LOFAR SOURCES IDENTIFICATION WITH MACHINE LEARNING

Credit: Cyril Tasse and the LOFAR surveys team.

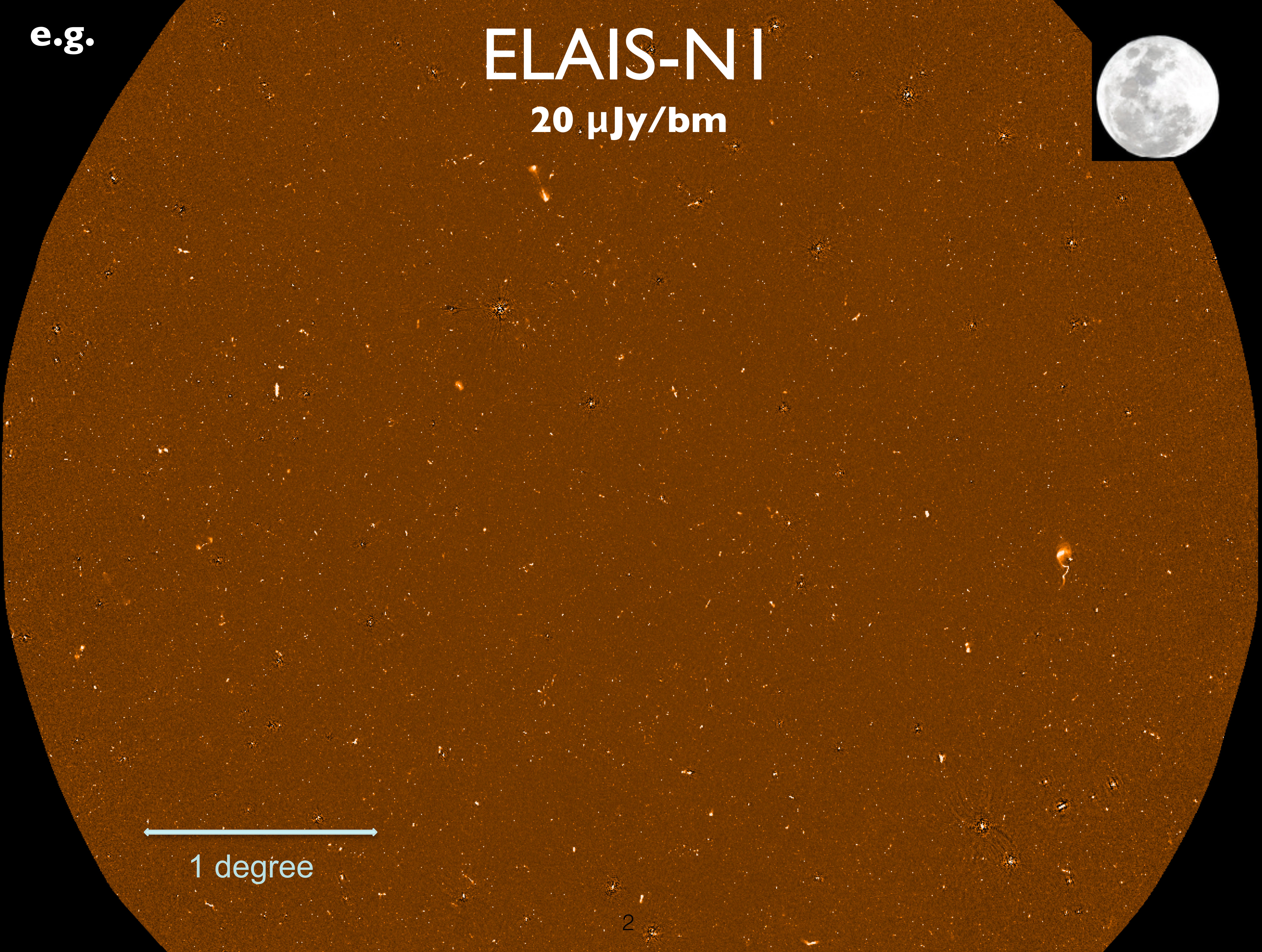
www.lofar-surveys.org

LARA ALEGRE (SHE) - PHD STUDENT
SUPERVISORS: PROF. PHILIP BEST & DR. JOSE SABATER

e.g.

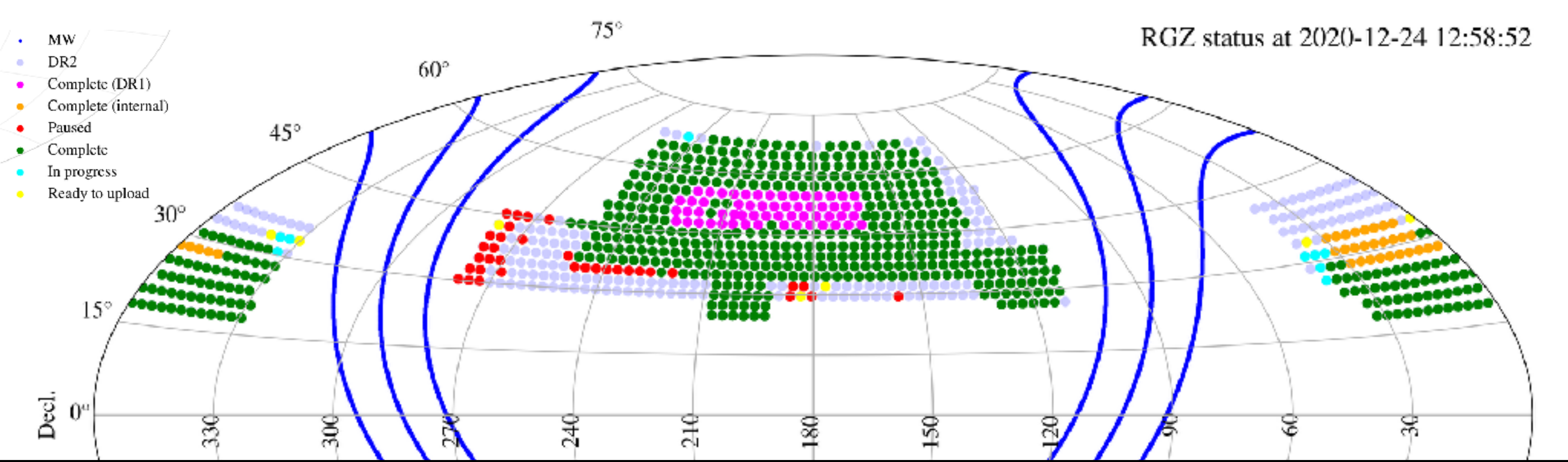
ELAIS-N1

20 μJy/bm



LoTSS

LOFAR Two-metre Sky Survey



LoTSS-DR1

- HETDEX
- 424 deg^2 (2% LoTSS)
- 58 pointings
- Radio sources: 318542
- Optical counterparts: 71% of the radio sources (PanSTARSS, WISE), Williams et al., 2019

