Main Astronomical Observatory
of NAS of Ukraine

Machine learning technique for morphological classification of galaxies from the SDSS

Dobrycheva D.V., Vavilova I.B., Melnyk O.V., Elyiv A.A.

The aim and tasks

Aim: To apply the machine learning technique for morphological classification of galaxies from the SDSS DR5 and DR9

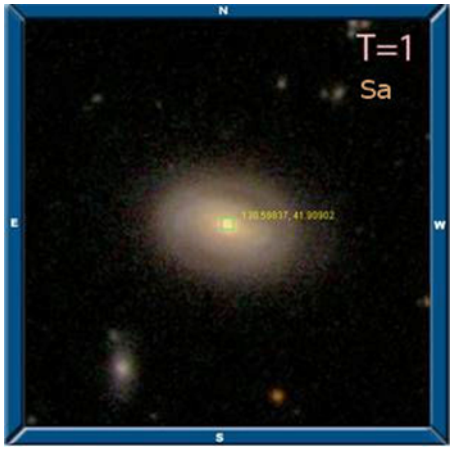
Tasks:

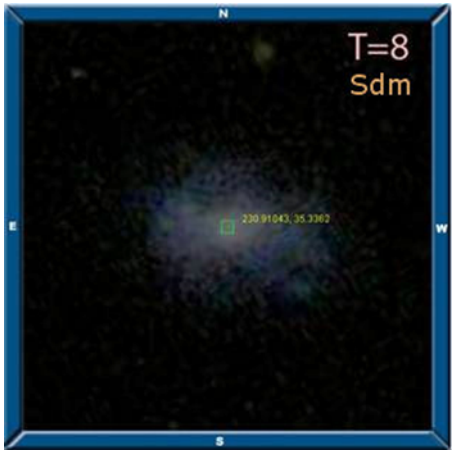
- To analyze the existing criteria for automatic morphological classification of galaxies using parameters, which are most correlated with morphology of galaxies: the inverse concentration index, absolute magnitude, de Vaucouleurs radius and scale radius
- To use various machine learning methods for classifying 60561 galaxies from the SDSS DR9 with unknown morphologies

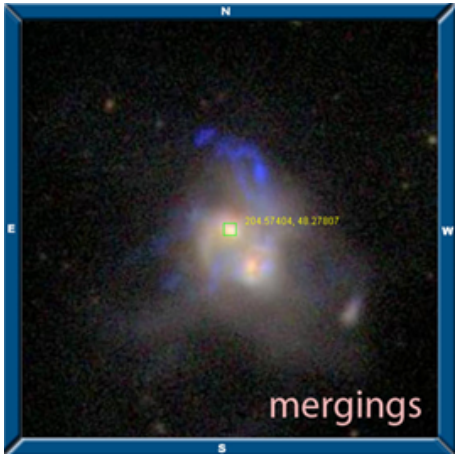
Sample of galaxies from SDSS DR5 for analysis of existing criteria of the automatic classification

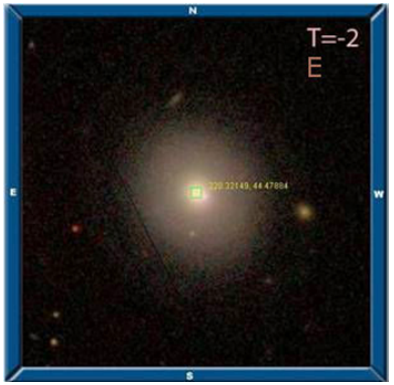
Sample contains of 730 galaxies

$3000 < V < 9500$ km/s,
V – radial velocity,
$H_0 = 75$ km/s/Mpc.

