B4 Seminar #1

Nao Sakai

18.4.2018

Improving galaxy morphologies for SDSS with Deep Learning

H. Domínguez Sánchez,[1,2]⋆ M. Huertas-Company,[1,2,3] M. Bernardi[1], D. Tuccillo[2,4] and J. L. Fischer[1]

[1] *Department of Physics and Astronomy, University of Pennsylvania, 209 South 33rd Street, Philadelphia, PA 19104, USA*
[2] *LERMA, Observatoire de Paris, PSL Research University, CNRS, Sorbonne Universités, UPMC Univ. Paris 06, F-75014 Paris, France*
[3] *University of Paris Denis Diderot, University of Paris Sorbonne Cité (PSC), 75205 Paris Cedex 13, France*
[4] *MINES Paristech, PSL Research University, Centre for Mathematical Morphology, Fontainebleau, France*

ABSTRACT

We present a morphological catalogue for ∼ 670,000 galaxies in the Sloan Digital Sky Survey in two flavours: T-Type, related to the Hubble sequence, and Galaxy Zoo 2 (GZ2 hereafter) classification scheme. By combining accurate existing visual classification catalogues with machine learning, we provide the largest and most accurate morphological catalogue up to date. The classifications are obtained with Deep Learning algorithms using Convolutional Neural Networks (CNNs).

We use two visual classification catalogues, GZ2 and Nair & Abraham (2010), for training CNNs with colour images in order to obtain T-Types and a series of GZ2 type questions (disk/features, edge-on galaxies, bar signature, bulge prominence, roundness and mergers). We also provide an additional probability enabling a separation between pure elliptical (E) from S0, where the T-Type model is not so efficient. For the T-Type, our results show smaller offset and scatter than previous models trained with support vector machines. For the GZ2 type questions, our models have large accuracy (> 97%), precision and recall values (> 90%) when applied to a test sample with the same characteristics as the one used for training. The catalogue is publicly released with the paper.

Key words: Galaxies – Morphology – Machine learning

0.Abstruct

This Work

- a morphological catalogue for ~670,000 galaxies in SDSS

  the largest and most accurate morphological catalogue up to date
- better T-Type classification
- separation between pure E from S0

Method

Deep Learning - Convolutional Neural Network(CNNs)

Data Sets for Training

Galaxy Zoo 2 (GZ2)
Nair & Abraham 2010 (N10)

Classification

GZ2

T-Type to the Hubble sequence

1.Introduction

Motivation

morphology is related to the physical property of the galaxy

↓

accurate morphological classification for large samples

- time consuming
- not obvious   i.g. Galaxy Zoo

↓

Deep Learning

Dieleman et al. 2015 (D15)

SDSS DR7 Main Galaxy Sample   reproduce the GZ2

Problem

galaxies with uncertain classifications used for training

This Time

~ improved version ~

galaxies with robust GZ2 classification used for training

simplify the galaxy decision tree

complement the GZ2 classification scheme with a T-Type

2.Data Sets

for Training

GZ2

N10 (visual classifications)

- T-Type classification
- separate pure E from S0

for Testing

GZ2

N10 (visual classifications)

