Automating Rare Star Classification Using Hubble Space Telescope Photometry and Machine Learning Techniques

Team members: Boglarka Ecsedi, Dillon Greer, Jonathan Eubanks, Pranav Gudapati, Zachary Elis
External advisor: Rohit Raj

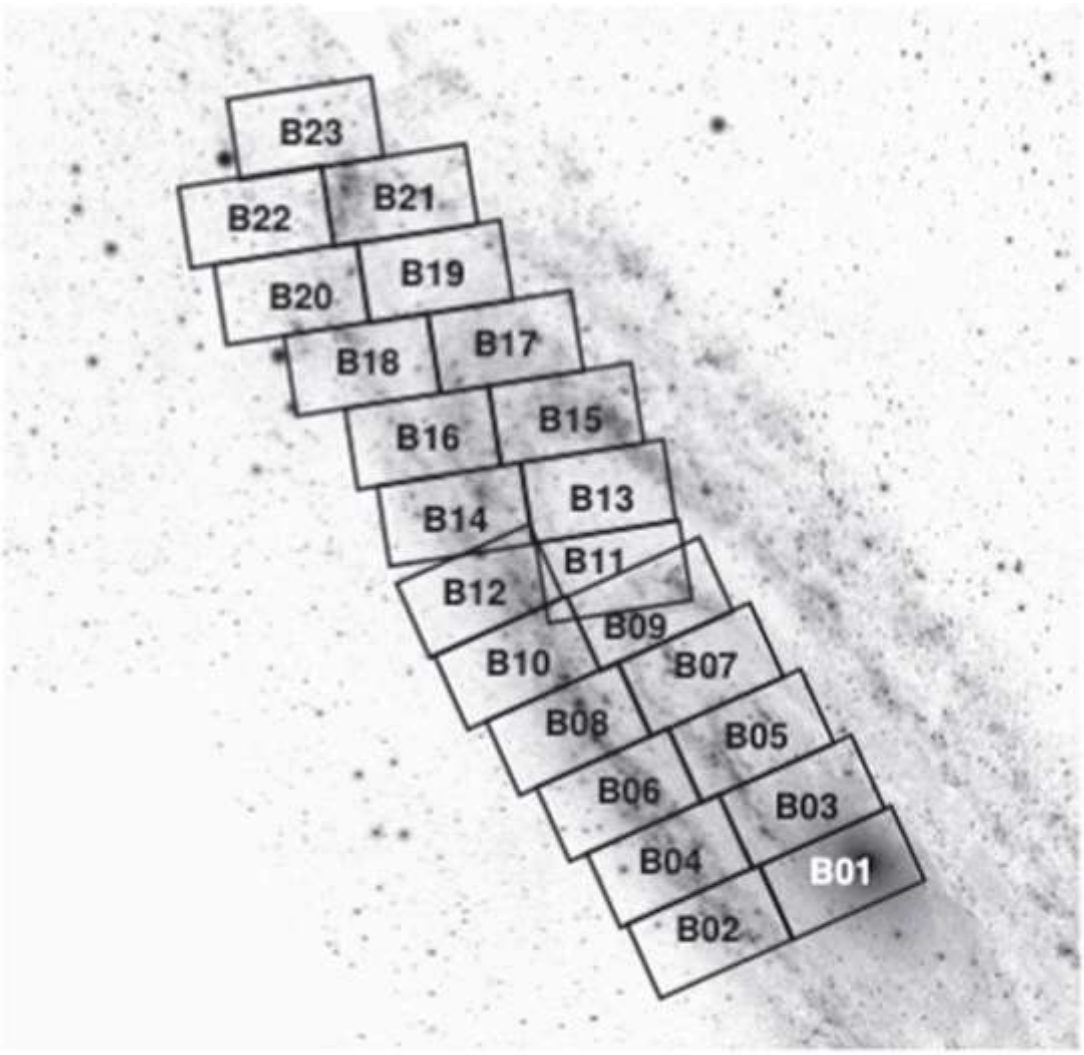


Figure 1. *The location of the 23 bricks of the PHAT survey. Each of these bricks consists of 18 HST pointings, each of which contains data in all three HST cameras. Figure adapted from [5].*

1. Introduction/Background

Wolf-Rayet (WR) stars are a rare type of massive star (more than 15-25x the size of the Earth's sun [8]), vital for understanding stages of galactic development. These stars burn rapidly and eject large portions of their mass through stellar winds, hence they have a short lifespan compared to typical stars. These winds are uniquely made of heavier elements like carbon, nitrogen and oxygen. Both the quantity and mass of the elements ejected from Wolf-Rayet stars directly contribute to the formation of new stars, planets and nebulae in their locality. WRs' larger sizes also mean that they die in violent supernovae that leave behind black holes or neutron stars. These properties make finding and studying Wolf-Rayet stars critical for understanding the heavy metal content and formation history of the galaxy.

The relative scarcity (670 confirmed, ~2000 estimated in the Milky Way [8] out of billions of stars) and locations of Wolf-Rayet stars in complex dusty regions make detecting them quite difficult. WRs are often located in areas highly obstructed by intergalactic dust which can block large frequency domains including much of the visible light range. This makes traditional techniques like searching for the unique emission spectra of nitrogen or oxygen difficult if not impossible for a large portion of WRs [8]. Photometry instead measures the total intensity of light across a frequency band which has proven to be an effective technique for locating WRs. With good quality photometry like from Hubble helps us make stellar tracks and isochrones which helps us estimate their mass and current evolutionary stage. These stars are also frequently located together in clusters, helping to identify massive regions of star formation [8].

Using machine learning techniques for analyzing stellar populations is a relatively new concept. Traditional machine learning techniques such as random forests and XGB classifiers have shown promise at identifying WR and other star types, claiming a 98% accuracy but struggle to classify WR subtypes claiming 78% accuracy [8]. These findings are so recent however that neither code nor data from this study has been published. Currently published studies have also primarily focused on data sets from the Milky Way galaxy.

Our dataset is a tabular, labeled dataset where each row represents a star, with features including location, brightness, photometric measurements across six filters, and color indices (differences in intensity across adjacent filters). These stars are from the Andromeda (M31) and M33 galaxies from the PHAT [6] and PHATTER [7] Hubble data sets respectively.