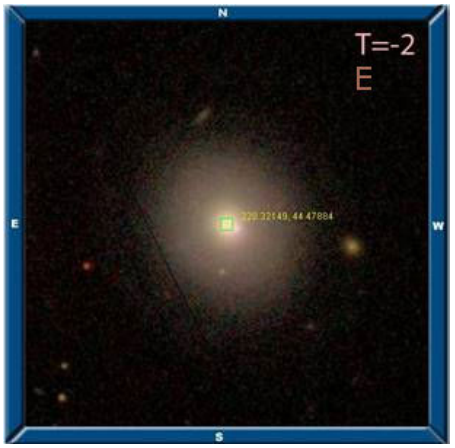


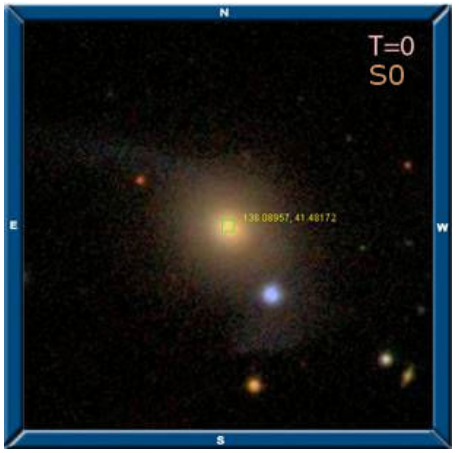


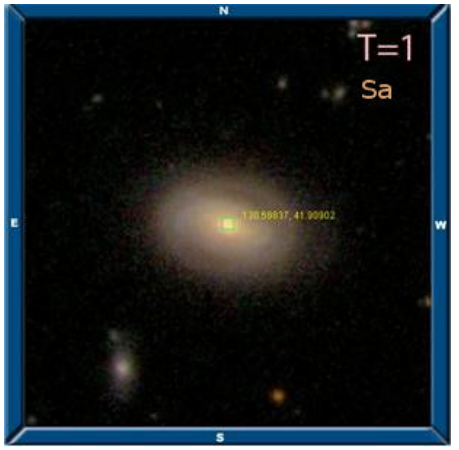


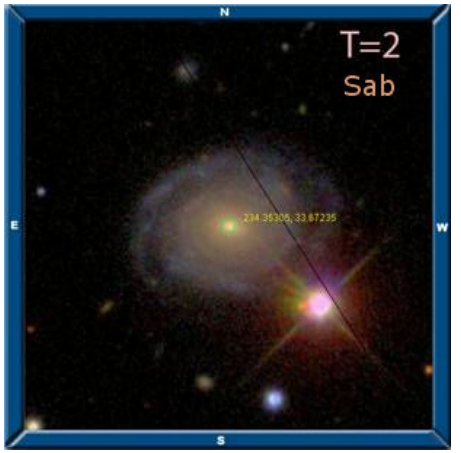


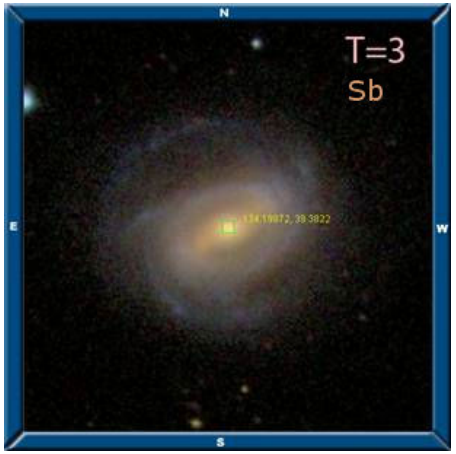
Morphological types of galaxies (modified classification)

