भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

B.Tech. Project - Fall 2022

Star-Galaxy classification using ALHAMBRA Photometry and Machine Learning Algorithms

PRATYAKSH RAJ
EP20BTECH11017
Engineering Physics
Indian Institute of Technology, Hyderabad

Supervisor
Dr. Shantanu Desai
Department of Physics
Indian Institute of Technology, Hyderabad

Contents

1 Introduction

It is known that classifying an object as a star or galaxy, based on their morphology, is sometimes a difficult task from earth-based telescopes as their images get blurred due to plenty of reasons.

The first classification methods were morphology based and they consisted of the estimation of an optimal cut on the space of observable image properties. However, these methods perform poorly when classifying faint objects as morphological information contained in noisy measurements is limited. Other classification methods use Bayesian-based approaches. The application of a Bayesian classification approach to multiband data must consider information coming from their morphologies and colour: the morphological features of a galaxy will be correlated with its magnitude. Hence, as the number of photometric bands increase, this approach gets more and more complicated.

Therefore, here we have used ML algorithms (Convolutional NN, XGBoost, Artificial NN, Random forest, Ada Boost) to classify an object as star or galaxy over ALHAMBRA-4 field. We have used ALHAMBRA survey-isophotal magnitudes (20 Narrow band, 3 NIR broad band, F814, FWHM) as features over COSMOS field, taking Hubble's Morphology based classification as ground truth. There are a total 31870 objects matched to our reference catalog- 29,615 galaxies and 2247 stars for f814w$\leq$26. And for f814w$\leq$ 22.5, there are 5998 galaxies and 1765 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 and XG Boost comes out to be best, with former one being much faster. We also checked the performance of each feature (Narrow bands, broad bands, f814w, fwhm) by gradually adding them. Then we try to reduce overfitting and CPU time significantly by feature reduction. After that we have used colors instead of magnitudes to see that they are better classifiers. We also showed the rankings of features at each stage. At last, we have showed the AUC scores using best and reduced features for CNN and XGBoost respectively.

2 ALHAMBRA survey

The main goal of the ALHAMBRA survey is to probe Cosmic Evolution. In order to achieve this goal it is necessary to cover cosmological meaningful volumes at all redshifts for which a large area coverage and good depth are needed. Besides, good spectral resolution is fundamental to identify the different populations of objects and large spectral coverage is important to sample enough redshift range and allow easier identifications.This survey aims at covering a total of 8 square degrees.
We have used the ALHAMBRA catalogue [1] over the ALHAMBRA-4 field, which overlaps with COSMOS. It contains 31870 objects matched to our reference COSMOS catalogue with a maximum separation of 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$=26.5$. In the ALHAMBRA catalogue whenever a source was not detected in a given band, its magnitude was set to a 'sentinel' value of 99. And for objects which can not be classified using their sextractor (photometric fluxes and morphology based), specially fainter than 22.5, they have assigned a value of 0.5 in their Gold catalog.

3 Hubble Telescope survey

Here we will use Hubble's morphology based classification as ground truth. COSMOS space-based imaging catalogue [2], which provides a morphology-based classification (MU_CLASS) for the objects to train and test our methods on. It contains 1.2×106 objects to a limiting magnitude of F814W $= 26.5$ from images observed with the Hubble Space Telescope (HST) using the Advanced Camera for Surveys (ACS)4, therefore its image quality (very deep and unaffected by the atmosphere) can be used as a 'truth' reference. Images were taken through the wide F814W filter (I).

The catalogue contains, roughly, 1.1×106 galaxies, most towards the faint end, 30 000 stars and the rest are fake detections.

4 Data Wrangling and Metrics

We will take only those objects for which number of unknown magnitudes are less than 5. Here, missing magnitude of an object is filled with median(rounded off to 3 decimal places) of that magnitude's column. Also, as ALHAMBRA do not provide any classification for many objects fainter than 22.5 (they assign a probability of 0.5), its a good practice to consider all those objects as galaxies as they are very few (150-250)stars in this range as compared to galaxies but here we will also ran our algorithms on different ranges of F814w, without doing any changes.
Here we have used AUC scores and ROC curves for measuring the performances of our algorithms. AUC - ROC curve is a performance measurement for the classification problems at various threshold settings (here 0.5). ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1. in our case 0 is galaxy and 1 is star. ROC is made between False positive rate along x-axis and True positive rate along y-axis. And AUC score is the area under this curve.