LoTSS-DR2

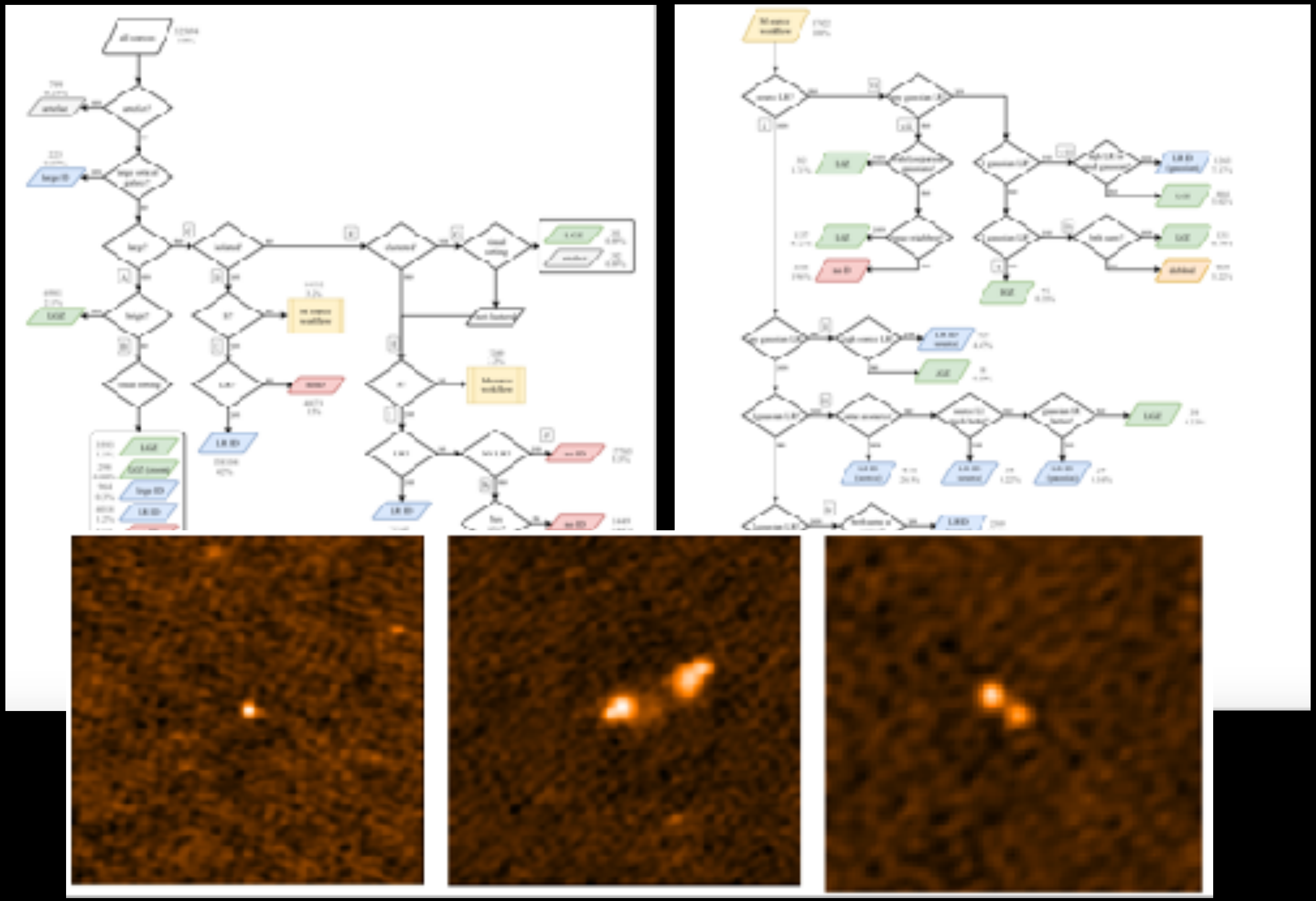
- 13h and 0h fields
- 5700 deg^2 (27% LoTSS)
- Radio sources: 4.3M
- in prep
- Status of DR2 (observations + LGZ)

LoTSS-DR1

CROSS-IDENTIFICATION

LIKELIHOOD RATIO TECHNIQUE
&
VISUAL ANALYSIS

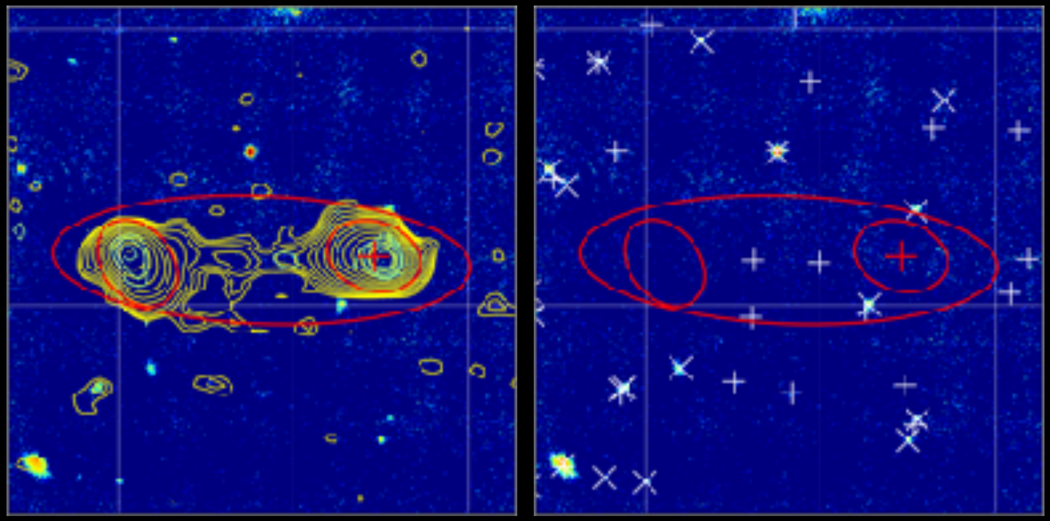
WILLIAMS ET AL., 2019 "FLOWCHART"
ONE BIG DECISION TREE



SUITABLE FOR STATISTICAL CROSS-MATCH

NOT SUITABLE FOR STATISTICAL ANALYSIS

LOFAR GALAZY ZOO



Source name ILTJ133142.18+503610.6 (RA 202.936 DEC 50.603)

PyBDSF gaussians

LOFAR radio (150 MHz)

FIRST radio (1.4-GHz)

+ WISE (IR. W1 band)