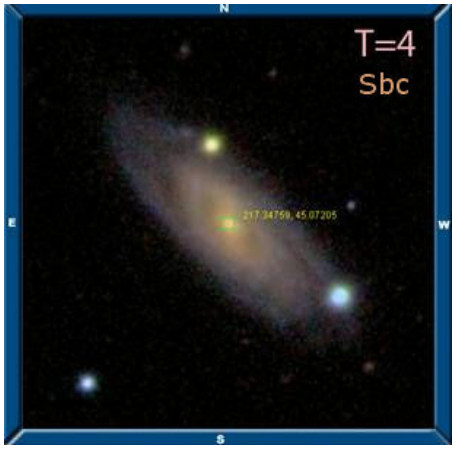


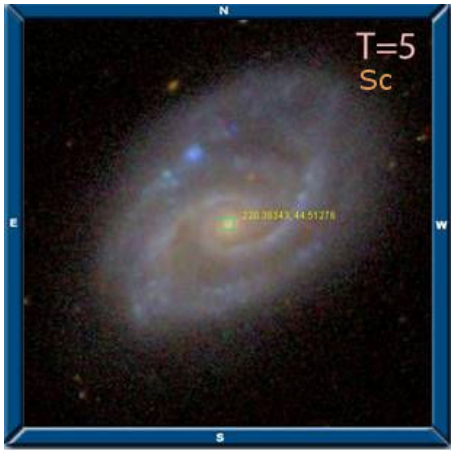


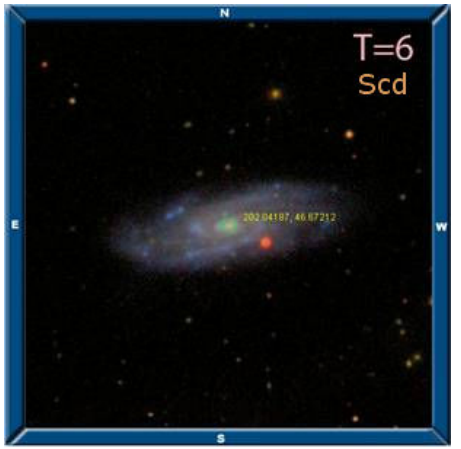


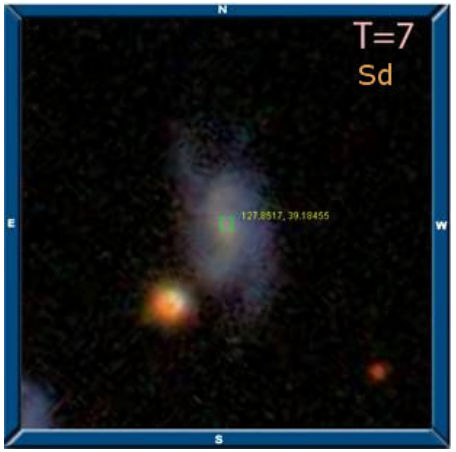


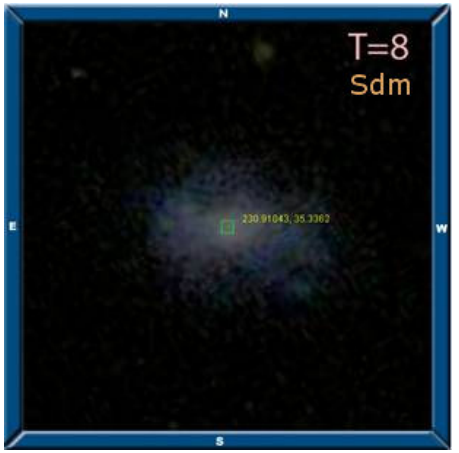


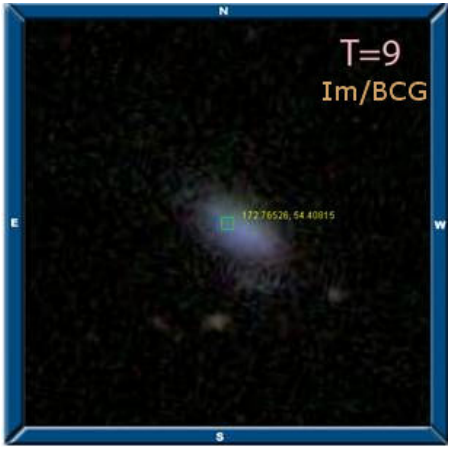


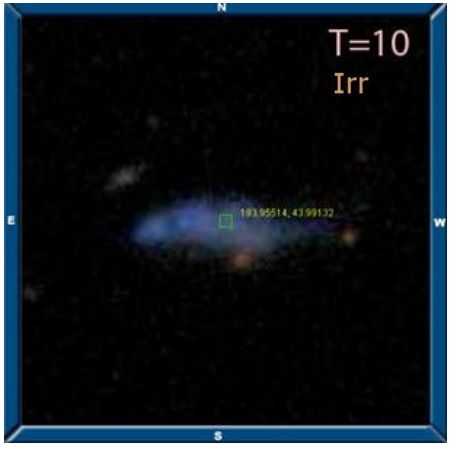












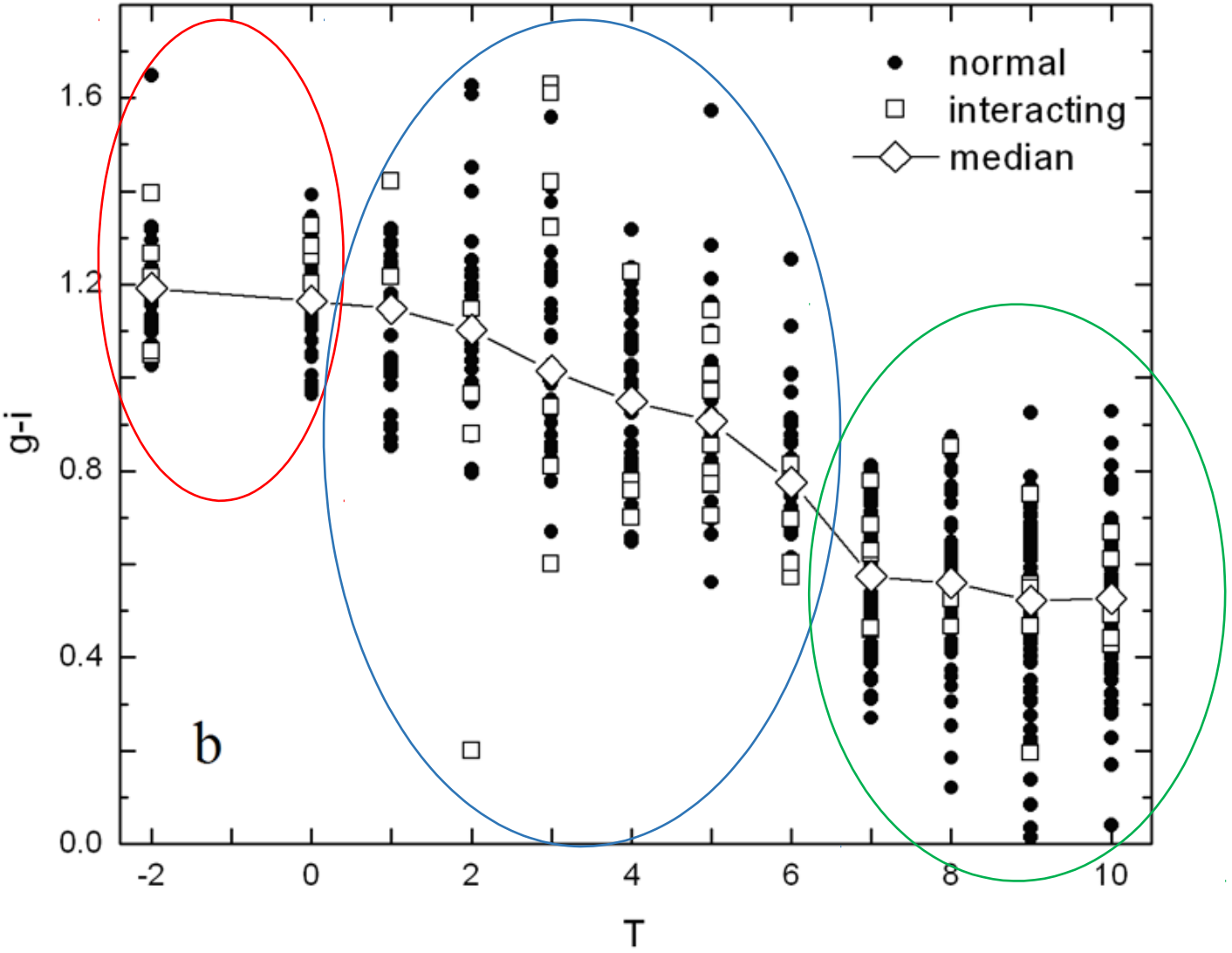
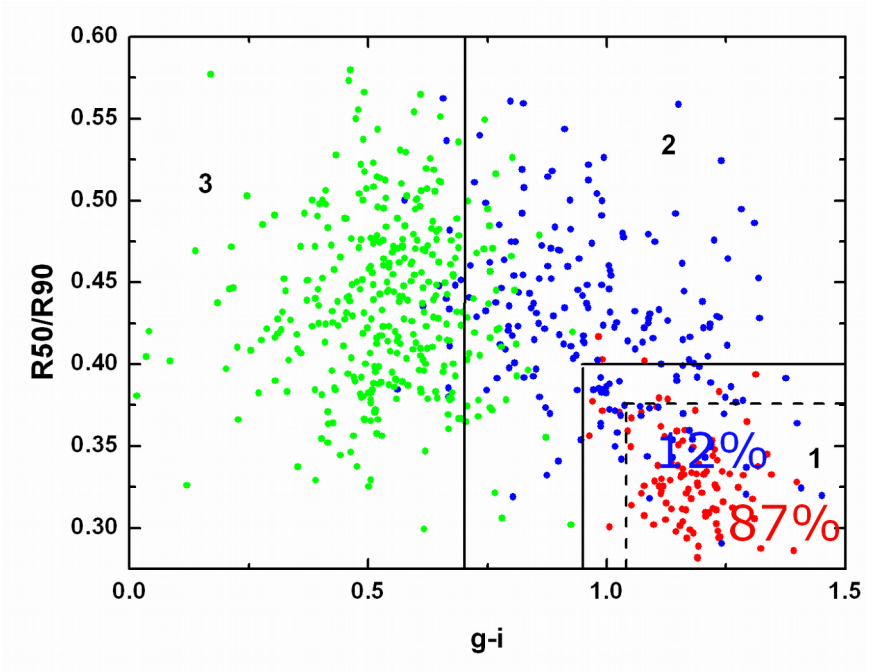
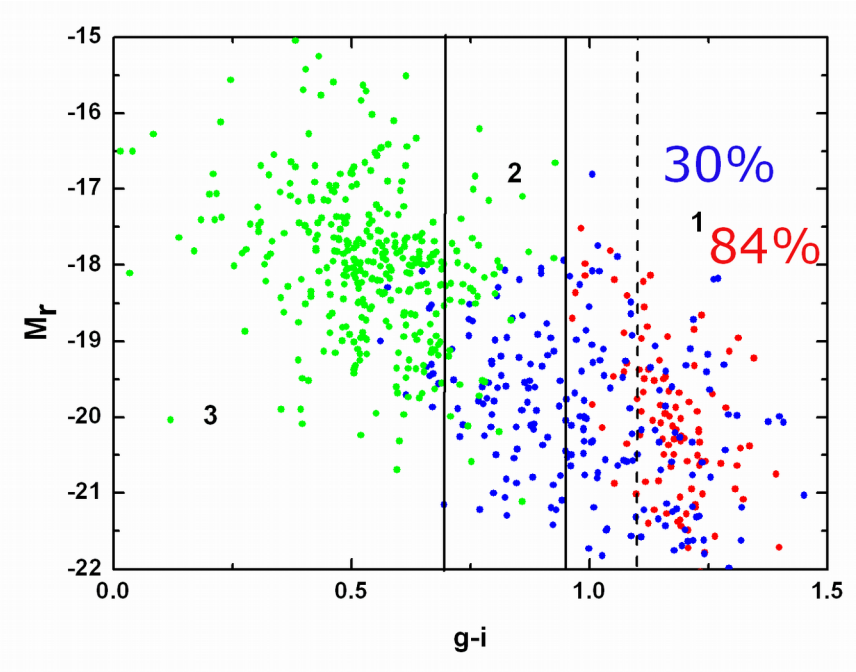


