DUSTY STELLAR SOURCES CLASSIFICATION BY IMPLEMENTING MACHINE LEARNING METHODS BASED ON SPECTROSCOPIC OBSERVATIONS IN THE MAGELLANIC CLOUDS

Sepideh Ghaziasgar,[1] Mahdi Abdollahi,[1] Atefeh Javadi,[1] Jacco Th. van Loon,[2] Iain McDonald,[3]
Joana Oliveira,[2] Amirhossein Masoudnezhad,[4] Habib G. Khosroshahi,[1,5] Bernard H. Foing,[6] and
Fatemeh Fazel[6]

[1]School of Astronomy, Institute for Research in Fundamental Sciences (IPM), P.O. Box 1956836613, Tehran, Iran
[2]Lennard-Jones Laboratories, Keele University, ST5 5BG, UK
[3]Jodrell Bank Centre for Astrophysics, Alan Turing Building, University of Manchester, M13 9PL, UK
[4]Department of Physics, Sharif University of Technology, P.O. Box 11155-9161, Tehran, Iran
[5]Iranian National Observatory, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran
[6]Leiden Observatory, Leiden University, P.O.Box 9513, 2300 RA Leiden, Netherlands

ABSTRACT

Dusty stellar point sources are a significant stage in stellar evolution and contribute to the metal enrichment of galaxies. These objects can be classified using photometric and spectroscopic observations using color-magnitude diagrams (CMD) and infrared excesses in spectral energy distributions (SED). We employed supervised machine learning spectral classification to categorize dusty stellar point sources, including young stellar objects (YSOs) and evolved stars comprising oxygen- and carbon-rich asymptotic giant branch stars (AGBs), red supergiants (RSGs), and post-AGB (PAGB) in the Large and Small Magellanic Clouds, based on spectroscopic labeled data derived from the Surveying the Agents of Galaxy Evolution (SAGE) project, which involved 12 multiwavelength filters and 618 stellar objects. Despite dealing with missing values and uncertainties in the SAGE spectral datasets, we achieved accurate classifications of these sources. To address the challenge of working with small and imbalanced spectral catalogs, we utilized the Synthetic Minority Oversampling Technique (SMOTE), which generates synthetic data points. Subsequently, among all the models applied before and after data augmentation, the Probabilistic Random Forest (PRF) classifier, a tuned Random Forest (RF), demonstrated the highest total accuracy, reaching **89**% based on the recall metric in categorizing dusty stellar sources. In this study, using the SMOTE technique does not improve the accuracy of the best model for the CAGB, PAGB, and RSG classes; it stays at **100**%, **100**%, and **88**%, respectively. However, there are variations in the OAGB and YSO classes. Accordingly, we collected photometrically labeled data with properties similar to the training dataset and classified them using the top four PRF models with an accuracy of more than **87**%. We also collected multiwavelength data from several studies to classify them using our consensus model, which integrates four top models to present common labels as the final prediction.

1. INTRODUCTION

Dusty stellar objects are significant in the chemical enrichment of the interstellar medium (ISM) and galaxies, which contain heavy elements required for star

Corresponding author: Atefeh Javadi
atefeh@ipm.ir

and planet formation. These objects are point sources surrounded by gas and dust that are important for a wide range of processes, including the loss of mass from evolved stars (Höfner & Olofsson 2018) and the formation of these stars. These types of stars are easily detectable due to their high luminosity ($\sim 10^{3.5-5.5}$ L$_\odot$) and red colors (Karakas & Lattanzio 2014), making them stand out in galaxies, especially in the infrared (IR) observations. The Magellanic Clouds (MCs), com-

prising the Large and Small Magellanic Clouds (LMC and SMC), are nearby dwarf irregular galaxies, respectively 50 kpc and 60 kpc away (Pietrzyński et al. 2013; Subramanian & Subramaniam 2009, 2011) with metallicities of 0.5 and $0.2Z_\odot$ (Russell & Dopita 1992) is an excellent location to investigate stellar contributions to dust emission (Ruffle et al. 2015a). In addition, since the MCs are close to the Milky Way, their stellar populations and dusty stellar sources can be distinguished with considerable accuracy. The point sources in these galaxies can be identified photometrically according to their color-magnitude diagrams (CMDs). However, photometric identification alone may not provide sufficient information to distinguish between classes of dusty stars. To enable more accurate stellar classification, infrared spectra obtained by the Infrared Spectrograph (IRS) aboard the *Spitzer* Space Telescope (Houck et al. 2004) have been used to classify approximately 1000 point sources across the LMC and SMC (Kemper et al. 2010; Woods et al. 2011; Ruffle et al. 2015a; Jones et al. 2017b). Dusty stellar point sources are classified according to their evolutionary stage and chemical enrichment, comprising young stellar objects (YSO), asymptotic giant branches (AGB; oxygen- and carbon-rich), red supergiants (RSG), and post-asymptotic giant branches (PAGB). While other dusty stellar sources, such as AGNs (Hony et al. 2011; Pennock et al. 2022) and OB stars (Adams et al. 2013; Sheets et al. 2013), may potentially have IR excesses, this study's scope is centered around the five dusty stellar classes discussed due to the availability of the IRS spectral labeled data.

The early-evolutionary star is known as a YSO. It can be a valuable probe for studying recent star-formation activity in galaxies (Kokusho et al. 2023), along with their physical conditions and processes. Due to the star formation process, YSOs are surrounded by protostellar envelopes containing gas and dust disks (Suh 2016). Therefore, YSOs contribute to our understanding of star formation and allow us to characterize star-forming regions. In addition, the YSOs can have varying luminosities based on their mass and evolution stage, which can be classified based on their spectral class from optical to IR and luminosity (Seale et al. 2009; Oliveira et al. 2013; Miettinen 2018).
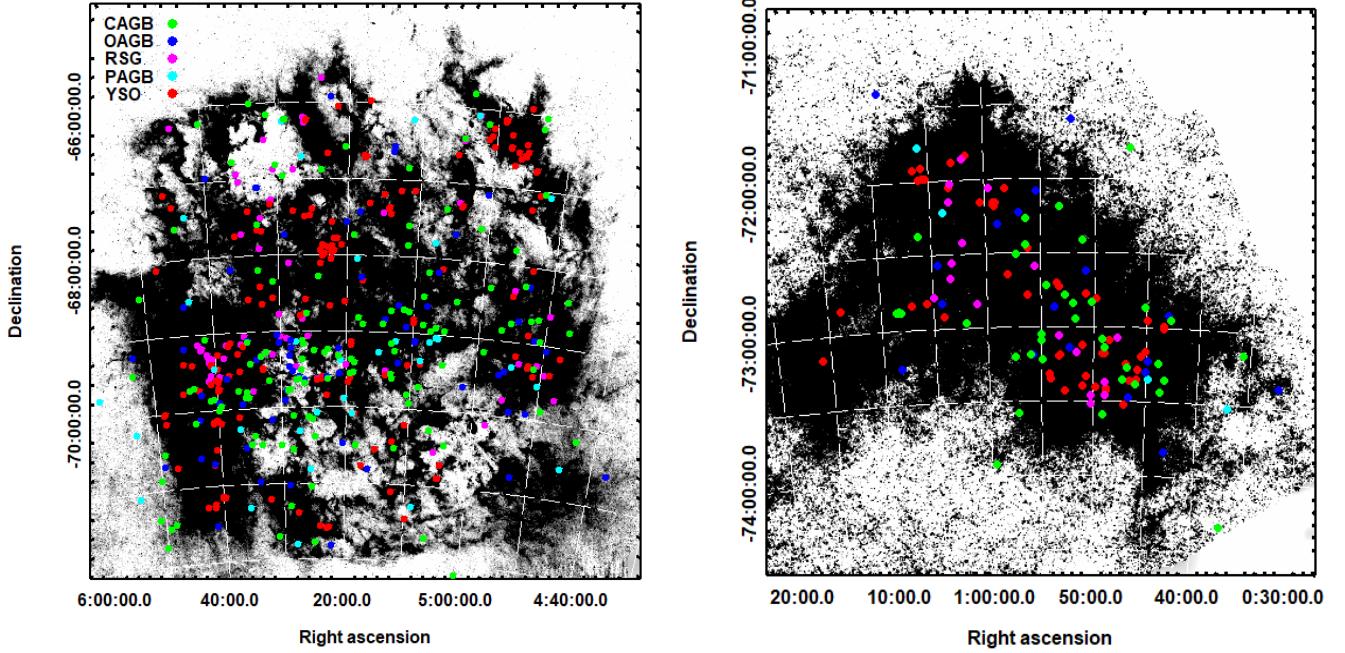
The evolved stars (Boyer et al. 2011) are dust producers, and their role in enriching galaxies is crucial. In the final stage of stellar evolution, low– and intermediate mass (0.8-8 M$_\odot$) stars evolve into the Asymptotic Giant Branch (AGB) phase, and high mass (M$\geq$ 8 M$_\odot$) stars enter the Red Supergiants (RSG) phase (Herwig 2005; Höfner & Olofsson 2018). During the AGB phase, a fraction of the mass of the stars is lost at

rates of up to $\sim 10^{-4}$ $M_\odot$ yr^{-1} (Wood et al. 1992; Srinivasan et al. 2009), leading to the chemical enrichment of the interstellar medium (ISM). Most AGB stars are distinguished as long-period variables (LPVs) with pulsation periods in the range of months to years (Wood et al. 1983; Whitelock et al. 2003; Vijh et al. 2009; Javadi et al. 2011a; Riebel et al. 2012), and they have circumstellar dust envelopes with high mass-loss rates (Javadi et al. 2011b; McDonald & Zijlstra 2016; Goldman et al. 2017; Höfner & Olofsson 2018; Javadi & van Loon 2022). According to the chemistry of the photosphere and the outer dust envelope and their carbon-to-oxygen abundance ratios, carbon-rich AGB stars (C-rich or CAGB stars; $C/O > 1$), oxygen-rich AGB stars (O-rich or OAGB stars; $C/O < 1$), and S-type stars ($C/O \sim 1$) can be categorized. Hence, the OAGB stars have more oxygen than carbon, forming silicates and oxides, while CAGB stars, where carbon exceeds oxygen due to the third dredge-up process (Kamath 2020), produce carbon-based molecules. As a result, these stars differ in spectral characteristics and dust production due to their initial mass and metallicity (Javadi et al. 2011a,b, 2013; Suh 2020, 2021).

The RSGs evolved from the Main sequence O-types or B-types with initial masses ($\sim$8-30 M$_\odot$) in the helium-burning evolutionary phase. During this phase, there is a significant increase in size and a decrease in surface temperature. Therefore, they are at a key stage in stellar evolution, and lastly, their core collapses to produce supernovae, neutron stars, or black holes. (Massey & Olsen 2003; Levesque 2010; Yang et al. 2020).

The PAGB stars, in the late stages of their evolution, are luminous, low- or intermediate-mass stars that conclude their AGB branch evolution with a phase of mass loss. Their transition from giant branches to white dwarfs after the oxygen-rich and carbon-rich AGB phases (Kamath et al. 2014, 2015). Having shed their outer layers, these stars are characterized by unique chemical compositions. These stars provide information about nucleosynthesis processes in AGB stars and how galaxies evolve chemically (van Winckel 2003; Kamath 2020).