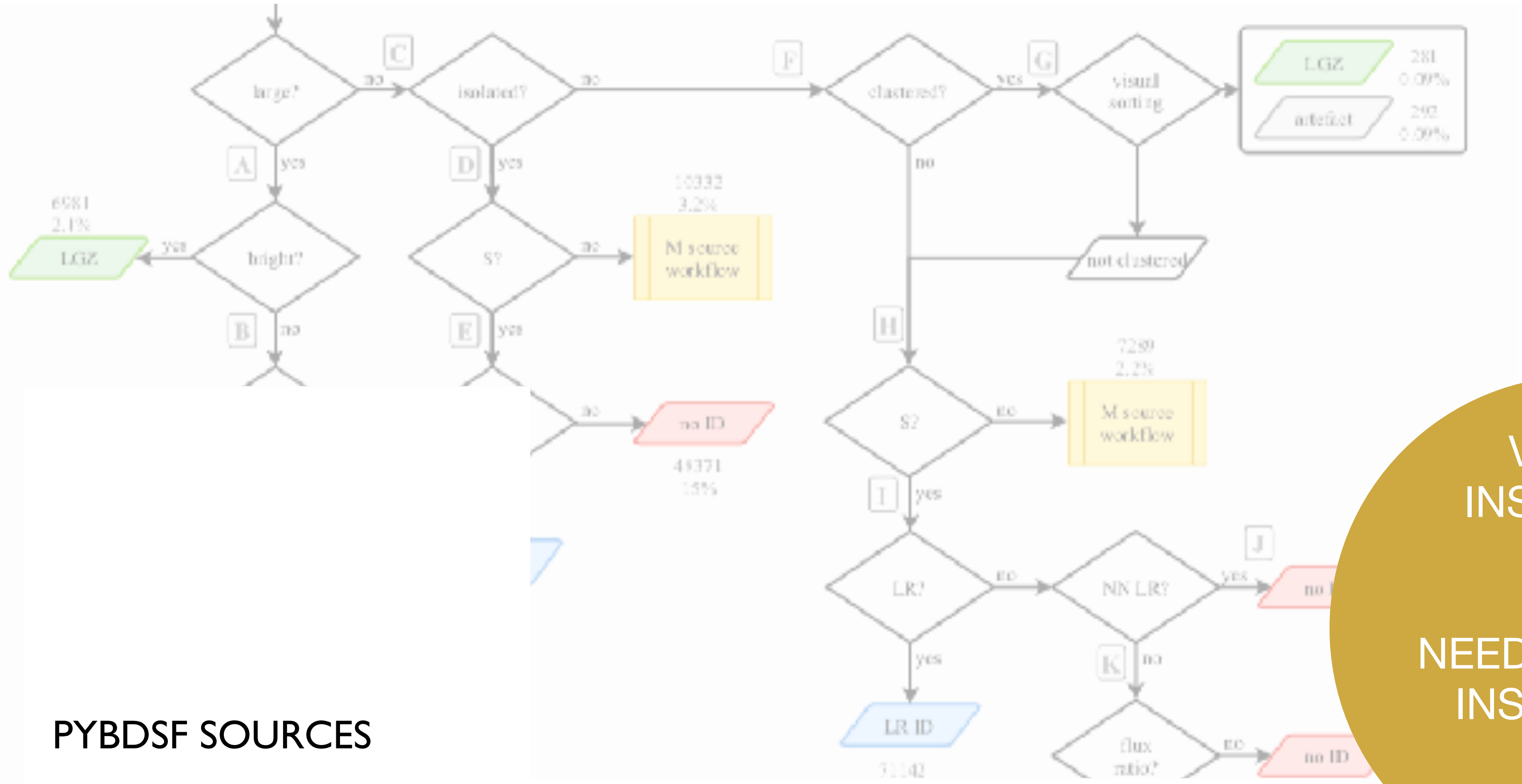
× PANSTRARRS (optical, r band)

RADIO PYBDSF SOURCES

- MULTIPLE RADIO COMPONENTS
- EXTENDED EMISSION
- BLENDED

LoTSS-DR1

Cross-identification



PYBDSF SOURCES

	Decision tree outcome	Suitable for LR	Need visual inspection	Artefacts
LR	295364	294129	1096	139
LGZ	8432	3143	5051	238
Prefilter	21099	10079	9604	1416
Artefacts	799	1	N/A	798
Total	325694	307352	15751	2591

> 99 %
~ 60%
~ 45%

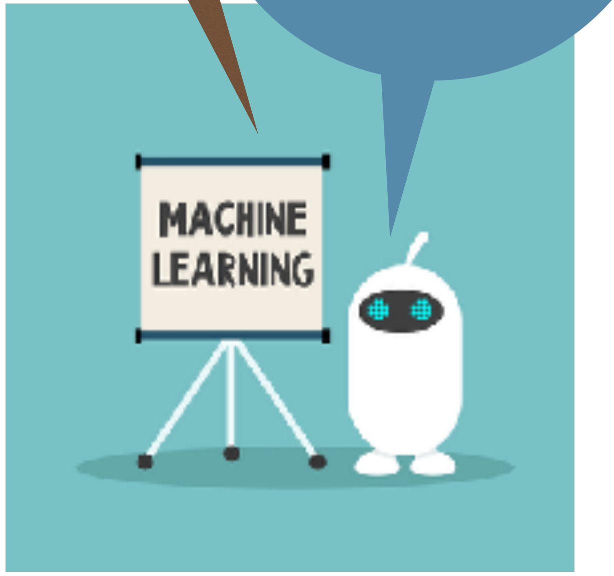
CORRECT DECISION TREE OUTCOME

VISUAL INSPECTED 29531

NEEDED VISUAL INSPECTION 15751

- KEEP VISUAL INSPECTION LOW
- KEEP NUMBER OF SOURCES WRONGLY ACCEPTED BY LR LOW

USE THE CHARACTERISTICS OF THE SOURCES AS INPUT FEATURES



MACHINE LEARNING

DATASET CREATION

BINARY CLASSIFIER

CLASSES

CLASS 1
LR

- Pybdsf sources that were not associated with other PyBDSF sources
- were not deblended
- sources for which LR gave correct optical ID (or correctly lack of ID)

CLASS 0
LGZ

- PyBDSF sources that were associated with other sources in LGZ
- deblended into more than one source
- LR obtained incorrect ID

DATASET

- Number of sources in class 0: 15751
- Number of sources in Class 1: 307352
- Exclude the artefacts: 2591

