Star-Galaxy classification using ALHAMBRA photometry and Machine Learning Algorithms

PRATYAKSH RAJ[1]

[*] *Corresponding author: email@my-email.com*

Compiled October 31, 2022

http://dx.doi.org/10.1364/ao.XX.XXXXXX

1. INTRODUCTION

It is known that classifying an object as star or galaxy, based on their morphology, is a difficult task from earth based telescopes(sometimes even for space based) as their images get blurred due to plenty of reasons. Therefore, Here we have used ML algorithms (Convolutional NN, Artificial NN, Random forest, Ada Boost) to classify an object as star or galaxy over ALHAMBRA-4 field. We used ALHAMBRA survey- isophotal magnitudes (20 Narrow band, 3 NIR broad band, f814, FWHM) over COSMOS field, taking Hubble's Morphology based classification as ground truth. There are a total 31870 objects- 29615 galaxies and 2247 stars. For f814w<22.5, there are 5991 galaxies and 1764 stars. We divide our train and test data in 85:15 ratio.

Here we will find out the best algorithm with least overfitting and best AUC score. CNN performed best on both datasets- f814w <22.5 (AUC score: 0.986) and f814w <26 (AUC score: 0.962). Then making correlation heat map of features to find that f814 is highly correlated with 10 other magnitudes (correlation coefficient >0.94).Therefore number of parameters reduced to 13, giving even better AUC score (0.993), much less overfitting and less CPU time. Here,Performance of all algorithms and for different brightness ranges are visualized via ROC Curve

We also ran all algorithms on only narrow optical bands, than gradually adding NIR, f814w and FWHM. We found that all algorithms give bad results on narrow bands except for CNN (0.983 for f814w<22.5 and 0.961 for f814w<26) which maintained its performance. Also, it is a good practice to consider all objects fainter than 23 as Galaxies, as these objects are mostly galaxies (9000-15000) and very few stars (100-250).

Now we calculated the ranking of features towards classifyinng the object (using Randomforest and ADa boost) and found that except FWHM, f814w,K and j magnitude, all other magnitudes are not a good classifiers, even after making them good learners using Ada Boost (although CNN captured the relation very well). Therefore instead of 20 narrow band magnitudes, we ran CNN on 19 colors(differnce between consecutive narrorw band magnitudes) + 2 colors(3 broad band) + FWHM + F814w, resulting in AUC score increasing from 0.9860 to 0.9902, for f814w<22.5. Also from ranking of features we see that these colors (f365-f396 , f396-f427 , f458-f459) become equally important as f814w and FWHM in classification.Therefore now using only these 5 magnitudes on CNN we get AUC score 0.9902 for f814w<22.5, and much less CPU time.

Therefore we can conclude that using CNN we don't need all 25 magnitudes, we can do even better classification by using much less features, overfitting and CPU time. We need only sex features- FWHM, f814w, (f365-f396), (f396-f427), (f458-f459), (f768-f799). Also we can see that most of these colors are composed of magnitudes of low wavelengths.

Therefore it can also be concluded (visualised) that CNN was able to perfectly capture the spectrum of stars and galaxies resp. It give very good result with 96% accuracy, with nearly no wrongly classified galaxy (only stars fainter

than 23 are miss classified as they are very few, can be contamination as these objects are mostly galaxies) and correctly able to identify objects for which ALHAMBRA do not provide any classification.

