Final Reflection Report

I. General Information

Project Title: Applying Machine Learning/deep learning to Astronomy
Date submitted: March 15th

II. Discuss your accomplishments and experience in the project. You should comment on the following areas.

a) Describe your most important accomplishments in this project. What milestones have been achieved. What remains to be done.

Our most important accomplishments in this project are solving data imbalance problems, while preprocessing the image datasets; and making the decision to implement One Hot Encoding after plenty of research and discussions. Our dataset has 3 minority classes with very few images which contains tens of examples, while there are hundreds or thousands of examples in other classes, thus making it a difficult endeavor. To counter this, we used a traditional strategy—oversampling through data augmentation and then merged some minority classes into one to prevent imbalanced data that can cause potential issues during the training. One Hot Encoding was also implemented to allow the representation of categorical data to be more expressive.

Construction of our Convolutional Neural Network (CNN) model was one of the milestones we have achieved. We built the CNN model with 5 convolutional layers, each followed with a maxpooling layer. Then, we added a flatten layer and a dense layer with 256 units; then another dense layer with 4 units since there are 4 categories to be predicted. After all the jobs we've done, we tuned the hyperparameters which resulted in the highest validation precision of the epochs at 0.85—we all considered this our most important milestone so far.

The jobs remaining to be done are to fine tune the CNN model to make its user interface more user friendly.

b) Reflect on the work breakdown structure and timeline given in your project description. Did your project progress according to what was indicated? Discuss any deviations from the WBS that was projected.

According to the WBS of our project description, we should gather datasets of galaxy images, determine one as the best fit dataset, preprocess the image, and then develop the classifier. The total time we planned to spend on the dataset was around 20 days, but we spent longer time on this because of the data imbalance issue. This resulted in the team to do extra research and decide what strategy to use to solve the problems. We even took a 7-day crash-course named "Imbalanced Classification with Python" through the internet to learn how to deal with the dataset properly. Fortunately, the strategies we learned helped a lot in minimizing the influence of data imbalanced issue and hence improved the precision of our model.

c) Discuss your experience working on the project. What went right? What went wrong? Any other comments.

One of the most important decisions we had to make was about the methodology whether to use Random Forest Algorithm or Convolutional Neural Networks to implement the classifier. We decided to choose CNN after some research. This decision was considered as a proper decision by all of us since the procedure followed went smoothly. We did not use Random Forest Algorithm due to our research, the Random Forest Algorithm may fail to classify some images. Therefore, we chose CNN due to higher accuracy.

Unfortunately, when we decided to combine under sampling and oversampling this caused a decrease of precision since the model could not recognize this type of galaxies due to lack of samples. After disabled the under-sampling part, the precision of our model improved significantly.

d) Discuss the key takeaways from this experience.

One of our important lessons from this experience is to implement the practical application of our project as soon as possible to begin troubleshooting early. Instead of being in a planning phase inferring possible outcomes, before doing anything, it is better to act, fail, re-evaluate the outcome, and try again in order to find the best possible method for our project.

If we could go back to the very beginning of this project, we would probably do more research in the area of astronomic images analysis. Although the main part of the project is implementing the classifier through Machine Learning, figuring out the meaning and details of the galaxy features could definitely make us more confident when facing problems as which feature should we pay more attention or what class should we imposes an additional cost on during training if we want to improve the model to be more specific for some certain category of galaxies. We believe that this thought we gained would benefit us in the future—when it comes to the intersection of Computer science and other subjects.