5 Machine learning algorithms and hyperparameterization

We have used XG Boost, CNN, ANN, RandomForest, AdaBoost. 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. They both are highly important in getting good ranking of features. Also Ada Boost and XG Boost are helpful in getting the complex relations as they make weak learners strong. We will see that CNN will work best and is least overfitted. Then we will also check XG Boost, it is based on gradient boosting and is known to outperform neural nets and Randomforest, quite commonly.
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 and XG Boost via GridSearchCV with 3 fold cross-validation. These are:-

XG Boost- γ= 0.2, learning rate= 0.3, max depth= 4, n estimators =80, reg alpha= 1, scale pos weight= 1

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)

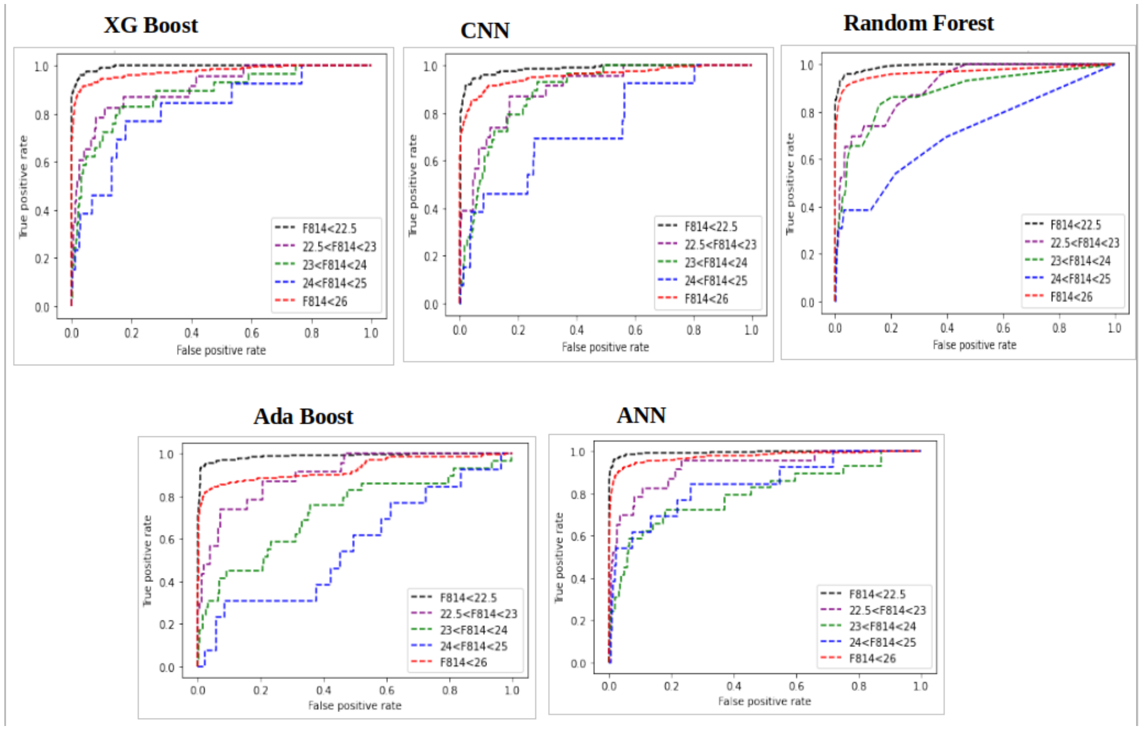


Figure 1: ROC curves for different ranges of F814W, using all 25 magnitudes as features. In magnitude ranges 22.5-23, 23-24 and 24-25, there are 4000 to 9000 galaxies and 100 to 250 stars in each range

6 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 fig2 . For only optical bands as features, CNN gives highest AUC 0.983 for f814w$\leq$ 22.5 and 0.96 for f814w$\leq$ 26. Followed by XG Boost with an AUC of 0.97 for f814w $<$ 22.5 .

Also, we can even see that CNN is not overfitted whereas ANN is highly overfitted, in fig3 .

RandomForest also give high AUC scores but performance drops with only optical bands(AUC of 0.957). Also ranking of features shows that it depends highly on FWHM (very less on other magnitude, See fig5), therefore not as useful as CNN . And Ada boost shows the least AUC in every case.

Therefore it can be concluded that all algorithms except CNN and XG Boost are highly dependent on FWHM. CNN is consistent without being overfitted.

Therefore CNN comes out to be the best algorithm without FWHM. And XG Boost while taking fwhm in consideration with AUC of 0.994. AUC scores for different ranges of F814w, for CNN are shown in fig4.
Also in fig2 and fig8 we can see that NIR Broad bands are making a high improvement in classification as already discussed in [3] and [4], but not in case of CNN as classification rate is already too high to improve.