In recent years, the growth of astronomical datasets, including photometric and spectroscopic catalogs, has increased thanks to ground- and space-based telescopes and surveys. In this way, machine learning methods (Ball & Brunner 2010; Baron 2019) have emerged as powerful tools for classifying astronomical objects due to their significant time-saving capabilities. Therefore, the stellar classification using artificial intelligence techniques, including machine learning and deep learning, is becoming increasingly important, and there are

Figure 1. Left panel: Locations of dusty sources in the Large Magellanic Cloud (LMC) based on *Spitzer*/Infrared Spectrograph (IRS) data, including sources from both the SAGE-Spec program and archival observations. Right panel: Locations of dusty sources in the Small Magellanic Cloud (SMC) selected from IRS staring-mode archival data within the SAGE-SMC footprint. The background images of the Magellanic Clouds were taken from the *Herschel* Space Observatory (Pilbratt et al. 2010).

many studies regarding this trend (Kuntzer et al. 2016; Pashchenko et al. 2018; Miettinen 2018; Brice & Andonie 2019; Cornu & Montillaud 2021; Dorn-Wallenstein et al. 2021; Kinson et al. 2021; Ghaziasgar et al. 2022; Maravelias et al. 2022; Kinson et al. 2022; Abdollahi et al. 2023; Smith & Geach 2023; Wilson et al. 2023; Cody et al. 2024; Zeraatgari et al. 2024; Ghaziasgar et al. 2024; Pennock et al. 2025; Li et al. 2025).

Machine learning algorithms are commonly categorized into supervised and unsupervised methods, each offering advantages and limitations. In supervised learning, models are trained using labeled datasets to learn patterns from known classifications. This approach can provide accurate results, mainly when reliable labeled data is available (Baron 2019; Li et al. 2025). Supervised learning is limited in identifying new objects because its performance depends on the quality and completeness of the training data (Pennock et al. 2025). Additionally, the classifier's performance may be affected if the training dataset is imbalanced, meaning that some classes contain significantly more samples than others. Furthermore, supervised learning requires a large amount of labeled data, which is difficult since labeling stellar sources is time-consuming (Pennock et al. 2025; Li et al. 2025). In contrast, unsupervised learning is employed when no predefined labels exist, making it a powerful tool for discovering hidden structures within the data

(Ball & Brunner 2010; Ivezić et al. 2014; Baron 2019; Sen et al. 2022; Djorgovski et al. 2022). However, a significant limitation of unsupervised learning is its lack of astrophysical interpretability (Fotopoulou 2024), making it less suitable for our dataset. Given the labeled nature of the dataset, supervised learning is the most appropriate choice. Furthermore, since our objective is to refine existing classifications rather than identify new subpopulations, a supervised approach aligns better with the goals of this study.

The dataset used in this study consists of spectroscopically confirmed classifications of LMC and SMC point sources, initially identified photometrically using *Spitzer* imaging (see Section 2) (Meixner et al. 2006). Using the multiwavelength infrared survey, these sources were first identified through the Surveying the Agents of Galaxy Evolution (SAGE) program (Gordon et al. 2011). Spectroscopic follow-ups were performed through the targeted SAGE-Spec *Spitzer* Legacy program and additional archival IRS observations compiled into the broader SAGE-Spec (Kemper et al. 2010) database. While not all photometrically identified sources in SAGE have spectroscopic follow-up, this study exclusively used sources with confirmed spectroscopic classifications. Ongoing and future spectroscopic observations are expected to further improve the classification and understanding of stellar objects and dusty

stars by increasing the volume of labeled data. This will enable more accurate machine learning analyses on dusty stellar populations, allowing models to identify such sources better and enhance classification performance (Jones et al. 2023). Mid-infrared stellar population studies are typically limited to nearby galaxies; however, the James Webb Space Telescope (JWST) will extend these capabilities to galaxies across the Local Volume (Jones et al. 2017a).

In this study, we implemented supervised learning algorithms on the IRS spectral dataset to distinguish young stellar objects (YSOs) and evolved stars. The models were trained on spectroscopically confirmed data and evaluated for classification performance.
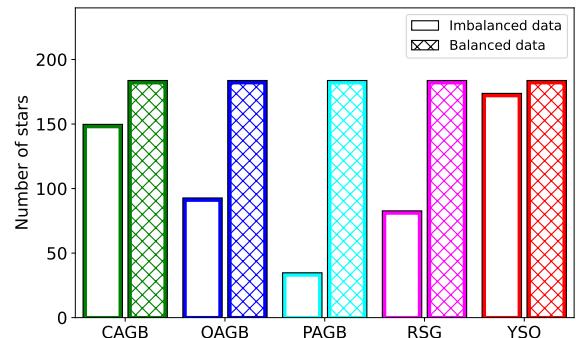
The organization of the study is as follows: we delineate the archival dataset for training and testing in Section 2. In Section 3, the method designed for preprocessing is proposed. Section 4 details some models for multiclassification described. We present classification results in Section 5. The metallicity effect on classifying dusty stellar sources in the MCs is given in Section 6. In Section 7, we compare the classification results obtained from the newly collected dataset with its corresponding photometric classifications. Finally, in Section 8, our findings are summarized.

2. DATA

The Surveying the Agents of Galaxy Evolution (SAGE) collaboration (Meixner et al. 2006) used the *Spitzer* Space Telescope (Werner et al. 2004) to obtain infrared imaging and spectroscopic data for the Magellanic Clouds, using the Infrared Array Camera (IRAC; Fazio et al. (2004) 3.6, 4.5, 5.8, and 8.0 μm bands) and the Multiband Imaging Photometer (MIPS; Rieke et al. (2004) 24, 70, and 160 μm bands). Observations were conducted in two epochs, three months apart.

The spectroscopic data in this study come from the SAGE-Spec (Kemper et al. 2010) database, which includes all *Spitzer*/IRS staring-mode observations within the regions covered by SAGE-Spec (for the LMC) and the SAGE-SMC survey. This database combines data from the SAGE-Spec Legacy program (Kemper et al. 2010), which targeted 197 point sources (Woods et al. 2011) in the LMC, and additional archival IRS observations from various *Spitzer* programs. Nearly 800 point sources were observed in the LMC, resulting in over 1000 IRS spectra (Jones et al. 2017b). For the SMC, 209 point sources within the SAGE-SMC footprint were observed using IRS staring-mode (Ruffle et al. 2015a).

We used the spectroscopically labeled sources published in the classification catalogs of the LMC and SMC (Woods et al. 2011; Ruffle et al. 2015a; Jones et al.
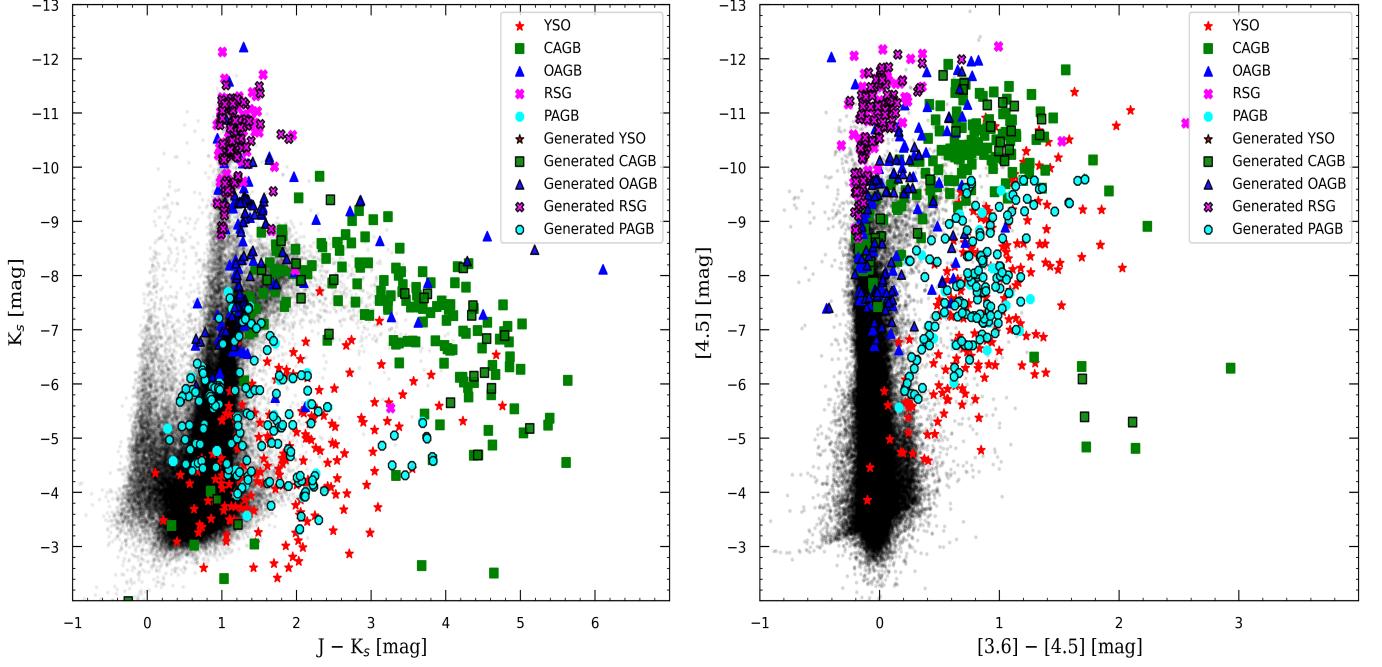

Figure 2. Dusty stellar sources distribution for the training dataset before and after data augmentation within the Synthetic Minority Oversampling Technique (SMOTE).

2017b), whose spatial distribution is shown in Fig. 1.

The spectral classes (SpClass) of dusty stellar objects in both the LMC and SMC catalogs were derived using a structured binary decision tree (yes/no) algorithm, as illustrated in the classification flowcharts by (Woods et al. 2011; Ruffle et al. 2015a; Jones et al. 2017b). In multiclassification studies, decision trees expand binary decisions (yes/no branches) into multi-way decisions, and objects are classified starting at the root node and sorted according to their feature values (Murthy 1998; Rokach & Maimon 2005). This algorithm consists of many steps that bring us closer to the final label, such as redshift, Spitzer infrared spectral features, continuum, spectral energy distribution shape, and bolometric luminosity. For instance, in the top part of the SAGE-Spec classification flowchart, the galaxy and star branches are distinguished according to redshift features. To continue, the stars are subdivided into multiple branches by considering different features and thresholds. The SAGE spectral catalog contains 1071 objects, including stars, galaxies, and planetary nebulae (Woods et al. 2011; Jones et al. 2017b). Among these, 618 objects are classified as dusty stellar sources (see Table 1), with 486 belonging to the LMC and 132 to the SMC. The remaining objects, such as galaxies and planetary nebulae, are present in the catalog but excluded from our analysis, since this study focuses exclusively on dusty stellar sources.

3. DATA PREPROCESSING AND EXPLORATORY DATA ANALYSIS

In preprocessing, we perform a series of steps to prepare the data for model training. For classification, we have collected two datasets from LMC and SMC (Jones et al. 2017c; Ruffle et al. 2015b) that include several columns of data. These columns consist of optical to far-

Figure 3. Color-magnitude diagram (CMD) plots the magnitude of stars against their color, representing the locations of each dusty stellar class of object. As shown in this figure, dusty stellar categories are shown along with those generated with the SMOTE method as discussed in Section 3.1, in the near and mid-infrared bands. The left panel shows near-infrared data from 2MASS (J, K_s), and the right panel shows mid-infrared data from *Spitzer* ([3.6], [4.5]). The black lines around stellar types indicate data generated by SMOTE, and the black dots in the background represent all the stars that belong to the Magellanic Clouds.