↓

CREATE A BALANCED DATASET

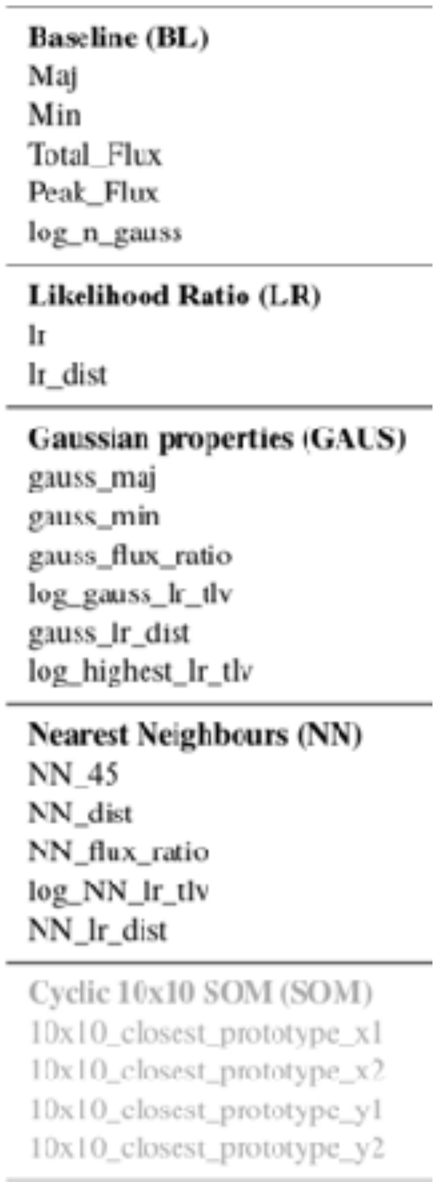
↓

DOWNSAMPLING THE MAJORITY CLASS

75% TRAIN 25% TEST

MACHINE LEARNING

FEATURES



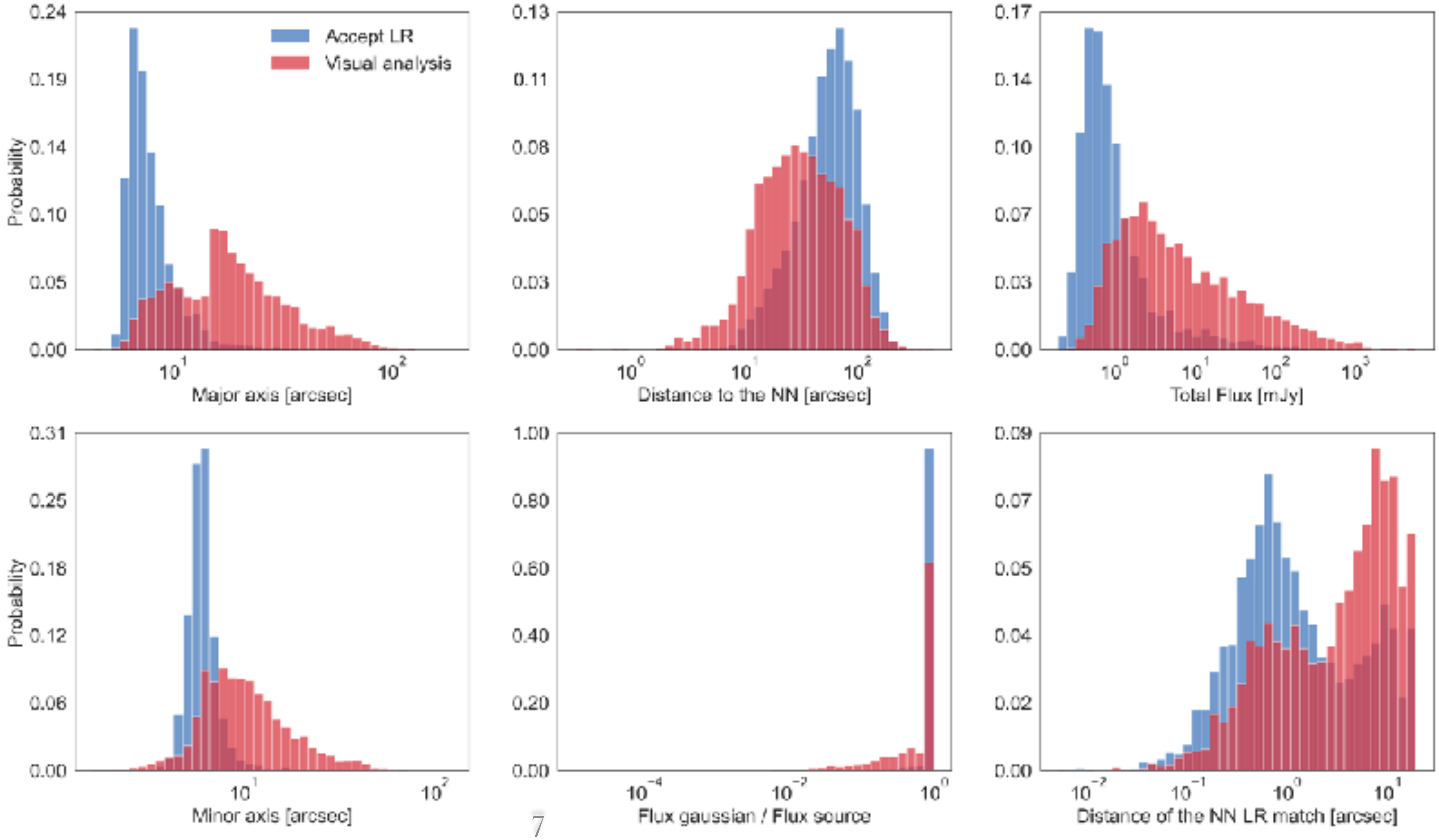
Baseline (BL)
Maj
Min
Total_Flux
Peak_Flux
log_n_gauss

Likelihood Ratio (LR)
lr
lr_dist

Gaussian properties (GAUS)
gauss_maj
gauss_min
gauss_flux_ratio
log_gauss_lr_tlv
gauss_lr_dist
log_highest_lr_tlv

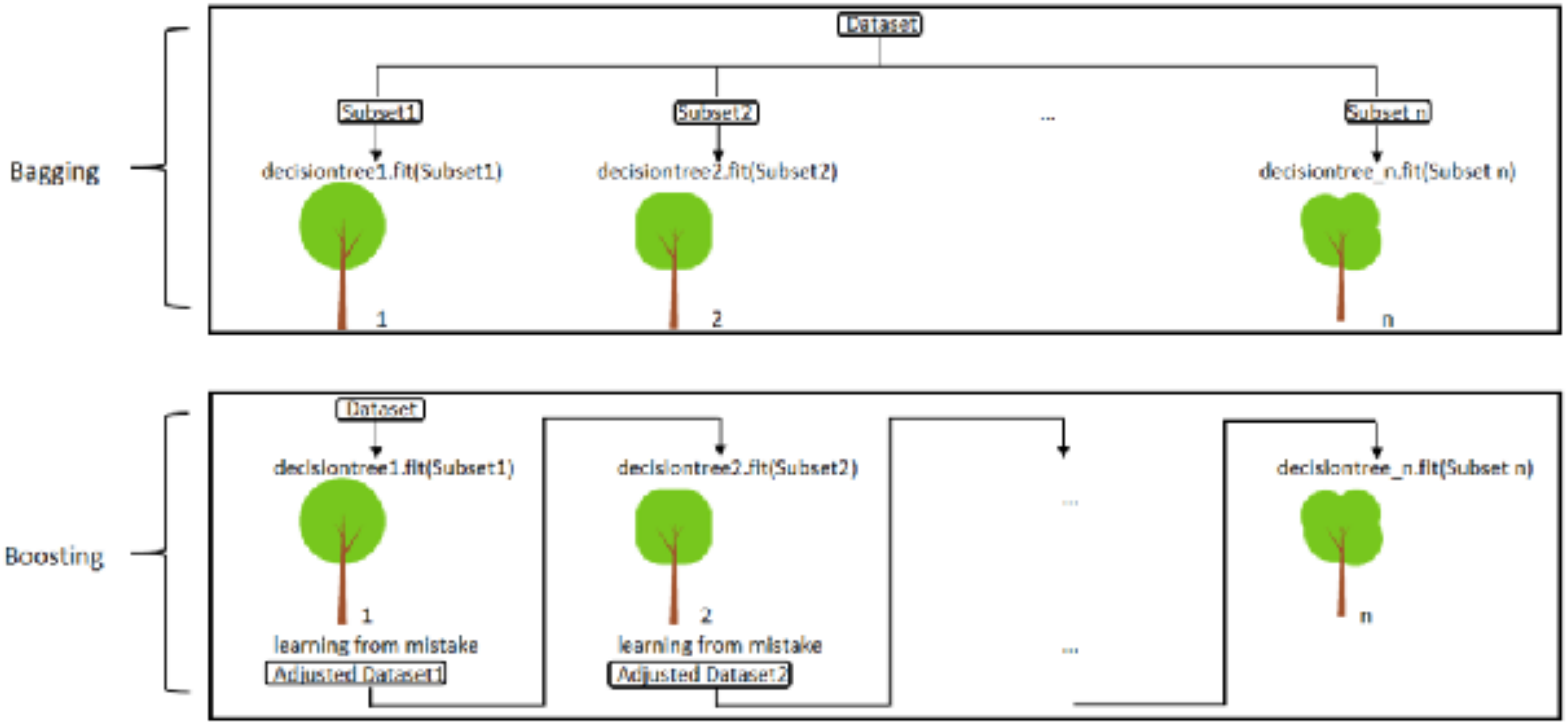
Nearest Neighbours (NN)
NN_45
NN_dist
NN_flux_ratio
log_NN_lr_tlv
NN_lr_dist