Table 1
Simplified Table of Our Photometry for Easy Use

R.A. (J2000)	Decl. (J2000)	F275W	S/N	G	F336W	S/N	G	F475W	S/N	G	F814W	S/N	G	F110W	S/N	G	F160W	S/N	G
10.57619538	41.24338318	99.999	−0.2	0	26.952	2.1	0	24.988	23.3	1	23.254	28.5	1	99.999	0.0	0	99.999	0.0	0
10.57620236	41.24336347	29.530	0.1	0	27.954	0.9	0	25.190	18.7	1	23.422	24.5	1	99.999	0.0	0	99.999	0.0	0
10.57623847	41.24342833	26.121	1.5	0	99.999	−0.4	0	26.978	4.5	1	24.184	12.9	1	99.999	0.0	0	99.999	0.0	0
10.57623862	41.24345726	99.999	−0.4	0	99.999	−2.4	0	26.256	11.0	1	24.611	11.5	1	99.999	0.0	0	99.999	0.0	0
10.57627747	41.24343760	99.999	−0.4	0	99.999	−0.7	0	25.951	13.4	1	24.117	15.5	1	99.999	0.0	0	99.999	0.0	0
10.57630315	41.24360886	26.079	1.3	0	99.999	−0.8	0	26.549	7.9	1	24.276	11.5	1	99.999	0.0	0	99.999	0.0	0
10.57631263	41.24340316	99.999	−1.5	0	29.410	0.2	0	26.995	5.6	0	25.168	6.8	0	99.999	0.0	0	99.999	0.0	0
10.57631284	41.24333622	99.999	−1.0	0	27.143	1.6	0	26.957	6.1	1	24.279	10.9	1	99.999	0.0	0	99.999	0.0	0
10.57631873	41.24347745	99.999	−0.8	0	99.999	−1.1	0	25.653	19.0	1	22.850	47.2	1	99.999	0.0	0	99.999	0.0	0
10.57631976	41.24358853	26.863	1.0	0	99.999	−0.0	0	26.260	8.2	1	24.131	13.2	1	99.999	0.0	0	99.999	0.0	0

Note. Given are the positions and Vega magnitudes. The first 10 lines are shown here. The full 138 million line version is available in a machine-readable format on Zenodo at doi:10.5281/zenodo.8147574.

Table 1. *Simplified Table of Photometry for Easy Use. Each row represents a star with corresponding features of location, brightness, and photometric measurements. Table adapted from [5].*

2. Problem Definition

The small population size, short lifespan, and frequent location within regions dense with galactic dust makes identifying Wolf-Rayet stars with spectroscopy alone extremely difficult if not impossible. Classical machine learning models have already shown promise at identifying and classifying these rare star types [8], however this is a relatively novel approach that has only been tested on a small number of datasets very recently. We intend to apply traditional machine learning techniques such as PCA to better understand our dataset, and models like XGBoost, Random Forest, SVM, and a simple MLP to our dataset representing stellar populations of the Andromeda [6] and M33 [7] galaxies with the goal of identifying and classifying Wolf-Rayet stars. To enhance performance, we could leverage ensemble and or deep learning approaches. We also believe that we can use similar techniques to analyze the locations of WR stars and predict possible locations of unidentified stars in the same survey, as well as in other surveys involving photometric measurements.

3. Methods

Dataset Curation

The PHATTER survey contains 21 980 690 data points, but it's all unlabeled. It was our task to curate a dataset for Wolf-Rayet classification and label the dataset. With the help of our external advisor, Rohit Raj, we identified the WR stars and sampled stars for the negative class ourselves. Initially, we sampled 500 non-WR stars randomly from a spectral region that doesn't contain WR stars at all, and performed our analysis. After seeing the results, we increased the non-WR class sample size to 1000, and finally, we decided to change our sampling approach because we realized that the bias we introduce by manually selecting very distinct data points results in a very easy task. To increase the task complexity, we sampled non-WR stars from a similar spectral region as the WR stars (we could do this because the chance that we accidentally sample a WR star from this region after excluding the known and verified WR stars is very low). This way, we curated a dataset with a little more ambiguity and we were curious to see how our models performed on this new, updated dataset which is more realistic for our proposed application.

Exploratory Data Analysis

To learn more about our dataset, we utilized Principal Component Analysis and t-SNE plots. We first plotted the full dataset, then after performing feature selection, we performed data visualization again.

We started with a 2-component PCA. When we ran the PCA after the feature selection (on all of the photometry data, including the FX_VEGA features, and their derived lightband features), we saw that our data was linearly separable [Figure 2]. We also visualized our dataset with scaling our features [Figure 3].

PCA Analysis

We decided on using a Principal Component Analysis to reduce the dimensionality of our data and learn more about the feature space. We looked into how using different features impact PCA, and since we received very low mutual information scores (see Feature Ranking) for the location features, we decided to do the PCA analysis with [Figure 4] and without them [Figure 2, 3]. The results of PCA mirrored the feature ranking findings, including the location features made our feature space more complex. Leaving them out, however, resulted in clearly separable classes. We performed a 2-component PCA. We also experimented with scaling, and visualized our data with [Figure 3, 4] and without scaling [Figure 2]. These findings carried over into the new dataset which created a better barrier for us to test with [Figure 5].We performed all of this using Python's sklearn library. This library allowed us to control the number of components as well as use Pandas Data Frames. This was helpful in combining our labeling and features.

We also used t-distributed stochastic neighbor embedding (t-SNE) for a better visualization of the preprocessed data. This along with PCA is a dimensionality reduction technique but has randomness in its finality. The t-SNE visualizations led us to similar conclusions and revealed that the dataset was perfectly linear separable.

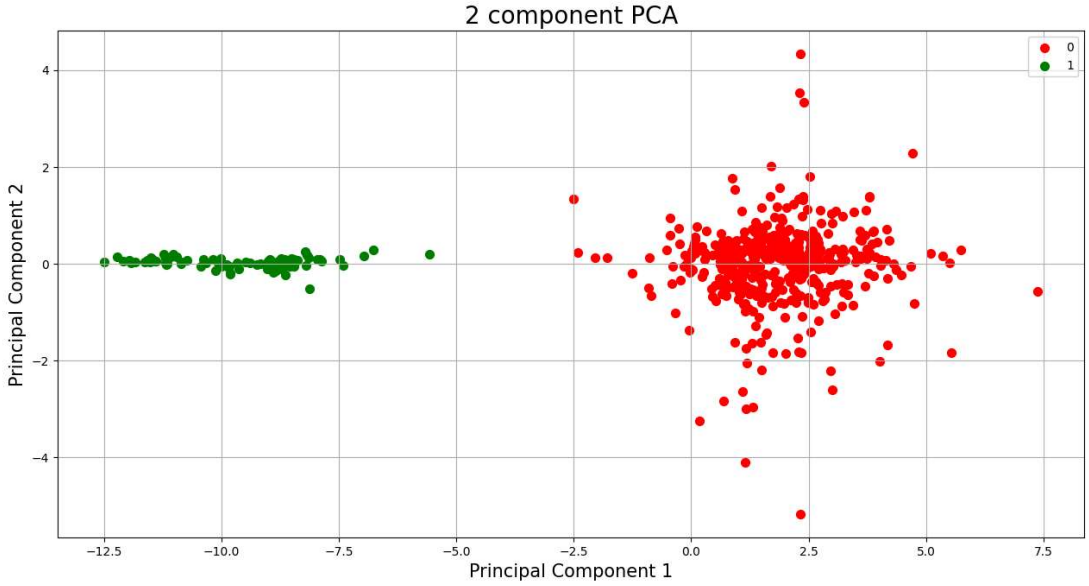
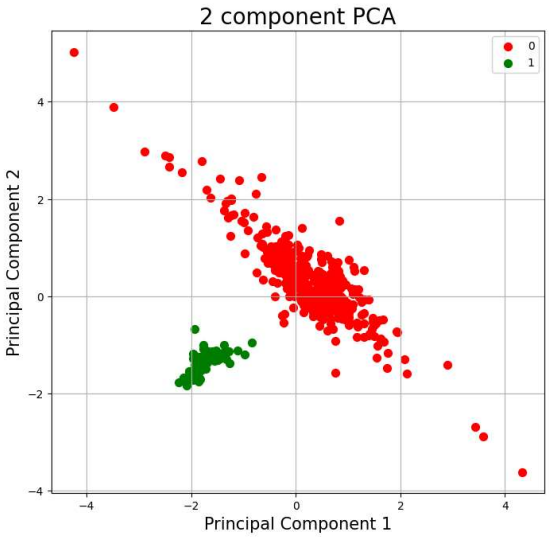