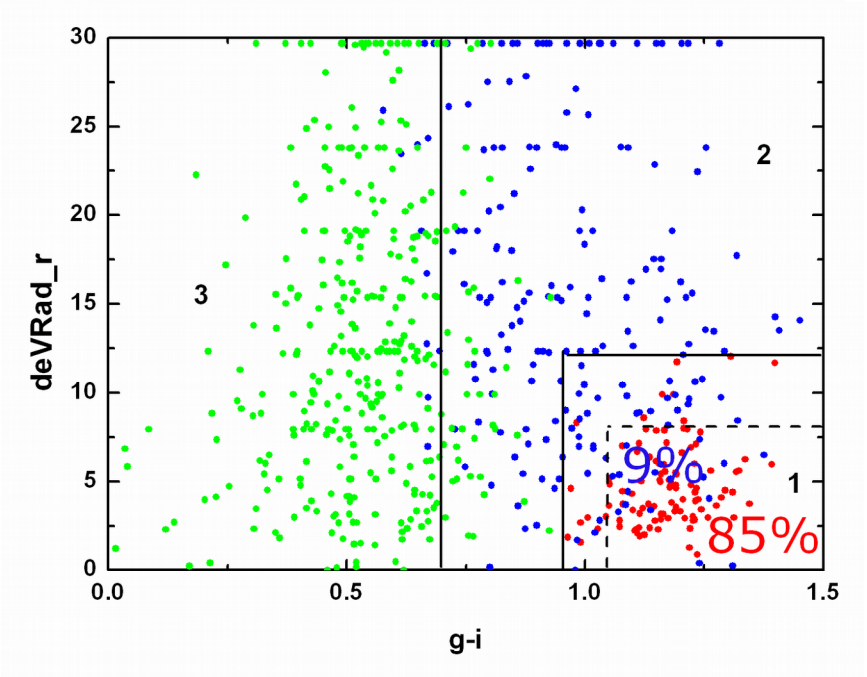
Figure shows the dependence of morphological type T on color index g-i for 730 galaxies. There is a correlation between the color indices and morphological type like for normal galaxies (N=666) and for merging galaxies (N=64), which is indicated by white squares. We can see that the overlap of the color indices is big but the median is essentially constant for (-2-0/1) and (7-10) types. Thus, we have divided the morphological types into three groups: early (-2-0), spirals (1-6), and late spiral and irregular (7-10).

