Identifying YSO and their stages from Spitzer field of view; an ML approach

B. Crompvoets

July 25, 2022

1 Introduction

This work will address the issue of classifying young stellar objects (YSOs) given input from a field of view from the Spitzer telescope. YSOs are defined as any stellar object which has not yet reached Main Sequence evolution. In this regime, there are five stages known as the Class 0, Class I, Flat-Spectrum, Class II, and Class III stages.

Class 0 is the label given to the youngest YSOs which have only just begun to collapse from their natal dust cloud, and are still heavily obscured. They are actively accreting dust and gas and thus gaining mass. Their luminosity peaks deep in the infrared (IR), a wavelength regime not well characterized by Spitzer. Class I YSOs are those slightly older than Class 0. Their dust and gas envelope has begun to thin, but they are still obscured. Due to this, they also are located in the IR, but not to the same extent as the Class 0s, and are thus within the detectability range of Spitzer. Flat-spectrum YSOs, so called because the slope of their spectral energy distribution is approximately zero, have an actively dissipating envelope. Class II YSOs no longer have an envelope, but instead are surrounded by a disk of material. These are the stars often imaged when searching for planet formation in their debris disks. Finally, Class III YSOs are the eldest. Their disk has dissipated, and they are condensing as they near the point of Hydrogen ignition in their cores.

The objects of greatest interest in this study are the youngest of these YSOs, as probing their formation can give us information on the kinematics and dynamics of star formation. In order to get a full sense of the process, we must be able to rely upon several samples of YSOs across all stages. Thus, the initial objective must be to obtain a reliable categorization of objects identified in surveys. This objective was fulfilled by Gutermuth et al. (2008) and updated by Gutermuth et al. (2009).

Gutermuth et al. (2009) provided a prescription for categorizing objects given the four Spitzer IRAC bands (located at 3.6 μm, 4.5 μm, 5.8 μm, and 8.0 μm), the 24 μm Spitzer MIPS band, and the J, H, and K_S 2MASS bands. They apply empirical cuts onto a range of colour-colour and colour-magnitude diagrams in order to separate stellar objects into a number of classes. Although a very popular method to use, this method is time-expensive and requires heavy supervision. In the era of big data science, such methods are becoming archaic, and thus new methods must be developed to handle the large input of data. The modern solution to this problem is the use of machine learning, or ML. Several groups have previously worked to identify YSOs using ML, and we summarize their works here.

Miettinen (2018) use a compilation of various ML methods in order to classify YSOs into three categories: Class 0, Class I, and Flat-Spectrum. They test eight different methods, seven classical methods, and one multi-layer perceptron (MLP). They found that the best method to separate the classes from one another was a gradient boosting model. As they used supervised models, they had to have a classification pre-applied to their data, and this was done by using the data-set from the Orion Protostar Survey, as classified by Furlan et al. (2016). Although providing fair results, the major shortcoming of this work is that they only consider proto-stars, and thus they can only classify objects into their YSO category if they already know the object is a YSO.

A more realistic approach is to create a method which can classify YSOs based off of the unsorted input from a telescope. This is performed by Cornu & Montillaud (2021), who use an MLP to classify objects from Spitzer as either Class I YSOs, Class II YSOs, or contaminant objects. The reason for using only the three classes is that (1) Class 0 objects are not visible in Spitzer data (the only data used in this survey); and (2) Class III objects have a signature very similar to regular main sequence stars at these wavelengths, and thus the neural network would lose performance in trying to correctly classify Class III. As the most important YSOs to star

formation research is the youngest of the categories, it is then a reasonable sacrifice to allow any Class III or flat spectrum YSOs to be classified as contaminants. They test their network on the full Orion Survey (Megeath et al., 2012), as well as on a survey of NGC 2264 (Rapson et al., 2014). These surveys include a number of different object types, and the training and validation sets are specifically chosen to increase model performance while still being representative of the night sky. In choosing such active star-forming regions, they put themselves at risk of training a network which is not applicable to more quiescent regions.

