Using Machine Learning to Identify P-Cygni Profiles

Anthony Pearce

May 2025

Abstract

P-Cygni profiles are distinct spectral features that indicate high velocity ejecta in astrophysical events such as novae. In this study, a machine learning algorithm was developed to identify these profiles automatically in spectral data. A cross-correlation algorithm was first applied to locate candidate features by comparing observed spectra with a P-Cygni template. 100 data point snapshots of these candidates were then labeled and used to train a Random Forest classifier. The resulting model achieved a testing accuracy of 97.6% and an AUC of 0.99. When applied to new spectra, the trained model successfully located P-Cygni features using a sliding window prediction method within an accuracy of 99.7%.

1 Background

P-Cygni profiles are distinct spectral line features that are characterized by a combination of blueshifted and redshifted emission and absorption components. They serve as evidence of the rapid expansion of stellar envelopes in astrophysical phenomena such as novae and heavy stellar wind.

These P-Cygni profiles develop due to the mechanisms in which the light reaches the observer from one of these rapid expansion events. During a nova, a dense envelope of gas will expand from the origin star. The portion of this envelope that is moving toward the observer will absorb some of the light from the origin star and scatter it in all directions. This causes a portion of the light that would have originally traveled to the observer to be sent elsewhere, thus causing the absorption line as well as the blueshift effect on the photon's wavelength. Similarly, the portion of the envelope that is moving away from the observer absorbs light from the origin star and emits it in all directions. This causes the redshifted emission peak after the absorption profile. The portions of the envelope that move across the observer's line of sight do not get Doppler shifted, but still contribute to the overall P-Cygni profile toward the center[**1**].

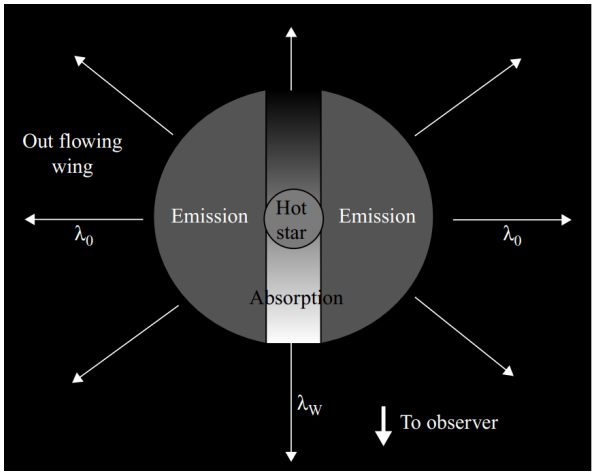

Figure 1: Emission/Absorption from a rapidly expanding envelope.[**1**]

We can use the shape of these P-Cygni profiles to determine properties of the ejected envelope. By measuring the distance between the middle of the figure and either the emission or absorption feature, we can use the standard Doppler effect formula to determine the velocity of the envelope at that moment[**1**].

$$v = c\frac{(\frac{\lambda_{abs/emi}}{\lambda_{profile}}) - 1}{(\frac{\lambda_{abs/emi}}{\lambda_{profile}}) + 1} \quad (1)$$

Multiple measurements over the course of a few days may be useful as well to compare the rate of change of this ejection velocity.

2 Identifying P-Cygni Profiles

In order to create a machine learning model, a dataset of P-Cygni profiles has to be collected to train the model with. A cross-correlation algorithm was chosen for this purpose. Cross-correlation allows a wave to be compared to a model to determine how similar they are. Because we do not care for the entire spectrum, a template model for a P-Cygni profile was created and then effectively "slid" across the spectrum. This allowed each point to be checked for correlation, and labeled as potential profiles.

For something to be considered a candidate profile, two criteria were set to weed out false positives. The correlation factor was required to be above 30. This value was determined by hand while gradually adjusting it. This worked pretty well, but there were many false positives toward the beginning and end of the spectrum due to the instrumental noise. To combat this, the a second requirement was put in place. Once a candidate profile is identified via the threshold, it is then checked for a local minimum to the left, and a local maximum to the right. If both were present, a ± 50 data point "snapshot" of the profile was taken and saved.