Figure 2. *Results of PCA without scaling.*



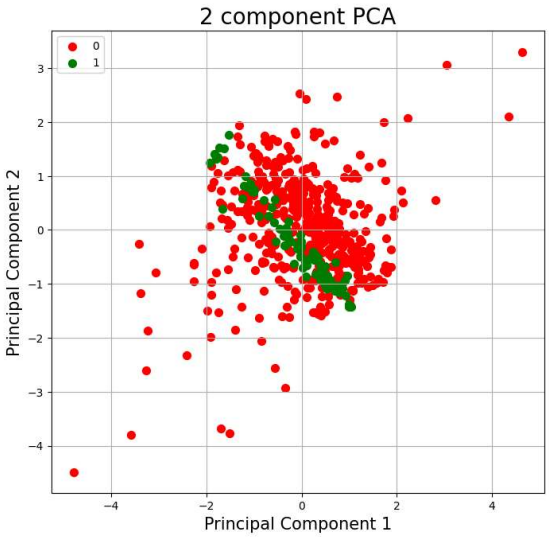


Figure 3. (Left): Results of PCA on scaled dataset.

Figure 4. (Right): Results of PCA on the scaled dataset with location features.

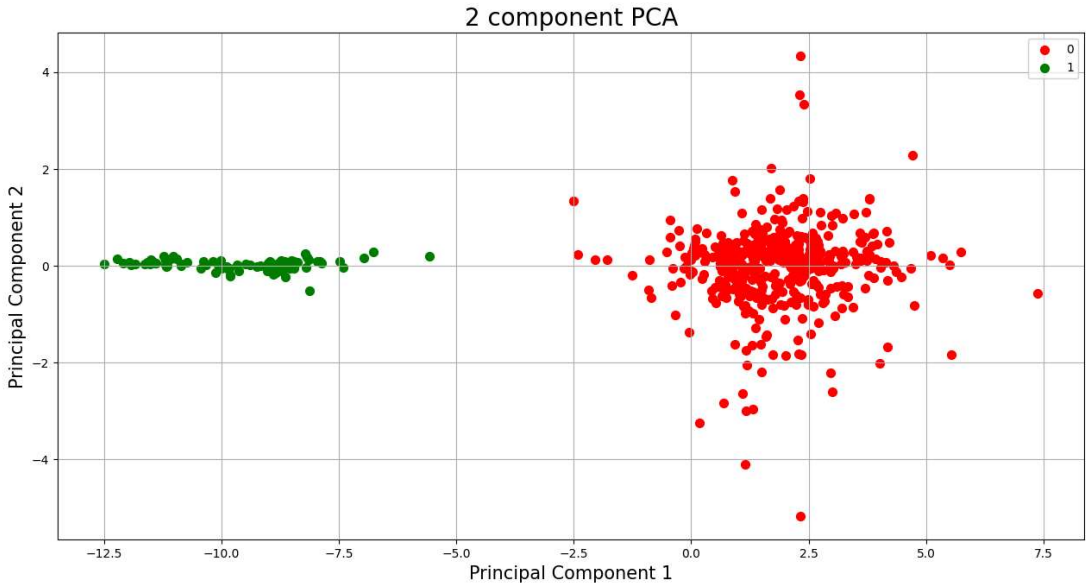


Figure 5. *Results of PCA on New Dataset.*

t-SNE Analysis

After our testing with PCA, we were able to apply this dataset to t-SNE to get a clearer picture of how linearly separable our data was [Figure 6]. As we had expected, it looked very promising as there were clear lines which separated the Wolf-Rayet stars from other ones. We additionally tried to look at the location data interweaved with the photometry data, however found similar results to our PCA. This help

of visualizing the data persisted with the new data set, while also showing which stars had some common features with WR stars.