Cyclic 10x10 SOM (SOM)
10x10_closest_prototype_x1
10x10_closest_prototype_x2
10x10_closest_prototype_y1
10x10_closest_prototype_y2



METHOD - SUPERVISED ML

ENSEMBLES OF DECISION TREES



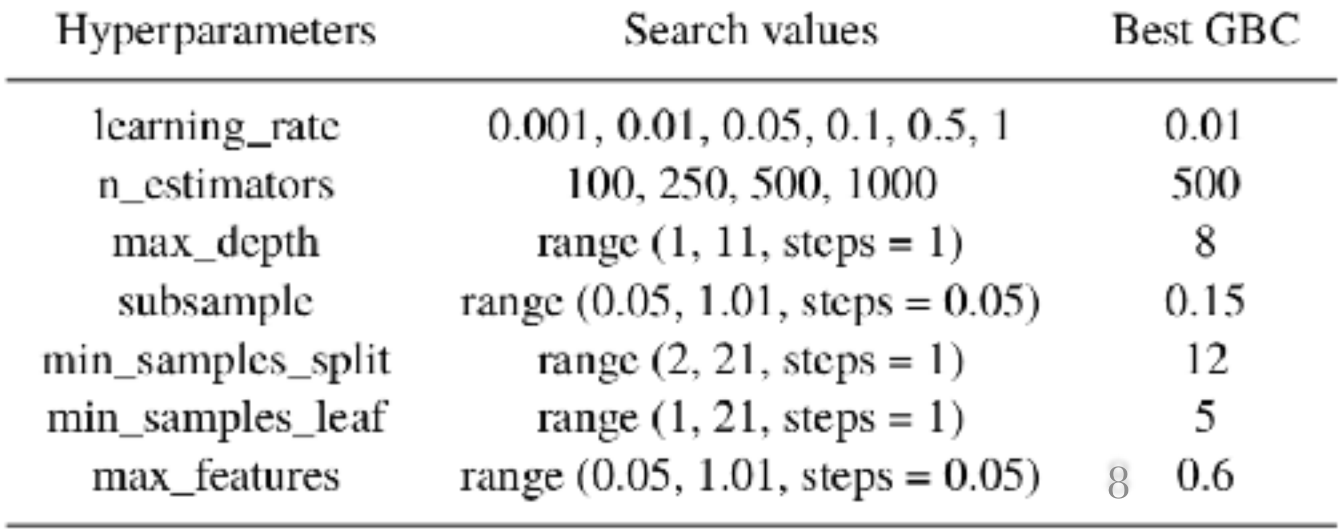
Bagging
Dataset
Subset1
Subset2
...
Subset n
decisiontree1.fit(Subset1)
decisiontree2.fit(Subset2)
decisiontree_n.fit(Subset n)
1
2
n
Boosting
Dataset
decisiontree1.fit(Subset1)
decisiontree2.fit(Subset2)
...
decisiontree_n.fit(Subset n)
1
2
n
learning from mistake
Adjusted Dataset1
learning from mistake
Adjusted Dataset2
...

MAJORITY OF THE VOTES

RANDOM FOREST

- MINIMIZATION OF TOTAL LOSS
- MORE WEIGHT TO MODELS WITH BETTER PERFORMANCE

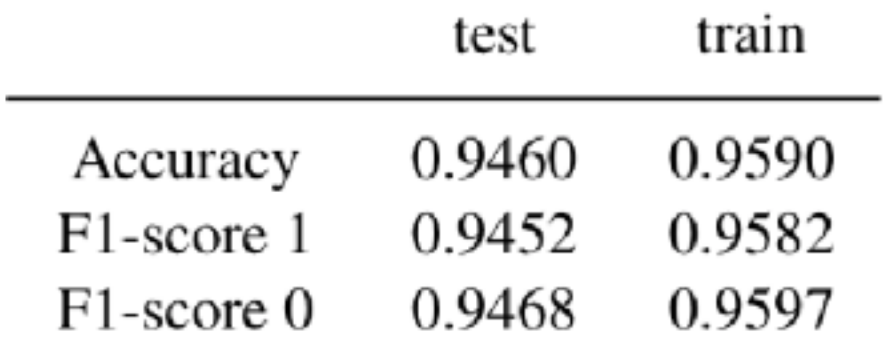
GRADIENT BOOSTING CLASSIFIER



Hyperparameters	Search values	Best GBC
learning_rate	0.001, 0.01, 0.05, 0.1, 0.5, 1	0.01
n_estimators	100, 250, 500, 1000	500
max_depth	range (1, 11, steps = 1)	8
subsample	range (0.05, 1.01, steps = 0.05)	0.15
min_samples_split	range (2, 21, steps = 1)	12
min_samples_leaf	range (1, 21, steps = 1)	5
max_features	range (0.05, 1.01, steps = 0.05)	0.6