Table 1. This is a list of dusty stars within the Magellanic Clouds, categorized by their spectral classes. *The "Augmented Data" column represents the population after data augmentation using the SMOTE method. Each number indicates the sum of the train and test datasets in each class, with SMOTE applied only to the train datasets to equalize their population to the most populated class.

Classes	LMC	SMC	Total	*Augmented Data
CAGB	136	38	174	200
OAGB	88	19	107	193
PAGB	33	4	37	183
RSG	72	22	94	190
YSO	157	49	206	206
Total	486	132	618	972

infrared magnitudes gathered from different telescopes, such as 2MASS and Spitzer. Additionally, the dataset consists of other features calculated from various studies. In general, many features provide more information for model training but may increase the complexity of training classifiers. It can also be challenging to handle NaN records due to the many missing values, which require oversampling algorithms to fill the gaps. Hence,

we faced these issues and solved them step by step, as presented in the following:

1. The primary dataset contains 14 features, including UMmag, BMmag, VMmag, IMmag, J2mag, H2mag, Ks2mag, IRAC1, IRAC2, IRAC3, IRAC4, [24], [70], [160].

2. Exploring the data, we found some columns in our dataset containing many NaN records. Consequently, columns with more than 500 NaNs (more than 70% of all records) have been removed, which includes columns [70] and [160]. As a result, 12 features remain.

3. Drawing the correlation matrix (presented in Appendix D), we used this matrix to identify correlated features and removed those with a high correlation. Therefore, correlated features do not assist the model in learning and can sometimes complicate the learning process, potentially leading to instability in model parameters (Hastie et al. 2009). According to Fig. D.1, closely related features (filters) demonstrate high correlation and overlap during preprocessing. This is expected due to the behavior of blackbody radiation of stars, which rises at shorter wavelengths and falls at longer ones (Rybicki & Light-

Table 2. Sample of spectral properties of detected dusty stellar point sources in the Magellanic Clouds, derived from *Spitzer*/IRS observations. The data for the LMC include sources from both the SAGE-Spec *Spitzer* Legacy program and archival IRS observations within the SAGE-LMC footprint (Jones et al. 2017c). For the SMC, sources were selected from the complete IRS archive based on 209 IRS staring-mode observations of IRAC point sources located within the SAGE-SMC region (Ruffle et al. 2015b). This table presents five representative dusty stellar sources from the Magellanic Clouds across 12 observational filters, with distance modulus values adopted from (Scowcroft et al. 2016; Bhardwaj et al. 2016) and extinction corrections from Schlafly & Finkbeiner (2011). The RA and DEC are included for positional reference and are not used as classification features.

RA	DEC	U	B	V	I	J	H	K_s	[3.6]	[4.5]	[5.8]	[8.0]	[24]	SpClass
(deg)	(deg)	(mag)	(mag)	(mag)	(mag)	(mag)	(mag)	(mag)	(mag)	(mag)	(mag)	(mag)	(mag)	
71.6131	-68.7963	1.63	2.23	0.53	-2.14	-3.50	-5.24	-6.84	-8.67	-9.39	-9.81	-10.24	-11.19	CAGB
71.8277	-69.7057	-0.73	-2.88	-4.8	-7.22	-8.42	-9.25	-9.51	-9.65	-9.48	-9.65	-9.76	-10.97	RSG
72.1573	-69.3936	NaN	3.45	2.47	1.14	NaN	NaN	-5.60	-8.15	-9.04	-9.84	-10.65	-13.66	YSO
16.0398	-72.8377	-2.2	-3.75	-5.56	NaN	-8.64	-9.32	-9.59	-9.7	-9.71	-9.92	-10.05	-10.8	OAGB
15.9264	-72.2284	-5.30	-5.16	-5.06	-5.39	-5.89	-6.37	-7.20	-8.99	-9.75	-10.36	-11.42	-13.68	PAGB

man 1986; Carroll & Ostlie 2017), resulting in lower correlation where the difference between two features is greater. As a result, all selected features exhibit an acceptable correlation matrix, indicating that no features need to be eliminated.

4. There are still many NaNs in these remaining features, and we need to find a solution. Thus, we neglected stellar sources with more than six NaNs. This threshold is chosen based on the size of the dataset and the improvement in training performance. Finally, after removing some dusty objects, we selected 618 stars with some remaining NaN values. Table 1 presents the final population of each class in the used dataset. In detail, the sample of each class with its features is shown in Table 2.

5. To fill in the remaining missing values, we applied an IterativeImputer technique (Pedregosa et al. 2011) from the `scikit-learn` library in Python to handle them. For each missing feature in the dataset, the IterativeImputer constructs a regression model utilizing the remaining features as predictors. This model approximates the missing values for each feature, and the process is repeated until no further approximations can be made. Applying the approximated missing values to the subsequent iteration makes the approximations more precise with each iteration (Pedregosa et al. 2011).

6. Finally, we applied extinction (Schlafly & Finkbeiner 2011) and distance modulus (DM) (Scowcroft et al. 2016; Bhardwaj et al. 2016) to all 12 filters at the Magellanic Clouds.

After these preprocessing steps to handle and fill in the missing data, the dataset should be split into training and testing sets. Typically, different ratios are used for training and testing to evaluate the model's performance, and the choice of split ratio depends on the dataset and its distribution. In our case, we partitioned the preprocessed dataset by allocating 85% of the data to training and 15% to testing.

As an optional step, we can use the Exploratory Data Analysis (EDA) approach, which includes statistical and visual tools, to understand the data better. One of these tools is the pairplot (presented in Appendix E), which visualizes features in various colors according to their relationship. In consequence, Fig. E.1 is drawn based on the multiwavelength features selected for training the model. Since dusty point sources overlap with different classes, they cannot be distinguished by a single or two particular features, making classification challenging.

3.1. *Data Augmentation*

The data we used does not have an equal distribution of stars among the five spectral classes. This indicates that we have imbalanced classes of dusty stellar point sources, as illustrated in Fig. 2. We can expect the class distribution to be imbalanced and the skewness to be present when we have actual data (Liu et al. 2021).

The imbalanced distribution can pose challenges during model training, as it may bias the models toward more populated classes. This bias occurs because the models learn more from classes with larger populations (Sun et al. 2009). Thus, classifier models may find it more difficult to effectively learn from the minority class, which contains significantly fewer samples compared to the majority class.

It is possible to solve the problem of imbalanced data distribution in each class and balance samples among classes using a variety of resampling techniques, including undersampling (reducing samples from majority classes), oversampling (adding additional samples to minority classes) (Liu et al. 2021; Zhang et al. 2023), and

hierarchical classification (Bader-El-Den et al. 2016; Hosenie et al. 2020; Abdollahi et al. 2023).

Typically, oversampling is performed with the Synthetic Minority Oversampling Technique (SMOTE), which generates random synthetic data from the nearest neighbors of a minority class based on Euclidean distance (Chawla et al. 2011). Through SMOTE, a few samples are linearly interpolated into their neighbors, and a certain number of artificial minority samples are generated to reduce the data imbalance ratio (Wang et al. 2021).

We applied this data augmentation method to the training dataset to address imbalanced distributions within stellar classes, as shown by bars with black diagonal cross lines in Fig. 2. Also, Table 1 shows the final population of dusty stellar objects with imbalanced training and test data in a column named total and balanced data in an augmented column. To understand how SMOTE works and how it increases the number of stars in each stellar class, see the CMDs comparison in near-IR (J, K_s) and mid-IR [3.6], [4.5] passbands before and after applying SMOTE shown in Fig. 3. Therefore, the graph shows that the SMOTE method generates the minority class, namely PAGB stars, based on the population of the YSOs, which is the majority class. As a result, the augmented PAGB stars are generated near the previous PAGB stars, where the YSOs are located. After generating the data, it is clear that there is an overlap between these two classes, YSO (red stars) and PAGB (cyan circles), which may cause issues with the model training.

4. MODELS

In machine learning, the selection of models depends on the nature of the problem, data characteristics, and especially data size (Smith & Geach 2023). For instance, neural networks typically require large datasets (with tens of thousands of samples) to learn effectively, due to their high complexity. Since this study deals with a low stellar population, we chose classical algorithms that can provide reliable classification even with sparsely populated data.

The models we used as described (see Appendix A) included Probabilistic Random Forest (PRF) (Reis et al. 2019; Baron 2019; Kinson et al. 2021, 2022; Pennock et al. 2025), Random Forest (RF) (Breiman 2001; Carliles et al. 2010; Baron & Poznanski 2017; Baron 2019), K-Nearest Neighbor (KNN) (Altman 1992), C-Support Vector Classification (SVC) including SVC-poly and SVC-rbf (Vapnik 1995; Baron 2019), and Gaussian Naive Bayes (GNB) (Wilson et al. 2023).

When dealing with training models, each model has

hyperparameters, such as the number of estimators in the RF model and neighbors in the KNN model. Therefore, we employed manual and automatic adjustments using the "Grid Search" (Pedregosa et al. 2011) from the `scikit-learn` library to find the optimal hyperparameters of the model. The grid search assigns possible values to each one to determine which hyperparameter performs best.

The PRF, which is a developed version of Random Forest (see Appendix A.1), generates a forest of decision trees where each tree is trained using a different subset of the data and a random set of features. It assigns probability distributions to each output class, as illustrated in Reis et al. (2019), and can handle noisy datasets. Despite constant hyperparameters, different answers are obtained in the PRF model. By selecting various hyperparameters and finding the best model obtained Pennock et al. (2025), we explore several models to address the uncertainty in this problem.
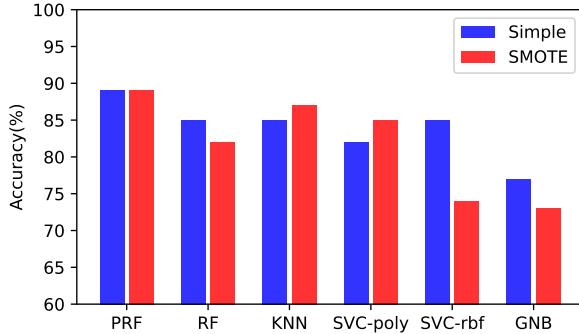
In this study, we train and test the models with two kinds of datasets, imbalanced data and balanced data augmented by the SMOTE method, as mentioned in Section 3.1. In the next section, we will present the results in two approaches, Simple (referred to as imbalanced data) and SMOTE (referred to as balanced data using the augmented technique), named based on the distribution of each of the datasets and shown in Fig. 2.
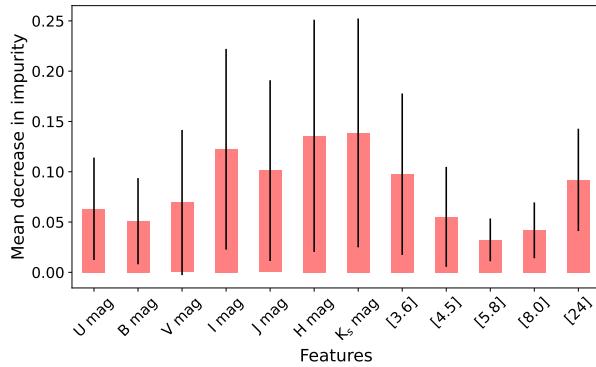
5. CLASSIFICATION RESULTS

This section compares the classification of dusty stellar objects before and after using the SMOTE method, and the classifiers in each approach are evaluated using the performance metrics (Grandini et al. 2020) and the confusion matrix presented in Appendix B and C.