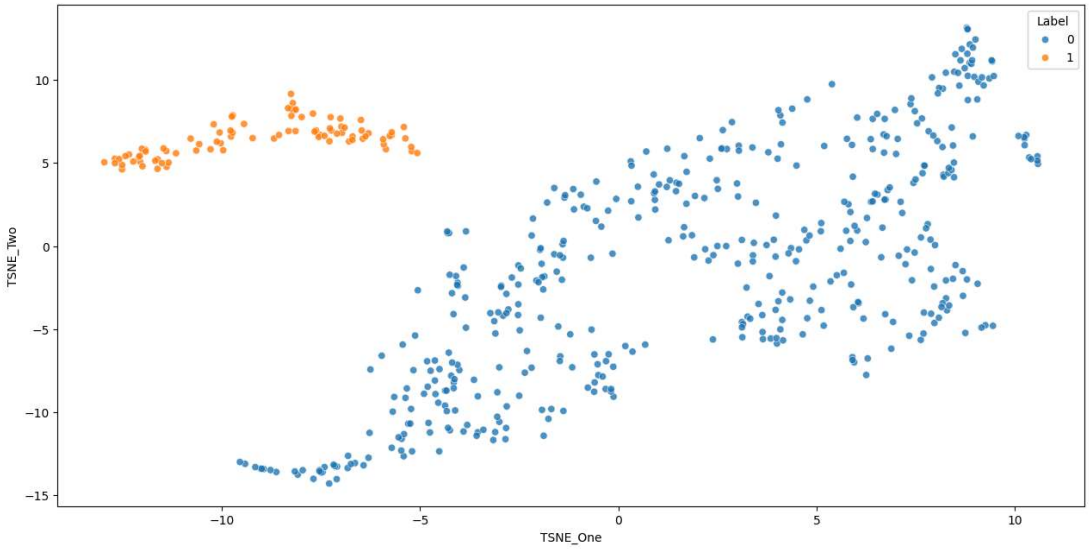


Figure 6. *t-SNE using Scaled Dataset*

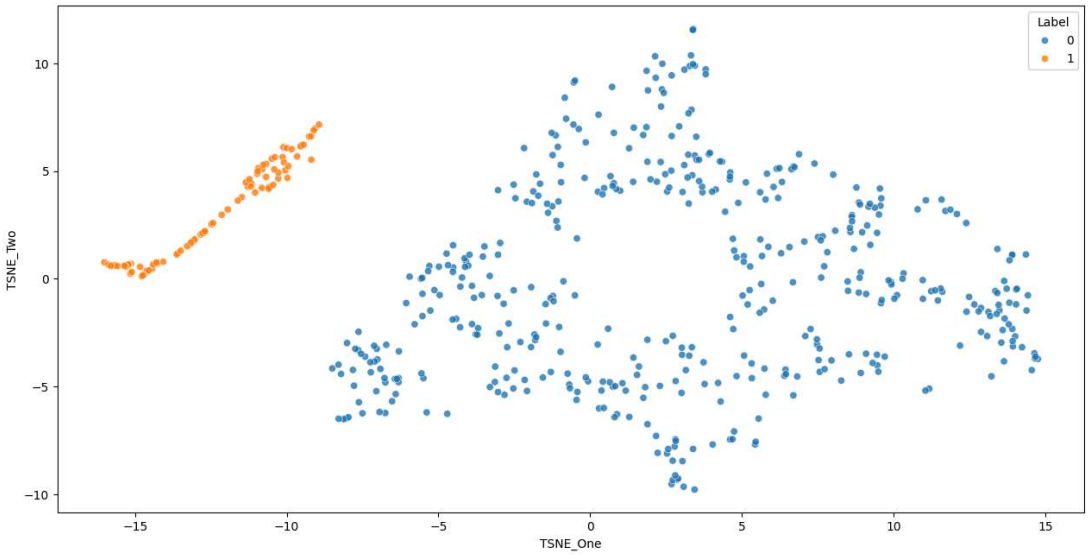


Figure 7. *t-SNE using Scaled Dataset and Location Data*

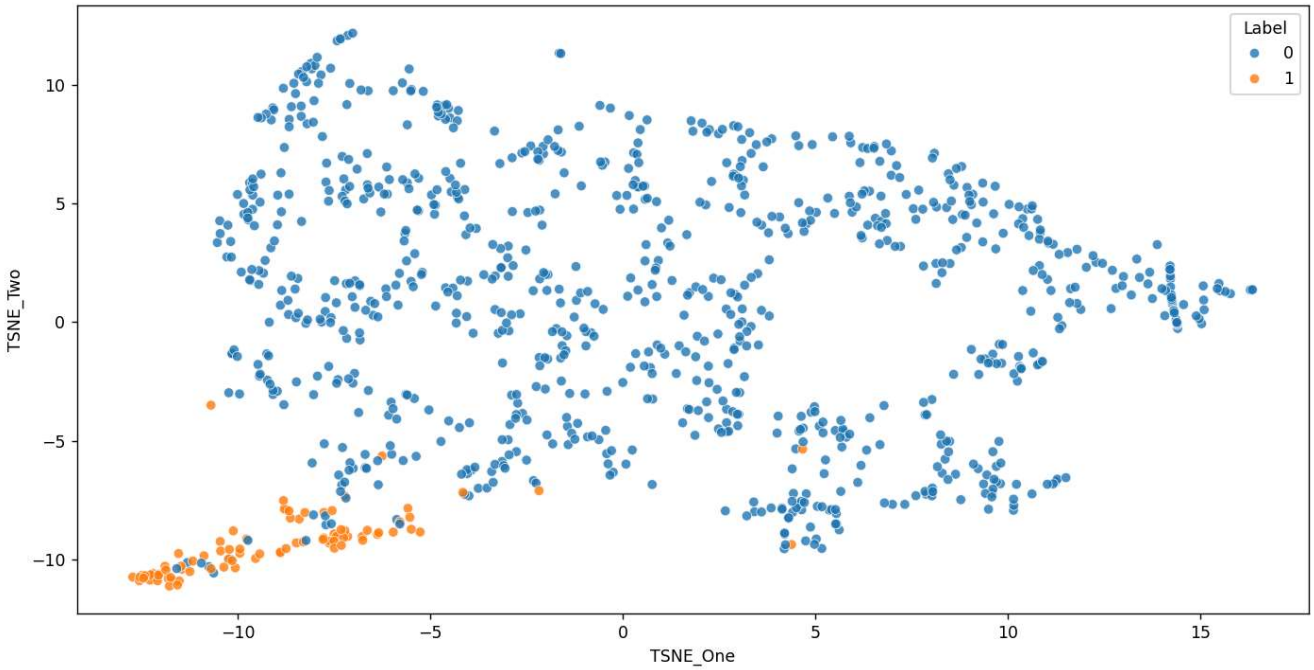


Figure 8. *t-SNE on New Dataset*

Data Preprocessing

We chose to implement scaling/recentering, and feature ranking in our preprocessing because our data sample had many different features and we were unsure of how these varying features would meaningfully contribute to our classification model. By conducting these two processes, we're able to ensure our model was able to more effectively focus on features that more distinctly contribute to determining whether a data point was a Wolf-Rayet star or not. These are both standard machine learning techniques that result in more stable computations, increase efficiency in the training of our models, and can contribute to reducing the negative effects of noise in data samples.

Feature Ranking

Pretty consistently, mRMR ranked "F275W_VEGA," "UV_color," and "F336W_VEGA" as the most important features while it ranked "RA" and "DEC" as the least important features as seen in Table 2. "RA" and "DEC" are purely coordinate values that give the locations of stars within a grid and no other information about that location Thus their relative lack of importance is not surprising. The rankings of the filters and color values was interesting however. For reference, the number in the filter names corresponds to the wavelength of light in nanometers at the center of the range of wavelengths that the filter observes. The UV filters and colors were all highly ranked and so was the infrared filter. There was a pretty sharp drop off in mutual information in the visible light and near visible light filters. While we expected the infrared filters to be more significant than the UV filters, the low rank of the visible light filters was expected based on background research [8]. Mutual information values correlated exactly with the mRMR ranks.

Fold 1		Fold 5		Fold 10	
Feature	Mutual Info.	Feature	Mutual Info.	Feature	Mutual Info.
F275W_VEGA	0.443923	F275W_VEGA	0.443923	F275W_VEGA	0.446472
UV_color	0.443923	F336W_VEGA	0.443923	F336W_VEGA	0.446472
F336W_VEGA	0.442682	UV_color	0.443923	UV_color	0.446472
F475W_VEGA	0.428338	F475W_VEGA	0.430178	F475W_VEGA	0.436351
F814W_VEGA	0.392416	F814W_VEGA	0.354913	F814W_VEGA	0.411435
F110W_VEGA	0.353697	F110W_VEGA	0.354913	F110W_VEGA	0.373571
F160W_VEGA	0.314127	F160W_VEGA	0.320387	F160W_VEGA	0.334601
UV_visible_color	0.274256	UV_visible_color	0.277652	UV_visible_color	0.266716
NearIR_color	0.194418	NearIR_color	0.212560	NearIR_color	0.205128
Visible_color	0.168344	Visible_color	0.168611	Visible_color	0.156961
Visible_nearIR_color	0.139885	Visible_nearIR_color	0.125312	Visible_nearIR_color	0.131826
RA	0.039201	DEC	0.035330	DEC	0.034606
DEC	0.023738	RA	0.031424	RA	0.032643

Table 2. *Feature rankings and mutual information scores for folds 1, 5, and 10 of the data.*

SMOTE

We also decided to apply SMOTE on the dataset. Our dataset was rather imbalanced, therefore we opted into using SMOTE with oversampling of the minority class in order to result in more balanced classification results.