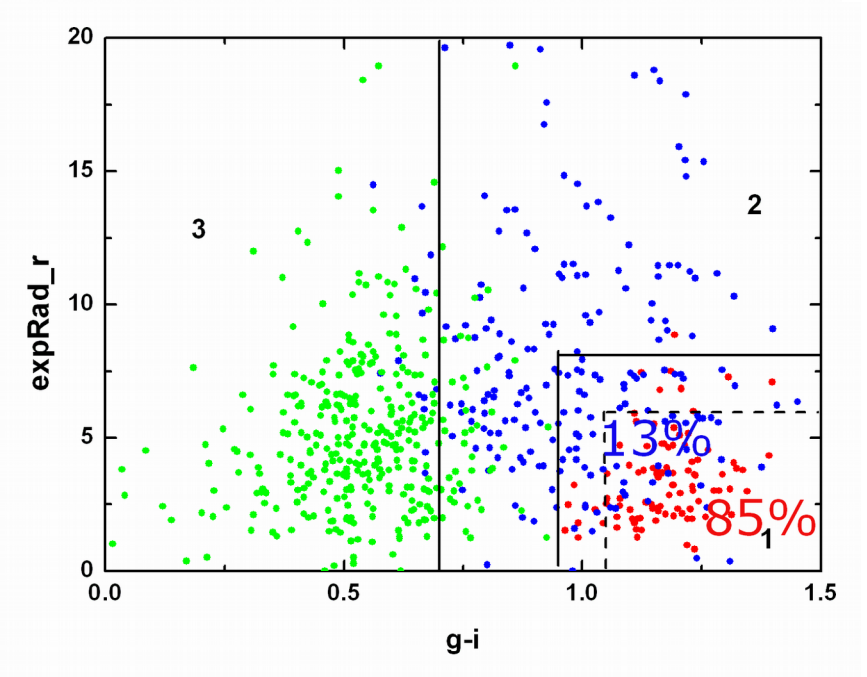
Figure shows the plots of color index as a function of the: inverse concentration index R50/R90 toward the galactic center; galactic radii: the de Vaucouleurs fit scale radius (deVRad_r) and exponential fit scale radius (expRad_r) in the r band; absolute stellar magnitude Mr.

Region 1 – early galaxies and lenticulars (-2-0), Region 2 – spirals (1-6), Region 3 – late spirals and irregulars (7-10).









Parameters	Morph. type	N	Region 1		Region 2		Region 3	
			N	%	N	%	N	%
R_{50}/R_{90}	-2-0	102	98	96.1	4	3.9	-	-
	1-6	207	53	25.6	**140**	**67.6**	14	6.8
	7-10	357	-	-	41	11.5	**316**	**88.5**
Mr	-2-0	102	102	100.0	-	-	-	-
	1-6	207	117	56.5	**76**	**36.7**	14	6.8
	7-10	357	-	-	41	11.5	**316**	**88.5**
deVRad_r	-2-0	102	100	98.0	2	2.0	-	-
	1-6	207	55	26.6	**138**	**66.7**	14	6.8
	7-10	357	-	-	41	11.5	**316**	**88.5**
expRad_r	-2-0	102	100	98.0	2	2.0	-	-
	1-6	207	74	35.7	**119**	**57.5**	14	6.8
	7-10	357	-	-	41	11.5	**316**	**88.5**