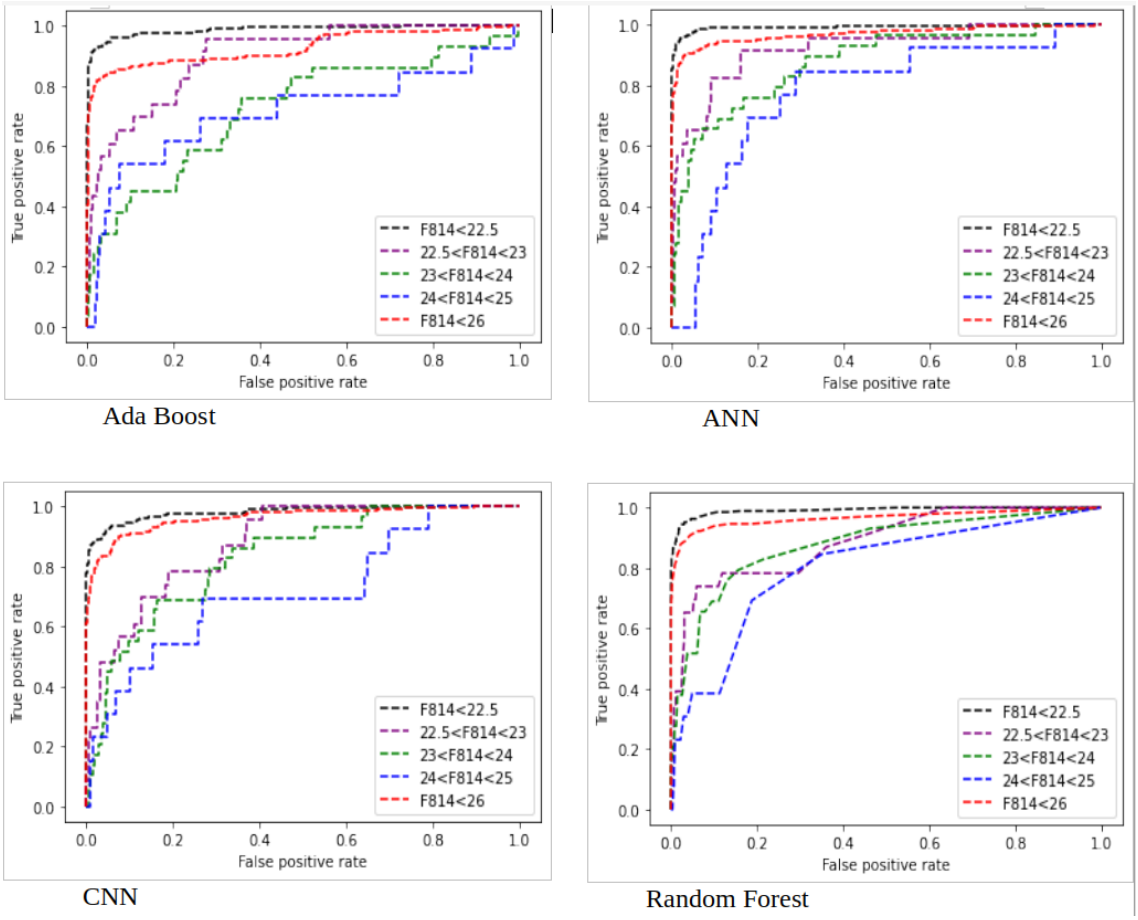


Fig. 1. ROC curves for different ranges of F814W, using all 25 magnitudes as features. In magnitude ranges 22.5-23, 23-24 and 24-25, there are 4000 to 9000 galaxies and 100 to 250 stars in each range

2. ALHAMBRA SURVEY

We have used the ALHAMBRA catalogue (Molino et al. 2014) over the ALHAMBRA-4 field, which overlaps with COSMOS. It contains 31870 objects matched to our reference COSMOS catalogue with a maximum sepration on 3 arcsecond. Out of which 29615 are galaxies and 2247 are stars. The ALHAMBRA photometric system is characterized by 20 constant width (31 nm), non-overlapping medium band filters covering a wavelength range from 350 to 970 nm. The images were taken using the Calar Alto 3.5m telescope using the wide field optical camera LAICA and the NIR instrument Omega-2000, which are equipped with 20 intermediate width bands and three NIR broad-bands: J, H, K. The catalogue presents multicolour PSF-corrected photometry detected in synthetic F814W images with objects up to a magnitude of F814W 26.5.

3. DATA WRANGLING

We will take only those objects for which less than 5 band magnitudes is unknown. Here median (rounded off to three decimal places) is used to fill the missing data of respective column. Also, as ALHAMBRA do not provide any classification for most of the objects fainter than 22.5, its a good practice to consider all those objects as galaxies as they generally are. There are very few (100-200)star in this range which produces contamination but here we ran our algorithms on different ranges of F814w >22.5, without doing any changes.