As shown in Fig. 4, the comparison of both Simple and SMOTE approaches in six different classification models indicates that the SMOTE approach outperforms the Simple ones in two of them. In comparison, the situation is reversed for the three other models, including RF, SVC-rbf, and GNB. Using the SMOTE method can sometimes result in misclassification, especially if the original sample point is near the edge of the minority sample distribution, and may not always produce an improved classification (Wang et al. 2021; Last et al. 2017).

According to Fig. F.1 and Fig. F.2, we also find that increasing the population of the sparsely populated star class results in misclassification in the class that overlaps with them. In the case of the OAGB star class, which is a third less populated class with 96 stars, applying data augmentation through the SMOTE method results in increased overlaps with the CAGBs (as shown

Figure 4. This chart compares the performance of a Simple classifier with a SMOTE classifier based on their respective accuracy scores.

Figure 5. The feature importance diagram ranks the significance of different variables (or features) using the RF model's Mean Decrease in Impurity (MDI) algorithm A.1. This diagram illustrates how important each of the filters in the SAGE catalog is for making accurate stellar classification predictions. As shown, the importance of the stellar classes and the selected features (12 filters) are all the same, and no feature is preferred.

in Fig. 3). Based on the comparison of SVC-rbf and GNB models before and after data generation, it can be seen that some CAGBs are classified as OAGBs by models that use augmented data. These misclassifications are reversed among OAGB, PAGB, and RSG stars. As a result, some OAGB stars are classified as PAGBs and RSGs, as shown in the confusion matrix of RF and SVC-rbf models.

Comparing the models in each approach, the results obtained from PRF models were superior to others, with the best record of 89% accuracy. As shown in Table 3, the PRF model was trained with different values for the hyperparameters, keep_probability, and number of estimators due to the uncertainty we found in them. The best PRF model classification reports in each approach are shown in Tables 4 and 5. Considering two models

Table 3. PRF models comparison for Simple and SMOTE in terms of accuracy.

Number of estimator	Keep-probability	Accuracy (Simple)	Accuracy (SMOTE)
10	0.01	89	87
10	0.1	89	89
10	0.3	89	87
10	0.5	87	89
10	0.8	89	85
10	0.9	89	85
100	0.1	85	87

with the highest accuracy in different approaches, we find that the Simple approach shows better results in terms of macro average accuracy at 83%. The other comparison method utilizes the confusion matrix, as illustrated in F.1. The accuracy of the three classes, including CAGB, PAGB, and RSG, does not change when switching from the Simple to the SMOTE approach. However, the accuracy of OAGB decreases from 73% to 64% due to classifying some stars as RSGs, which occurs more frequently than in the Simple approach. This could result from data augmentation of other classes, especially the RSG class, which is close to OAGB, as shown in the CMDs in Fig. 3. In contrast, the accuracy of YSOs increases from 88% to 92%, indicating that data augmentation could work better for certain classes. It could be the result that data augmentation does not work well for all classes, and its effectiveness varies depending on the class and the specific problem. As shown in Table 3, several PRF models have the same results. Therefore, we selected four of them from both approaches for classification and comparison in Section 7.

Regarding accuracy, the RF model is one of the accurate models. This model performed well in the Simple approach but not in the SMOTE one. In addition, this model provides a report showing the importance of each feature. Indeed, identifying the most important features in machine learning helps streamline the problem by eliminating useless ones. The RF classifier can employ the Mean Decrease in Impurity (MDI) algorithm (Breiman 2001; Li et al. 2019) to assess the importance of features. The MDI for each feature is calculated by averaging the decrease in impurity over all trees in the ensemble. Features with high MDI values are more important for prediction. As Fig. 5 demonstrates, based on the MDI, on the Y axis, there is no significant difference in the importance of used features; we did not remove any feature based on this calculation. However, it does reveal that the significance of some infrared passbands

Table 4. Classification report, considering the number of estimators: 10, Keep probability: 0.8, Simple PRF.

Class	Precision	Recall	F1-score
CAGB	0.95	1.00	0.97
OAGB	0.80	0.73	0.76
PAGB	0.50	1.00	0.67
RSG	0.78	0.88	0.82
YSO	0.95	0.88	0.91
accuracy			0.89
macro avg	0.80	0.90	0.83
weighted avg	0.89	0.89	0.89

Table 5. Classification report, considering the number of estimators: 10, Keep probability: 0.5, SMOTE PRF.

Class	Precision	Recall	F1-score
CAGB	0.86	1.00	0.92
OAGB	1.00	0.64	0.78
PAGB	0.50	1.00	0.67
RSG	0.70	0.88	0.78
YSO	1.00	0.92	0.96
accuracy			0.89
macro avg	0.81	0.89	0.82
weighted avg	0.91	0.89	0.89

Table 6. Classification report for the SMC catalog. The classification was performed using the following settings: number of estimators = 10, Keep probability = 0.8, and Simple PRF.

Class	Precision	Recall	F1-Score
CAGB	1.00	1.00	1.00
OAGB	1.00	1.00	1.00
PAGB	0.00	0.00	0.00
RSG	1.00	1.00	1.00
YSO	0.80	1.00	0.89
accuracy			0.93
macro avg	0.76	0.80	0.78
weighted avg	0.88	0.93	0.90

Table 7. Classification report for the LMC catalog. The classification was performed using the following settings: number of estimators = 10, Keep probability = 0.8, and Simple PRF.

Class	Precision	Recall	F1-Score
CAGB	0.94	0.88	0.91
OAGB	0.72	0.90	0.86
PAGB	0.67	0.67	0.67
RSG	1.00	0.90	0.95
YSO	0.80	0.80	0.88
accuracy			0.88
macro avg	0.84	0.85	0.84
weighted avg	0.80	0.88	0.88

is greater than that of others, as expected, given that dusty stellar objects exhibit excess in this wavelength range.

6. METALLICITY IMPACT ON THE CLASSIFICATION OF DUSTY STELLAR SOURCES IN THE MAGELLANIC CLOUDS

The Large and Small Magellanic Clouds have different metallicity (Russell & Dopita 1992; Gordon et al. 2011), as mentioned in Section 1. To assess the impact of these metallicity variations on machine learning-based classification models, we conducted separate training and testing experiments for each galaxy. The PRF model, identified earlier as the best-performing classifier, was used to evaluate classification performance under these conditions. The dataset utilized in this work consists of a stellar sample from the LMC and SMC, as summarized in Table 1. In general, we consider them as a single dataset. However, in this section, we analyze them separately to assess the impact of metallicity.

Beginning with the SMC, the PRF successfully classified four out of five stellar classes. However, PAGB classification failed entirely, with most samples misclassified as YSO. Despite achieving an overall accuracy of 93%, as shown in Table 6 and Appendix F.3, the inability to classify PAGB sources correctly limits result

interpretability. The primary challenge in SMC classification was the limited number of PAGB samples. With only four PAGB sources available, manual test allocation was necessary, using three for training and one for testing. Additionally, the total sample size in SMC was limited to 132. The SMOTE algorithm, commonly used for data augmentation, was not applicable due to its requirement of at least six samples per class.

In contrast, LMC contained 486 dusty stellar sources, providing a more balanced dataset across classes. This reduced some classification difficulties and enabled more stable training compared to SMC. For the LMC, the PRF achieved an accuracy of 88%, as presented in Table 7 and Appendix F.3. This result aligns closely with those obtained from the combined LMC and SMC data, indicating that metallicity does not affect the classification. This outcome was anticipated, as the combined sample is predominantly composed of LMC data. However, challenges remained for the classification of less populated classes. While 67% of PAGB stars were correctly classified, some were misclassified as YSO.

As the results indicated, PAGB stars were misclassi-

fied due to their low population. To address this and to ensure a stable and fair comparison, we conducted additional assessments excluding PAGB stars. We trained the best models for these assessments using a dataset containing only the four remaining classes.
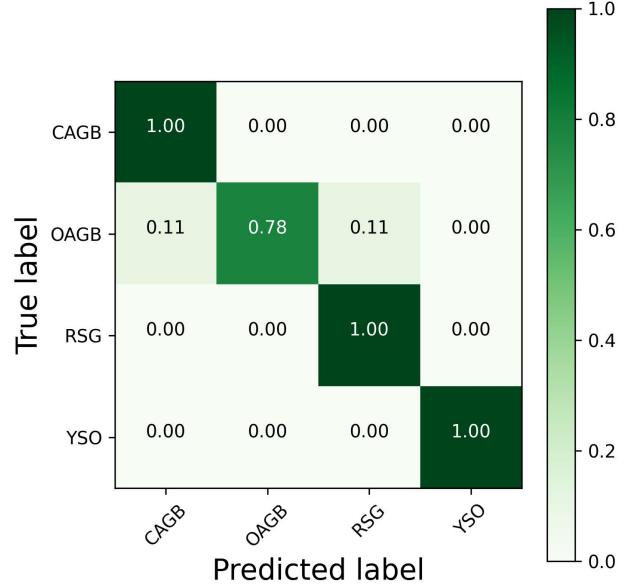
In the first approach, the master dataset was divided into two datasets for the SMC and LMC, and models were trained separately for each. The results, presented in Table F.1 and Table F.2, showed identical classification accuracies of 92% for the SMC and LMC. Next, we trained and tested the best model using the master dataset, combining data from four LMC and SMC classes. The results remained consistent at 92%. The close agreement between these results suggests that stellar objects in the four classes can be accurately classified in the MCs, whether the data are analyzed separately or combined. The corresponding confusion matrices are presented in Figure F.4. Additionally, the findings indicate that separating the LMC and SMC datasets based on metallicity differences does not significantly impact the classification of dusty stellar sources.

In the second approach, we again used the four-class dataset. We trained a model using the LMC training dataset and evaluated it with the SMC test data. The results, presented in Table F.4 and Fig 6 as a confusion matrix, were compared with those obtained from the model trained on the SMC dataset in the previous approach (Table F.1). In this case, the training datasets came from two locations with different metallicities, which should have led to differences in training and, consequently, in evaluation results. Yet, the results remained unchanged, suggesting that metallicity has no significant impact on model training for the LMC and SMC.

Finally, we assessed the impact of metallicity on classification and found it insignificant when using the four-class dataset. However, our master dataset includes five classes, one of which—PAGB—has a low population. To address this, we combined the two datasets, as discussed in Section 3, to enhance the sample size. As an additional verification step, we examined the absolute CMDs of the SMC and LMC to investigate potential differences in stellar populations. The strong agreement observed between their stellar populations suggests that combining data from these two sources is justified.

7. COMPARISON WITH PHOTOMETRIC CLASSIFICATION

Dusty stellar sources can be classified using photometric techniques (Dorn-Wallenstein et al. 2021; Maravelias et al. 2022) that distinguish stellar properties from circumstellar dust. For the photometric classifica-

Figure 6. The confusion matrix obtained from the PRF model trained on the LMC dataset and tested on the SMC dataset, including four populated stellar classes.

tion of such sources, observations are conducted within IR wavelengths because of the presence of infrared excess, a characteristic of dusty stellar environments due to dust grains absorbing ultraviolet and visible light and re-emitting it in the infrared. In this way, it is possible to detect dust disks and envelopes surrounding stars by comparing the observed IR luminosity to what is expected from the star. Therefore, this technique could be useful for identifying young stellar objects (YSOs) and evolved stars.

In addition to employing spectral features and emissions from dusty stellar sources, classifying such objects utilizes techniques such as the CMD, narrow-band filters, and blackbody deviations. These methods detect infrared excess evidence of circumstellar dust and distinguish between intrinsic stellar properties and dust influence. The following is a brief explanation of these techniques. CMD is one of the most popular methods (Blum et al. 2006), which plots stars by brightness and color to display their evolutionary stages, which vary among different types of stars. As shown in Fig. 2 and Fig. 3, it is clear that each class gathers in a distinct region of the CMD; however, there is some overlap.

