Classifying Foreground Dwarfs and Background Giants in the Perseus Cluster using Machine Learning

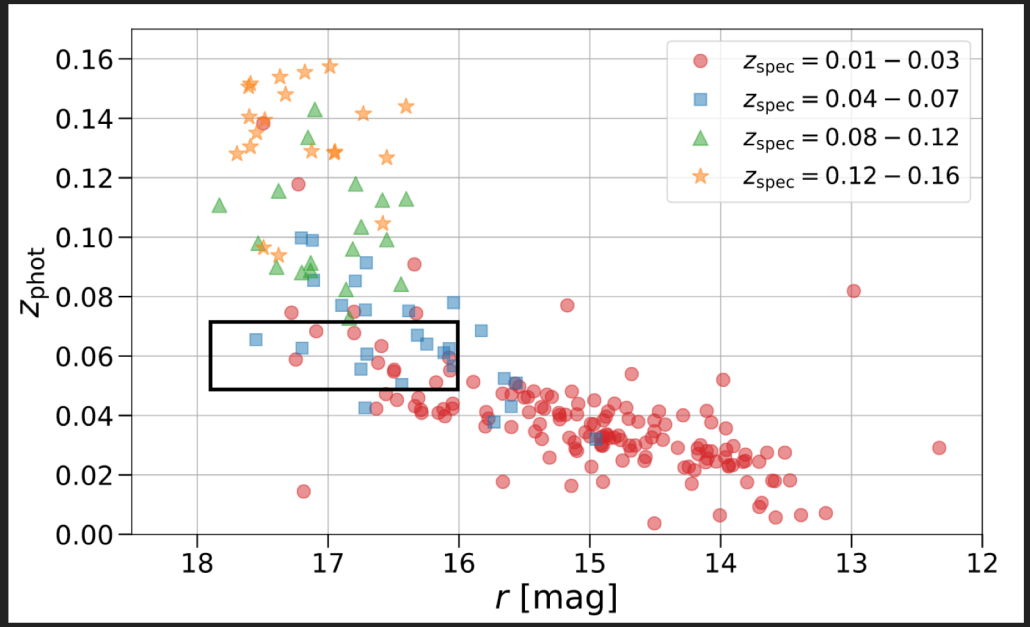
Steven Li & Dan Rachou

Why?

- When looking at a galaxy within a cluster, can we know if it’s actually a part of the cluster or not visually?
- No, it’s often very ambiguous because it can either be a foreground dwarf and a background giant looking visually the same relevant to our perspective.

So, how do we tell them apart?

- Within the Perseus cluster, we can sometimes tell the difference between foreground dwarfs from background galaxies by knowing the true redshift.
- That is only true for photometric redshift above 0.08 and below 0.04.
- Z-spec is currently the only true way to tell what they are.



Without Z-spec, it’s not that simple.

- We wondered if there are any other parameters or combination of parameters that can give the same concrete classification like Z-spec?
- Is there a single parameter or a combination of parameters that are more abundant, easily accessible that does not have that gray area where it could be either one?

Can you visually discriminate the two from one another?

Background Giant Galaxies

Foreground Dwarf Galaxies

Why so Machine Learning?

- Spot patterns in in data
- Computationally Intensive
- More efficient
- Can detect patterns
- Lastly, data is increasing while computational processing is get cheaper and more efficient.

Types of Machine Learning

- Supervised Learning: We feed the algorithm labeled data, and the algorithm tries to find patterns to create an overall model.
- Unsupervised Learning: We feed the algorithm unlabeled data and allow the algorithm to infer a natural structure within the inputs without giving it any hints.
- Both types require a good amount of training and testing data.