4. MACHINE LEARNING ALGORITHMS AND HYPERPARAMETERIZATION

We have used ANN, RandomForest, AdaBoost and CNN. We have measured the performance of ANN and CNN by making ROC, overfitting curves and AUC scores. For RandomForest and AdaBoost, we can check AUC score and ranking of features (Ada Boost is helpful in getting the complex relations which randomforest might not able to get. It makes weak learners strong). They both are highly important in getting good ranking of features
Firstly we hyperparameterized all our algorithms for objects brighter than 22.5, to compare them. ANN, CNN are hypertuned via KerasClassifier and Random forest, AdaBoost via GridSearchCV. These are:-

CNN- Convolution2D(32, kernel size=(5, 1), input shape=(25,1,1), activation= 'LeakyReLU') + (MaxPool2D(pool size=(3,1)))
Convolution2D(32, kernel size=(3, 1),activation= 'LeakyReLU') + MaxPool2D(pool size=(2,1))
Convolution2D(64, kernel size=(2, 1),activation= 'LeakyReLU') + MaxPool2D(pool size=(1,1))
Dropout(0.3) + Flatten() + Dense(1 ,activation="sigmoid")
Loss Function: Binary Cross Entropy, optimizer= Adam(learning rate=0.03)

Random Forest- estimators: 200, min samples split: 2

Ada Boost- base estimator: Decision Tree, n estimators: 100, max depth: 4

ANN- dense layer(neurons: 25, activation func: Relu, BatchNormalization())
dense layer(neurons: 25, activation func: Relu)
dense layer(neurons: 25, activation func: Relu, dropout: 0.5)
dense(neuron: 1, activation func: sigmoid)
Loss Function: Binary Cross Entropy, optimizer= Adam(learning rate=0.03)

5. SELECTION OF BEST ALGORITHM

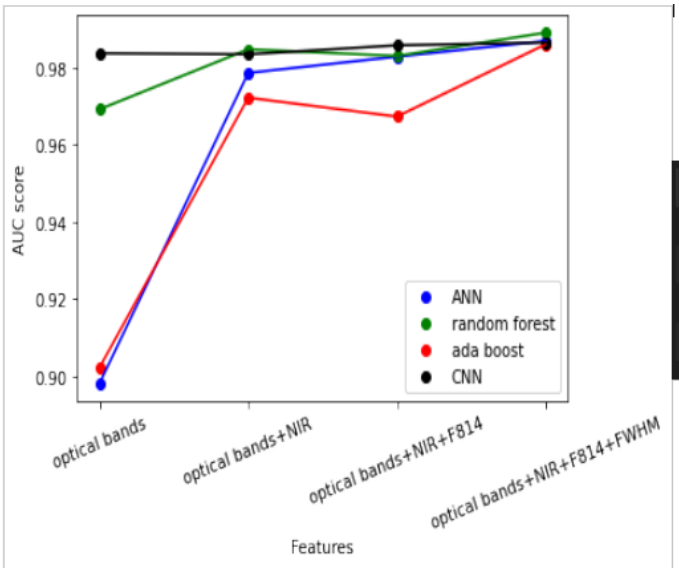
we ran all parameterized algorithms on data set(15% testing and 85% training) for various ranges of F814w and also by gradually adding features, See Fig. 1 . We can see that CNN gives AUC 0.986 for f814w<22.5 and 0.983 for only optical bands as features whereas other algos are going as low as 0.898 (ANN), See Fig. 2 . Also, we can even see that CNN is not overfitted where as ANN is highly overfitted, See Fig. 3 . RandomForest also give high AUC scores but performance drops with only optical bands(AUC 0.969). Also ranking of features shows that it depends highly on FWHM (much more than any other magnitude, See Fig. 4), therefore not as useful as CNN . Ada boost also shows same results as of ANN. Therefore it can be concluded that all algorithms except CNN are highly dependent on FWHM for f814w <22.5, but CNN is consistent without being overfitted.

Therefore CNN comes out to be the best algorithm with outstanding results, AUC 0.986. AUC scores for different ranges of F814w are shown in fig. 5.