TPOT

Results
Model performance



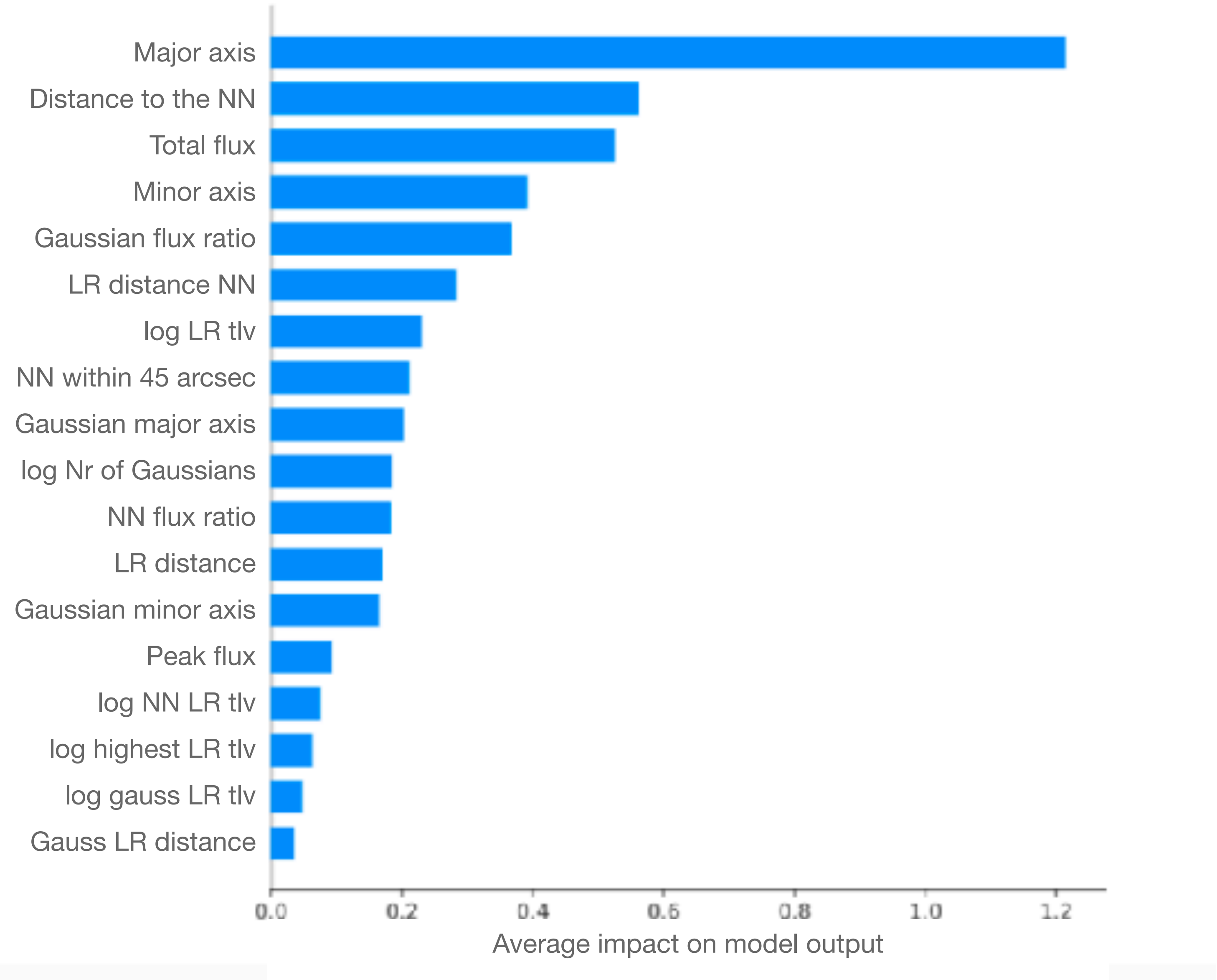
	test	train
Accuracy	0.9460	0.9590
F1-score 1	0.9452	0.9582
F1-score 0	0.9468	0.9597

- Train vs test performance: avoid overfitting
- F1-score: performance on the different classes

- 96.4% of the sources that need visual inspection are sent to LGZ (but a different component of the same source may be sent to visual inspection)