Why we choose Supervised Learning

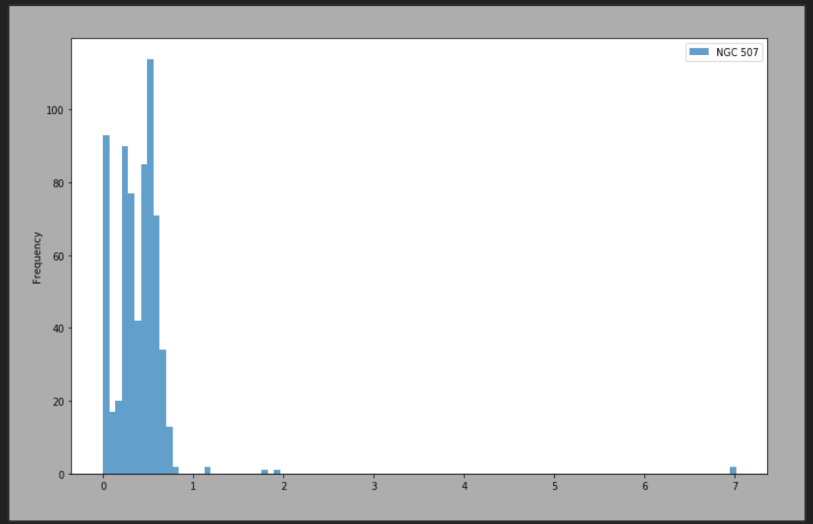
- We didn’t know what unsupervised learning would spit out as a result.
- Unsupervised learning is very technical.
- We would be able to label data from Perseus and from two nearby clusters with similar redshifts, NGC 383 & 507.

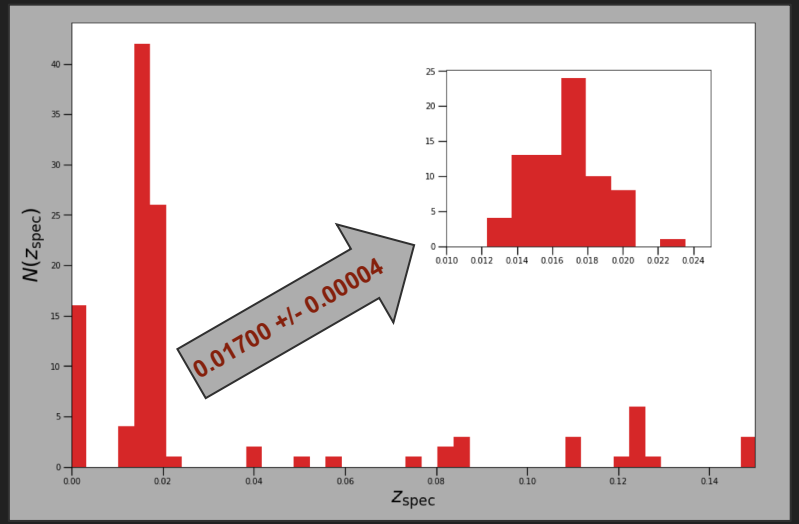
Choosing our training dataset

- Need to find the foreground and background of each cluster.
- NED has more complete library of redshifts than SDSS. Both spectroscopic or photometric.
- We take all the galaxies within 60 arcminutes of each cluster from NED and take only the entries with spectroscopic redshifts as a dataset.

Each clusters' redshift range

- To find the foreground and background, to plot a histogram and zoom into each clusters' center redshift.





- Perseus: **[0.0153-0.0187]**
- NGC 383: **[0.0125-0.0225]**
- NGC 507: **[0.012-0.022]**