Narrow-band filters have become useful for classifying dusty stellar objects. By isolating specific wavelengths of light, these filters provide information about the properties of such objects. Mainzer et al. (2004) explores this method, highlighting its effectiveness in discerning

crucial details about the composition, temperature, and distribution of dust surrounding these objects.

Dusty stellar sources typically show deviations from blackbodies at IR wavelengths that indicate dust emission; the stellar properties of the source can be determined by fitting observed SEDs to models (McDonald et al. 2012; Boyer et al. 2015; McDonald et al. 2017, 2024); as a result dusty sources can be classified.

In summary, photometric methods provide a more straightforward and cost-effective approach compared to spectroscopic methods for classifying stars and galaxies. However, dusty star catalogs and stellar labels are typically photometric, which may not be as reliable as spectroscopic labels. Nonetheless, photometric labeling can provide valuable information for identifying targets, which can be verified using spectral labeling. Additionally, spectroscopic data offer accurate details about the structure and composition of the stars (Riebel et al. 2010; Jones et al. 2017a).

In this section, we collected the LMC and SMC photometric catalogs discussed in the context of dusty stellar categories from various works of literature, which are listed in Table 8 and Table 9. The common characteristic among these catalogs is the data structure, similar to what we used for training models, as explained in Section 2. However, while the labels in our work are identified by a spectroscopic method, the labels in these catalogs are obtained from the photometric methods explained earlier.

Subsequently, we relabeled these data using models trained with spectroscopic labels and compared the new labels with the original photometric labels based on a redefined confusion matrix.

In the following, we briefly review the photometric classifications presented in each catalog.

- Whitney et al. (2008)
Whitney et al. (2008) identified ∼1,000 YSO candidates using *Spitzer* photometry (SAGE Point Source Catalog) in the LMC, focusing on young high- or intermediate-mass objects. The study built a color–magnitude grid based on radiative transfer models to select YSO candidates in the LMC using filters including *J*, *H*, *Ks*, [3.6], [4.5], [5.8], [8.0], and [24]. Therefore, the selection criteria were refined to focus on regions of CMD space that are less confused with other IR-bright populations. In addition, the selected list was cross-correlated with other stellar population catalogs and compared with them, resulting in a final list of 1,197 YSO candidates, among which 207 were identified as non-YSOs. The study also found that these YSOs strongly correlate with 24 μm emission.

Physical parameters were derived for 299 YSOs, and their SEDs fit well with radiative transfer models.

- Srinivasan et al. (2009)
Srinivasan et al. (2009) used the SAGE survey to classify evolved stars in the LMC, which is combined with the optical point source catalogs from 2MASS and the Magellanic Cloud Photometric Survey (MCPS; *U*, *V*, *B*, *I*) to construct of spectral SEDs. In total, 16,000 O-rich, 6300 C-rich, and 1,000 extreme sources with 8 μm excesses were identified, whereas with 2MASS and IRAC, 4500 O-rich, 5300 C-rich, and 960 extreme sources with 24 μm excesses were identified. Their results indicate a distinct increase in infrared excess with luminosity, more notably at 8 μm, for both oxygen-rich and carbon-rich AGB populations, a sign of circumstellar dust influence. There was also a correlation between greater optical depth and greater infrared excess in extreme AGB candidates. This study suggested the contribution of AGB stars to the LMC's mass loss and dust production, with extreme AGB stars being significant dust contributors. The study quantifies these observations with empirical relations, highlighting the role of AGB stars in the interstellar dust and gas lifecycle.

- Gruendl & Chu. (2009)
Gruendl & Chu (2009) conducted a study to discover young stellar objects (YSOs) with high and intermediate masses in the LMC using Spitzer Space Telescope data. They employed IRAC and MIPS observations to assemble a photometric catalog. This effort led to the identification of 1172 probable YSOs. Their approach included examining mid-infrared photometric data, source shapes, and their surrounding interstellar environment. By comparing their findings with those of the SAGE survey reported (Whitney et al. 2008), they noted differences, such as SAGE's omission of YSOs in more complex regions. Their results underline that both catalogs have strengths and weaknesses, but together, they offer a thorough listing of YSOs in the LMC, which is crucial for studying star formation.

- Riebel et al. (2010)
Riebel et al. (2010) focused on ∼ 30,000 AGB candidates comprising oxygen-rich, carbon-rich, and extreme AGB that were identified using photometry. An infrared photometry archive available through the SAGE project and a survey of LMC variability by Massive Compact Halo Objects (MACHO) were combined to create a dataset of variable red sources. The study found that, whereas oxygen-rich and red giant branch stars display a wavelength-

Table 8. The photometric distribution of dusty stars from literature reviews of the Magellanic Clouds is presented herein and in Table 9 to make the comparison with our trained classification model.

Name	Location	CAGB	OAGB	PAGB	RSG	YSO	Total	Label Type
Whitney et al. (2008)	LMC					360	360	Photometry
Srinivasan et al. (2009)	LMC	6609	34584				41193	Photometry
Gruendl & Chu (2009)	LMC					1090	1090	Photometry
Riebel et al. (2010)	LMC	116	909				1025	Photometry
Boyer et al. (2011)	SMC	1718	2457		3271		1025	Photometry
Kamath et al. (2014)	SMC	43		14		38	95	Photometry
Kamath et al. (2015)	LMC	51		31		154	236	Photometry
Yang et al. (2018)	LMC				126		126	Photometry
Yang et al. (2019)	SMC				88		88	Photometry
Yang et al. (2020)	SMC				180		180	Photometry
Yang et al. (2021)	LMC				2467		2467	Photometry

independent PL slope, extreme AGB stars display a wavelength-dependent PL slope by exploring the period-luminosity (PL) relationship across multiple infrared wavelengths (J, H, Ks and mid-IR bands). The Ks band emerged as the optimal wavelength for PL relationship characterization due to its consistent results for oxygen-rich and carbon-rich AGB stars.

- Boyer et al. (2011)
  Boyer et al. (2011) examined the SMC's infrared properties of cool evolved stars, focusing on RGBs, RSGs, and AGBs. Utilizing observations from the SAGE-SMC, the survey provides IR coverage [3.6]-[160] of the SMC regions. By combining near-IR and mid-IR photometry, they identified evolved stars and discovered a feature in the mid-IR CMD likely associated with particularly dusty oxygen-rich AGB stars. RSG and AGB stars contribute about 20% of the total SMC flux at [3.6], highlighting their importance to the integrated flux of distant metal-poor galaxies. Comparisons with the high-metallicity Large Magellanic Cloud (SAGE-LMC) show that SMC's RSG stars produce less dust, as indicated by their [3.6] - [8] color. A higher fraction of carbon-rich stars in the SMC suggests efficient C-rich dust formation. Initial estimates indicate that extreme C-rich AGB stars dominate dust production in both galaxies, whereas oxygen-rich stars may play a more significant role in the LMC.

- Kamath et al. (2014)
  Kamath et al. (2014) focused on optically visible candidate sources selection, including PAGB/RGB stars and YSOs in the SMC, through mid-IR observations from the SAGE survey and then assessed the 801 candidates using the low-resolution optical spectra taken through the AAOmega double-beam multifibre spectrograph mounted on the 3.9 m Anglo Australian Telescope (AAT). The final sample comprised 63 post-AGB/RGB candidates of A and F spectral class, of which 42 of these 63 sources were classified by their luminosity as post-red giant branch (post-RGB) candidates, and the remaining 21 were post-AGB candidates. This study also resulted in a new sample of 40 YSOs of A–F spectral type.

- Kamath et al. (2015)
  Kamath et al. (2015) classified post-AGBs, post-RGBs, and YSO source candidates in the LMC using photometric and spectroscopic analyses based on mid-IR excess, using the same filters used in (, Kamath2014MNRAS) then obtained their optical spectra to confirm the existence of dusty post-RGB stars. This study identified 162 YSO candidates and 35 post-AGB candidates, including 69 hot objects with UV continuum that might be post-AGB or luminous YSO candidates. The classification contributed to understanding stellar objects in the LMC and their evolution.

- Yang et al. (2018)
  Yang et al. (2018) examined RSG stars' infrared characteristics and variability in the LMC, employing data from the ALLWISE and NEOWISE-R projects and a literature review. Based on multiwavelength data analysis, a sample of 773 RSG candidates was refined to 744, revealing a correlation between mid-infrared variability, mass loss rate, and warm dust. This paper provides insights into mass loss mechanisms, the importance of variability and luminosity at mass loss, and the evolutionary stages of RSGs in the LMC. In this study, variable and extinction factors were considered when comparing the identified RSG sample with the theoretical evolutionary models, and reduced differences were observed between observations and models.

- Yang et al. (2019)

  Yang et al. (2019) compiled the magnitude-limited (IRAC1 or WISE1 ≤ 15.0 mag) multiwavelength catalog for the SMC, counting 45,466 massive stars with low metallicity. The catalog provides broad spectral coverage from ultraviolet to far infrared, which combines data from SEIP, VMC, IRSF, AKARI, HERITAGE, and others. The study identified RSG populations using the evolutionary tracks, synthetic photometry from MESA Isochrones and Stellar Tracks, and theoretical $J - K_s$ color cuts. As a result, candidates were ranked according to the intersection of five CMDs. Therefore, comparing the models with observations demonstrated that RSGs were separated from AGBs and that their initial mass limit was 6-7 $M_\odot$. Out of the total number of stars in this catalog, 1405 were identified as RSGs, and the rest as massive supergiants.

- Yang et al. (2020)

  Yang et al. (2020) centered on the RSGs in the SMC, identifying 1239 potential RSGs through a selection process based on a combination of spectroscopic data from the literature and 2MASS CMDs. The study examined infrared color-magnitude diagrams to identify around 1800 potential RSGs by distinguishing them from AGB stars. RSGs were classified based on their variability and brightness, with higher variability associated with greater mass loss. The study also estimated the total output of gas and dust produced by the RSG population and found that the temperature of the RSG population was directly proportional to the color of the $J - K_s$, corrected for reddening.

- Yang et al. (2021)

  Yang et al. (2021) created the magnitude–limited (IRAC1 or WISE1 ≤ 15.0 mag) multiwavelength source catalog for the LMC derived from the *Spitzer* Enhanced Imaging Products (SEIP) and Gaia Data Release 2. This catalog contains ∼ 19700 sources in 52 different bands, including 21 optical and 29 infrared bands. The study used the catalog to identify and classify ∼2974 red supergiants in the LMC using modified magnitude and color cuts (which represent equivalent evolutionary phases (EEPs) from core helium burning to carbon burning with 7–40 $M_\odot$) Yang et al. (2019) in six CMDs. This work significantly contributes to understanding massive star evolution in nearby galaxies.

After data collection, we did a cross-match with the spectroscopically labeled data regarding removing repetitive stars. Following that, we applied all steps of the

Table 9. According to the selected list (see Table 8), dusty stars within the Magellanic Clouds are categorized by their photometric labeled class after cross-matching all catalogs, thereby excluding those stars having repetition across catalogs.

Stellar Type	Number
CAGB	8,537
OAGB	37,950
PAGB	45
RSG	6,132
YSO	1,642
Total	54,306

preprocessing method mentioned in Section 3, providing them as the input of our models.

In this section, the best model is applied to the data to determine the labels. According to Section 4, considering the uncertainty and different results obtained in the PRF model with various tuned parameters, we concluded that this model has the best results. By presenting this new label as an output of machine learning models, we can compare it with photometric labels and determine how well they match.