Huertas-Company et al. 2011 T-Type

Cheng et al. 2011 ETGs and spiral galaxies

Parent Sample of the Catalogue Presented in This Work

Meert et al. 2015, 2016 ~670,000 galaxies SDSS DR7

3.Deep Learning Model

Architecture

down-sample

SDSS DR7 424x424 pixels → 69x69x3

reduce computing time
avoid overfitting

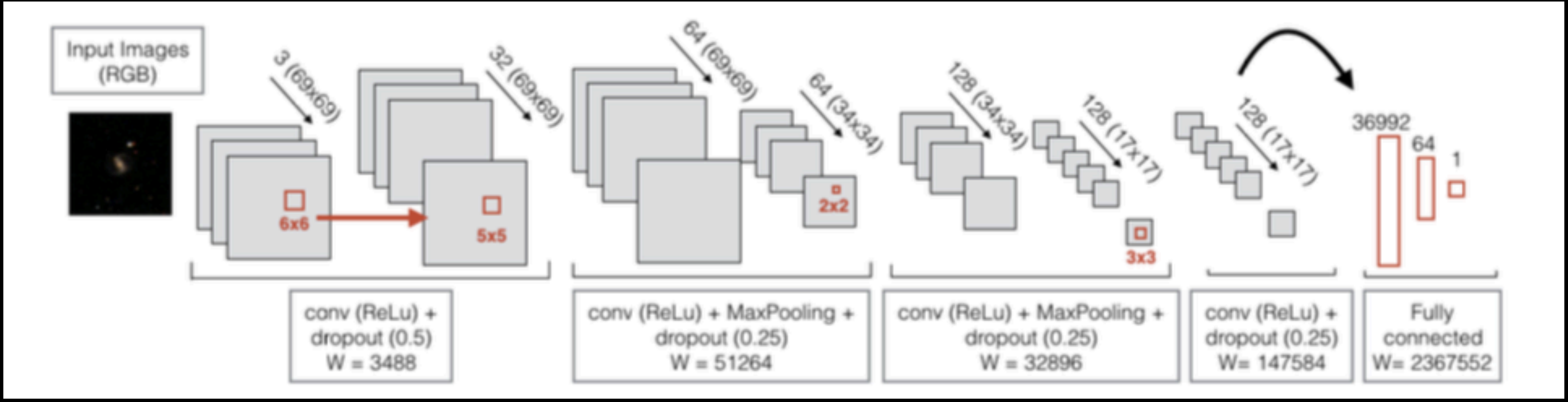


figure.1 network architecture

GZ2: binary classification mode
T-Type: regression mode

4.GZ2 based models

Questions binary classification mode

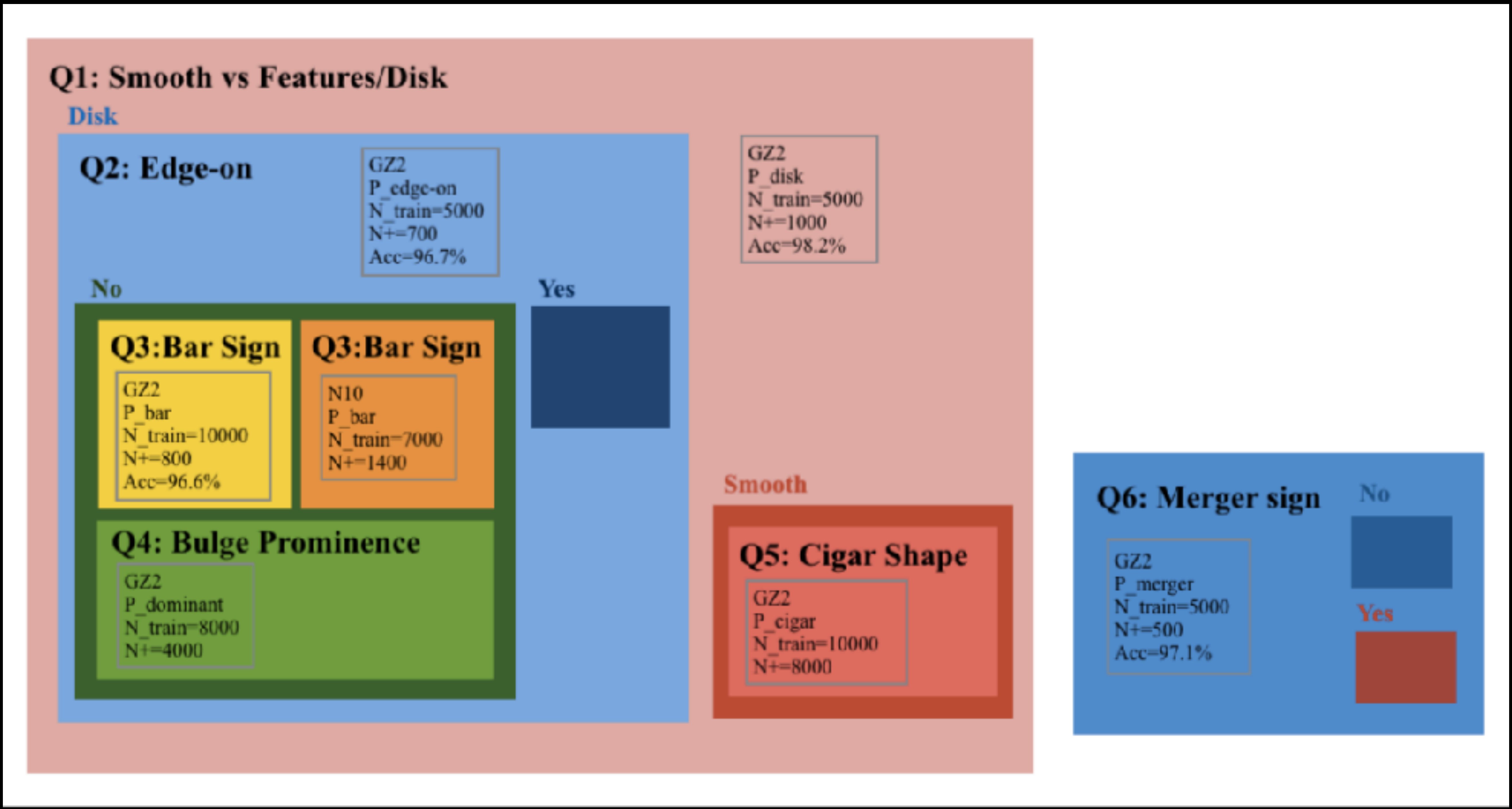


figure.3 decision tree

4.GZ2 based models

Certain Galaxies for Training

only use certain galaxies for training

Question	Meaning	N_{votes}	$N_{certain}$	N_{pos}
Q1	Disk/Features	239728 (99%)	134475 (56%)	28513 (21%)
Q2	Edge-on disk	151560 (63%)	123201 (81%)	17631 (14%)
Q3	Bar sign	117262 (48%)	76746 (65%)	6595 (8%)
Q4	Bulge prominence	117245 (49%)	49345 (42%)	27185 (55%)
Q5	Cigar shape	180223 (75%)	124610 (70%)	28230 (23%)
Q6	Merger signature	239669 (99%)	110079 (46%)	1399 (1%)