Machine Learning Algorithms:

Our task was to perform binary classification on our dataset. For our machine learning algorithm, we employed the supervised learning algorithm of XGBoost. We specifically selected this model for its time efficiency and ability to handle complex patterns such as trends in photometry data. Additionally, this star

classification problem is known to be relatively prone to overfitting, and XGBoost implements regularization techniques that help to prevent this overfitting issue- especially since we have a small dataset and are even more susceptible to this problem. In our implementation, we split 90% of our data points into our training data and 10% into our test data. Given our smaller dataset, we compromised having a large test set in order to afford more data points for training the model in hopes of further alleviating problems with overfitting. After training the model, we calculated multiple quantitative scoring metrics and created visualizations to assess the model's success classifying Wolf-Rayet stars in our data set.

Another model we decided to implement was Support Vector Machines (SVM). SVM is another supervised learning model that is effective at binary classifications like ours (WR star/ Not WR star). We decided SVM would be a good solution for our problem due to the high dimensionality of our dataset. SVM is specifically designed to find the hyperplane boundary for data with higher dimensions. The use and implementation of different kernels also allows for classification of data that may not be linearly separable. Similar to XGBoost, SVM is also fairly robust against overfitting. Again, we decided to use 90% of our data as training data, and 10% as testing data due to the small size of our labeled data. For specific implementation, we decided to use Scikit-learn's SVM initialized with a Radial Basis Function (RBF) kernel.

The last supervised model we decided to implement was a Random Forest (RF). We decided to use RF because of its ability to classify data with high dimensionality and resistance to overfitting. During our pre-processing (specifically PCA) we noticed that our data could potentially be linearly separable. We decided on a tree based model given that it could produce accurate results given linear data. Again, we decided to use 90% of our data as training data, and 10% as testing data due to the small size of our labeled data. For our specific implementation, we used Scikit-learn's random forest classifier with 100 classifiers.

4. Results and Discussion

XGBoost:

XGBoost is one of the best-performing classical machine learning models, this is why we chose it as our primary model. For all of our models, we generated a classification report, a confusion matrix, and an ROC curve, as well as calculated the specificity, sensitivity, and the ROC AUC score.

Accuracy: 0.972
Sensitivity: 0.9
Specificity: 0.98

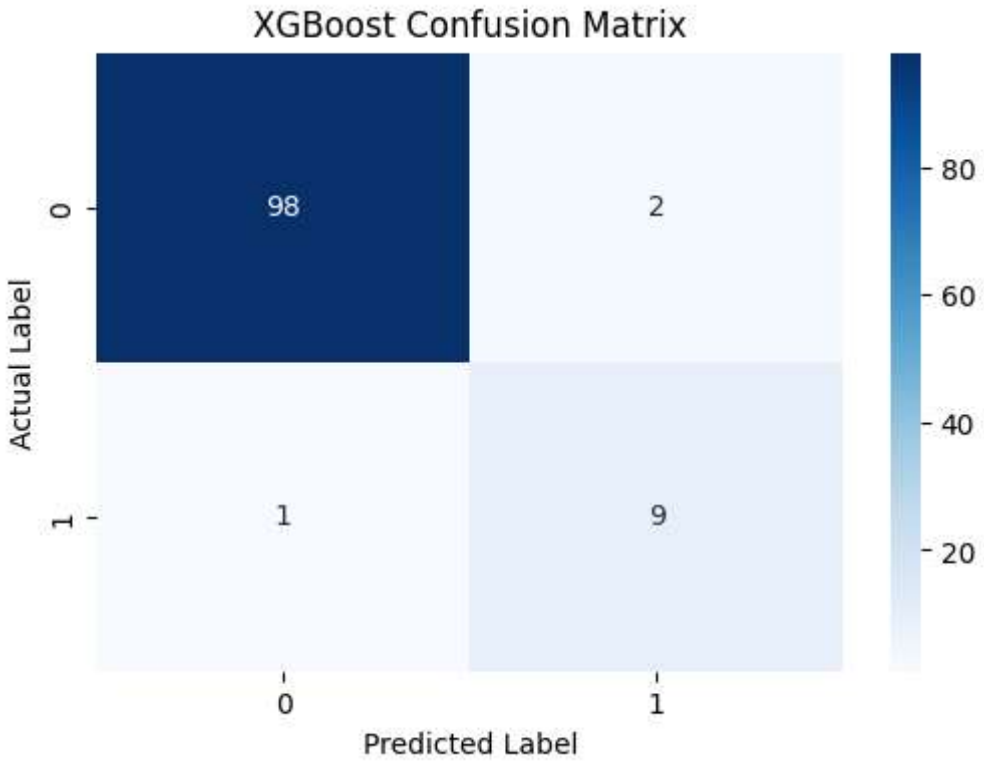


Figure 9. *Confusion matrix for XGBoost*

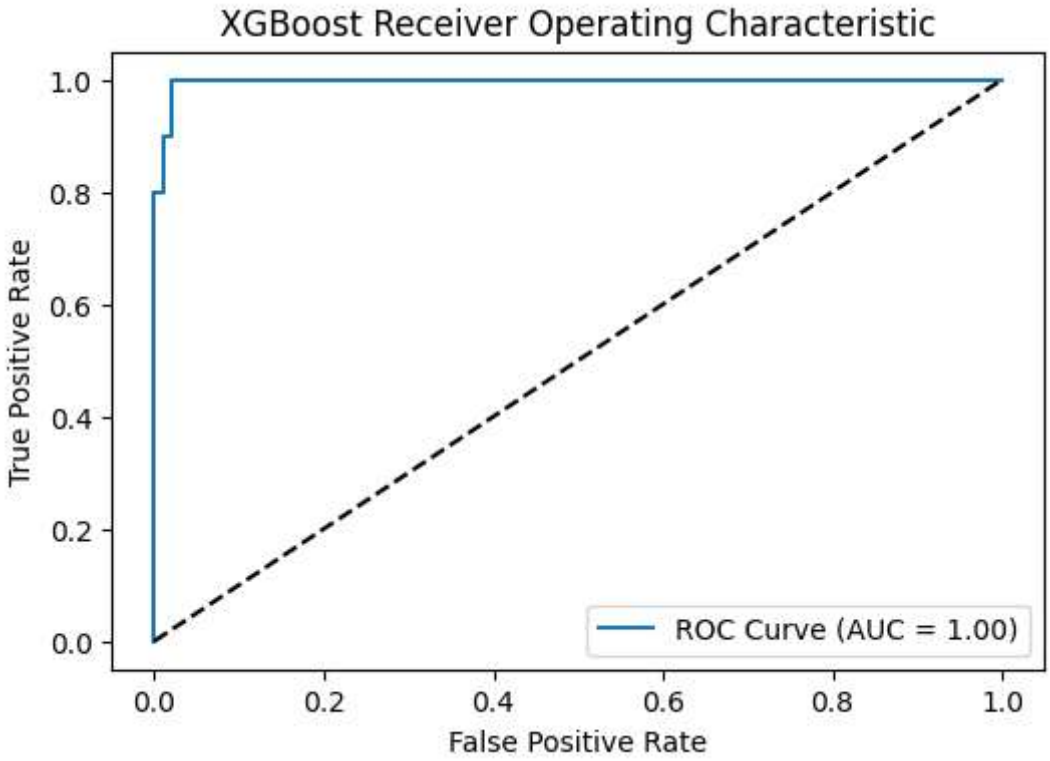


Figure 10.*XGBoost ROC Curve*

As shown in Figure 9, our XGBoost model was 97.2% accurate on our test set. With two false positive classifications and one false negative classification. XGBoost was very accurate at both identifying WR

stars and non-WR stars with very high sensitivity and specificity scores. This could be a case of incredible luck in the manner in which the fold was randomly selected as WR stars are always on the extreme end of the UV color, but further exploration needs to be done.