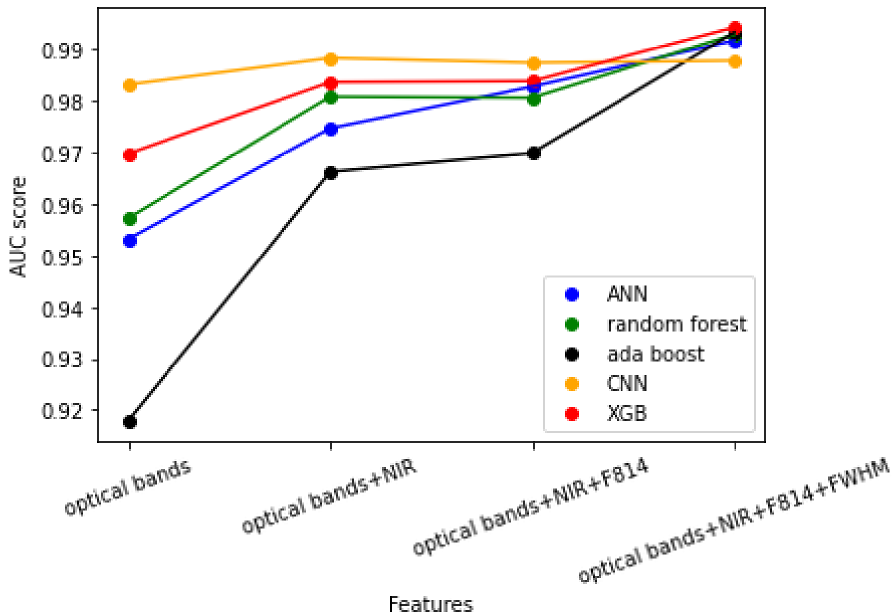
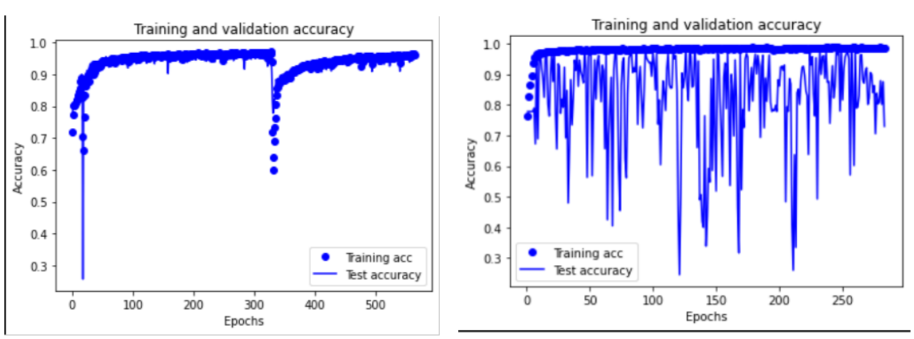
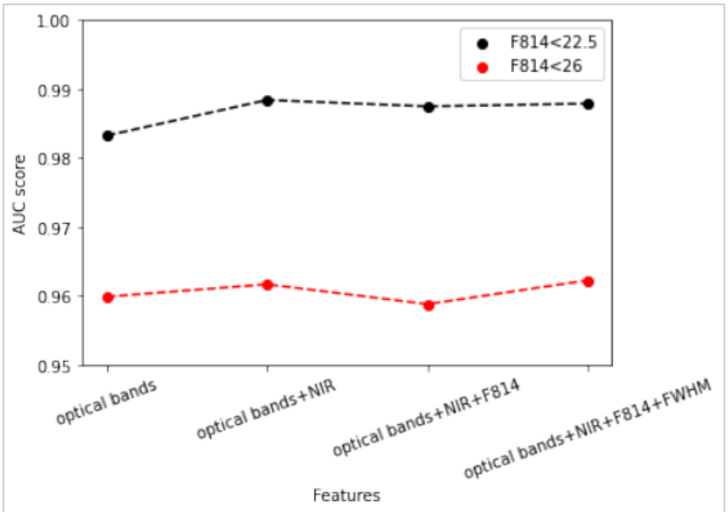


Figure 2: this graph is for f814w$\leq$ 22.5. CNN shows highest AUC 0.983 with only optical band input and XG boost shows highest AUC 0.994 when FWHM is added. We can see that CNN captured the spectrum successfully, much more than any other algo. Also, with all features and objects brighter than 26, CNN gives an accuracy of 95.7% and XG Boost gives 97.8%