table.1 questions in each tier

$P > 0.8$ or $P < 0.2$

4.GZ2 based models

Test

$$Acc = \frac{TP + TN}{(P + N)}$$

$$Prec = \frac{TP}{TP + FP};\quad R = \frac{TP}{TP + FN} = TPR$$

Question	Meaning	P_{thr}	TPR	Prec.	Acc.
		0.2	0.97	0.91	
Q1	Disk/Features	0.5	0.95	0.96	0.98
		0.8	0.90	0.99	
		0.2	1.00	0.67	
Q2	Edge-on	0.5	0.99	0.83	0.97
		0.8	0.92	0.95	
		0.2	0.93	0.48	
Q3	Bar sign	0.5	0.79	0.80	0.97
		0.8	0.58	0.92	
		0.2	0.98	0.54	
Q6	Merger signature	0.5	0.96	0.82	0.97
		0.8	0.90	0.97	

table.2 precision and TPR value for different P_{thr}

TP: true positive

FP: false positive

FN: false negative

P_{thr}: threshold

user can optimize this value

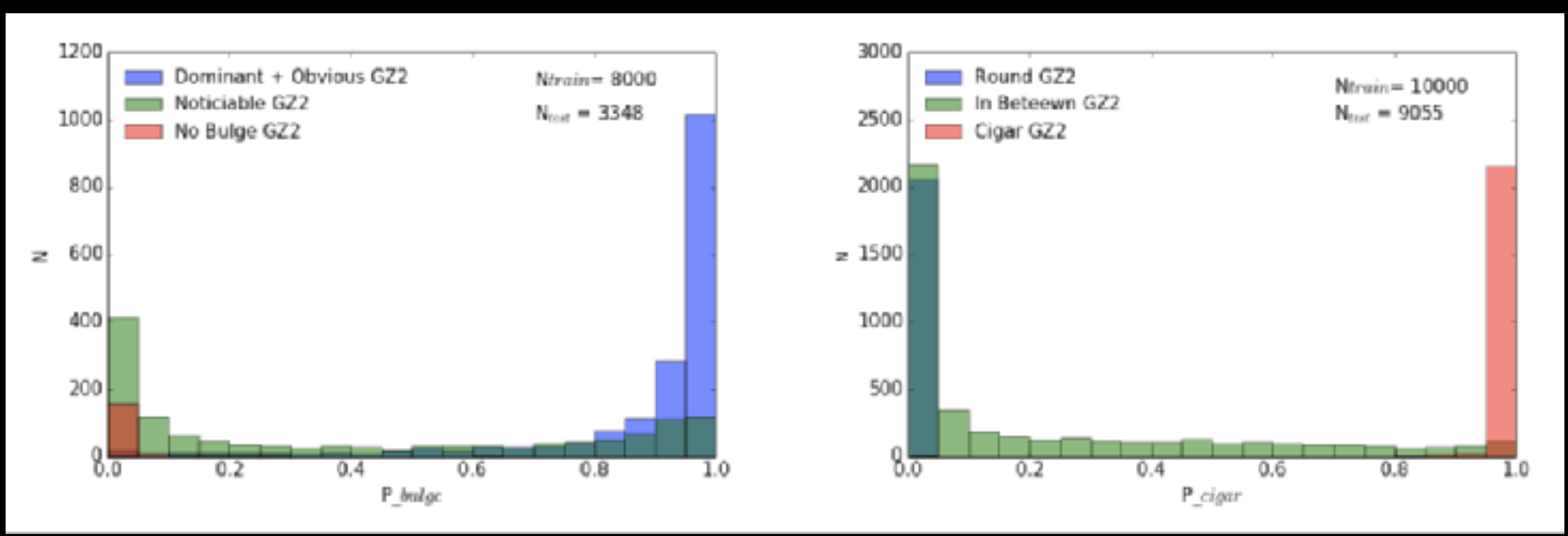


figure.10 probability distribution obtained by applying the model to a sample well classified

5.N10 based models

N10 very detailed morphological catalogue

T-Type regression mode

T-Type < 0 ETGs
T-Type = 0 S0
T-Type > 0 spiral galaxies
T-Type = 10 irregular galaxies

