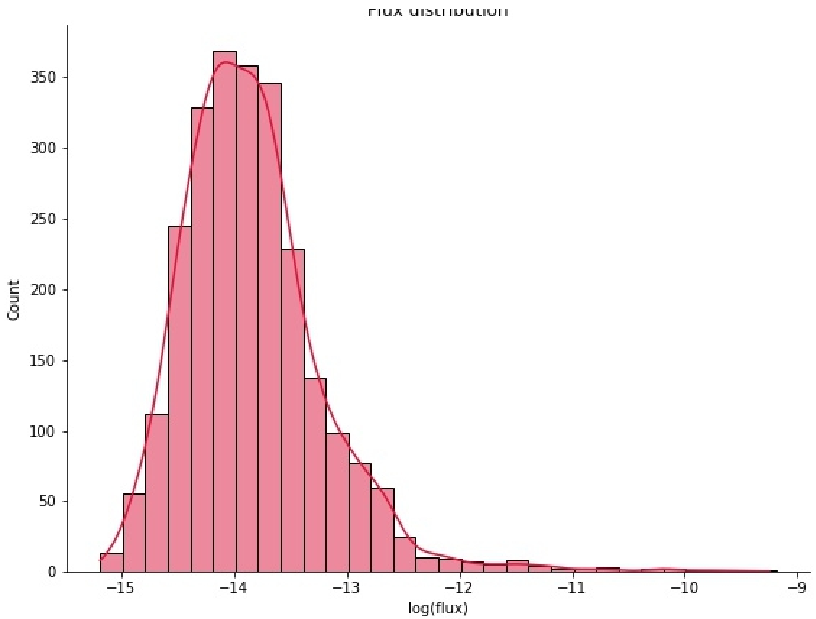
Data Imputation

- **Correlation Imputation**
  - Feature-wise imputation
  - Correlation between features
  - Fill in missing value using highest correlated feature

Limitations of SMOTE

- Can not infinitely upsample
- Available data should be able to represent parent distribution

How flux threshold is decided



Dist $\mid L_x$	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}

Data Imputation

- **Similarity Imputation using RF**

AdaBoost

- Ensamble classifier
- Can we improve further
- No further improvement.

- RF is able to capture as much information as possible

Globular Cluster

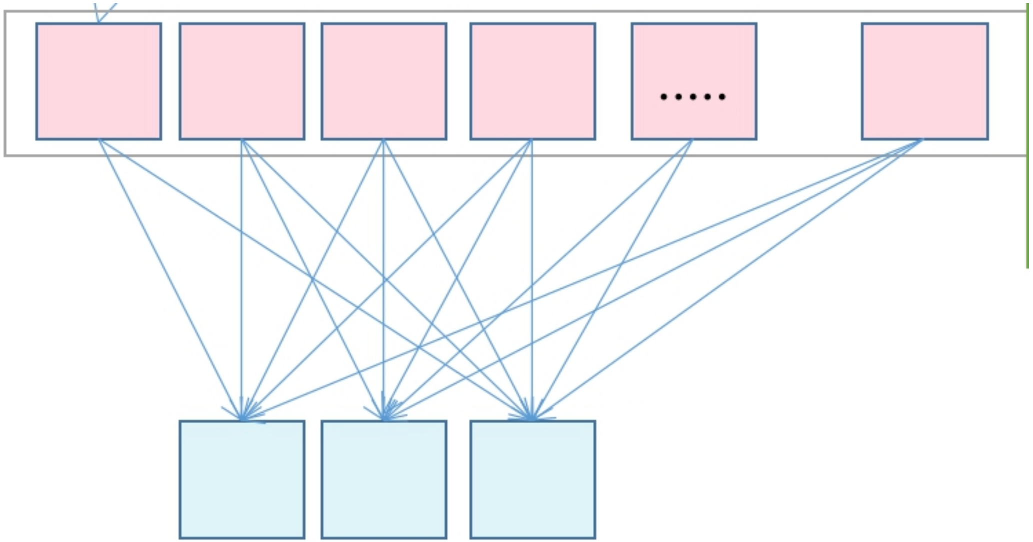
- **Improved Simulation**
  - Included XRB
  - Mean time scale matches
- **Hypothesis -**
  - XRB helps in stability against gravitational collapse

- **Dynamical evolution process with XRB[1]**
  - Core contraction phase
  - binary burning phase - collapse halts
  - After max-possible binaries formed, collapse restarts
  - binary burning phase restarts

[1] .Pooley, D. (2009). Globular cluster x-ray sources. PNAS

Classifier : Fully Connected Net

- Working
- Reason to try-on
- Caveats



Classifier : Convolution Neural Network

- **Working**
- **Reason to try-on**
  - We have correlated features
  - can take adantage if arranged feature-wise
- **Caveats**
  - Sensitive to missing values