Chiu et al. (2021) perform another classification using the full Planetary Cores to Stellar Disks (c2d) survey (Evans et al., 2014) which analyzes more quiescent regions than the Orion and NGC 2264 star-forming regions. They also use an MLP as well as testing random forest, XGBoost, KNN, and SVM classical methods to classify objects into one of three categories: YSOs, extra-galactic sources (EGs), and stars, They find that the MLP works the best, and further test the benefits of using all bands, bands plus errors, each object normalized such that distance effects are removed, and using only three bands. With full bands they utilize the four IRAC bands, the MIPS 24 band, and the HJK bands from 2MASS (adjusted to match the bands from UKIDSS). They obtain the highest recall and precision of any method yet. They provide a website to label new data available for use to anyone. This method could only be further improved by identifying the stages of the classified YSOs.

Kuhn et al. (2021) use a random forest classifier to sort objects into YSOs or others, and then use a series of cuts performed via the spectral index to further sort the YSOs into each of the aforementioned YSO classes. They further analyze the regions in which each class of YSO is found. Although performing an interesting and robust scientific analysis, their performance metric is heavily influenced by data imbalance, an issue they did not address within their work. Furthermore, the use of spectral index to specify the YSO stage may lead to confusion in which stage an object actually belongs in, as their spectral index is not the sole indicator of class.

In the following sections we will make use of each of the scientific benefits of these methods, and provide a new method of classification which will classify objects into one of six classes: Class I YSOs, Flat-Spectrum YSOs, Class II YSOs, Class III YSOs, Extra-Galactic sources (EGs), and Stars. At this stage we will focus only on utilizing Spitzer data, which is incapable of finding Class 0 objects. In Section 2 we explain our methodology and how it differs from previous works. Section 3 describes the results of this methodology for the Spitzer Space Telescope. Finally, discussion of these results takes place in Section 4.

2 Methodology

Our mission is to perform a series of classifications, following on each of the aforementioned works. First, we will create a network which is capable of sorting objects into Young Stellar Objects (YSOs), Extra-Galactic sources (EGs), and contaminating Stars, as in Chiu et al. (2021). Secondly, we will follow Kuhn et al. (2021) and make use of the spectral index values to further sort the YSOs into their respective stage (Class I through Class III). We will train on a balanced subset of the c2d Survey, validate on NGC 2264, and further test our network on the c2d CORES survey, which contains several star-forming cores which are more quiescent than NGC 2264. We initially classify our YSO targets using the spectral index, as this is a good approximation for the YSO stage. However, since we train on these stages, it allows the network to draw it's own conclusions about which object belongs where, which allows for the objects classification to be determined more thoroughly than purely via the spectral index.

As different clouds occur at different distances, their extinction values could shift the phase space of the different classes and hamper classification. Furthermore, a danger in using multiple text sources for training and validation is that the labels assigned by each team may not be universal. That is to say an object classified as a YSO candidate by one team may be classified as a field star by another, and these cuts in parameter space will affect how the network trains. This first source of concern is explored by using a large multi-cloud survey as the training source, and testing it on a smaller single cloud source. It is important to note that the hope of this study is to be able to analyze new data, and as such this method will allow us to review the applicability of our methods. The second concern would be solved if we were to perform a single analysis on all data, however that is not within the scope of this work. Instead we train and test on large in-depth data-sets that cover wide expanses, and take this classification as ground truth. We validate on a separate test set to ensure that this is widely applicable as well. All Training, Validation, and Test set sources are located in the Appendix under Table 4.