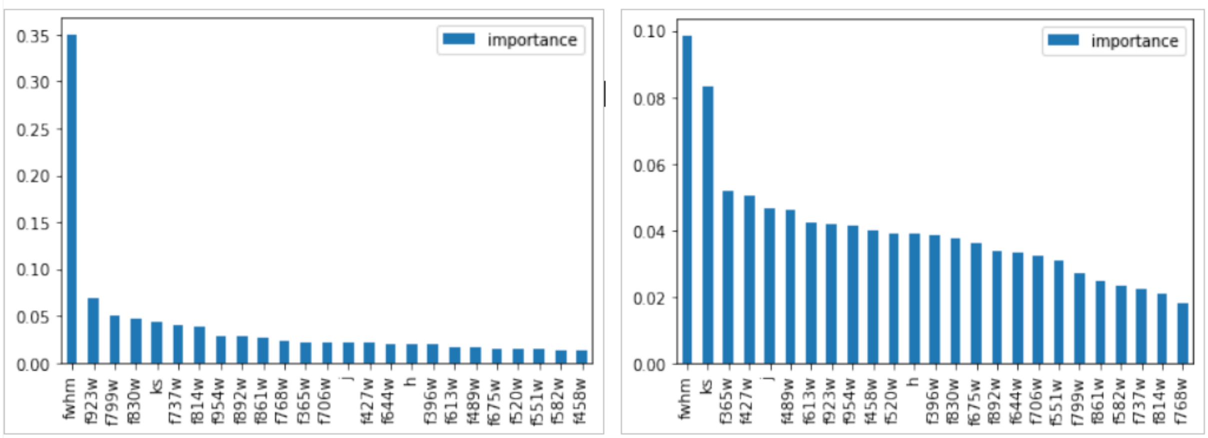
CNN ANN

Figure 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



F814W	AUC SCORE
F814W≤ 22.5	0.989
22.5 <F814W≤ 23	0.902
23 <F814W≤ 24	0.89
24 <F814W≤ 25	0.735
F814W≤ 26	0.96

Figure 4: Roc Auc for CNN. At right side, we see that AUC scores for fainter objects are low, it can be because there are not enough stars (100-250) in these ranges, they act as contamination to our model as these objects are generally galaxies. Also, as FWHM is most important feature but CNN captured the spectrum well and give the same AUC with nearly no wrongly classified galaxy.



Ranking via Random Forest

Ranking via AdaBoost

Figure 5: 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 complexity and CPU time 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

7 Feature Reduction

Training 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 fig6, 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 correlation coefficient > 0.94. Therefore dropping these magnitudes (also in fig5. we can see that their importance is too low and almost same), we are left with only 13 features.

Now running CNN, AUC increases to 0.997 for objects brighter than 22.5, and for XG Boost 0.9958.

Therefore we can conclude that (FWHM, f814w, ks, h, j, f954w, f551, f520, f458, f427, f396, f365) these 13 magnitudes of which 8 are of narrow bands are enough to classify star or galaxy with even high AUC score and much less CPU time for CNN. XG Boost is already quiet fast.