To better understand the two kinds of labels, photometric and spectroscopic, we use the confusion matrix to compare the types of dusty stellar objects classified by the photometric method and models trained by spectral types. Therefore, we consider photometric labels as actual labels in the confusion matrix, named in the following comparison matrix. In this comparison, we are not concerned about the correctness or incorrectness of the labels but about whether the labels match the photometric data.
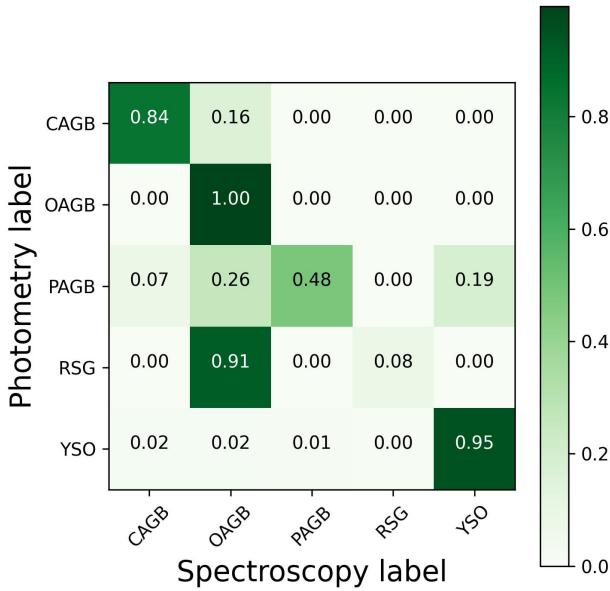
This section's results are provided in the appendix through comparison matrices (Fig. F.5). These matrices show that the photometric data do not match some stellar classes across all four models.

In machine learning, the best-performing models are introduced as the final model to label new input data based on classification metrics. After tuning the parameters in four modes, we found that PRF was our highest-performing model. There is a slight difference between the four PRF models, and most stars have been assigned the same label. Consequently, we constructed the consensus model, which considers the common results and excludes the rest from the classification, leading to more reliable predictions. For example, when all models consistently classify a star as OAGB, we present OAGB as its label. In contrast, if one predictive label differs from the others, it is excluded.

After applying this procedure to the stars, the classifi-

Figure 7. This figure illustrates the distribution of dusty stellar sources throughout the CMDs in the near and mid-infrared bands. As can be seen, there is an overlap between stellar classes if they are located in the same place; therefore, their classification cannot be easily achieved by using just two near-coupled filters, such as near-infrared (J, Ks) and mid-infrared ([3.6], [4.5]). In this regard, more filters or different methods may be required for better classification.

Figure 8. Comparison matrix of common labels derived from four selected classifiers.

- As much as 16% of CAGBs labeled by models based on the photometric method were oxygen-rich.

- 100% of OAGB stars labeled based on photometric data have been identified with spectroscopic models.

- 48% of PAGBs are correctly identified, whereas 19% and 26% are classified as YSOs and OAGBs, respectively. These stars cannot be accurately classified because of their small populations and overlap with other classes, as mentioned in Section 5.

- Only 8% of the stellar class of RSGs are confirmed by the spectroscopic model, and 91% are predicted as OAGBs.

- Around 95% of YSOs have been confirmed by models trained using spectroscopic labels, with the rest being subsets of others.

cation resulted in 23,601 objects, categorized as follows: 4,689 CAGBs, 17,124 OAGBs, 38 PAGBs, 409 RSGs, and 1,341 YSOs. We can now redraw and compare the comparison matrix with the photometric data, as illustrated in Fig. 8, which can be interpreted as follows.

The results indicate differences between labels predicted by models trained with the spectroscopic label and the photometric method. Thus, having a sufficient population compared to others is crucial for ensuring reliable, trained machine learning models.

The comprehensive dataset has been labeled by the best-trained models and released as a catalog in the supplementary material of this paper. The columns and their

Table 10. Description of the dataset containing columns from the comprehensive catalogs, including a total of 54,306 stars, as presented in Table 8 and Table 9. This dataset includes positional information based on Right Ascension (RA) and Declination (Dec), host galaxy classification (LMC and SMC), and multi-band photometric magnitudes. Additionally, it provides photometric predictions derived from other works mentioned in Table 8 and predictions from four of the best-trained models developed using the Simple and SMOTE approaches. The comparison matrices of these models are presented in Fig. F.5. The comprehensive catalog of this table is available electronically.

Column No.	Descriptor
1	Right Ascension (J2000)
2	Declination (J2000)
3	Host_Galaxy (LMC and SMC)
4	U–band magnitude
5	B–band magnitude
6	V–band magnitude
7	I–band magnitude
8	J–band magnitude
9	H–band magnitude
10	K_s–band magnitude
11	[3.6] band magnitude
12	[4.5] band magnitude
13	[5.8] band magnitude
14	[8.0] band magnitude
15	[24] band magnitude
16	Photometric prediction (Type)
17	Model #1 prediction
18	Model #2 prediction
19	Model #3 prediction
20	Model #4 prediction

descriptions are presented in Table 10. Specifically, this catalog includes 54,306 stars from the works presented in Table 8, along with their photometric magnitudes used for dusty stellar classification, as described in Section 3 and Table 2. These stars are classified using the four best models, with their labels provided as columns in the catalog and used in the preceding results, as shown in Fig. F.5.

8. CONCLUSION

We trained and tested a model for classifying dusty stellar objects using spectrally labeled data derived from the SAGE survey's multiwavelength filters, including infrared bands.

We identified the most accurate classifier, the PRF model, with accuracies exceeding 89%. We found that using the augmentation method to balance the dataset does not improve the results in this study, and it shows different behavior when dealing with various classes.

In Section 6, we found that metallicity differences did not impact the classification process when datasets from the SMC and LMC were combined. This was evaluated through three approaches, two involving classification using a dataset with four well-populated classes.

We collected photometrically labeled data with the same features as spectroscopically labeled data (as explained in Section 7). After feeding them into the models, they were classified, and a comprehensive catalog was presented as supplementary material for this paper.

In the future, multiwavelength data such as that used in training can be the input of models to determine each object's label. In addition, more multiwavelength and spectroscopic observations, for instance, the JWST, Gaia, WEAVE, 4MOST, MOONS, and ELT MOSAIC catalogs in terms of object classification, are needed to address having more accurate dusty stellar classifiers, significantly in less populated classes, including PAGBs and RSGs.

REFERENCES

Abdollahi, M., Torabi, N., Raeisi, S., & Rahvar, S. 2023, arXiv e-prints, arXiv:2301.08497, doi: 10.48550/arXiv.2301.08497

Adams, J. J., Simon, J. D., Bolatto, A. D., et al. 2013, ApJ, 771, 112, doi: 10.1088/0004-637X/771/2/112

Altman, N. S. 1992, The American Statistician, 46, 175, doi: 10.1080/00031305.1992.10475879

Bader-El-Den, M. B., Teitei, E., & Adda, M. 2016, 2016 International Joint Conference on Neural Networks (IJCNN), 3584. https://api.semanticscholar.org/CorpusID:2457143

Ball, N. M., & Brunner, R. J. 2010, International Journal of Modern Physics D, 19, 1049, doi: 10.1142/S0218271810017160

Baron, D. 2019, arXiv e-prints, arXiv:1904.07248, doi: 10.48550/arXiv.1904.07248

Baron, D., & Poznanski, D. 2017, MNRAS, 465, 4530, doi: 10.1093/mnras/stw3021

Bhardwaj, A., Kanbur, S. M., Macri, L. M., et al. 2016, AJ, 151, 88, doi: 10.3847/0004-6256/151/4/88

Blum, R. D., Mould, J. R., Olsen, K. A., et al. 2006, AJ, 132, 2034, doi: 10.1086/508227

Boyer, M. L., McDonald, I., Srinivasan, S., et al. 2015, ApJ, 810, 116, doi: 10.1088/0004-637X/810/2/116

Boyer, M. L., Srinivasan, S., van Loon, J. T., et al. 2011, AJ, 142, 103, doi: 10.1088/0004-6256/142/4/103

Breiman, L. 2001, Machine Learning, 45, 5, doi: 10.1023/A:1010933404324

Brice, M. J., & Andonie, R. 2019, AJ, 158, 188, doi: 10.3847/1538-3881/ab40d0

Carliles, S., Budavári, T., Heinis, S., Priebe, C., & Szalay, A. S. 2010, ApJ, 712, 511, doi: 10.1088/0004-637X/712/1/511

Carroll, B. W., & Ostlie, D. A. 2017, An Introduction to Modern Astrophysics, 2nd edn. (Cambridge University Press)

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. 2011, arXiv e-prints, arXiv:1106.1813, doi: 10.48550/arXiv.1106.1813

Cody, S., Scher, S., McDonald, I., et al. 2024, Open Research Europe, 4, doi: 10.12688/openreseurope.17023.1

Cornu, D., & Montillaud, J. 2021, A&A, 647, A116, doi: 10.1051/0004-6361/202038516

Djorgovski, S. G., Mahabal, A. A., Graham, M. J., Polsterer, K., & Krone-Martins, A. 2022, arXiv e-prints, arXiv:2212.01493, doi: 10.48550/arXiv.2212.01493

Dorn-Wallenstein, T. Z., Davenport, J. R. A., Huppenkothen, D., & Levesque, E. M. 2021, ApJ, 913, 32, doi: 10.3847/1538-4357/abf1f2

Fazio, G. G., Hora, J. L., Allen, L. E., et al. 2004, ApJS, 154, 10, doi: 10.1086/422843

Fotopoulou, S. 2024, Astronomy and Computing, 48, 100851, doi: 10.1016/j.ascom.2024.100851

Ghaziasgar, S., Abdollahi, M., Javadi, A., et al. 2024, Communications of the Byurakan Astrophysical Observatory, 71, 377, doi: 10.52526/25792776-24.71.2-377

Ghaziasgar, S., Masoudnezhad, A., Javadi, A., et al. 2022, arXiv e-prints, arXiv:2211.03403, doi: 10.48550/arXiv.2211.03403

Goldman, S. R., van Loon, J. T., Zijlstra, A. A., et al. 2017, MNRAS, 465, 403, doi: 10.1093/mnras/stw2708

Gordon, K. D., Meixner, M., Meade, M. R., et al. 2011, AJ, 142, 102, doi: 10.1088/0004-6256/142/4/102

Grandini, M., Bagli, E., & Visani, G. 2020, ArXiv, abs/2008.05756. https://api.semanticscholar.org/CorpusID:221112671

Gruendl, R. A., & Chu, Y.-H. 2009, ApJS, 184, 172, doi: 10.1088/0067-0049/184/1/172

Hastie, T., Tibshirani, R., & Friedman, J. 2009, The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Springer)

Herwig, F. 2005, ARA&A, 43, 435, doi: 10.1146/annurev.astro.43.072103.150600

Höfner, S., & Olofsson, H. 2018, A&A Rv, 26, 1, doi: 10.1007/s00159-017-0106-5

Hony, S., Kemper, F., Woods, P. M., et al. 2011, A&A, 531, A137, doi: 10.1051/0004-6361/201116845

Hosenie, Z., Lyon, R., Stappers, B., Mootoovaloo, A., & McBride, V. 2020, MNRAS, 493, 6050, doi: 10.1093/mnras/staa642

Houck, J. R., Roellig, T. L., van Cleve, J., et al. 2004, ApJS, 154, 18, doi: 10.1086/423134