2.1 General Set-Up

In each of the below cases, we perform the same tests. First, we train a neural network to classify YSOs, EGs, and Stars, varying layers, learning rate, momentum, and number of neurons within the hidden layers. We run a test of each set-up for 3000 epochs to determine the best fit. If the best network from each of our variations in layer shows a continuing decrease in loss rate, we proceed for another 3000 epochs in order to obtain more accurate predictions. We use the F1-score of the YSO class as a metric for the best fit in all cases. This is chosen as it is the harmonic mean between recall and precision, and thus ensures we are maximizing our number of correctly classified YSOs whilst minimizing the number of contaminating objects in this class. These three metrics are defined below, where $TP = True\ Positive$ (e.g. YSO classified as YSO), $FP = False\ Positive$ (e.g. EG classified as YSO), and $FN = False\ Negative$ (e.g. YSO classified as EG).

$$Precision = P = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = R = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = 2\frac{P \times R}{P + R} \tag{3}$$

Multiple classical methods such as random forest and gradiant boosting classifiers were tested, but they did not perform as well as MLPs and hence were not used.

Secondly, we take all those objects labelled as YSOs, and we perform a second classification, this time separating data into Class I, Flat-Spectrum, Class II, and Class III YSOs. As the survey we used to train on the first set will not include targets for the second set, we must pre-label our objects as belonging to each of these classes. This is done by utilizing the spectral index value α, Equation 4, which is the slope of the spectral energy distribution.

$$\alpha = \frac{\mathrm{d}\log(\lambda f)}{\mathrm{d}\log(\lambda)} \tag{4}$$

This equation can be simplified into Equation 5. Where the square brackets denote data from those bands in units of magnitudes.

$$\alpha_{[4.5]-[8]} \approx 1.64([4.5] - [8]) - 2.82 \tag{5}$$

The spectral index is an important input for the separation for YSOs from regular field stars as in Figure 1. It was first added, as one of the methods by which each class of YSO is diagnosed is by applying cuts via the spectral index. That is, if the spectral index is positive (i.e. a rising SED), the YSO is likely a Class I. If the spectral index is approximately 0 (or flat SED), then this is a flat-spectrum YSO. Finally, negative spectral indices are indicative of Class II and Class III YSOs. The more negative the spectral index, the more advanced in evolution the stellar object. Indeed, the relatively young MS stars and Class III objects overlap in this space. We may apply a cut, in this particular diagram, at $\alpha = -2$, in which $\sim 98\%$ of stars occur above this line, and $\sim 95\%$ of YSOs occur below this line, making the two classes separable at 95% confidence limits. EG sources, however, have a high contamination in this phase space with YSOs, and thus strict lines cannot be drawn between those classes, however the most positive spectral indices will still belong to EG sources, and thus this information still remains valuable for separating all three classes. Hence, we utilize the original classifications for EG and Stars, as well as defining each true YSO candidate with a new label specific to the sub-class of YSO it belongs to.

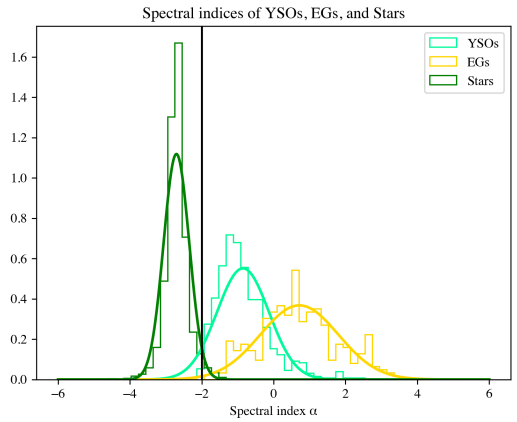

Figure 1: Histograms of the spectral index of the three over-arching classes, fitted with normal distributions. The area within each histogram is normalized.