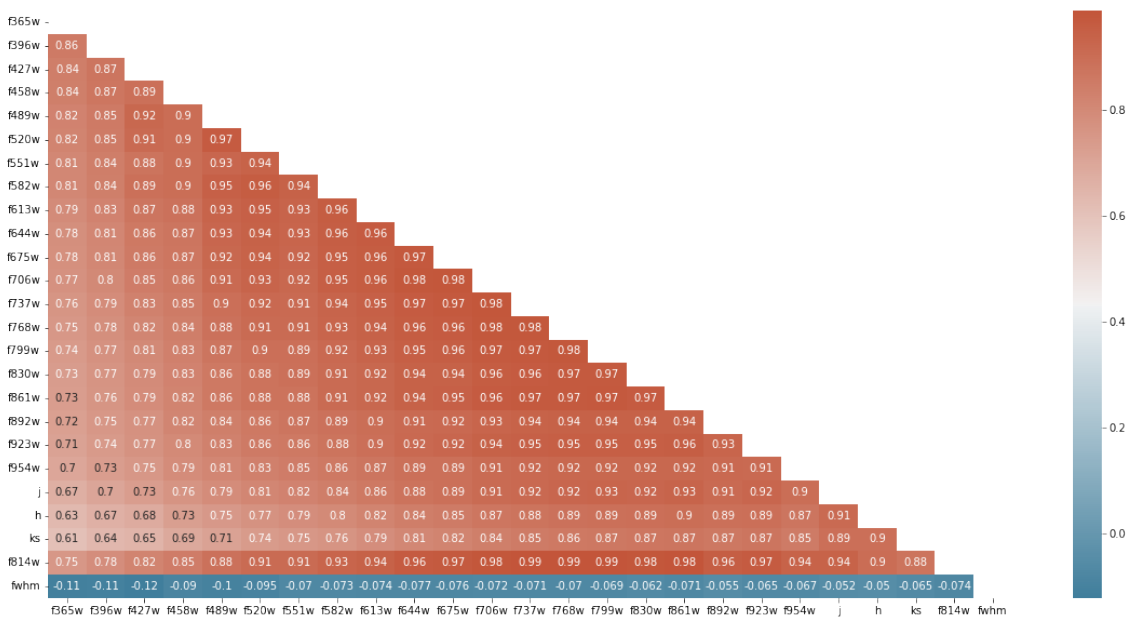


Figure 6: Correlation Heatmap of features (magnitudes). It shows that Magnitudes, especially of higher wavelengths, are highly correlated to F814w. Therefore, we will drop the features with correl. coefficient >0.94, resulting in just 13 magnitudes. Although, its a common practice to drop all features with corr coefficient > 0.9, but here it may lead to too less features, can result in underfitting

8 Introducing colors

8.1 Using colors on best algos. CNN and XG Boost

From ranking of features in fig5, 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 19 colors (difference of consecutive 20-bands)+ 2 colors(i-h, h-k) + f814w + fwhm on our best algorithms i.e. CNN and XG Boost
See fig7. For CNN, we get an AUC of 0.994 for objects brighter than 22.5 and 0.953 for objects brighter than 26. For XG Boost, we get an AUC of 0.995 and 0.978 resp. Again the later one in both cases a little low because of those 100-200 stars, producing contamination as they are very few in number.
In fig8, we see that optical bands (19 colors) are providing good classification, AUC= 0.994 and after using all features, 0.995, for XG Boost. Same results for CNN but is very slow.
Therefore colors are proven to be more important than magnitudes and Fwhm for classification. CNN gives an accuracy of 96.2% and XG Boost 98.6%, compared with Alhambra's classification. Such a high difference in accuracy is just because of those few stars in fainter ranges, which XG boost correctly classified.

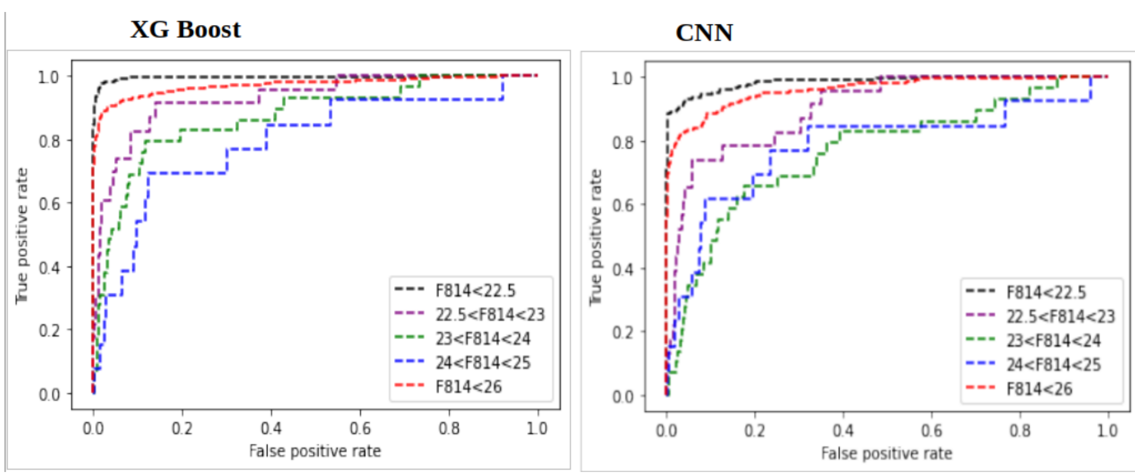


Figure 7: This time (with colors) Curves are more closer to one and better, corresponding to their Roc's in previous sections. Also XG Boost is performing almost equivalently to CNN for brighter objects but is better for fainter objects as it was able to classify those few stars better than CNN