SVM

Accuracy: 0.936
Sensitivity: 0.5
Specificity: 0.98

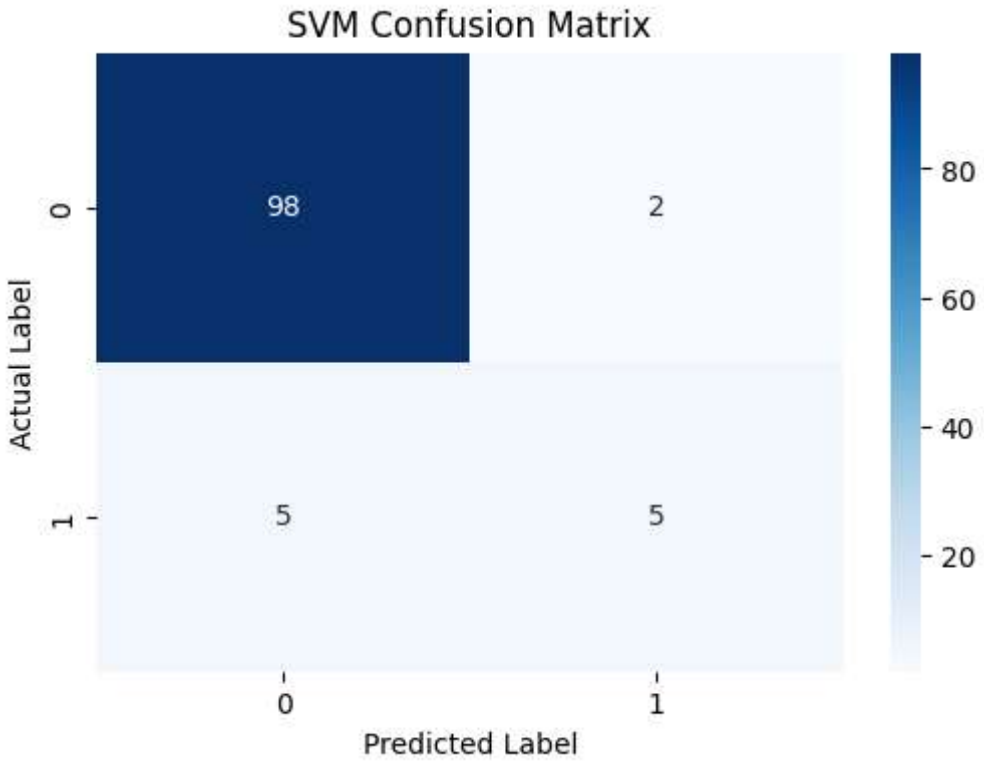


Figure 11. *Confusion matrix for SVM*

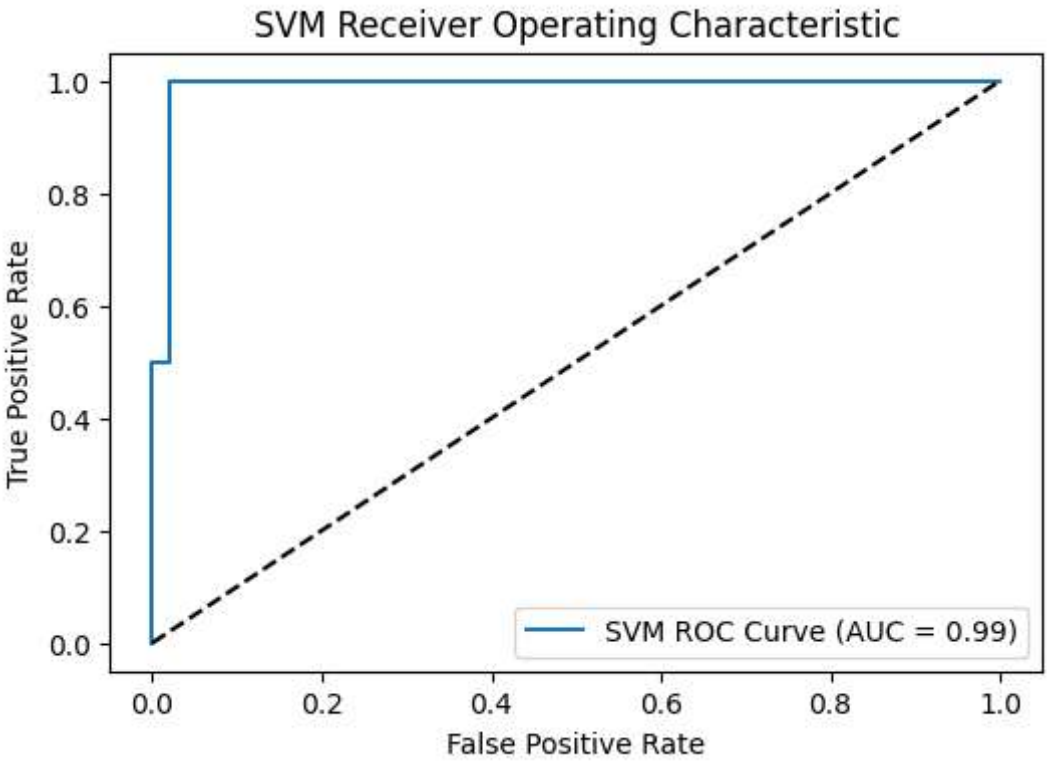


Figure 12.*SVM ROC Curve*

As shown in Figure 11, our SVM model was 93.6% accurate on our test set. With two false positive classifications and five false negative classifications. While initially this seems like a very high accuracy, the model only had five true positive classifications when there were 10 positive labels which produced a sensitivity of 50%. While SVM was very good at identifying non WR stars with high accuracy (98% specificity), it struggled to identify WR stars.