As the spectral index is not a hard-and-fast rule to which type of object the YSO is, we rely instead upon machine learning to sort out the most similar data across all bands. First, we attempted the use of unsupervised clustering algorithms to sort each YSO into four categories, however, when reviewed, these algorithms sorted each class nearly equally among all four options. We tested the clustering algorithms of DBSCAN, Spectral Clustering, Gaussian Mixture Models, K-Means, Mini-batch K-Means, and Agglomerative Clustering, all clustering methods available

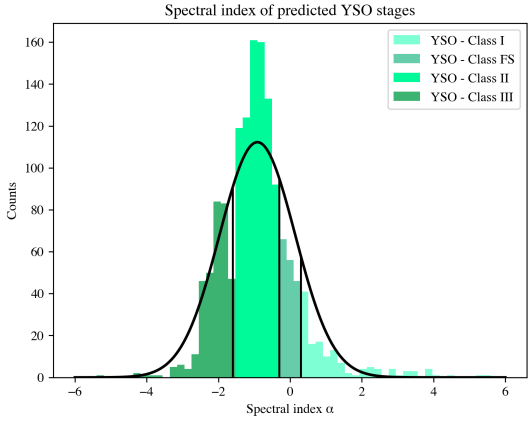

Figure 2: The distribution of YSO predictions via spectral index. Black vertical lines indicate where cuts were made for initial classification.

through the Python Sci-kit Learn library (Pedregosa et al., 2011). As these methods did not work, we instead chose to use both a trained MLP and three MLPs with cuts to determine the final YSO stage and find that the networks determine these cuts very clearly, as in Figure 2.

YSO Stage	Spectral Index
YSO - Class I	> 0.3
YSO - Class FS	≤ 0.3 and ≥ -0.3
YSO - Class II	< -0.3 and ≥ -1.6
YSO - Class III	< -1.6

Table 1: Cuts in spectral index to label YSO stages.

In order to have data which is comparable across our multiple surveys, we chose to convert all units of measurement into magnitudes, not mJy. To do this, we performed the standard calculations, using values provided via gemini.edu for the zero magnitude flux.

To begin, we make use of only the four IRAC bands, along with their errors. Numerous analysis (Cornu & Montillaud, 2021; Kuhn et al., 2021; Chiu et al., 2021), have found the IRAC bands to be the most useful for classifying objects. In our prescription, we also make use of the spectral index α, Equation 5.

We test several different data set-ups, testing different amounts of synthetic data, different methods of synthesizing data, different surveys to train upon, and finally settled on using the full c2d survey (Evans et al., 2014) to train, a survey of NGC 2264 (Rapson et al., 2014) to validate, and the c2d CORES (Evans et al., 2014) survey to test. Synthesized data via copulas is used to supplement the training data until 10k and 15k objects of each of our three classes (YSOs, EGs, and Stars) are obtained.

We iterate through a range of momentum values, learning rate values, layers, and hidden neurons until a best network is determined, which returns the highest F1-Score for the YSO class of the validation set. We then took the best network from those trained on 10k and 15k objects of each class, and saved their settings to be used later. Secondly, we find the best network including MIPS bands, and include this as another check. Lastly, we perform a similar iteration, now classifying the four YSO stages along with stars and extra-galactic sources using only IRAC data. The best network from these iterations is also saved.

2.2 Data Imbalance

A couple common concerns in machine learning relevant to this work are data imbalance between classes, and the amount of data available to train and validate on. If the training data is imbalanced, then the network will preferentially classify objects as the most often occurring object, which in our case are the stars. Hence, we adjust our training data such that we have an even amount of all classes. Our validation and test sets maintain observational proportions as we wish to ensure the model can perform correct classifications on imbalanced data.

Although using balanced data for the training set eliminates the possibility of over-training one particular set, it leaves us with not enough data to perform reliable classifications. Indeed, a widely used empirical rule (Cornu & Montillaud, 2021, e.g.) states that the size of the training data (N) must be a magnitude larger than the number of degrees of freedom (here translatable as the weights in the network), or:

$$N = [(n_i+1)n_h+(n_L-1)(n_h+1)n_h+(n_h+1)n_o]\times 10 \quad (6)$$

Where n_i is the number of input neurons, n_h is the number of neurons in the hidden layers, n_L is the number of hidden layers, and n_o is the number of output neurons. For the networks we test this can vary from 1330 to greater than 30 000 objects. It is not feasible to be able to obtain even amounts of YSOs, EGs and Stars from a single source to this extent of objects. Instead, we must consider the use of synthetic data.