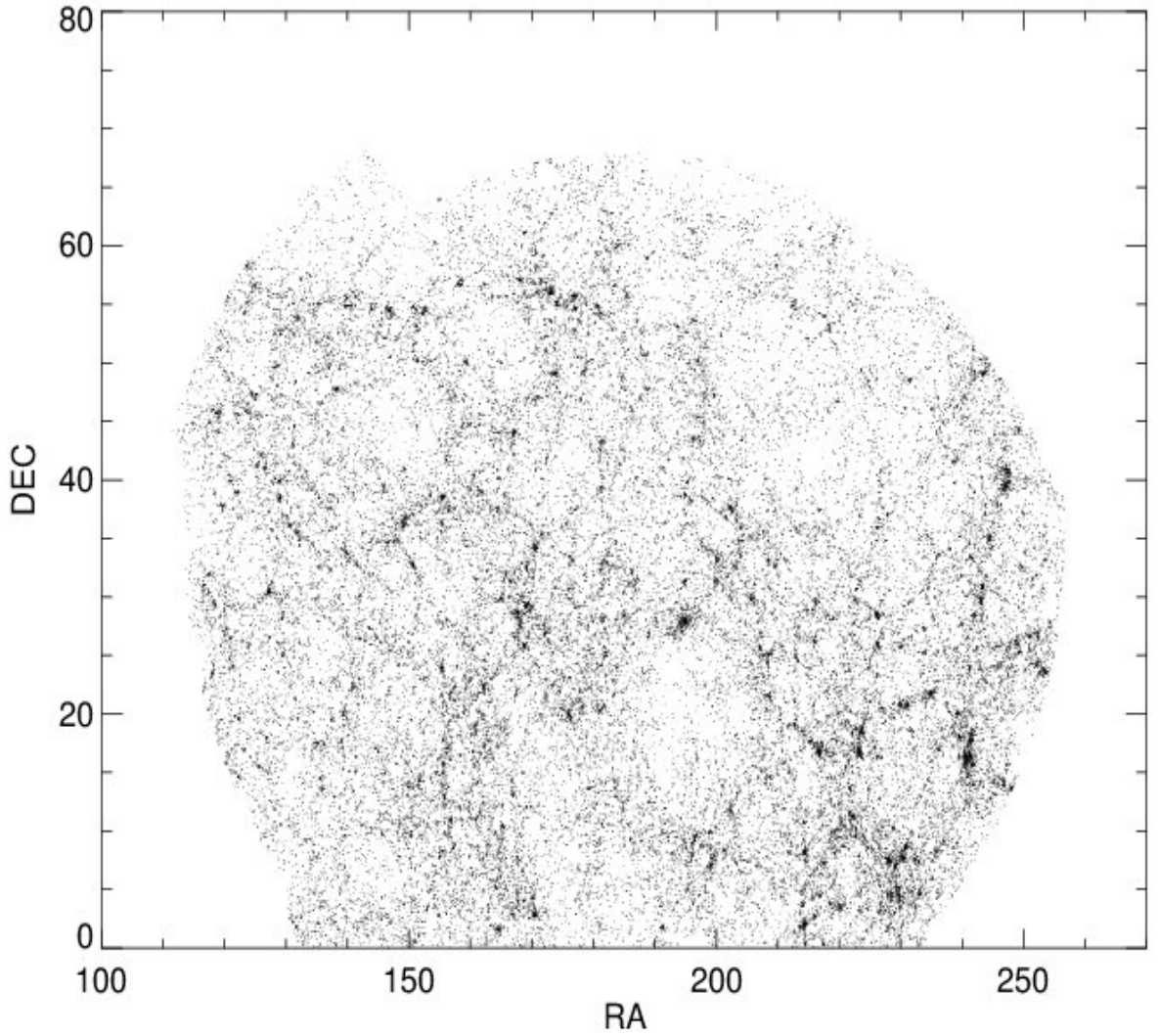
CRITERIA FOR THE CLASSIFICATION OF GALAXIES

T	u-r	g-i	r-z	R50/R90	deVRad	expRad	M_r
(-2-0) E+S0	2.2÷3.0	0.95÷1.5	0.6÷1	0÷0.4	0÷12	0÷8	-22÷-17.5
(1-6) Sa-Scd	1.6÷2.2	0.7÷0.95	0.45÷0.6	0.275÷0.6	0÷30	0÷30	-21,5÷-16.5
	2.2÷3.0	0.95÷1.5	0.6÷1	0.4÷0.6	12÷30	8÷30	
(7-10) Sd-Sdm+ Im/BCG	0÷1.6	0÷0.7	0÷0.45	0.275÷0.6	0÷0	0÷30	-21.5÷-15

Table lists the criteria for classification of galaxies according to the types (-2,0), (1-6), and (7-10) using the pairwise conditions for color indices and parameters R50/R90, Mr, deVRad_r, and expRad_r. These conditions correspond to the regions 1, 2, and 3.

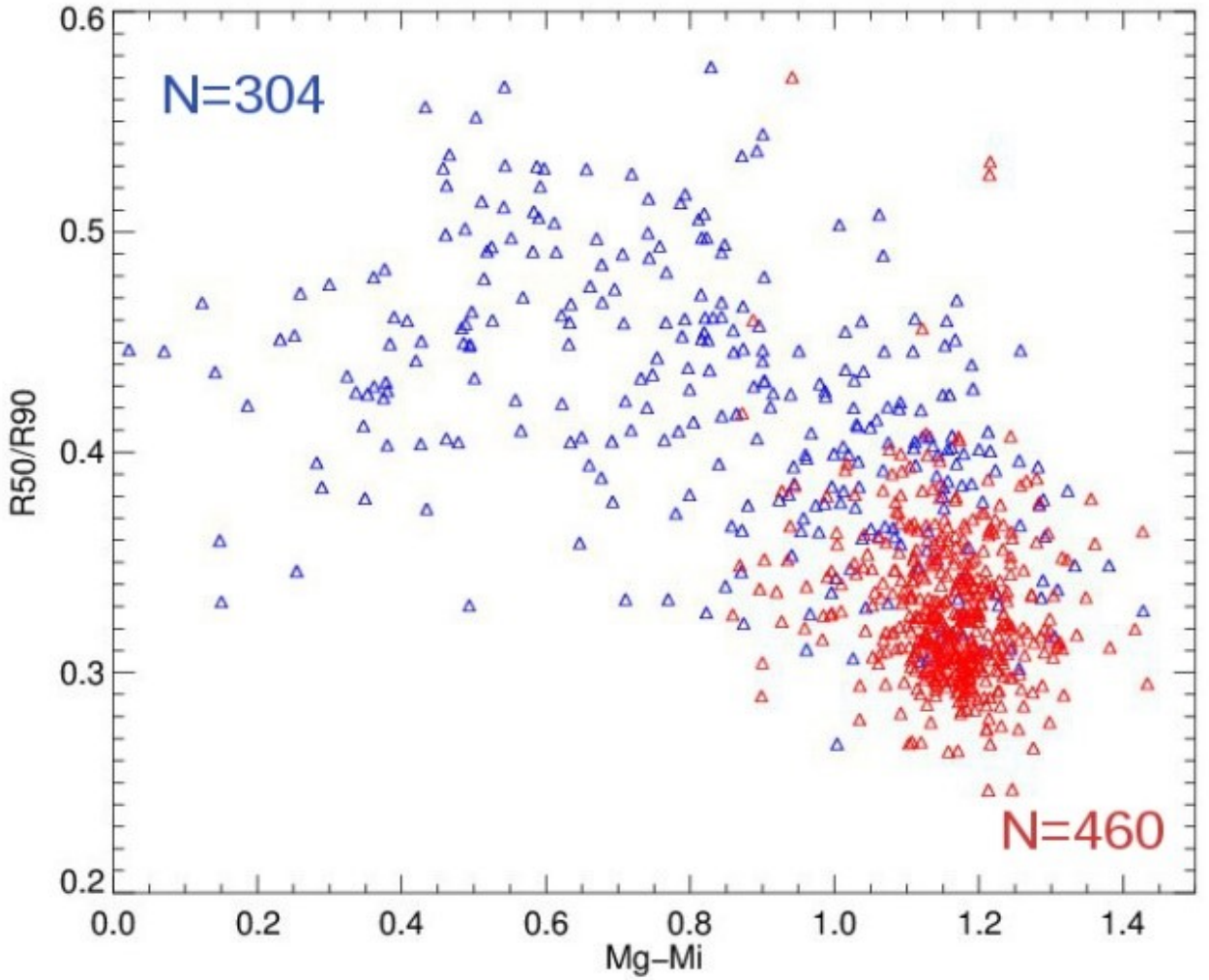
With one of color indices and such parameters as the inverse concentration index, absolute stellar magnitude, de Vaucouleurs radius, and scale radius, it is possible to conduct a reliable preliminary morphological classification without invoking visual inspection. Here, more than 90% of galaxies of the types (-2-0) and (7-10) are located in "their" regions of conformity to the morphological types.