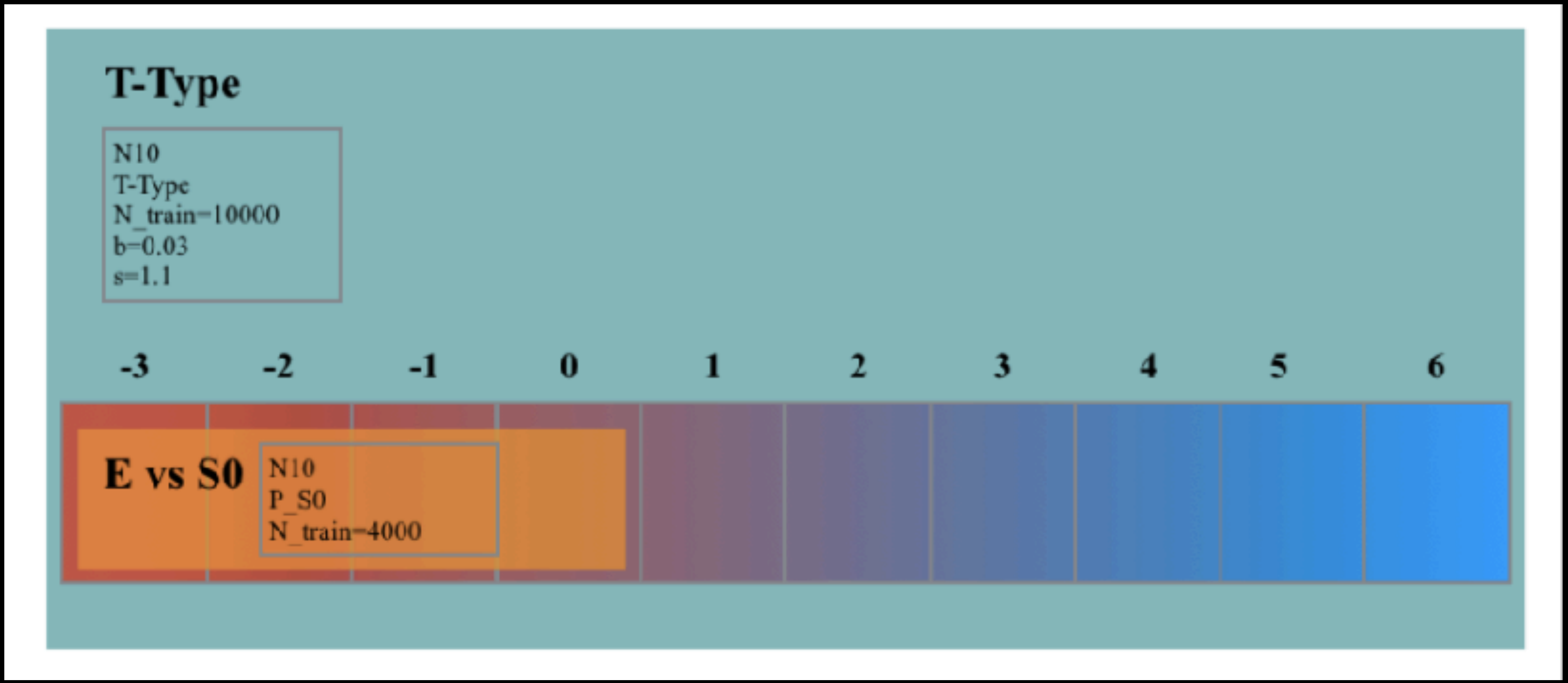


figure.13 T-Type questions scheme

5.N10 based models

E versus S0

apply 681 N10 galaxies not used for training

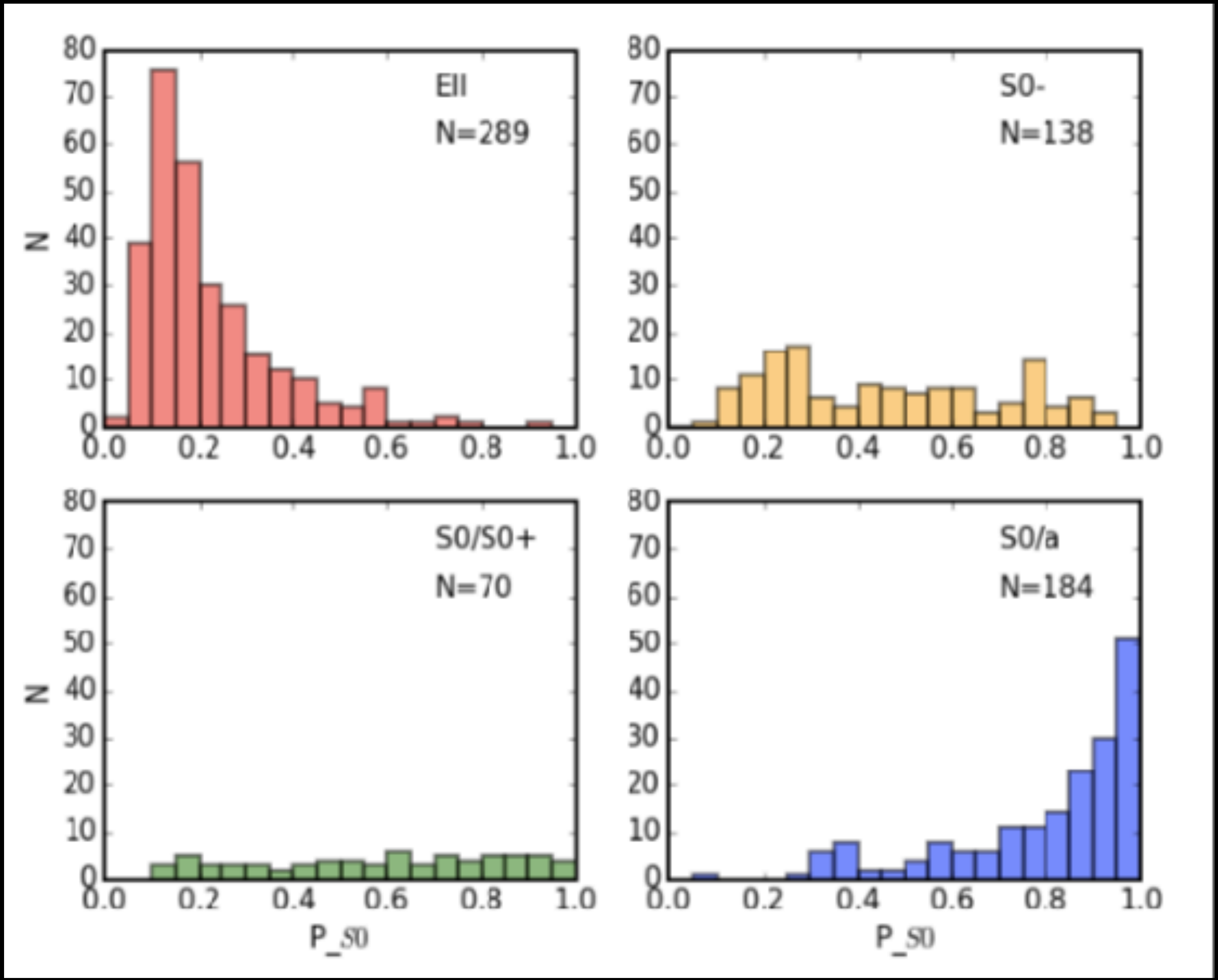


figure.14 probability distribution of being S0 rather than E

training N = 4000

positive

-3 <= T-Type <= 0

negative

T-Tye = -5

better performance than Cheng et al. 2011