We consider two different methods of synthetic data creation, and compare to not using synthetic data. First, we consider the use of copulas to produce synthetic data. Briefly, copulas are a method by which the cumulative distribution function of a pa-

rameter space is determined and used to create samples. We use a Gaussian copula within the package PYPI in the Python language to create our synthetic data, and find best settings for a balanced data-set with 5000, 10k, and 15k of each class.

Secondly, we produce synthetic data by taking classified data, grouping it by class, and taking the mean and standard deviation of each input band for each class. With this in hand, we produce synthetic data by varying within this mean and within the errors of these bands. We also produced 10k objects per class.

Finally, we use the data as is, and limit the number of EGs and Stars, such that they occur as frequently as YSOs in the training set. Since the number of stars greatly outweighed the number of YSOs and even EGs, and covered a wide parameter space, we had to carefully choose exactly which stars our networks should be trained on such that they were representative of all stars despite their small sample size. We do this by running the results of our previously determined "best" network 3000 times with different combinations of stars, saving the data which produced the highest F1-score on average over all classes.

3 Results

In order to ensure that the results presented are secure, we choose to cross-reference our four best networks. In doing so, we create three flags to identify the "security" of the classification. In particular, if the classification across at least two of our four networks is the same, we flag it as secure, else if none of the four methods agree, the object is flagged as insecure. Insecure objects are given the prediction from our MLP which generally performs the best.

Finally, we review the insecure classifications to see if they appear as outliers when manipulated via t-distributed stochastic neighbor embedding (t-SNE), 3. t-SNE is a method in which an unlabelled data set is convoluted into two-dimensional parameter space, which does not necessarily have a scientific equivalent. We find that there are very few objects which are labelled as insecure, and that the convolution performed by t-SNE shows outliers not necessarily labelled as insecure objects.

With both the full c2d survey and the survey of NGC 2264 we found that when the networks attempted to learn the Class III YSO stage, they invariably returned a large number of regular field stars and labeled them as Class III. The only way to avoid this while still retaining this class is to impose an empirical constraint. In this case, we previously noted that the spectral index of an object becomes more negative with it's evolution, and that Class III objects occur at this negative tail end of the YSO distribution. We hence chose to perform a secondary check on Class III YSOs to ensure their spectral index falls within the 95% confidence lower limits for YSOs. As there are many more stars than Class III YSOs, we choose this limit as the overabundance of stars will over-saturate the Class III YSO class in large samples. The final metrics for both the full c2d survey and the survey of NGC 2264 are located in Tables 2 and 3, respectively.

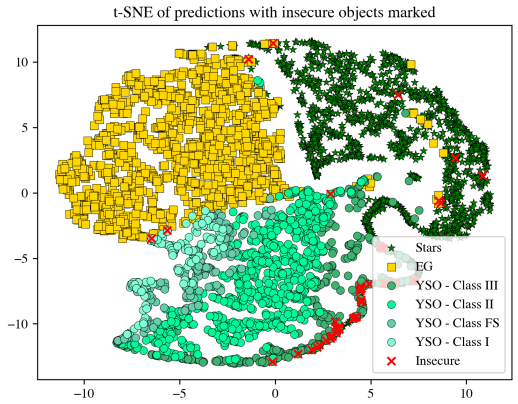

Figure 3: t-SNE representation of the data, with our predictions labelled by shape and/or colour. The locations of objects which none of the networks could agree upon are marked with red 'X's.

4 Discussion

Although the c2d survey provided visual extinction values for each of their objects, we choose to instead train on non-extinction corrected fluxes, as this allows a faster comparison when classifying unseen data. This results in a less precise network than other works (e.g. Chiu et al., 2021)

Despite the fact that t-distributed stochastic neighbour embedding provides no guarantee to group objects by class, we do see a very definite grouping by class. In particular, each of the YSO stages seem to bleed into the next, with Class III objects clumping along a line mixed with Class II objects and regular field stars which attaches the YSO class to the star class. This apparent evolutionary line indicates that, despite t-SNE not knowing the labels of the data, there is a strong correlation between all of the variables and their apparent age.