Ivezić, Ž., Connolly, A. J., VanderPlas, J. T., & Gray, A. 2014, Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data (Princeton University Press), doi: 10.1515/9781400848911

Javadi, A., & van Loon, J. T. 2022, in The Origin of Outflows in Evolved Stars, ed. L. Decin, A. Zijlstra, & C. Gielen, Vol. 366, 210–215, doi: 10.1017/S1743921322001326

Javadi, A., van Loon, J. T., Khosroshahi, H., & Mirtorabi, M. T. 2013, MNRAS, 432, 2824, doi: 10.1093/mnras/stt640

Javadi, A., van Loon, J. T., & Mirtorabi, M. T. 2011a, MNRAS, 411, 263, doi: 10.1111/j.1365-2966.2010.17678.x

—. 2011b, MNRAS, 414, 3394, doi: 10.1111/j.1365-2966.2011.18638.x

Jones, O., Nally, C., Habel, N., et al. 2023, Nature Astronomy, 7, 1, doi: 10.1038/s41550-023-01945-7

Jones, O. C., Meixner, M., Justtanont, K., & Glasse, A. 2017a, ApJ, 841, 15, doi: 10.3847/1538-4357/aa6bf6

Jones, O. C., Woods, P. M., Kemper, F., et al. 2017b, MNRAS, 470, 3250, doi: 10.1093/mnras/stx1101

—. 2017c, VizieR Online Data Catalog, J/MNRAS/470/3250

Kamath, D. 2020, Journal of Astrophysics and Astronomy, 41, 42, doi: 10.1007/s12036-020-09665-4

Kamath, D., Wood, P. R., & Van Winckel, H. 2014, MNRAS, 439, 2211, doi: 10.1093/mnras/stt2033

—. 2015, MNRAS, 454, 1468, doi: 10.1093/mnras/stv1202

Karakas, A. I., & Lattanzio, J. C. 2014, PASA, 31, e030, doi: 10.1017/pasa.2014.21

Kemper, F., Woods, P. M., Antoniou, V., et al. 2010, PASP, 122, 683, doi: 10.1086/653438

Kinson, D. A., Oliveira, J. M., & van Loon, J. T. 2021, MNRAS, 507, 5106, doi: 10.1093/mnras/stab2386

—. 2022, MNRAS, 517, 140, doi: 10.1093/mnras/stac2692

Kokusho, T., Torii, H., Kaneda, H., Fukui, Y., & Tachihara, K. 2023, ApJ, 953, 104, doi: 10.3847/1538-4357/ace10e

Kuntzer, T., Tewes, M., & Courbin, F. 2016, A&A, 591, A54, doi: 10.1051/0004-6361/201628660

Last, F., Douzas, G., & Bacao, F. 2017, arXiv e-prints, arXiv:1711.00837, doi: 10.48550/arXiv.1711.00837

Levesque, E. M. 2010, in Astronomical Society of the Pacific Conference Series, Vol. 425, Hot and Cool: Bridging Gaps in Massive Star Evolution, ed. C. Leitherer, P. D. Bennett, P. W. Morris, & J. T. Van Loon, 103, doi: 10.48550/arXiv.0911.4720

Li, G., Lu, Z., Wang, J., & Wang, Z. 2025, arXiv e-prints, arXiv:2502.15300, doi: 10.48550/arXiv.2502.15300

Li, X., Wang, Y., Basu, S., Kumbier, K., & Yu, B. 2019, arXiv e-prints, arXiv:1906.10845, doi: 10.48550/arXiv.1906.10845

Liu, Z., Wei, P., Wei, Z., et al. 2021, arXiv e-prints, arXiv:2111.12791, doi: 10.48550/arXiv.2111.12791

Mainzer, A. K., McLean, I. S., Sievers, J. L., & Young, E. T. 2004, ApJ, 604, 832, doi: 10.1086/382020

Maravelias, G., Bonanos, A. Z., Tramper, F., et al. 2022, A&A, 666, A122, doi: 10.1051/0004-6361/202141397

Massey, P., & Olsen, K. A. G. 2003, AJ, 126, 2867, doi: 10.1086/379558

McDonald, I., & Zijlstra, A. A. 2016, ApJL, 823, L38, doi: 10.3847/2041-8205/823/2/L38

McDonald, I., Zijlstra, A. A., & Boyer, M. L. 2012, MNRAS, 427, 343, doi: 10.1111/j.1365-2966.2012.21873.x

McDonald, I., Zijlstra, A. A., Cox, N. L. J., et al. 2024, RAS Techniques and Instruments, 3, 89, doi: 10.1093/rasti/rzae005

McDonald, I., Zijlstra, A. A., & Watson, R. A. 2017, MNRAS, 471, 770, doi: 10.1093/mnras/stx1433

Meixner, M., Gordon, K. D., Indebetouw, R., et al. 2006, AJ, 132, 2268, doi: 10.1086/508185

Miettinen, O. 2018, Ap&SS, 363, 197, doi: 10.1007/s10509-018-3418-7

Murthy, S. K. 1998, Data Mining and Knowledge Discovery, 2, 345. https://api.semanticscholar.org/CorpusID:207741345

Oliveira, J. M., van Loon, J. T., Sloan, G. C., et al. 2013, MNRAS, 428, 3001, doi: 10.1093/mnras/sts250

Pashchenko, I. N., Sokolovsky, K. V., & Gavras, P. 2018, MNRAS, 475, 2326, doi: 10.1093/mnras/stx3222

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825

Pennock, C. M., van Loon, J. T., Anih, J. O., et al. 2022, MNRAS, 515, 6046, doi: 10.1093/mnras/stac2096

Pennock, C. M., van Loon, J. T., Cioni, M.-R. L., et al. 2025, MNRAS, 537, 1028, doi: 10.1093/mnras/staf080

Pietrzyński, G., Graczyk, D., Gieren, W., et al. 2013, Nature, 495, 76, doi: 10.1038/nature11878

Pilbratt, G. L., Riedinger, J. R., Passvogel, T., et al. 2010, A&A, 518, L1, doi: 10.1051/0004-6361/201014759

Reis, I., Baron, D., & Shahaf, S. 2019, AJ, 157, 16, doi: 10.3847/1538-3881/aaf101

Riebel, D., Meixner, M., Fraser, O., et al. 2010, ApJ, 723, 1195, doi: 10.1088/0004-637X/723/2/1195

Riebel, D., Srinivasan, S., Sargent, B., & Meixner, M. 2012, ApJ, 753, 71, doi: 10.1088/0004-637X/753/1/71

Rieke, G. H., Young, E. T., Engelbracht, C. W., et al. 2004, ApJS, 154, 25, doi: 10.1086/422717

Rokach, L., & Maimon, O. 2005, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 35, 476. https://api.semanticscholar.org/CorpusID:14808716

Ruffle, P. M. E., Kemper, F., Jones, O. C., et al. 2015a, MNRAS, 451, 3504, doi: 10.1093/mnras/stv1106

Ruffle, P. M. E., Kemper, F., Jones, O. C., et al. 2015b, VizieR Online Data Catalog, J/MNRAS/451/3504

Russell, S. C., & Dopita, M. A. 1992, ApJ, 384, 508, doi: 10.1086/170893

Rybicki, G. B., & Lightman, A. P. 1986, Radiative Processes in Astrophysics (Wiley-VCH)

Schlafly, E. F., & Finkbeiner, D. P. 2011, ApJ, 737, 103, doi: 10.1088/0004-637X/737/2/103

Scowcroft, V., Freedman, W. L., Madore, B. F., et al. 2016, 816, 49, doi: 10.3847/0004-637X/816/2/49

Seale, J. P., Looney, L. W., Chu, Y.-H., et al. 2009, ApJ, 699, 150, doi: 10.1088/0004-637X/699/1/150

Sen, S., Agarwal, S., Chakraborty, P., & Singh, K. P. 2022, Experimental Astronomy, 53, 1, doi: 10.1007/s10686-021-09827-4

Sewilo, M., Carlson, L. R., Seale, J. P., et al. 2013, ApJ, 778, 15, doi: 10.1088/0004-637X/778/1/15

Sheets, H. A., Bolatto, A. D., van Loon, J. T., et al. 2013, ApJ, 771, 111, doi: 10.1088/0004-637X/771/2/111

Smith, M. J., & Geach, J. E. 2023, Royal Society Open Science, 10, 221454, doi: 10.1098/rsos.221454

Srinivasan, S., Meixner, M., Leitherer, C., et al. 2009, AJ, 137, 4810, doi: 10.1088/0004-6256/137/6/4810

Subramanian, S., & Subramaniam, A. 2009, A&A, 496, 399, doi: 10.1051/0004-6361/200811029

Subramanian, S., & Subramaniam, A. 2011, in Astronomical Society of India Conference Series, Vol. 3, Astronomical Society of India Conference Series, 144

Suh, K.-W. 2016, Journal of Astronomy and Space Sciences, 33, 119, doi: 10.5140/JASS.2016.33.2.119

—. 2020, ApJ, 891, 43, doi: 10.3847/1538-4357/ab6609

—. 2021, ApJS, 256, 43, doi: 10.3847/1538-4365/ac1274

Sun, Y., Wong, A. K. C., & Kamel, M. S. 2009, Int. J. Pattern Recognit. Artif. Intell., 23, 687. https://api.semanticscholar.org/CorpusID:27118324

van Winckel, H. 2003, ARA&A, 41, 391, doi: 10.1146/annurev.astro.41.071601.170018

Vapnik, V. N. 1995, The nature of statistical learning theory (Springer-Verlag New York, Inc.)

Vijh, U. P., Meixner, M., Babler, B., et al. 2009, AJ, 137, 3139, doi: 10.1088/0004-6256/137/2/3139

Wang, S., Dai, Y., Shen, J., & Xuan, J. 2021, Scientific Reports, 11, 24039, doi: 10.1038/s41598-021-03430-5

Werner, M. W., Roellig, T. L., Low, F. J., et al. 2004, ApJS, 154, 1, doi: 10.1086/422992

Whitelock, P. A., Feast, M. W., van Loon, J. T., & Zijlstra, A. A. 2003, MNRAS, 342, 86, doi: 10.1046/j.1365-8711.2003.06514.x

Whitney, B. A., Sewilo, M., Indebetouw, R., et al. 2008, AJ, 136, 18, doi: 10.1088/0004-6256/136/1/18

Wilson, A. J., Lakeland, B. S., Wilson, T. J., & Naylor, T. 2023, MNRAS, 521, 354, doi: 10.1093/mnras/stad301

Wood, P. R., Bessell, M. S., & Fox, M. W. 1983, ApJ, 272, 99, doi: 10.1086/161265

Wood, P. R., Whiteoak, J. B., Hughes, S. M. G., et al. 1992, ApJ, 397, 552, doi: 10.1086/171812

Woods, P. M., Oliveira, J. M., Kemper, F., et al. 2011, MNRAS, 411, 1597, doi: 10.1111/j.1365-2966.2010.17794.x

Yang, M., Bonanos, A. Z., Jiang, B.-W., et al. 2018, A&A, 616, A175, doi: 10.1051/0004-6361/201832833

—. 2019, A&A, 629, A91, doi: 10.1051/0004-6361/201935916

—. 2020, A&A, 639, A116, doi: 10.1051/0004-6361/201937168

Yang, M., Bonanos, A. Z., Jiang, B., et al. 2021, A&A, 646, A141, doi: 10.1051/0004-6361/202039475

Zeraatgari, F. Z., Hafezianzadeh, F., Zhang, Y., et al. 2024, MNRAS, 527, 4677, doi: 10.1093/mnras/stad3436

Zhang, J., Zhang, Y., Kang, Z., et al. 2023, PASA, 40, e037, doi: 10.1017/pasa.2023.35

APPENDIX

A. SUPPORTING INFORMATION FOR CLASSIFICATION MODELS

A.1. *Random Forest (RF)*

The Random Forest (RF) method combines multiple decision trees to classify data. In the algorithm, the prediction made by each tree is averaged or voted on by the majority (Breiman 2001; Carliles et al. 2010; Baron & Poznanski 2017; Reis et al. 2019).