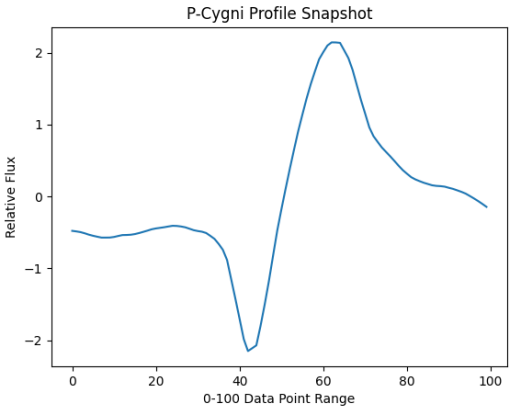

Figure 2: A 100 data point wide snapshot of a P-Cygni Profile collected for training.

To further clean the data, a second cross-correlation was run over the new 100 data point long snapshot. Once this method showed to be reliable, spectra was pulled from the ARAS nova database to have profile extracted from. This algorithm was run over 2,032 fit files of nova spectra, taking both verified P-Cygni profiles as well as a number of random samples to use for supervised machine learning.

3 Machine Learning

To automate the identification of P-Cygni profiles, machine learning in the form of Random Forest classification was used to reliably identify these target profiles from a provided spectral image.

The data set collected previously via cross-correlation was taken and labeled. A label of 0 was assigned to the poor profiles, and 1 was assigned to the good ones. The data was split into an 80/20 training and testing subset. The model training and testing accuracy was evaluated and determined to be 0.99997 and 0.97618 respectively. These numbers were then compared to a K Cross-Validation, and were found to be very similar. A Reciever Operating Characteristic (ROC) curve was plotted to compare the true and false positive rates. The Area Under the Curve (AUC) of the ROC was found to be 0.99, which implied an extremely high true to false positive ratio.

Using this model, a 100 data point window was moved along the provided spectrum, stopping at each data point to make a prediction on the presence of a P-Cygni profile.

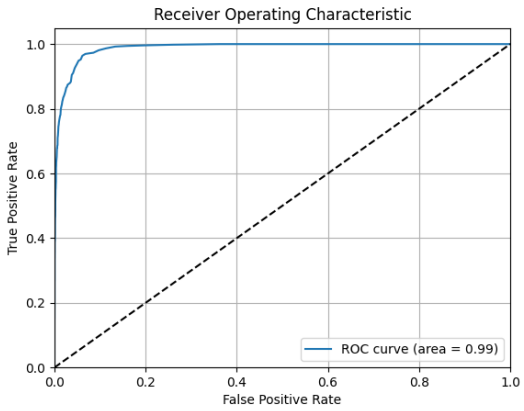

Figure 3: ROC curve showing a high rate of true positives compared to false ones.

4 Results

Using the sliding window prediction method, the model was able to successfully identify four P-Cygni profiles within the given spectrum with a few of the smaller profiles left

out. These profiles were likely not identified properly due to the relatively high threshold required for the cross-correlation portion of the profile collection step. The lines that were detected were placed with an accuracy of 99.7%.

While it often would neglect smaller possible features, the model had a very low false positive prediction rate, with none being found by general testing.

5 Discussion

While the model was able to accurately predict some of the more obvious P-Cygni profiles, it would often neglect some of the smaller profiles that were still fairly obvious to human analysis. This is likely due to having too rigid of a threshold for training data classification, as well as a low quantity of training data.

The threshold used when collecting the training data was determined manually via testing. While this works for the spectra that was used to determine the number, it was applied to over 2,000 different spectra. So it is reasonable to assume that it is not the best number for all of them. A variable threshold should be calculated for each spectrum individually based on the distribution of correlation coefficients. This would allow each spectrum to have its own unique threshold and make the collection process more efficient.

The total number of confirmed P-Cygni profiles was only around $\frac{1}{20}^{th}$ the size of the number of confirmed "bad" profiles. A robust collection of poor profiles is necessary to avoid false positives, but a collection of confirmed profiles this small will lead to poor identification of true profiles. It essentially made the model hesitant on less obvious profiles.

Additional features can be added to this model as well, such as calculating the velocity of the ejected envelope using the P-Cygni peak measurements. It may also be possible to include a database of known emission elements, and have the model identify not just the presence of a P-Cygni, but what element it is representing. The model could also be improved in such a way for it to identify transient heavy element absorption lines, a common and often tedious feature that can appear in these kinds of spectra.

6 Conclusion

By using a sliding window cross-correlation algorithm, P-Cygni profiles were efficiently identified in observation spectra using a template profile. These profiles were tagged and verified by checking for their emission peak and absorption dip. These tagged profiles were then extracted in 100 data point images, and then double checked by another round of cross-correlation verification.