Sample of galaxies from the SDSS DR9



The studied sample of galaxies is based on the SDSS DR9 ($0.02<z<0.06$, $m_r<17.7$, $-24<M_r<-17$) and contains of N = 60 561 galaxies.

Machine learning



First step for applying the machine learning was the formation of a training sample. We visually identified morphological types for 764 galaxies, witch were randomly selected from different redshift and luminosity, into two classes: E (including E, S0, S0a) and L (including from Sa to Irr)

Machine learning

- For training the classifier, we used the absolute magnitudes: Mu, Mg, Mr, Mi, Mz, all kinds of color performance and invert concentration index R50/R90 to the center.

- We spent a binary morphological classification using software with an open source **KNIME Analytics Platform ver. 2.11.3**, which is intended for prediction classification of data different machine learning methods and is actively used in the data science.

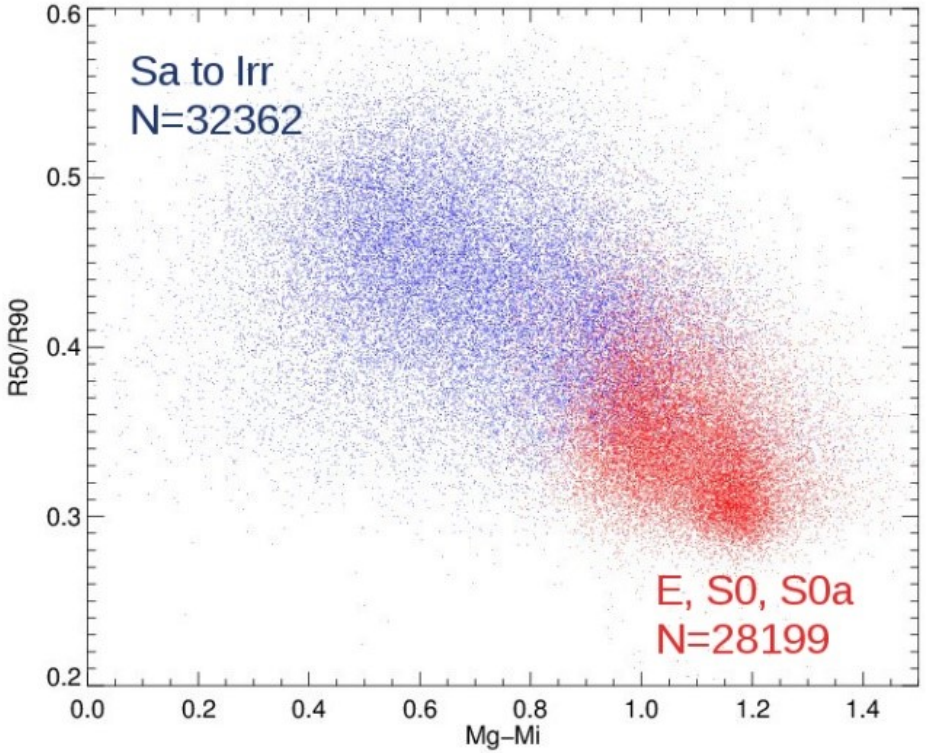
- First, we trained our classifiers methods: **Naive Bayes, Random Forest, Multi Class Classifier, Support Vector Classifier (SMO)**, based on WEKA 3.7 software, and neutral networks (**RProp MLP**).

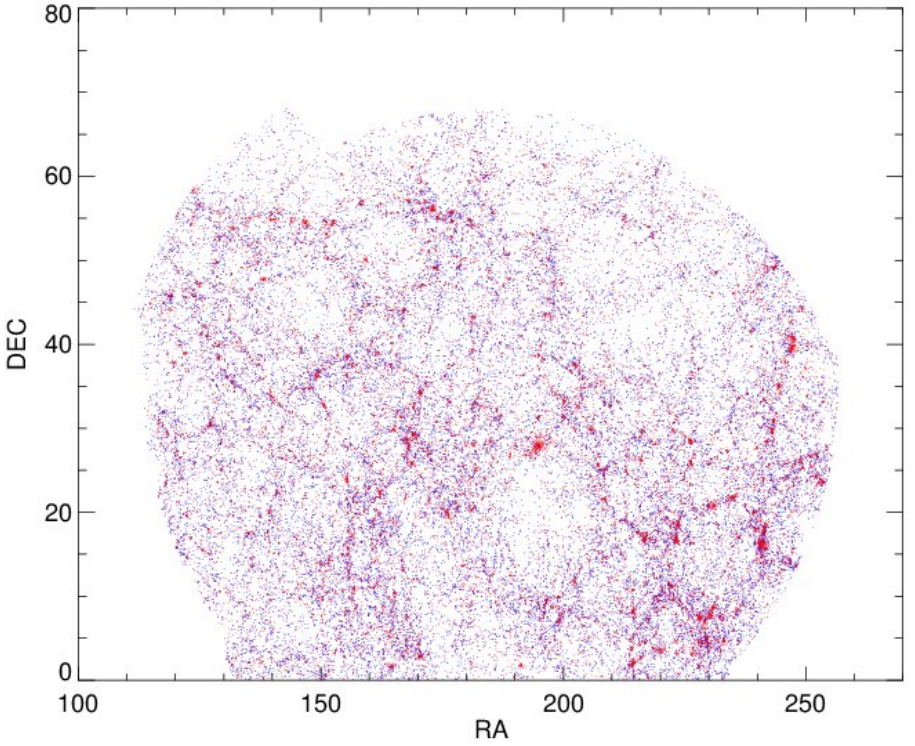
- Precision methods, we evaluated using the method of **k-folds validation**: we divided the sample into training randomly selected 5 pieces, one by one 4 of which served as a training and a test sample. Classification accuracy was defined as the average of the test samples.