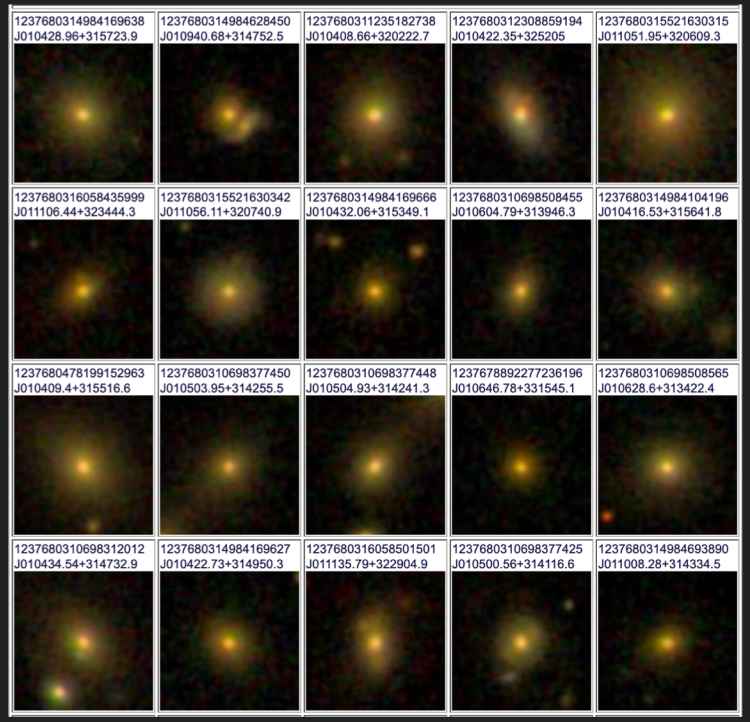
Filtering Foreground & Background

- Within redshift range: Foregrounds
- Outside the range: Backgrounds
- Separated into their own background & foreground dataset for each cluster.
- All background dataset for each cluster are combined into one dataset.
- All foreground dataset for each cluster are combined into one dataset.
- Limit the galaxies to be within 15 - 17.5 magnitude.

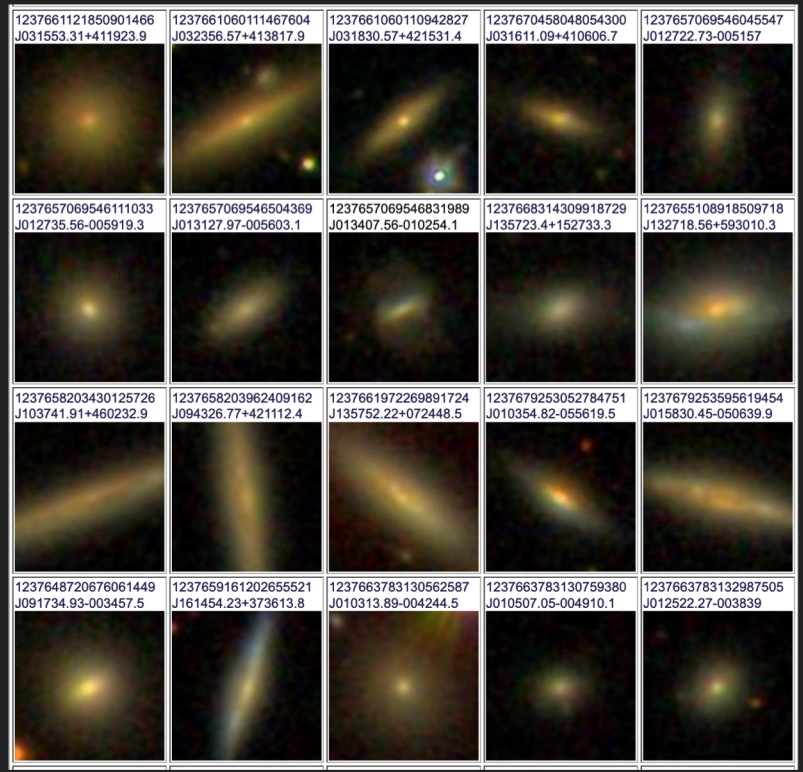
- Total Foreground Dwarfs: 414
- Total Background Giants: 162

Sample Thumbnail Images of Galaxy Datasets

Background Giants



Foreground Dwarfs



Adding the parameters

- Take the training data set’s object name, RA, and DEC into CrossID and grab all of PhotoObj’s parameters.
- From the PhotoObj parameters, we handpick the parameters that we believe to be relevant in the galaxy classifications removing junk and irrelevant parameters.