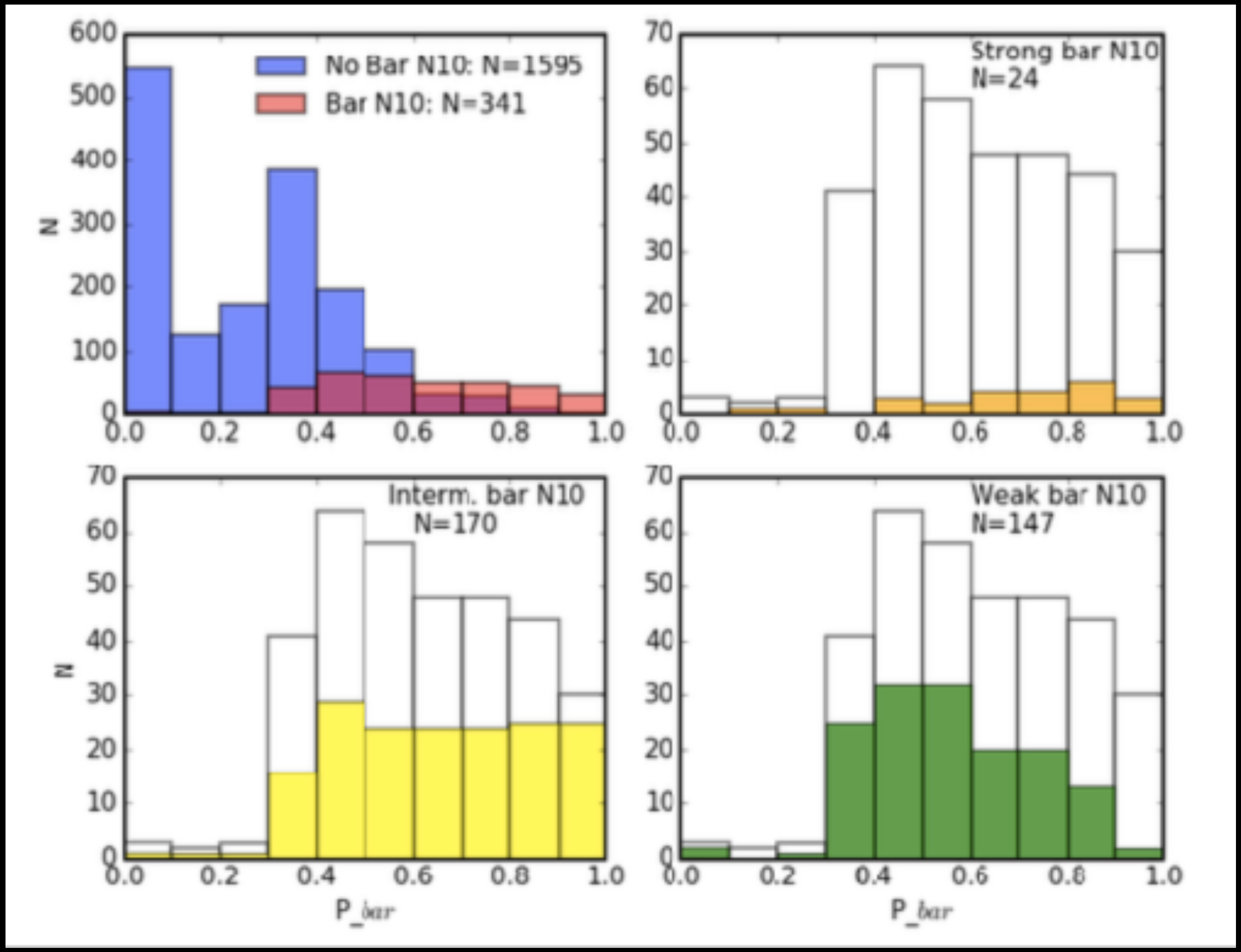
5.N10 based models

Barred Galaxies

alternative model to the GZ2 training N = 7000

the model trained with GZ2 is worst for barred galaxies

only exist (strong) or not → strong, intermediate, week



apply 1595 unbarred
314 barred galaxies
not used for training

figure.15 probability distribution of having bar signature

6.Details

Content

Col.	Name	Meaning	Train sample
1	dr7objid	SDSS ID	
2	galcount	Meert15 ID	
3	P_{disk}	Prob. features/disk	GZ2
4	$P_{edge-on}$	Prob. edge on	GZ2
5	$P_{bar-GZ2}$	Prob. bar signature	GZ2
6	$P_{bar-N10}$	Prob. bar signature	N10
7	P_{merg}	Prob. merger	GZ2
8	P_{bulge}	Prob. bulge prominence	GZ2
9	P_{cigar}	Prob. cigar shaped	GZ2
10	T-Type	T-Type	N10
11	P_{S0}	Prob. S0 vs E	N10