6. FEATURE REDUCTION

Training time of CNN can take a lot of time. Therefore feature reduction is important and removal of unnecessary and same kind of features increases the accuracy and lowers the overfitting.
In Fig. 6, We can see that these 12 magnitudes('f923w','f892w','f861w','f830w','f799w','f768w','f737w','f706w','f675w','f644w', 'f613w','f489w') are highly correlated with F814w, with corrl. coefficient > 0.94. Therefore dropping these magnitudes (also in Fig. 4. we can see that their importance is too low and almost same), we left with only 13 features. Now running CNN, AUC increases to 0.993 for objects brighter than 22.5, with even less overfitting. It gave an accuracy of 96%.

Therefore we can conclude that (FWHM, f814w, ks, h, j, f954w, f551, f520, f458, f427, f396, f365) these 13 magnitudes are enough to classify star or galaxy with 96% accuracy and 0.993 AUC score and much less CPU time. It can be interpreted that CNN was perfectly able to capture the spectrum of stars and galaxies resp.



	features	ANN	Random Forest	Ada boost	CNN
0	optical bands	0.898172	0.969176	0.902258	0.983675
1	optical bands+NIR	0.978597	0.984727	0.972175	0.983418
2	optical bands+NIR+F814	0.982811	0.982990	0.967277	0.985758
3	optical bands+NIR+F814+FWHM	0.986883	0.989031	0.986019	0.986477

Fig. 2. Here we can see that CNN is constantly giving AUC > 0.983, unlike others. And with increase in features (espesially FWHM), increase in AUC is quite significant

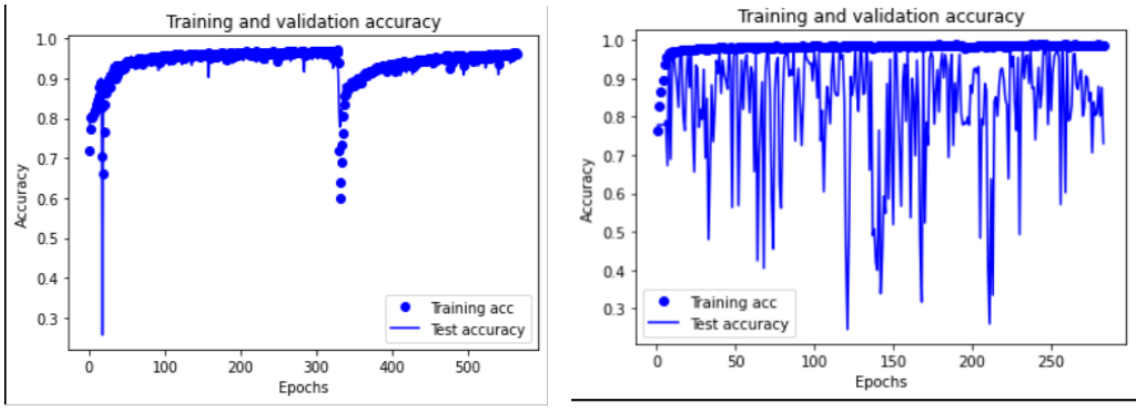


Fig. 3. These are for objects brighter than 22.5. In case of CNN, Training and Test accuracy is nearly same for all epochs, while this is not so in ANN (too many downfalls). Therefore it shows that CNN is not overfitted and captured the relation between feature and target variable very well

7. INTRODUCING COLORS

From ranking of features in Fig. 4, we see that fwhm is the most important feature followed by K band and 20 narrow bands are quiet low in importance. Therefore it will be good to see that whether colors be more suitable than magnitudes for classification. So now we will use 21 colors (difference of consecutive magnitudes) + f814w + fwhm on our best algorithm i.e. CNN.

See Fig. 7. We get an AUC of 0.99 for objects brighter than 22.5 and 0.953 for objects than 26. Again the later one is little low because of those 100-200 stars in 23 <f814w< 26, producing contamination.

In Fig. 8, we see that optical bands are providing good classification, AUC= 0.989 and after using all features, 0.99. CNN still captured the relation well without being overfitted and colors are proven to be more important than magnitudes for classification.