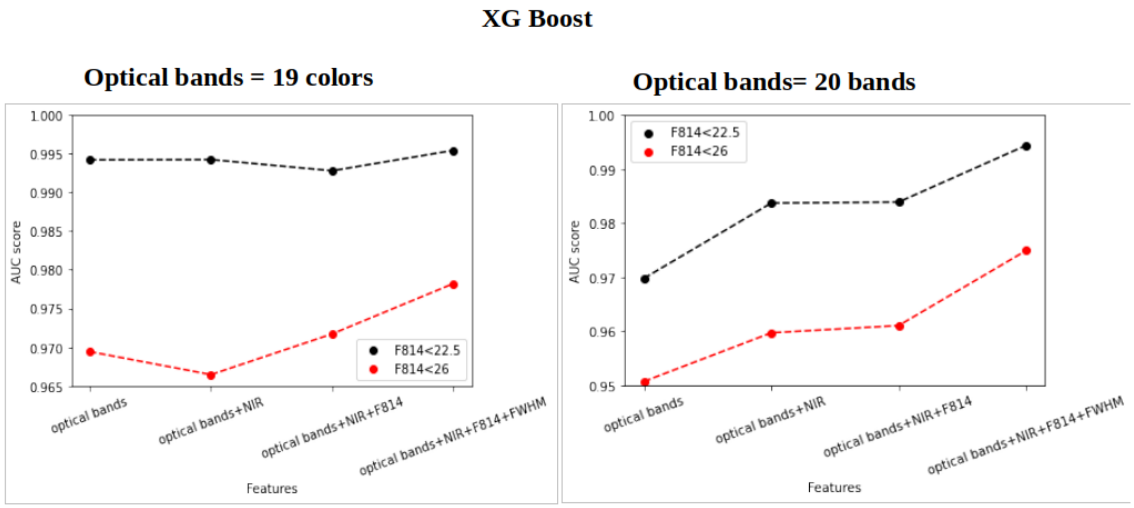


Figure 8: Here we can see that in right side, auc(0.995) which was obtained by adding FWHM can be obtained just by using colors, on left.

8.2 Feature importance and Reduction

In fig9, we can see that some colors are very 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 and XG Boost on these 6 features only and got an AUC score of 0.995 in both cases, same as when all features were used but overfitting has reduced to a greater extent. See fig10.
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.

9 Conclusions

As a result we see that only CNN was perfectly able to capture the spectrum (20-narrow bands) and classify an object as star or galaxy with a very high AUC score of 0.997 (after feature reduction) for f814w<22.5. But XG Boost performed better with colors with AUC of 0.995 and very fast.
Also, XG Boost was perfectly able to classify the objects fainter than 22.5 using colors, with an accuracy of 98.6% for f814w<26.

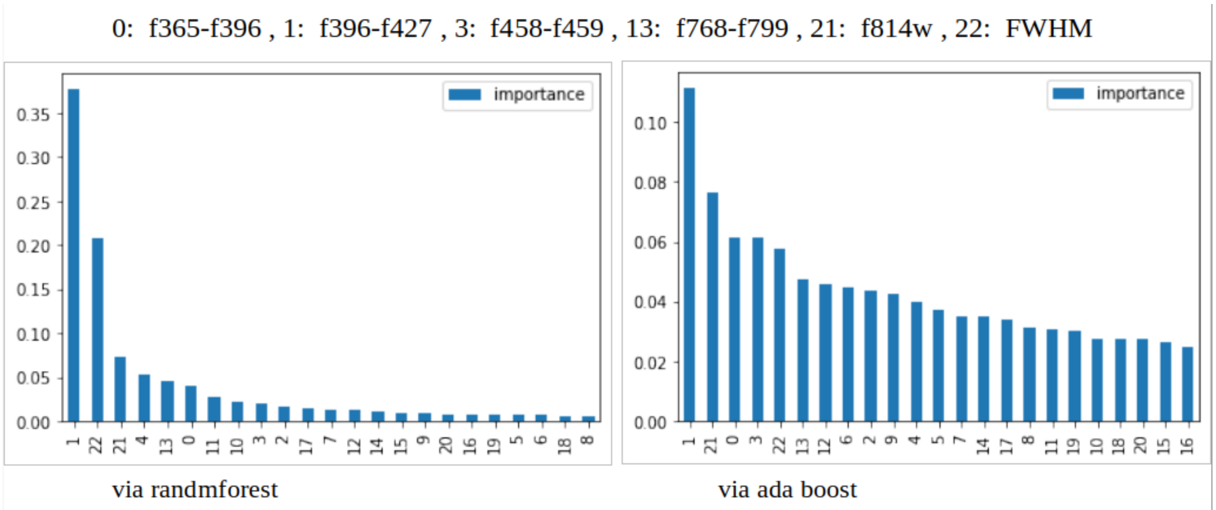


Figure 9: Now here we see that f396-f427 becomes the most important feature followed by FWHM, f814w, f365-f396, f458-f459 and then f768-f799 at last. we see that magnitudes of lower wavelength are more important in classification.