↓

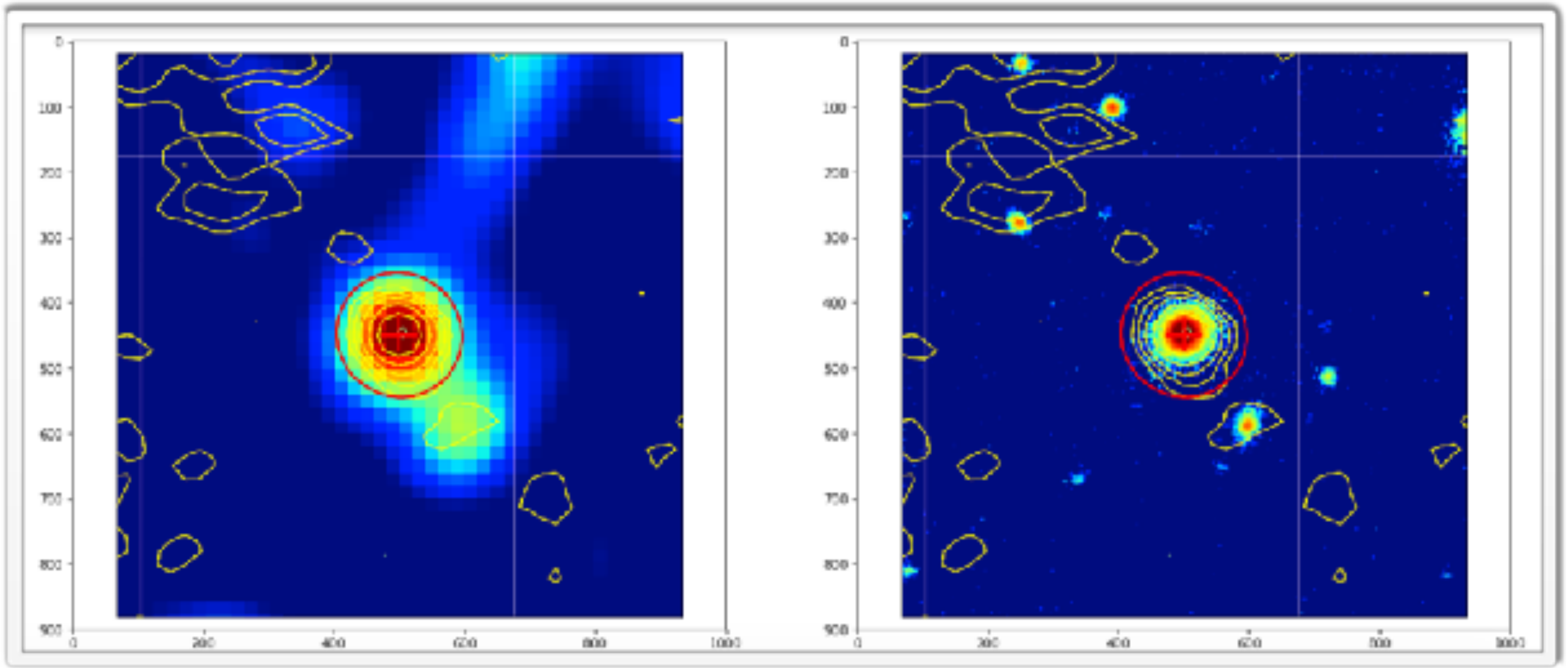
Apply corrections



Results

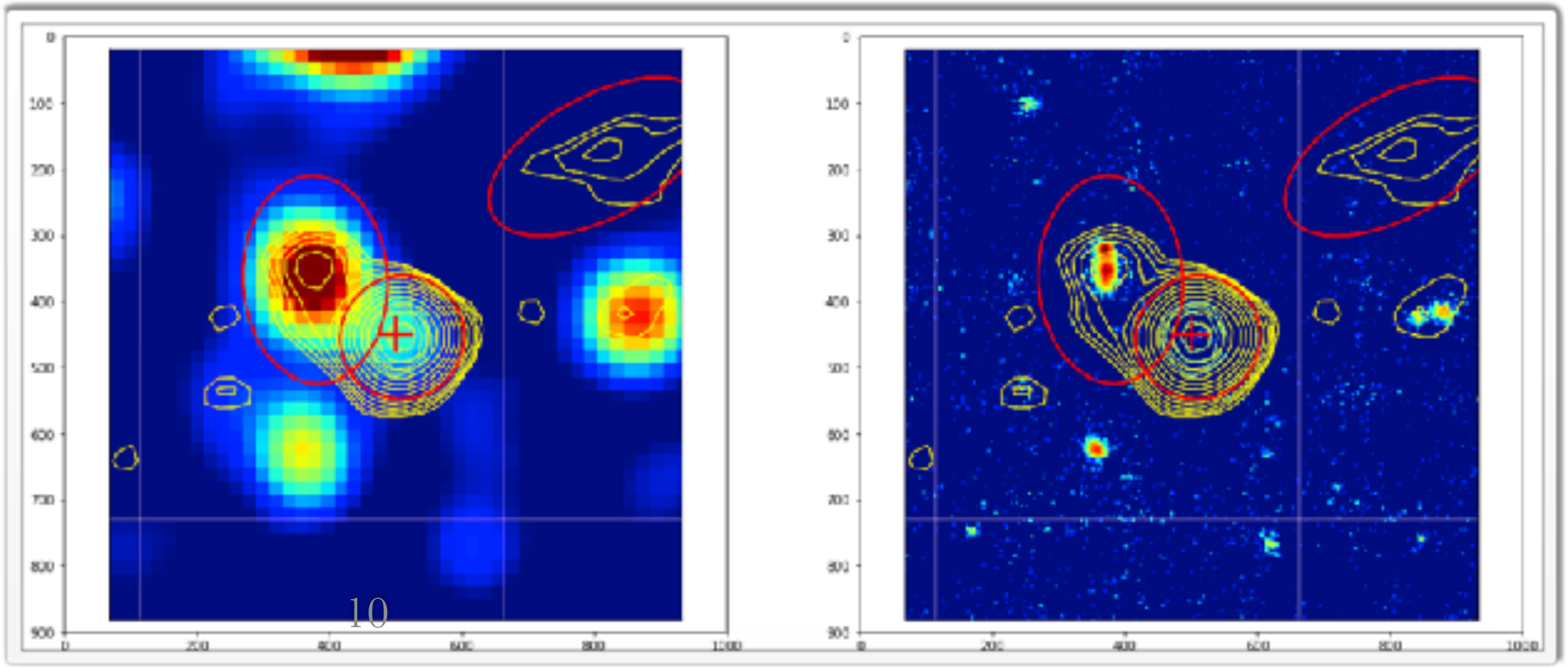
Fails and corrections

FALSE POSITIVES



MULTI COMPONENT SOURCE

ILTJ105709.24+484041.0

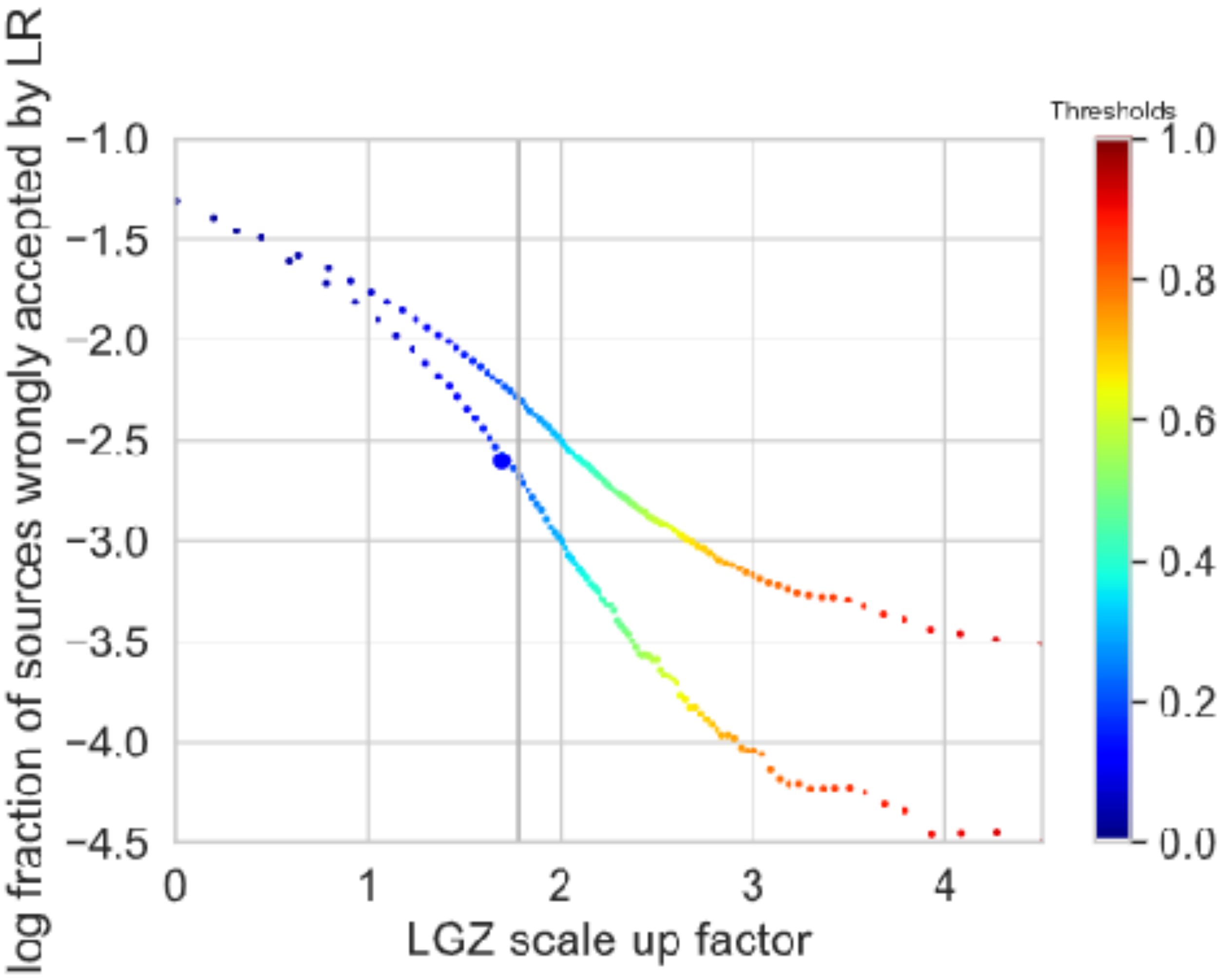


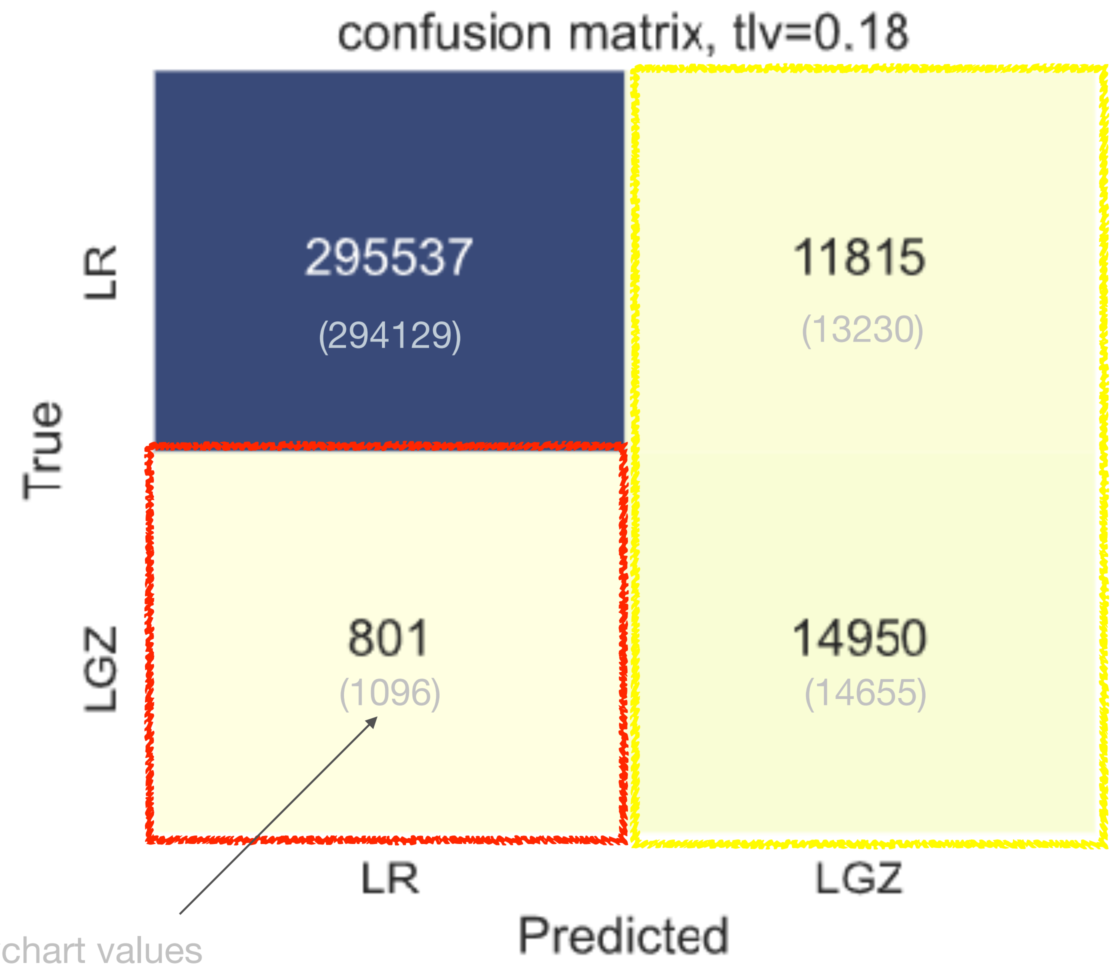
BLENDED SOURCE

ILTJ145409.19+503619.4

RESULTS

THRESHOLD VALUE AND CONFUSION MATRIX





confusion matrix, tlv=0.18

True \ Predicted	LR	LGZ
LR	295537 (294129)	11815 (13230)
LGZ	801 (1096)	14950 (14655)

Flowchart values lower limit

- Corrected and not corrected
- Threshold value of 18%

Visual inspections - 10%
False positives - 30%

SUMMARY
& WORK IN PROGRESS

ML THIS PROJECT