Random Forests

Accuracy: 0.972
Sensitivity: 0.8
Specificity: 0.99

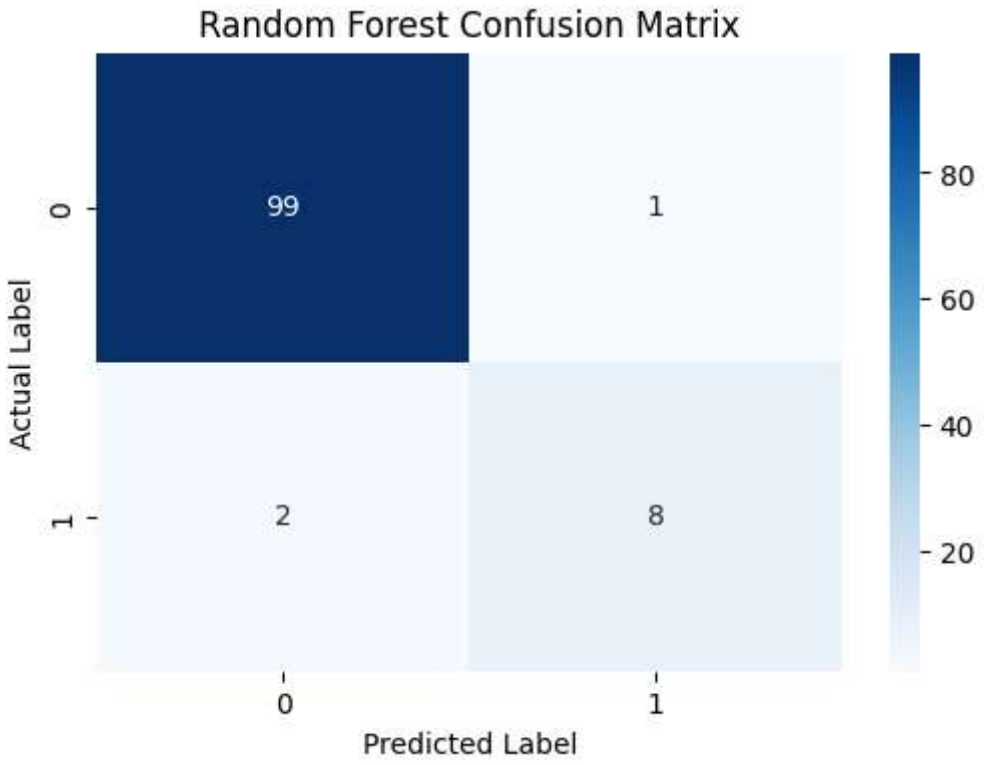


Figure 13. *Confusion matrix for RF*

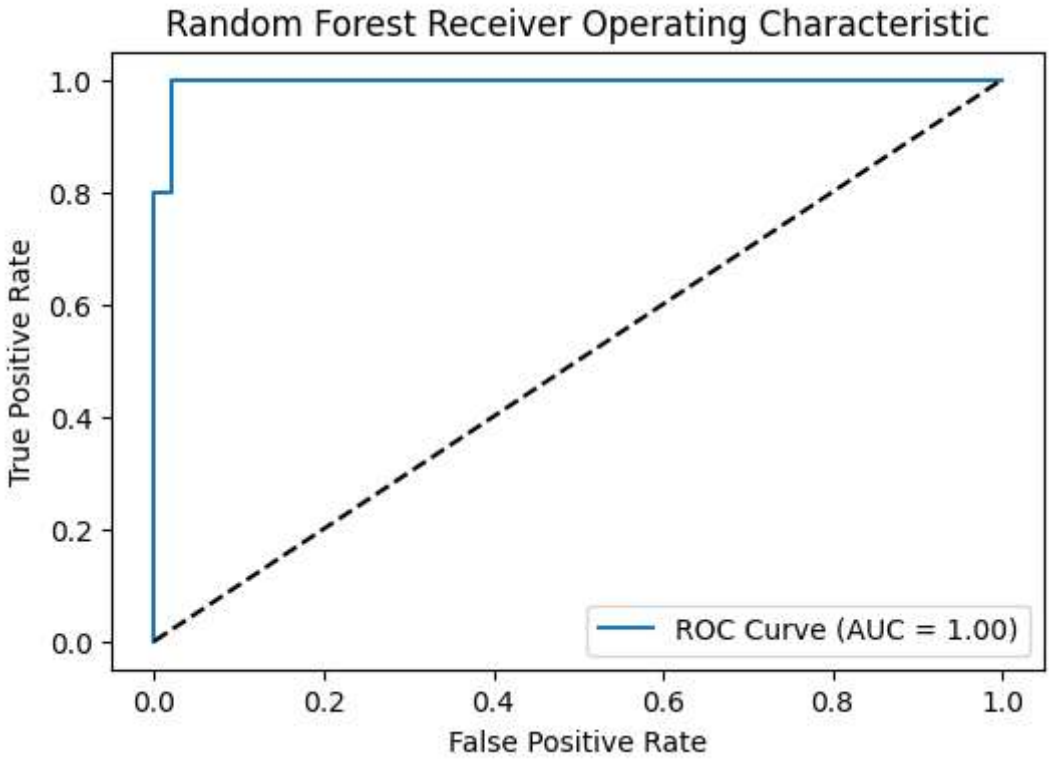


Figure 14.*RF ROC Curve*

As shown in Figure 13, our Random Forest model was 97.2% accurate on our test set. With one false positive classification and two false negative classifications. RF was very accurate at both identifying WR

stars and non-WR stars with very high sensitivity and specificity scores.

Comparison

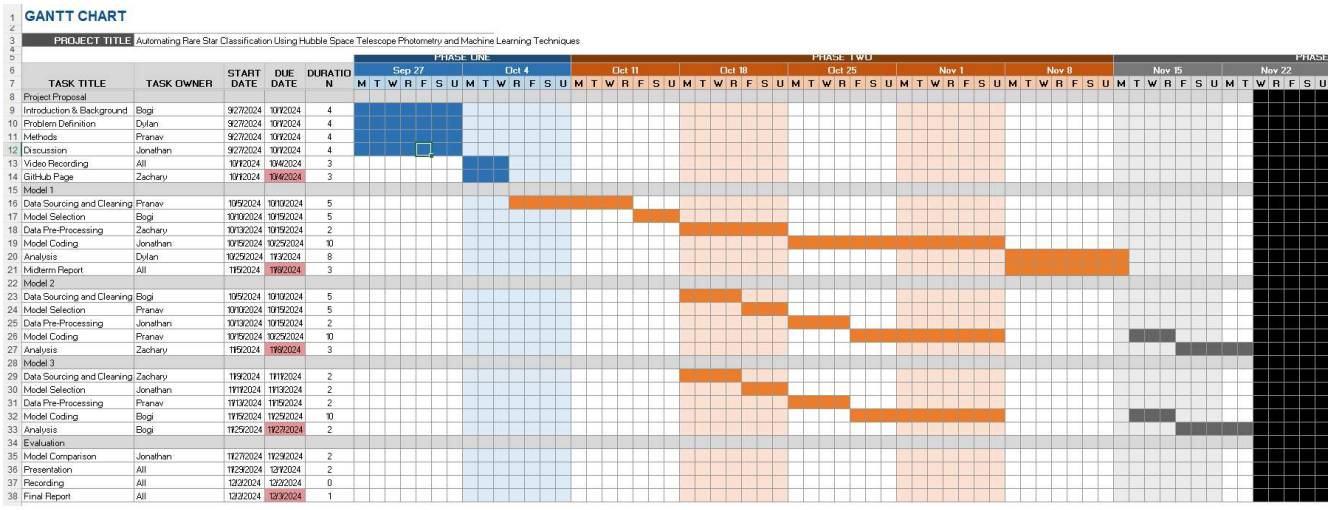
XGBoost and Random Forests models seemed to show the most promise for identifying WR stars. All of the models had very high accuracy on the testing data (>90%), however accuracy may not be the best way to evaluate models because our dataset was quite imbalanced (100 negative, 10 positive in the test). For example, SVM had ~93% accuracy but only had a sensitivity score of 50%. This may be more due to the size and unbalanced nature of our datasets than the effectiveness of SVM. The large number of negative classifications in our data set likely caused SVM to move the decision threshold closer to the WR data points leading to lower generalization for positive labels. XGBoost and RF both had very similar results with producing 90% and 80% sensitivity scores respectively.