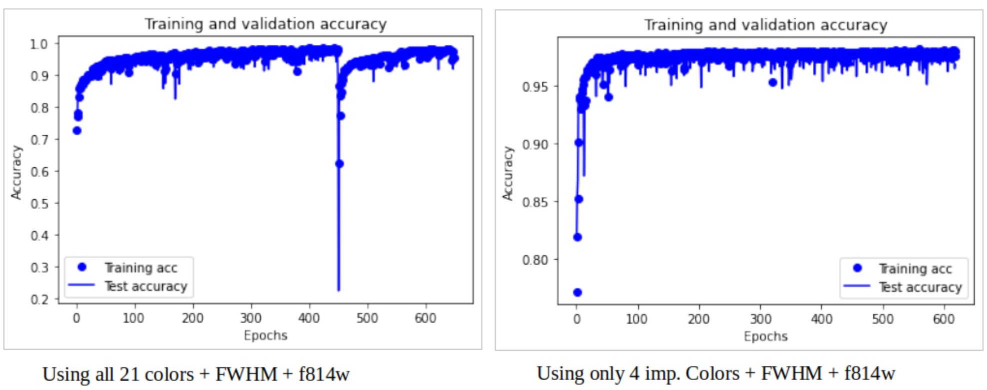
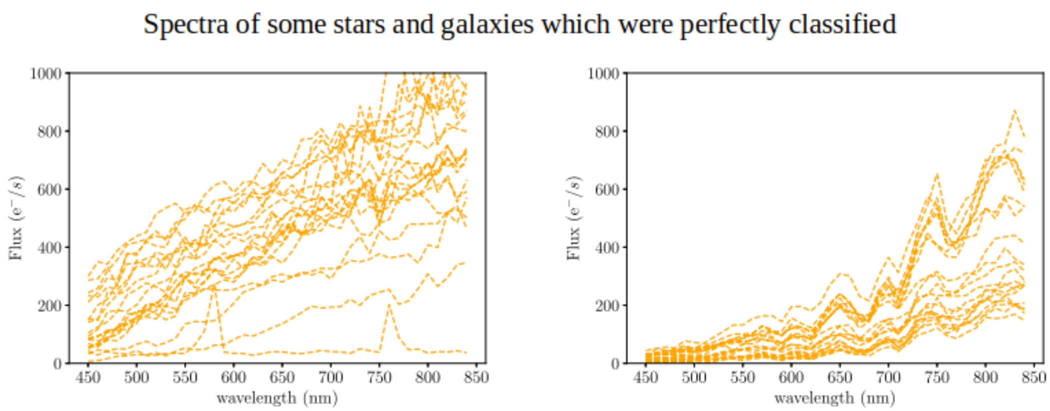


Figure 10: training and test accuracy for CNN. We see that overfitting has reduced significantly in right fig.

Therefore, here we were perfectly able to classify object as star or galaxy with much less overfitting and CPU time using only 6 features instead of 23. 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 (with nearly no misclassified galaxy).

Best models with features and AUC scores for different brightness ranges are given in fig12



(a) Galaxies. (b) Stars.

Figure 11: taken from [5] for PAU survey. Lines are more steep for galaxies in low wavelength ranges. Therefore, fluxes specially of low wavelengths differentiates stars and galaxies better. Similar to what our algorithms showed in above sections

CNN

F814W	AUC SCORE
F814W≤ 22.5	0.997
22.5 <F814W≤ 23	0.902
23 <F814W≤ 24	0.89
24 <F814W≤ 25	0.736
F814W≤ 26	0.962

XG Boost

F814W	AUC SCORE
F814W≤ 22.5	0.995
22.5 <F814W≤ 23	0.93
23 <F814W≤ 24	0.871
24 <F814W≤ 25	0.791
F814W≤ 26	0.978

Figure 12: Auc scores of CNN are for 8 narrow bands + 3 broad band + f814w + fwhm. Auc scores of XG Boost are for 4 imp. colors (narrow bands) + f814w + fwhm