509 parameters -> 137 parameters

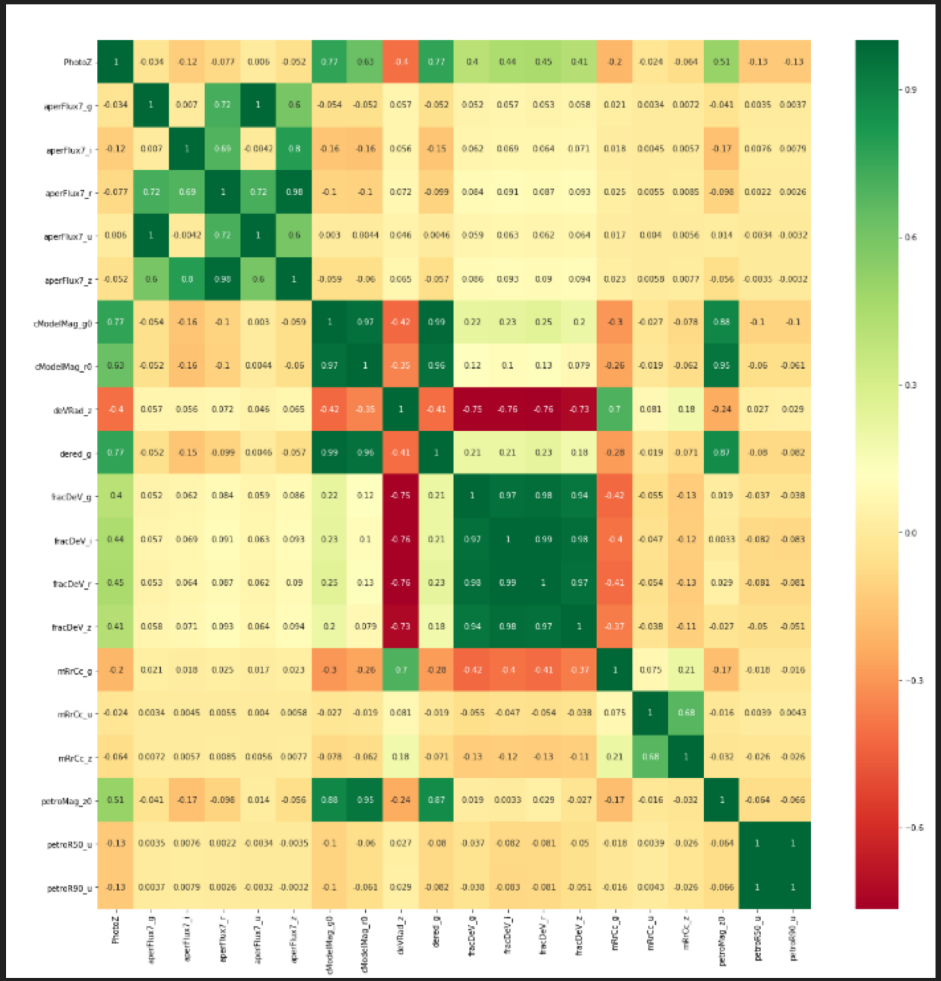
Finishing the final training dataset

- Last step is to grab the foreground and background training dataset and add a new column which denotes it’s label (what we know the classification is).
- Labeling foreground will all be 1’s and background will be 0’s. Then we join the datasets together to be our final training dataset.