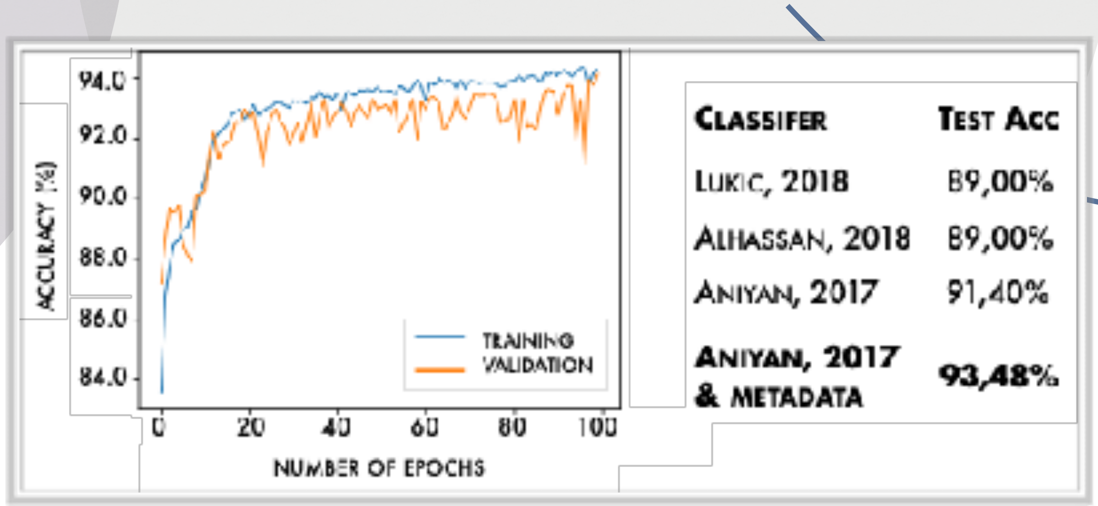
PERFORM THE SELECTION OF THE SOURCES FOR WHICH THE STATISTICAL MATCHING IS NOT RELIABLE

↓

- REDUCES THE NR OF SOURCES VISUALLY INSPECTED
- IMPROVES THE FLOWCHART PROCESS

DL IN PROGRESS

- SELECT THE MULTIPLE COMPONENTS AND BLENDED SOURCES
- ALLOWS TO IDENTIFY SINGLE COMPONENTS WERE THE LR FAILS



CLASSIFER	TEST ACC
LUKIC, 2018	89,00%
ALHASSAN, 2018	89,00%
ANIYAN, 2017	91,40%
ANIYAN, 2017 & METADATA	**93,48%**

LOFAR TEAM

CROSS-IDS FOR MULTIPLE COMPONENT SOURCES

AUTOMATE THE CROSS-MATCHING FOR THE SINGLE COMPONENT EXTENDED SOURCES

THANK YOU

LARA ALEGRE - PHD STUDENT
SUPERVISORS: PROF. PHILIP BEST & DR. JOSE SABATER