A. Feature importance and Reduction

In Fig. 9, we can see that some colors are important in classification unlike magnitudes, f369-f427 is even important than FWHM. Most important features are f396-f427, FWHM, f814w, f365-f396, f458-f459, f768-f799. Now we ran CNN on these 6 features only and got an AUC score of 0.9902, same as when all colors were used but overfitting has reduced to a greater extent. See Fig. 10.

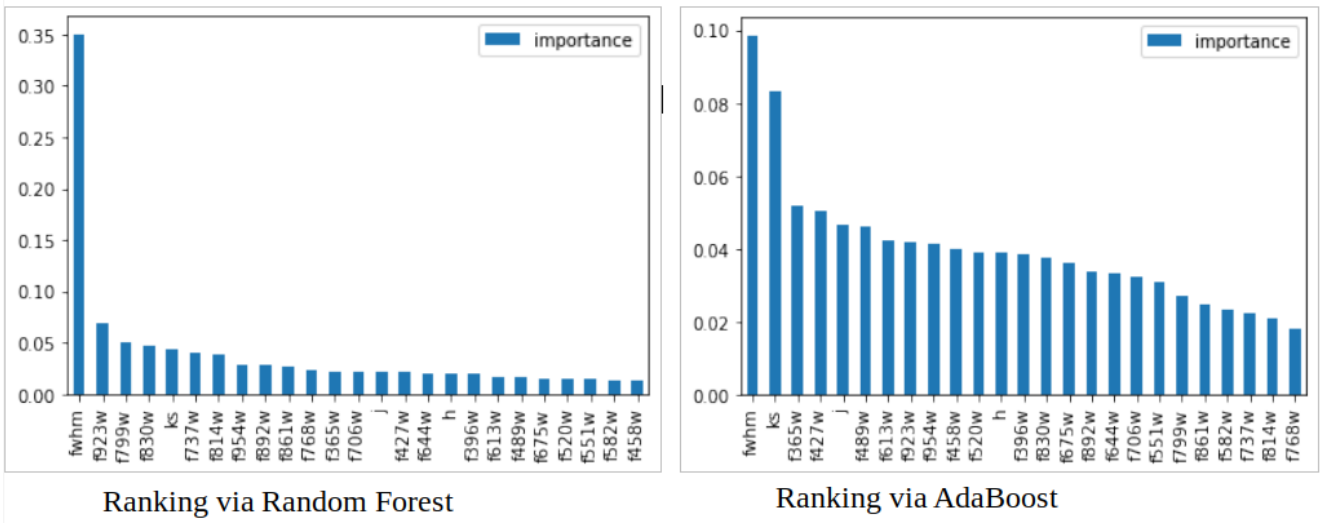
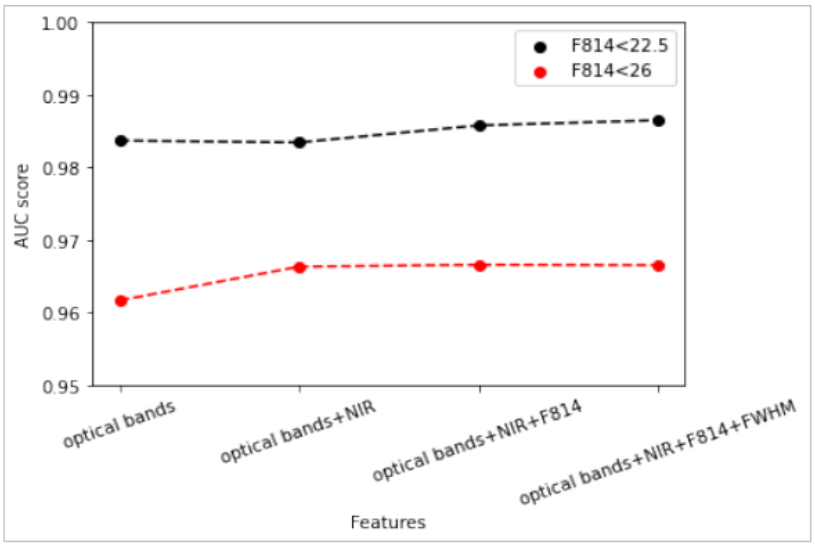