Feature Engineering

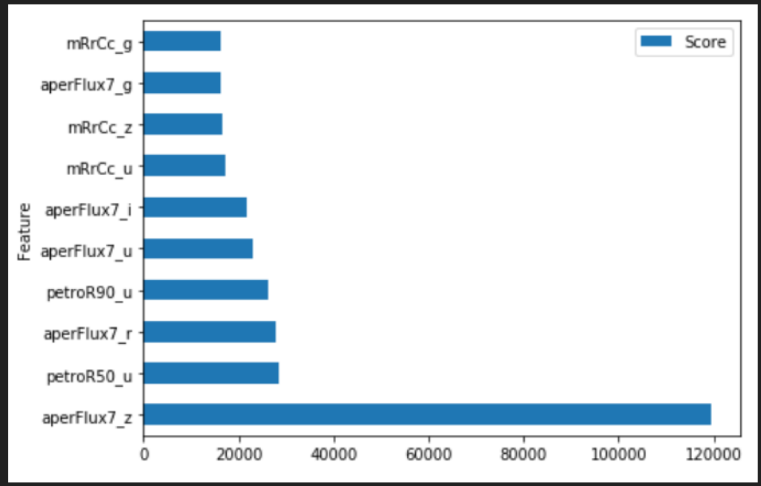
- Rather than only letting us pick the parameters, we will also give the computer a chance to pick the parameters through feature engineering.
- We will be utilizing Univariate Selection, Feature Importance and Correlation Matrix with Heatmap to visualize each parameter's correlation with one another.
- Why? Reduce training time, overfitting and increases accuracy.
- Univariate Selection selects the best features that have the strongest relationship with the output variable through scikit's selectKBest.
- Feature Importance selects the most important or relevant features toward our output variable through scikit's extra tree classifier.

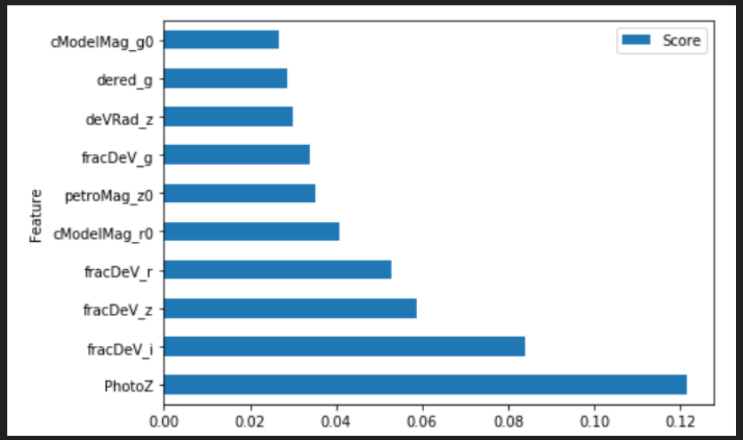
Correlation Matrix with Heatmap visualization of parameters



Resultant Parameter Variations

- Top 10 Best Features
- Top 10 Most Important Features
- Top 20 Most Important Features
- Top 10 Best Features & Important Features
- All Features