Next Steps

As next steps, we plan to validate our models on in domain and out of domain data to explore how well they generalize. We plan to incorporate more sophisticated feature extractors and other types of features (e.g. locality-based), or increase the complexity of features by incorporating imaging data/features that we can extract using image processing methods or a convolutional neural network or by using an astro foundation model for stars that might have been pre-trained on similar data, and could have the potential to extract more complex and expressive features. We can also use ensembling to yield a more stable classification. We plan to extend our classifier to other star types of subtypes of rare stars. Once we obtain a stable classifier, we can assign pseudo labels to a larger portion of data and train more complex deep learning models to perform other tasks that aid the discovery of new WR stars and understand their formation better, e.g. using unsupervised approaches to find potential locations where WR stars might have formed.

5. A Note on Changes from the Proposal

Since the submission of our project proposal, our group has decided to focus on identifying Wolf-Rayet stars in the Hubble datasets [6][7]. We believe that this will make for a more reasonable and manageable project while still providing meaningful scientific insights and tools in astronomy. For this midterm report, we focused on the curation of the dataset, data preparation and preprocessing for a binary classification task of stars as either WR or non-WR stars. Going forward, we intend to expand upon this by either classifying subtypes of WR stars or attempting to characterize the locations WR stars are found, as well as working on improving the generalizability and performance of our model(s).



GANTT CHART

PROJECT TITLE: Automating Rare Star Classification Using Hubble Space Telescope Photometry and Machine Learning Techniques

TASK TITLE	TASK OWNER	START DATE	DUE DATE	DURATION
Project Proposal				
Introduction & Background	Bogi	9/27/2024	10/1/2024	4
Problem Definition	Dylan	9/27/2024	10/1/2024	4
Methods	Pranav	9/27/2024	10/1/2024	4
Discussion	Jonathan	9/27/2024	10/1/2024	4
Video Recording	All	10/1/2024	10/4/2024	3
GitHub Page	Zachary	10/1/2024	10/4/2024	3
Model 1				
Data Sourcing and Cleaning	Pranav	10/5/2024	10/10/2024	5
Model Selection	Bogi	10/10/2024	10/15/2024	5
Data Pre-Processing	Zachary	10/13/2024	10/15/2024	2
Model Coding	Jonathan	10/15/2024	10/25/2024	10
Analysis	Dylan	10/25/2024	11/3/2024	8
Midterm Report	All	11/5/2024	11/6/2024	3
Model 2				
Data Sourcing and Cleaning	Bogi	10/5/2024	10/10/2024	5
Model Selection	Pranav	10/10/2024	10/15/2024	5
Data Pre-Processing	Jonathan	10/13/2024	10/15/2024	2
Model Coding	Pranav	10/15/2024	10/25/2024	10
Analysis	Zachary	11/5/2024	11/6/2024	3
Model 3				
Data Sourcing and Cleaning	Zachary	11/9/2024	11/11/2024	2
Model Selection	Jonathan	11/11/2024	11/13/2024	2
Data Pre-Processing	Pranav	11/13/2024	11/15/2024	2
Model Coding	Bogi	11/15/2024	11/25/2024	10
Analysis	Bogi	11/25/2024	11/27/2024	2
Evaluation				
Model Comparison	Jonathan	11/27/2024	11/29/2024	2
Presentation	All	11/29/2024	12/1/2024	2
Recording	All	12/2/2024	12/2/2024	0
Final Report	All	12/2/2024	12/3/2024	1

Figure 8.*Project Gantt Chart*

Contributions:

Name	**Final Proposal Contribution**
Bogi	Dataset curation, Data pre-processing, Feature Ranking, Methods, Results and Discussion, Project management, Presentation, GitHub pages
Jonathan	Revised Introduction/Problem Statement, Revised Models, Results and added comparison section, Literature review, References, Written Analysis, GitHub pages
Pranav	Proposal Intro, Implemented XGBoost, Gantt Chart, SMOTE preprocessing, Github Pages
Zachary	Conducted Quantitative Model Analysis and Created Plots, Results Section, Methods section, implemented SVM and Random Forest Classifier, GitHub Pages
Dillon	PCA and t-SNE Implementation and Analysis and Created Plots, Results, Github Pages

References

1. K. Neugent and P. Massey, “The Close Binary Frequency of Wolf-Rayet Stars as a Function of Metallicity in M31 and M33,” *The Astrophysical Journal*, vol. 789, no. 1, Jun. 2014. [Online]. Available: https://doi.org/10.1088/0004-637X/789/1/10. [Accessed: Oct. 4, 2024].
2. K. Neugent et al., “The Evolution and Physical Parameters of WN3/O3s: A New Type of Wolf–Rayet Star,” *The Astrophysical Journal*, vol. 841, no. 20, May 2017. [Online]. Available:

https://doi.org/10.3847/1538-4357/aa704b. [Accessed: Oct. 4, 2024].

3. J. Mikolajewska, N. Caldwell, and M. Shara, “First detection and characterization of symbiotic stars in M31,” Monthly Notices of the Royal Astronomical Society, vol. 444, no. 1, Oct. 2014. [Online]. Available: https://doi.org/10.1093/mnras/stu1463. [Accessed: Oct. 4, 2024].
4. M. Pettee et al., “Weakly supervised anomaly detection in the Milky Way,” *Monthly Notices of the Royal Astronomical Society*, vol. 527, no. 3, Jan. 2024. [Online]. Available: https://doi.org/10.1093/mnras/stab2993. [Accessed: Oct. 4, 2024].
5. B. F. Williams, M. Durbin, D. Lang, J. J. Dalcanton, A. E. Dolphin, A. Smercina, P. Y. Merica-Jones, D. R. Weisz, E. F. Bell, K. M. Gilbert, et al., "The Panchromatic Hubble Andromeda Treasury. XXI. The Legacy Resolved Stellar Photometry Catalog," *Astrophysical Journal Supplement Series*, vol. 268, no. 2, p. 48, 2023. [Online] Available: https://iopscience.iop.org/article/10.3847/1538-4365/acea61. [Accessed: Oct. 4, 2024].
6. Hubble Space Telescope PHAT Data Archive. [Online]. Available: https://archive.stsci.edu/hlsp/phat. [Accessed: Oct. 4, 2024].
7. Hubble Space Telescope PHATTER Data Archive. [Online]. Available: https://archive.stsci.edu/hlsp/phatter. [Accessed: Oct. 4, 2024].
8. S. Kar, R. Bhattacharya, R. Das, Y. Pihlström, and M. O. Lewis, "Classification of Wolf Rayet stars using Ensemble-based Machine Learning algorithms," arXiv preprint, arXiv:2410.14845, Oct. 18, 2024. [Online]. Available: https://arxiv.org/abs/2410.14845. [Accessed: Nov. 7, 2024].