Fig. 4. First graph (randomforest) shows that FWHM is the most important feature in classification, but not others (Narrow bands). Therefore second graph is also used, as Ada Boost make weak learners strong. Therefore, some of these magnitudes, based on correlation heatmap, can be removed as they do nothing except increasing error in classification. Also, it will be good to see whether colors instead of magnitudes be more useful and important in classification and comparable to fwhm



F814	AUC score
F814<22.5	0.986
22.5<F814<23	0.878
23<F814<24	0.83
24<F814<25	0.714
F814<26	0.962

Fig. 5. CNN results. At right side, we see that AUC scores for fainter objects are low, it can be because there are not enough stars(100-300) in these ranges to train our model

Therefore these six features (out of which 4 are colors) are enough to produce better results with same accuracy, less overfitting and much less CPU time.
Here we also saw that most important colors i.e. f369-f427, f396-f427, f365-f396 consists of magnitudes of low wavelengths.

8. CONCLUSION

As a result we see that CNN was perfectly able to classify an object as star or galaxy with 96% accuracy (can still be increased if we consider all objects fainter than 23 as galaxies- those 200 stars can be taken as contamination) and AUC score of 0.9902. Here we were perfectly able to classify object as star or galaxy with much less overfitting and CPU time using 19 colors instead of 20 magnitudes directly and then reducing features. We used only 6 important features, which consists of FWHM, f814w and four colors- f396-f427, f365-f396, f458-f459 and f768-f799, mostly of low wavelengths. This way, we can classify any object, even those for which morphology based classification is not possible

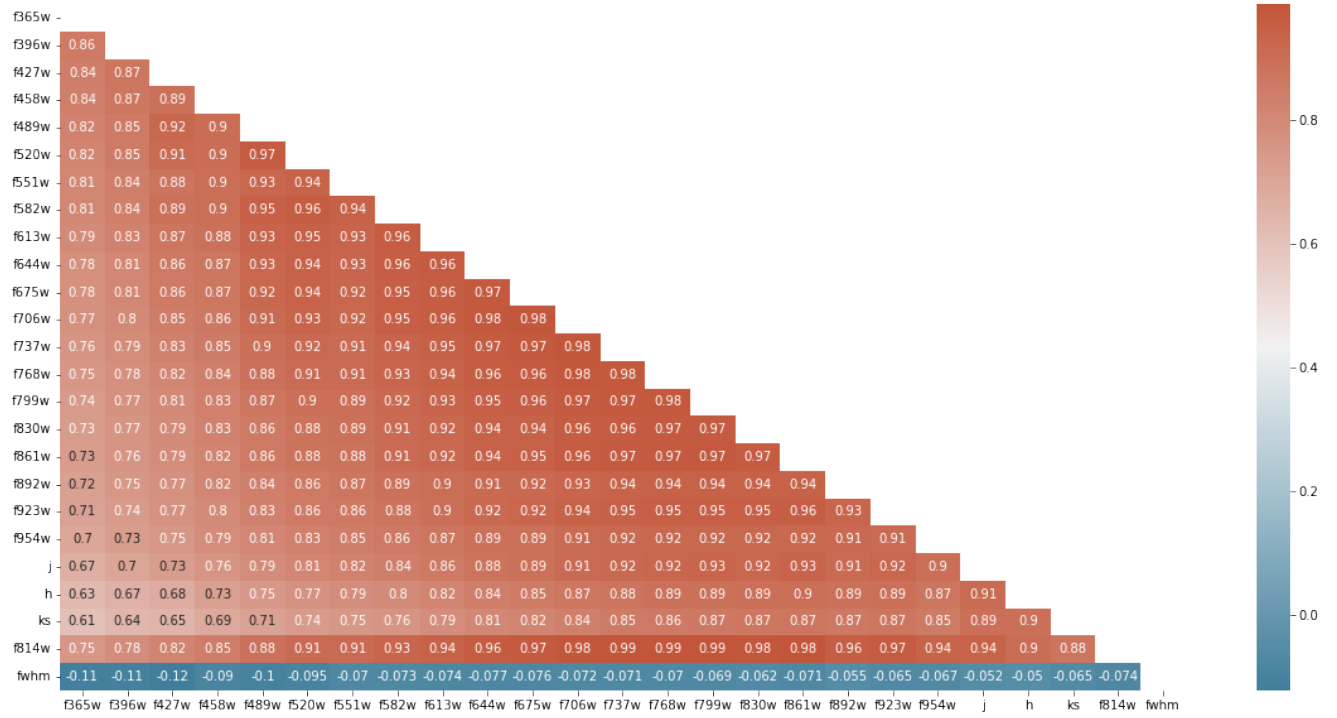
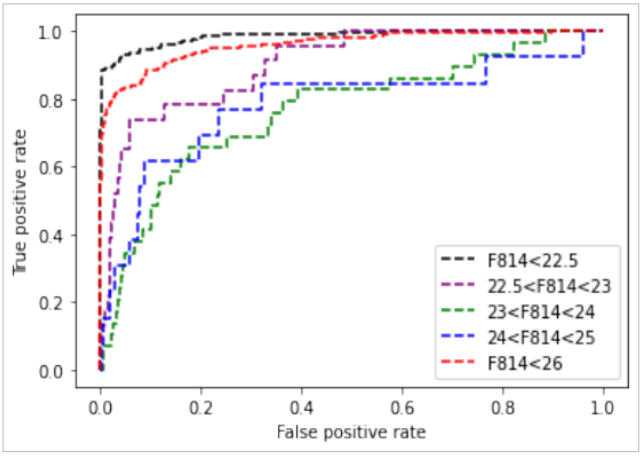
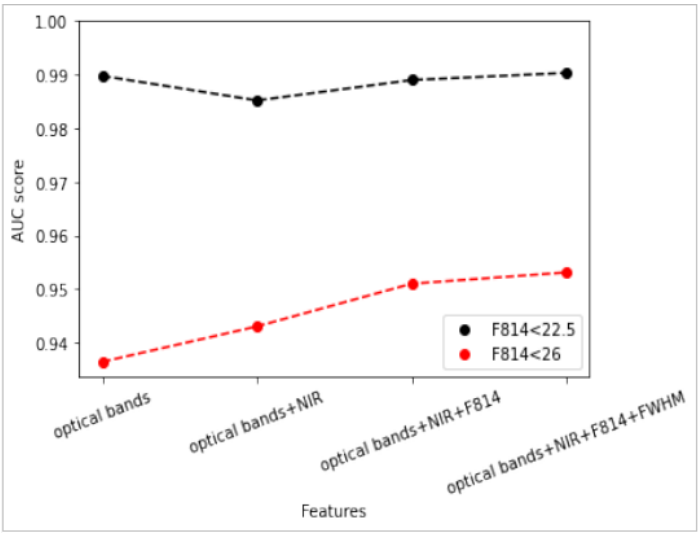