table.3 content of the catalogue released

6.Details
Unambiguous Classification

successful

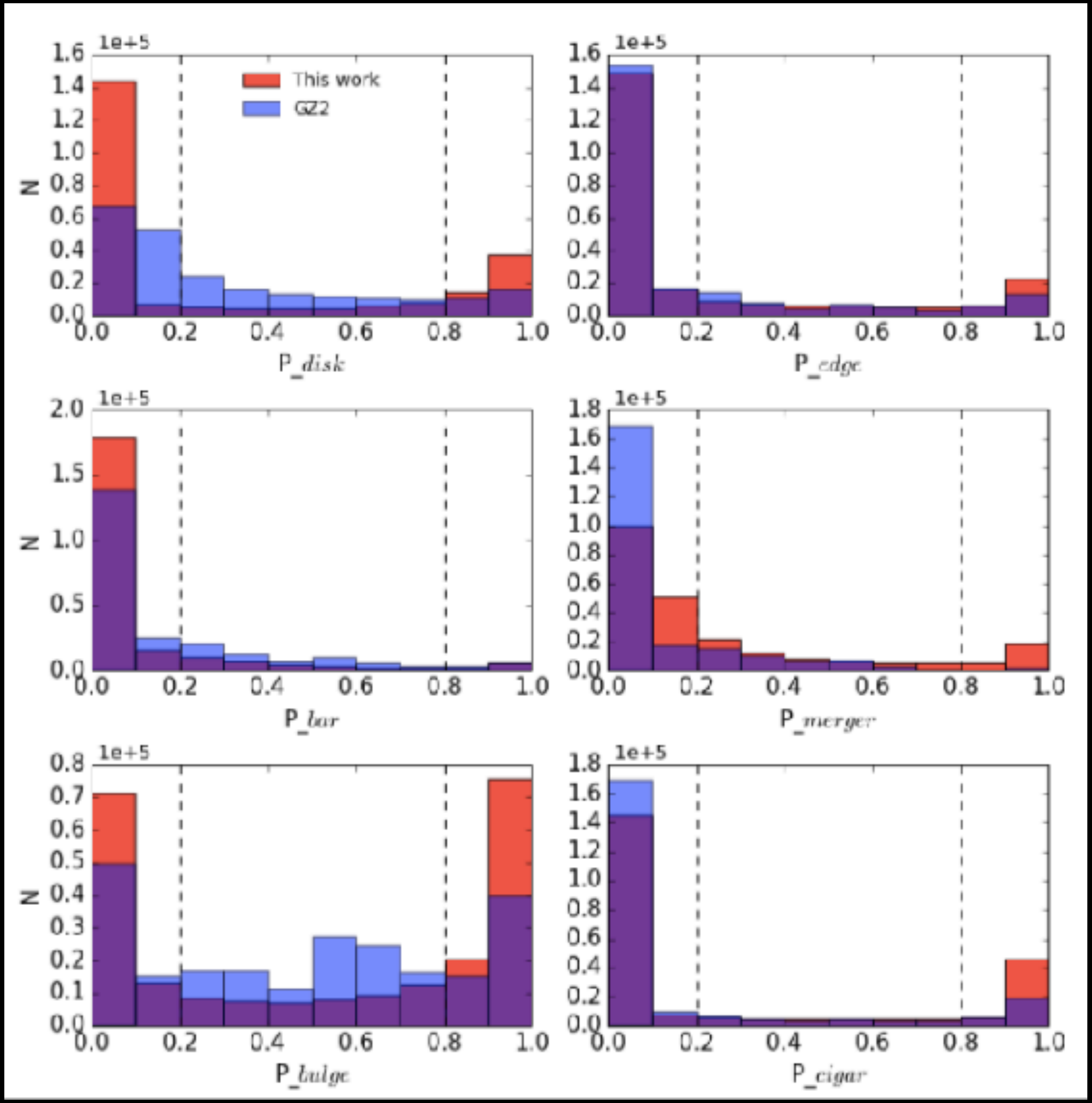


figure.17 probability distribution

different approach

$P_{dominant} + P_{obvious}$

6.Details

Correlation with Other Morphological Parameters

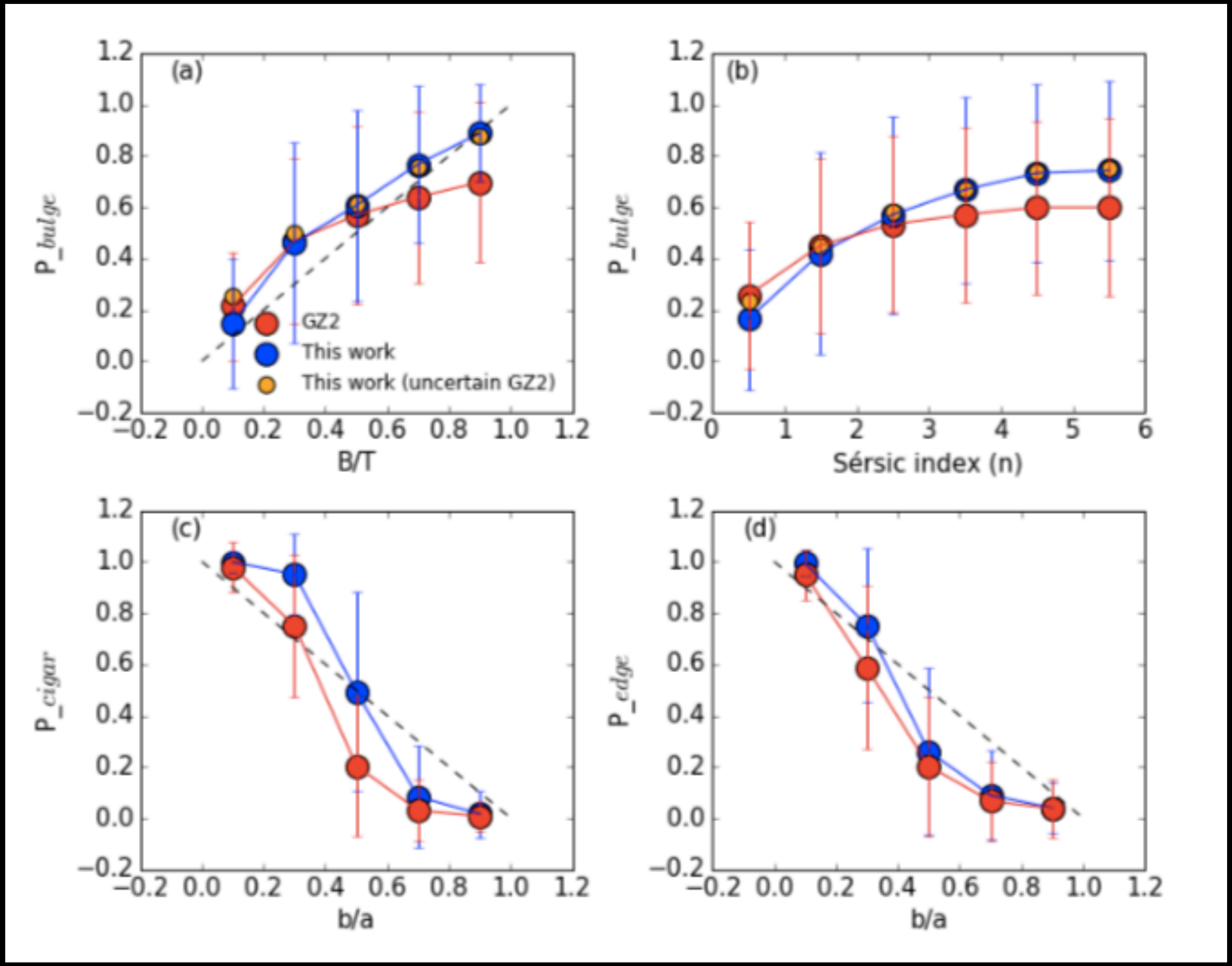


figure.18 mean probability values

6.Details

Visual Inspection

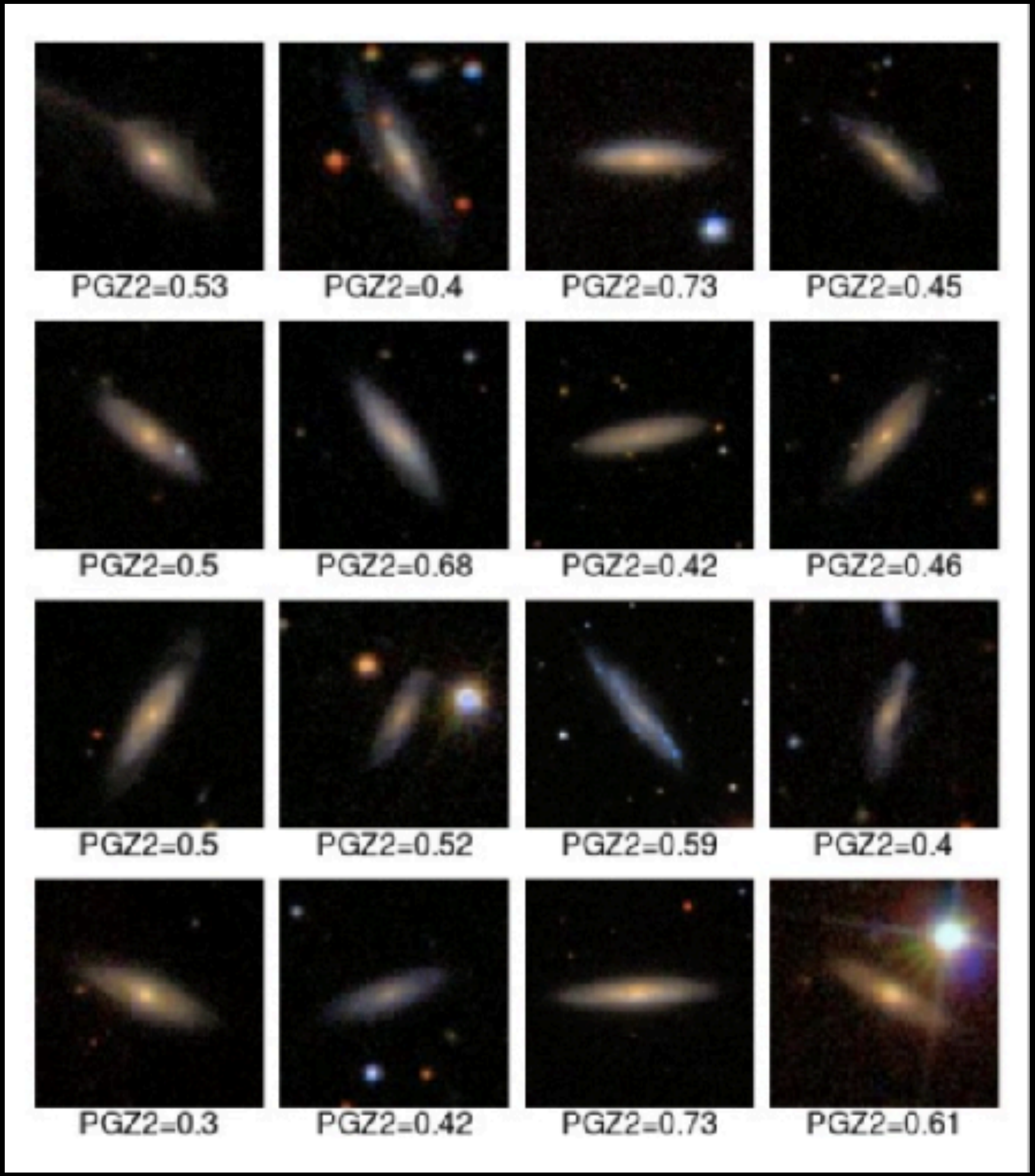


figure.20 edge on galaxies certain in this work though uncertain in GZ2

no “true reference” catalogue

7.Summaries

Improve D15

independently train each question from the GZ2 scheme

use only certain galaxies for training

binary classification mode

↓

670,722 galaxies
large accuracy
unambiguous classification (disk/features)

Complement the GZ2 Type Classification

T-Type ~50 times larger than N10 large accuracy

separation between E from S0 large accuracy

bar classification large accuracy

Forthcoming Work

apply the models to other SDSS samples
to complement the morphological classification catalogue