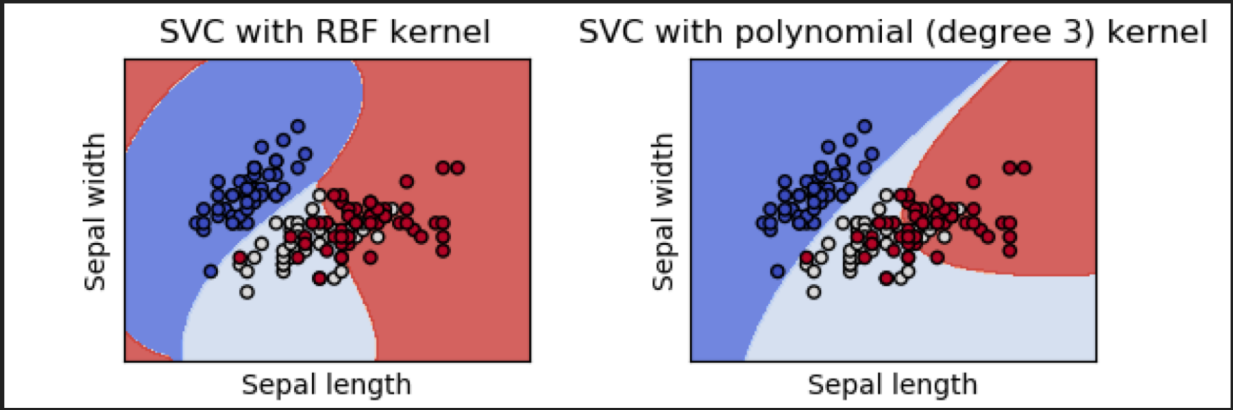
Getting our prediction dataset

- Perform an SDSS SQL search of 1.5x1.5 degree box around the perseus cluster of magnitude 15.5-16.5 and grab all the picked parameters for the returned galaxies.
- This is now our dataset we will use to test our machine learning model.

Prediction Dataset: 121 Galaxies

Performing the Machine Learning

- We will be performing the SVM (Support Vector Machine) algorithm because it is effective even in high dimensional spaces and we want to test all the parameters.
- SVM has 3 different models (kernel) types for classification: gaussian, sigmoid and polynomial.



Ref: https://scikit-learn.org/stable/modules/svm.html

Machine Learning accuracy from cross-validation for each model (kernel) type and parameter variation.

TABLE II. Model Type, Features and Accuracy

Kernel Model	All Features	10 Best Features & Most Important Features	10 Best Features	10 Most Important Features	20 Most Important Features
Gaussian	0.84 (±0.26)	0.87 (±0.27)	0.87 (±0.27)	0.91 (±0.15)	0.87 (±0.18)
Sigmoid	0.50 (±0.15)	0.49 (±0.37)	0.49 (±0.37)	0.66 (±0.02)	0.66 (±0.02)
Polynomial (n=1)	0.90 (±0.09)	0.87 (±0.06)	0.87 (±0.11)	0.93 (±0.16)	0.90 (±0.12)
Polynomial (n=2)	0.90 (±0.16)	0.89 (±0.10)	0.87 (±0.13)	0.94 (±0.09)	0.94 (±0.07)

Doing the prediction on the model

Classified Foreground: 84 -> 68
Classified Background: 37 -> 31

- Prediction dataset of Perseus cluster containing 121 galaxies

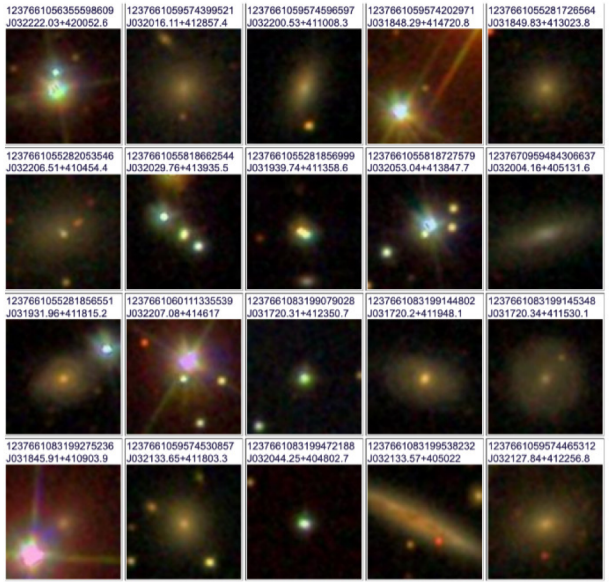


FIG. 9. Visual subset of the Machine Learning classified dwarf galaxies generated by a thumbnail search of the subset in SDSS image list format.