Fig. 6. Correlation Heatmap of features(magnitudes). It shows that Magnitudes, especially of higher wavelengths, are highly correlated to F814w. Therefore, we will drop the features with correl. coefficient >0.94, resulting in just 13 magnitudes as features



F814	AUC score
F814<22.5	0.990
22.5<F814<23	0.902
23<F814<24	0.766
24<F814<25	0.782
F814<26	0.959

Fig. 7. ROC using CNN with colors as features. Here we can see that these ROC curves are more closer to 1 and AUC scores are better than those obtained using magnitudes as features.



	information used	F814<22.5	F814<26
0	optical bands	0.989728	0.936417
1	optical bands+NIR	0.985151	0.943020
2	optical bands+NIR+F814	0.988958	0.950983
3	optical bands+NIR+F814+FWHM	0.990289	0.953087

Fig. 8. Here we can see that CNN is still consistent in results. This time it is giving AUC (0.989 with only optical bands) better than with only magnitudes in previous sections.

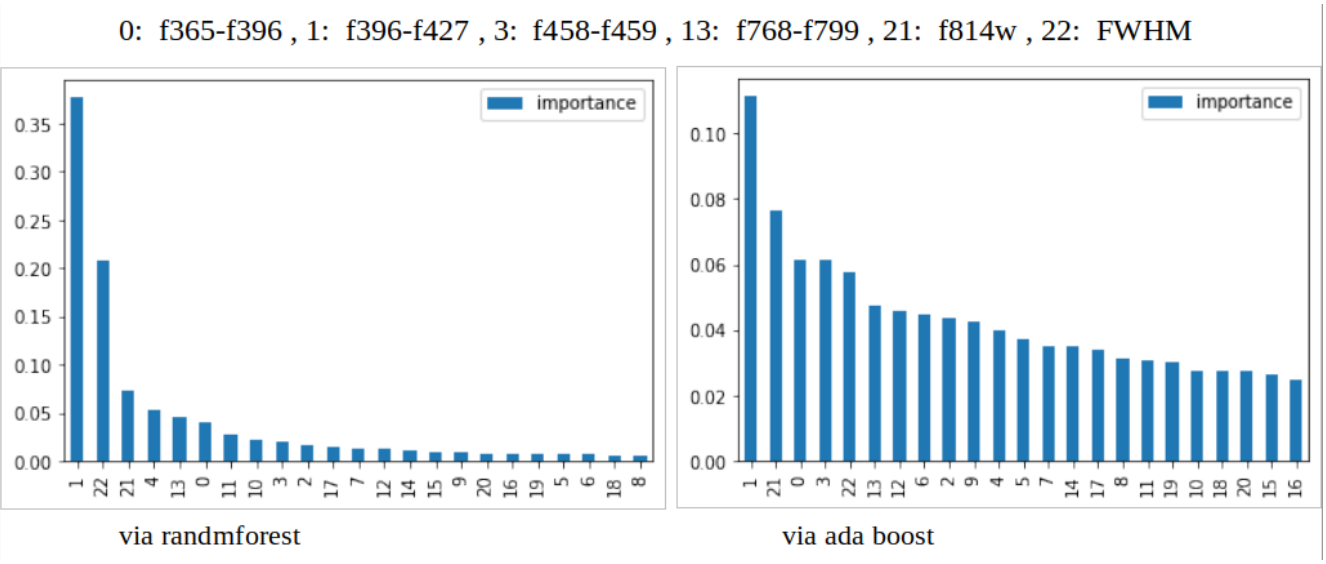


Fig. 9. Now here we see that f396-f427 becomes the most important feature followed by FWHM, f814w, f365-f396, f458-f459 and then f768-f799 at last. we see that magnitudes of lower wavelength are more important in classification.

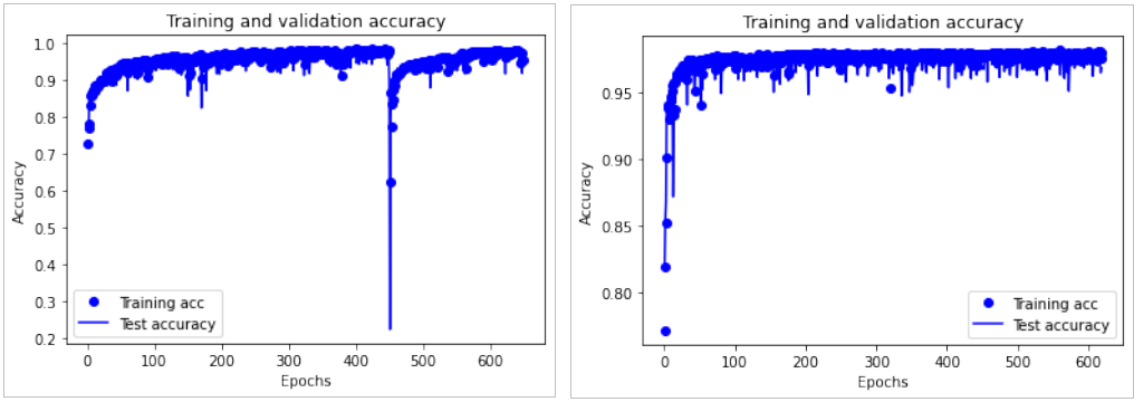


Fig. 10. training and test accuracy for CNN. Although not much difference but right one is still better

REFERENCES

ALHAMBRA and COSMOS data downloaded from- https://cosmohub.pic.es/catalogs
Research paper used as reference for this work- https://academic.oup.com/mnras/article/483/1/529/5188687?login=false
Molino et al. 2014- https://doi.org/10.1093/mnras/stu387