In regards to those objects that the t-SNE manip-

Class	Precision	Recall	F1-Score	# Objects
YSO - Class I	75	91	83	145
YSO - Class FS	84	98	90	166
YSO - Class II	77	99	87	809
YSO - Class III	32	47	38	343
YSO	66	86	75	1463
EG	64	92	76	3621
Stars	100	97	98	338497

Table 2: Metrics for the full c2d survey based upon our final predictions.

Class	Precision	Recall	F1-Score	# Objects
YSO - Class I	98	81	89	54
YSO - Class FS	94	99	96	67
YSO - Class II	78	99	87	358
YSO - Class III	57	57	57	74
YSO	78	92	84	553
EG	93	77	84	373
Stars	99	98	98	7518

Table 3: Metrics for NGC 2264 based upon our final predictions.

ulation places well outside of the domain of most of that class, this may be indicative of objects which our network has mislabelled, or it may be indicative of objects in those classes with interesting properties not considered within the design of the MLPs.

A Data-sets and network hyper-parameters

References

Chiu Y. L., Ho C. T., Wang D. W., Lai S. P., 2021, http://dx.doi.org/10.1016/j.ascom.2021.100470 Astronomy and Computing, 36, 100470

Cornu D., Montillaud J., 2021, http://dx.doi.org/10.1051/0004-6361/202038516 , 647, A116

Evans N. J. I., et al., 2014, VizieR Online Data Catalog, p. II/332

Furlan E., et al., 2016, http://dx.doi.org/10.3847/0067-0049/224/1/5 , 224, 5

Gutermuth R. A., et al., 2008, http://dx.doi.org/10.1086/524722 , 674, 336

Gutermuth R. A., Megeath S. T., Myers P. C., Allen L. E., Pipher J. L., Fazio G. G., 2009, http://dx.doi.org/10.1088/0067-0049/184/1/18 , 184, 18

Kuhn M. A., de Souza R. S., Krone-Martins A., Castro-Ginard A., Ishida E. E. O., Povich M. S., Hillenbrand L. A., COIN Collaboration 2021, http://dx.doi.org/10.3847/1538-4365/abe465 , 254, 33

Megeath S. T., et al., 2012, http://dx.doi.org/10.1088/0004-6256/144/6/192 , 144, 192

Miettinen O., 2018, http://dx.doi.org/10.1007/s10509-018-3418-7 , 363, 197

Pedregosa F., et al., 2011, Journal of Machine Learning Research, 12, 2825

Rapson V. A., Pipher J. L., Gutermuth R. A., Megeath S. T., Allen T. S., Myers P. C., Allen L. E., 2014, http://dx.doi.org/10.1088/0004-637X/794/2/124 , 794, 124

Bands	Class	Training	Validation	Testing
IRAC	All	Evans et al. (2014)	Evans et al. (2014)	Rapson et al. (2014)
	YSO	Evans et al. (2014)	Evans et al. (2014)	Rapson et al. (2014)
Full Spitzer[1]	All	Evans et al. (2014)	Evans et al. (2014)	Rapson et al. (2014)
	YSO	Evans et al. (2014)	Evans et al. (2014)	Rapson et al. (2014)

Table 4: The data-sets used to train, validate, and test for each network. See text for details.

	IRAC YSE	IRAC YSE 2	IRAC YSO	IRAC + 24 μm YSE
learning rate	0.001	0.01	0.1	0.1
momentum	0.9	0.75	0.6	0.75
epochs	3000	3000	3000	3000
layers	2	1	1	1
neurons	20	10	10	50

Table 5: Best network parameters for Spitzer. "YSE" indicates the network was trained to separate YSOs, Stars, and EGs. "YSO" indicates the network was trained to separate the stages of YSOs in addition to the other classes.