Once satisfied by the precision of these selected profiles, they were then labeled along side some randomly selected sections from the spectrum to create an array of labeled "good" and "bad" profiles. These profiles were then split and used to train a Random Forest machine learning algorithm, resulting in a testing accuracy of 97.6

7 Bibliography

[**1**] Robinson, Keith; Spectroscopy: The Key to the Stars (2007).

8 Appendix

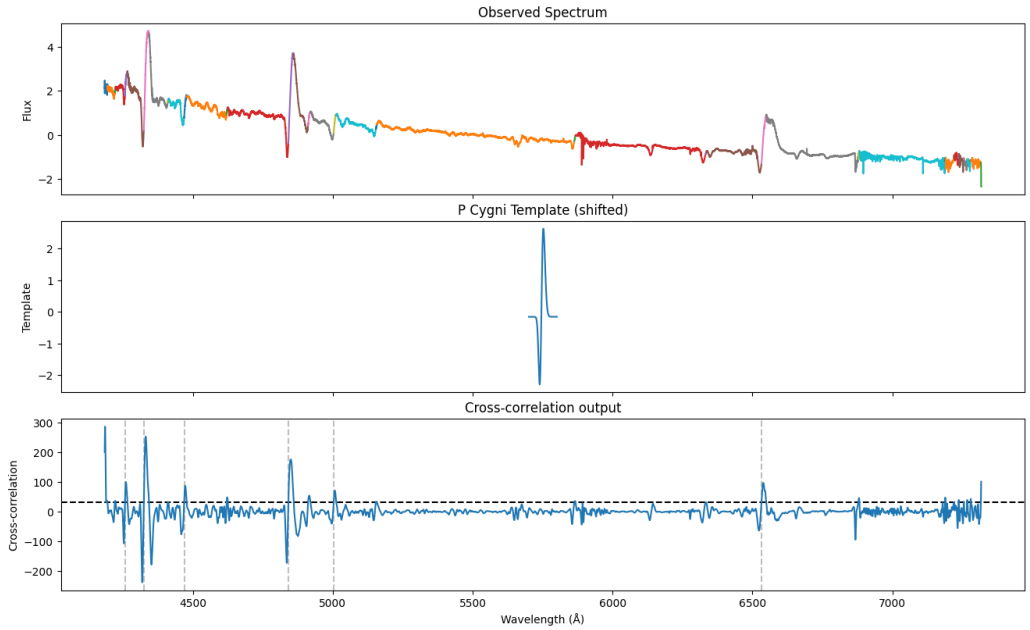

Figure 4: A comparison of an example spectrum (top), the P-Cygni Template (middle), and the cross-correlation factor (bottom)

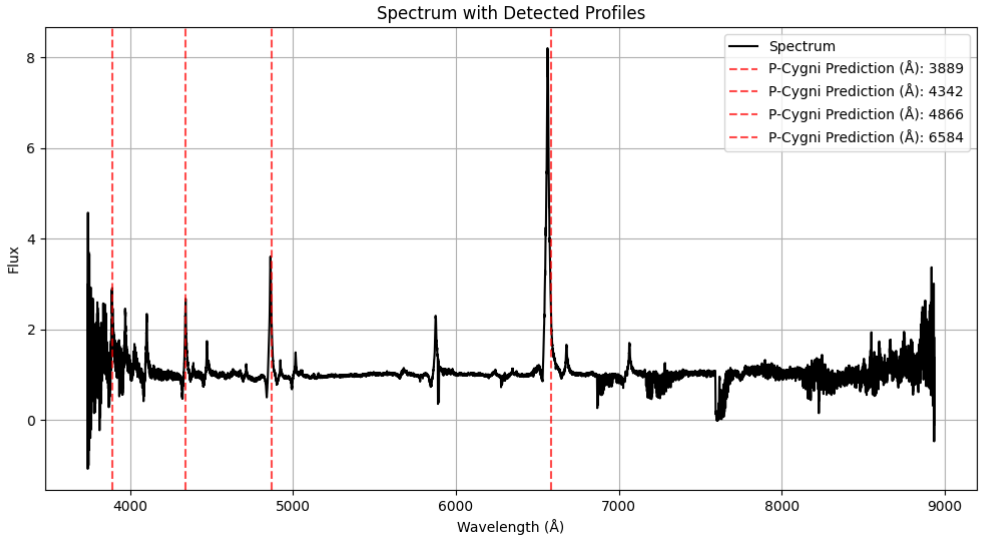

Figure 5: Spectrum with P-Cygni profiles identified by model.