- It turned out that the **Random Forest** method provides the highest accuracy - **91%** correctly classified (**96% E** and **80% L**). The accuracy of the rest by 85% to 90%.

- For comparison with our previous step, we used the classification criterion for color indices u-r, g-i, r-z and R50/R90. This criterion is determined visually on a graph the relationship between the two parametrs. The accuracy for the E type was 96% and for the L type --- 67%, however, the training sample in the time we have served as a test, which means that the real accuracy was at least a few percent lower.

Machine learning





Thus, using the information on the absolute values, color index, R50/R90 and coaching by Random Forest classifier to galaxies with visual morphological types, we applied the criteria to responsible 60,000 galaxies with unknown types, and got their classification.

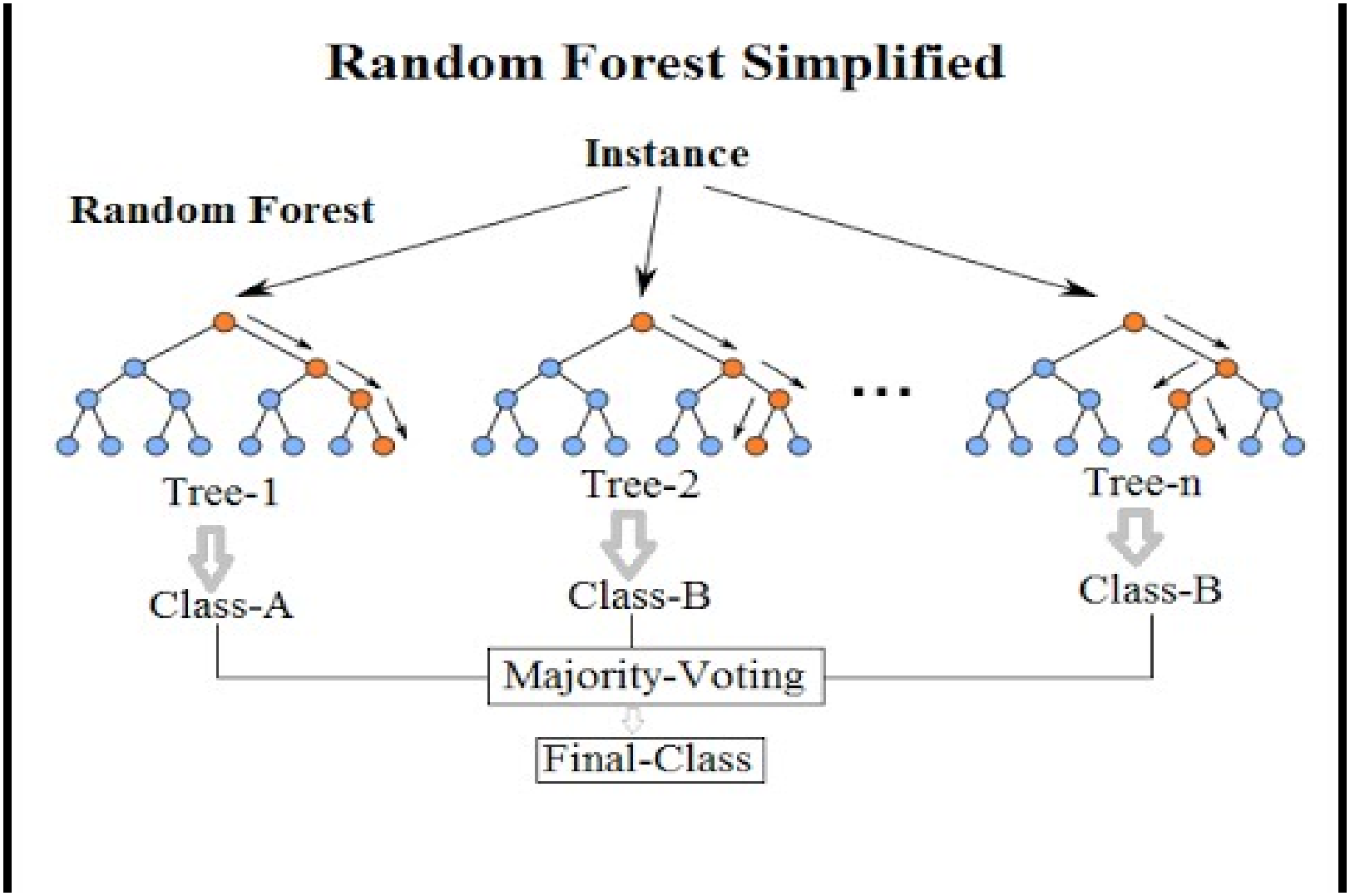
Conclutions :

- Using visual classification of galaxies and two-dimensional diagrams “color index-M_r”, “color index-R50/R90”, “color index-deVRad_r”, and “color index-expRad_r”, we found the possible criteria for separating the SDSS galaxies into three classes:
    1) early types— elliptical and lenticular;
    2) spiral Sa-Scd, and
    3) late spiral Sd-Sdm and irregular Im/BCG galaxies.

- Using the machine learning method with the Random Forest classifier and the data on the color indices, absolute magnitudes, inverse concentration index of galaxies with visual morphological types, we classified 60 561 galaxies from the SDSS DR9 with unknown morphologies.

Finally, we found 28 199 E and 32 362 L types among them.

Thank you for attention!



RF is an ensemble **learning method** for classification, that operate by constructing a multitude of decision trees at training time and outputting the class that is **mean prediction (regression) of the individual trees.**