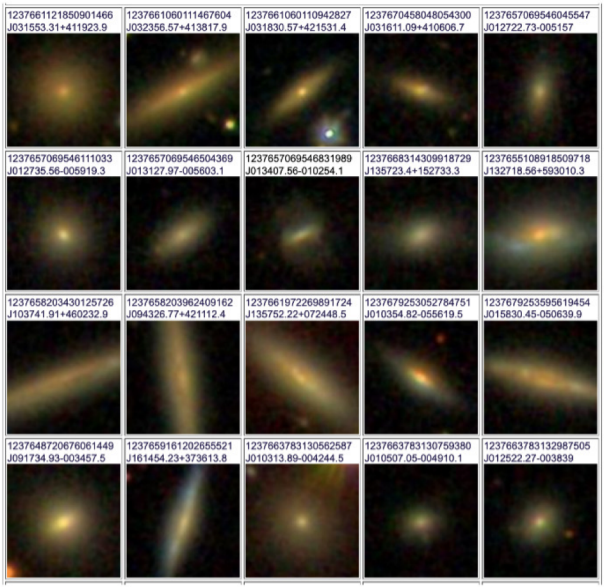
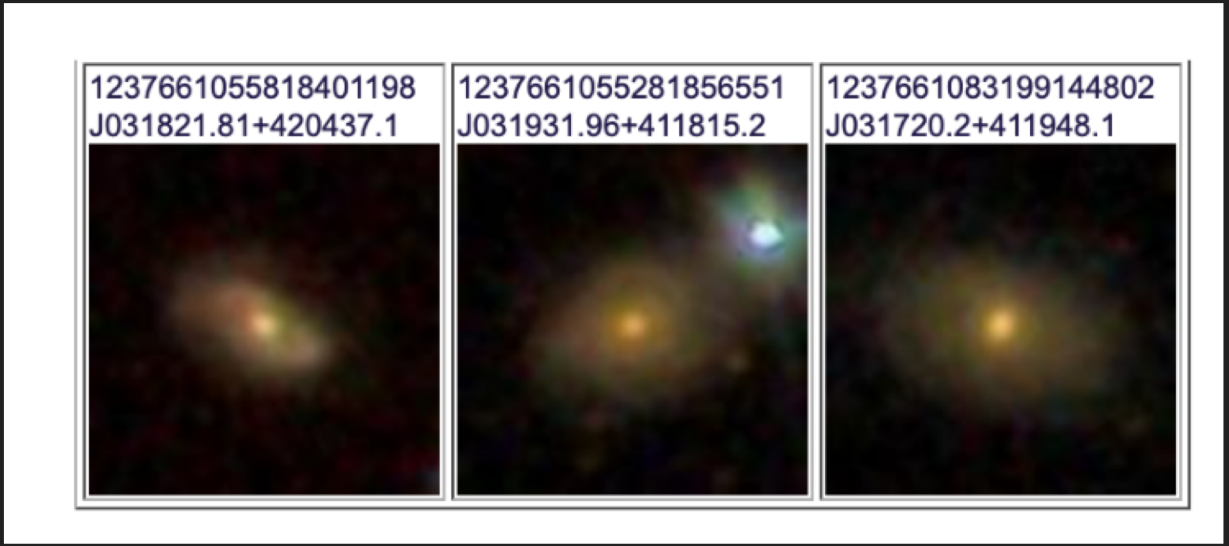


FIG. 10. Visual subset of known dwarf galaxies generated by a thumbnail search of the subset in SDSS image list format.

Testing for false-positives

Falsely classified foregrounds: **3**
Falsely classified backgrounds: **0**

Potential Foregrounds: **65**
Potential Backgrounds: **31**



Potential foreground classification accuracy ceiling: **95.58%**

Potential background classification accuracy ceiling: **100.00%**

Conclusion

- Yes, there are other parameters that can classify similar to z-spec.
- SVM is very accurate even without feature-engineering to reduce overfitting, reduce training time and improve accuracy.
- Test other supervised method such as Random Tree with Principal Component Analysis.
- Try using unsupervised machine learning such as k-clustering to see if they can produce similar results.
- Add a third class to the dataset.
- Change search radius of the prediction set to 50 arcmin.
- Increase background galaxy set by increasing search radius around Perseus cluster.