REFERENCES

[1] A. Molino, N. Benítez, M. Moles, A. Fernández-Soto, D. Cristóbal-Hornillos, B. Ascaso, Y. Jiménez-Teja, W. Schoenell, P. Arnalte-Mur, M. Pović, D. Coe, C. López-Sanjuan, L. A. Díaz-García, J. Varela, M. Stefanon, J. Cenarro, I. Matute, J. Masegosa, I. Márquez, J. Perea, A. Del Olmo, C. Husillos, E. Alfaro, T. Aparicio-Villegas, M. Cerviño, M. Huertas-Company, J. A. L. Aguerri, T. Broadhurst, J. Cabrera-Caño, J. Cepa, R. M. González, L. Infante, V. J. Martínez, F. Prada, and J. M. Quintana. The ALHAMBRA Survey: Bayesian photometric redshifts with 23 bands for 3 deg2. *Monthly Notices of the Royal Astronomical Society*, 441(4):2891–2922, 05 2014. ISSN 0035-8711. doi: 10.1093/mnras/stu387. URL https://doi.org/10.1093/mnras/stu387.

[2] Alexie Leauthaud, Richard Massey, Jean-Paul Kneib, Jason Rhodes, David E. Johnston, Peter Capak, Catherine Heymans, Richard S. Ellis, Anton M. Koekemoer, Oliver Le Fèvre, Yannick Mellier, Alexandre Réfrégier, Annie C. Robin, Nick Scoville, Lidia Tasca, James E. Taylor, and Ludovic Van Waerbeke. Weak gravitational lensing with cosmos: Galaxy selection and shape measurements. *The Astrophysical Journal Supplement Series*, 172(1):219, sep 2007. doi: 10.1086/516598. URL https://dx.doi.org/10.1086/516598.

[3] Manda Banerji, S. Jouvel, H. Lin, R. G. McMahon, O. Lahav, F. J. Castander, F. B. Abdalla, E. Bertin, S. E. Bosman, A. Carnero, M. Carrasco Kind, L. N. da Costa, D. Gerdes, J. Gschwend, M. Lima, M. A. G. Maia, A. Merson, C. Miller, R. Ogando, P. Pellegrini, S. Reed, R. Saglia, C. Sánchez, S. Allam, J. Annis, G. Bernstein, J. Bernstein, R. Bernstein, D. Capozzi, M. Childress, Carlos E. Cunha, T. M. Davis, D. L. DePoy, S. Desai, H. T. Diehl, P. Doel, J. Findlay, D. A. Finley, B. Flaugher, J. Frieman, E. Gaztanaga, K. Glazebrook, C. González-Fernández, E. Gonzalez-Solares, K. Honscheid, M. J. Irwin, M. J. Jarvis, A. Kim, S. Koposov, K. Kuehn, A. Kupcu-Yoldas, D. Lagattuta, J. R. Lewis, C. Lidman, M. Makler, J. Marriner, Jennifer L. Marshall, R. Miquel, Joseph J. Mohr, E. Neilsen, J. Peoples, M. Sako, E. Sanchez, V. Scarpine, R. Schindler, M. Schubnell, I. Sevilla, R. Sharp, M. Soares-Santos, M. E. C. Swanson, G. Tarle, J. Thaler, D. Tucker, S. A. Uddin, R. Wechsler, W. Wester, F. Yuan, and J. Zuntz. Combining Dark Energy Survey Science Verification data with near-infrared data from the ESO VISTA Hemisphere Survey. *Monthly Notices of the Royal Astronomical Society*, 446(3):2523–2539, 11 2014. ISSN 0035-8711. doi: 10.1093/mnras/stu2261. URL https://doi.org/10.1093/mnras/stu2261.

[4] András Kovács and István Szapudi. Star–galaxy separation strategies for WISE-2MASS all-sky infrared galaxy catalogues. *Monthly Notices of the Royal Astronomical Society*, 448(2):1305–1313, 02 2015. ISSN 0035-8711. doi: 10.1093/mnras/stv063. URL https://doi.org/10.1093/mnras/stv063.

[5] L Cabayol, I Sevilla-Noarbe, E Fernández, J Carretero, M Eriksen, S Serrano, A Alarcón, A Amara, R Casas, F J Castander, J de Vicente, M Folger, J García-Bellido, E Gaztanaga, H Hoekstra, R Miquel, C Padilla, E Sánchez, L Stothert, P Tallada, and L Tortorelli. The PAU survey: star–galaxy classification with multi narrow-band data. *Monthly Notices of the Royal Astronomical Society*, 483(1):529–539, 11 2018. ISSN 0035-8711. doi: 10.1093/mnras/sty3129. URL https://doi.org/10.1093/mnras/sty3129.

ACKNOWLEDGEMENTS

I express my deep gratitude to Shantanu Sir for his constant guidance and support.