A.2. *K-Nearest Neighbors (KNN)*

K-Nearest Neighbors (KNN) is a classification algorithm that assigns labels to data based on their nearest neighbors. The KNN algorithm calculates data points by computing their distance from all other data points. The data point can be predicted based on the average value of the K nearest neighbors or the most common classes (Altman 1992; Breiman 2001).

A.3. *C-Support Vector Classification (SVC)*

Support Vector Classification (SVC) is a method that utilizes a hyperplane to separate data into different classes and map data onto a high-dimensional feature space. Support vectors are part of the SVC model and are the closest data points to the hyperplane. Support Vector Machines (SVC) can be adjusted for some hyperparameters to achieve optimal performance. The polynomial kernel SVC (SVC-poly) is one of the most commonly used SVC models. In contrast, an SVC-rbf, according to the Radial Basis Function (RBF) kernel, is the other typically used SVC model (SVC-rbf) (Vapnik 1995; Baron 2019).

A.4. *Gaussian Naive Bayes (GNB)*

Gaussian Naive Bayes executes the Gaussian Naive Bayes algorithm for classification, assuming that the likelihood of the features follows a Gaussian distribution. It includes one key hyperparameter, `var_smoothing`, which is optimized using Grid Search.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \tag{A1}$$

B. PERFORMANCE METRICS

The performance metrics of classifiers can be used to predict all spectral classes. At the training stage, especially because the classes have unequal populations, accuracy can be misleading. In addition, the Classification report contains the model's precision, recall, and F1-score values for each class (Grandini et al. 2020).

Accuracy is a metric for evaluating the classification models, dividing the number of correct predictions by the total number of predictions.

$$\text{Accuracy} = \frac{\text{correct predictions}}{\text{all predictions}}, \tag{B2}$$

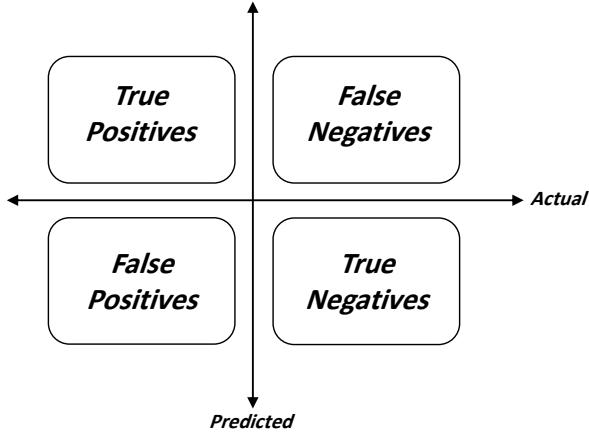Precision is the fraction of relevant instances among the retrieved ones.

$$\text{Precision} = \frac{\text{True Positive}}{\text{False Positive} + \text{True Positive}}, \tag{B3}$$

Recall is the fraction of retrieved relevant cases in the following formula.

$$\text{Recall} = \frac{\text{True Positive}}{\text{False Negative} + \text{True Positive}}. \tag{B4}$$

The F1-score is a combination of precision and recall. We used performance metrics to evaluate each classifier and defined them as follows (see Fig. C.1 for illustration).

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{B5}$$

Figure C.1. Illustration of the Confusion Matrix for Classification Outcomes. The diagram represents the possible outcomes in the classification task, distinguishing between actual and predicted classifications.

The macro average consists of precision, recall, and an F1-score average for all classes (The population of each class in this averaging is ineffective).

The weighted average is the average of precision, recall, and F1 scores based on the number of samples for each class. Precision and recall are weighted based on population weights computed separately for each subclass in the sample and averaged.

C. CONFUSION MATRIX

A confusion matrix is a visual representation of the performance of classification algorithms. This matrix displays the number of objects in each class based on the model's predictions. The diagonal elements represent each class's predicted and actual labels. As shown in Fig. C.1, the confusion matrix consists of true positives, false positives, true negatives, and false negatives. True Positive (TP): The number of samples predicted as positive, which is positive. False Positive (FP): The number of samples predicted as positive but negative. True Negative (TN): The number of samples predicted as negative and actual negative. False Negative (FN): The number of samples predicted to be negative but positive.
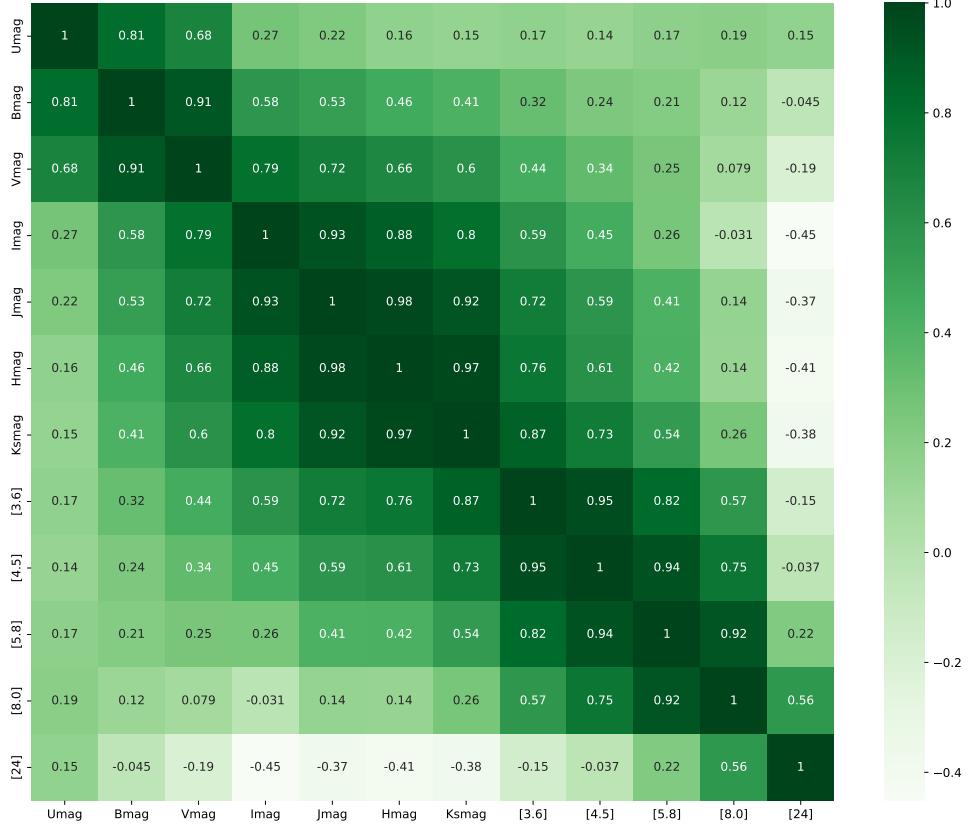
D. CORRELATION MATRIX

The correlation matrix shows the correlation between each feature and the other features. In this matrix, features with high correlation can be removed. A high correlation between two features indicates similar behavior, which can complicate the model. This matrix aims to reduce complexity as much as possible.
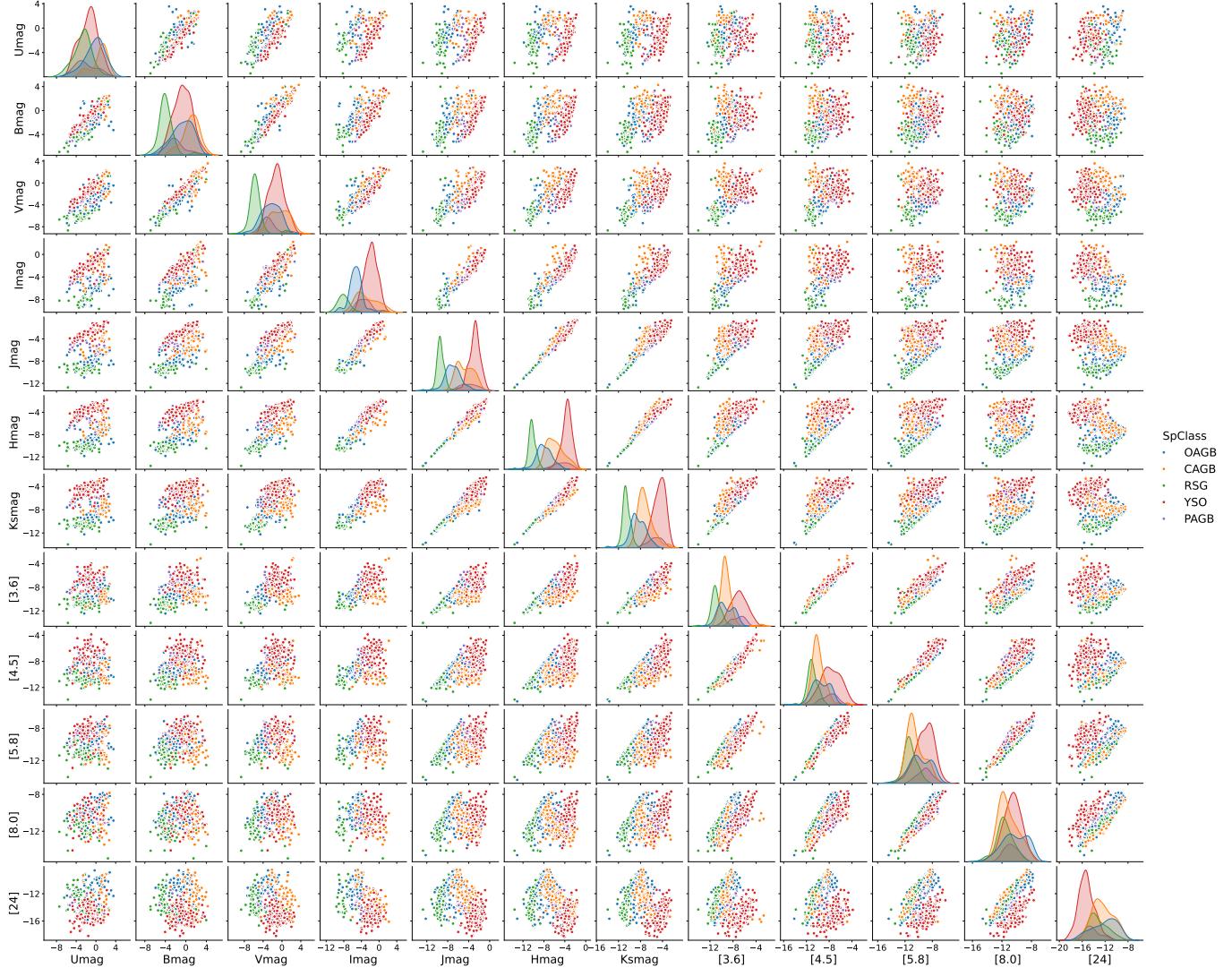
E. PAIRPLOT

The pairplot is a function that creates a grid of scatterplots and histograms for a given dataset derived from a data frame. This visualization allows for the examination of relationships between variables and data pairs. Additionally, it can identify correlations and outliers, which are needed for further analysis.
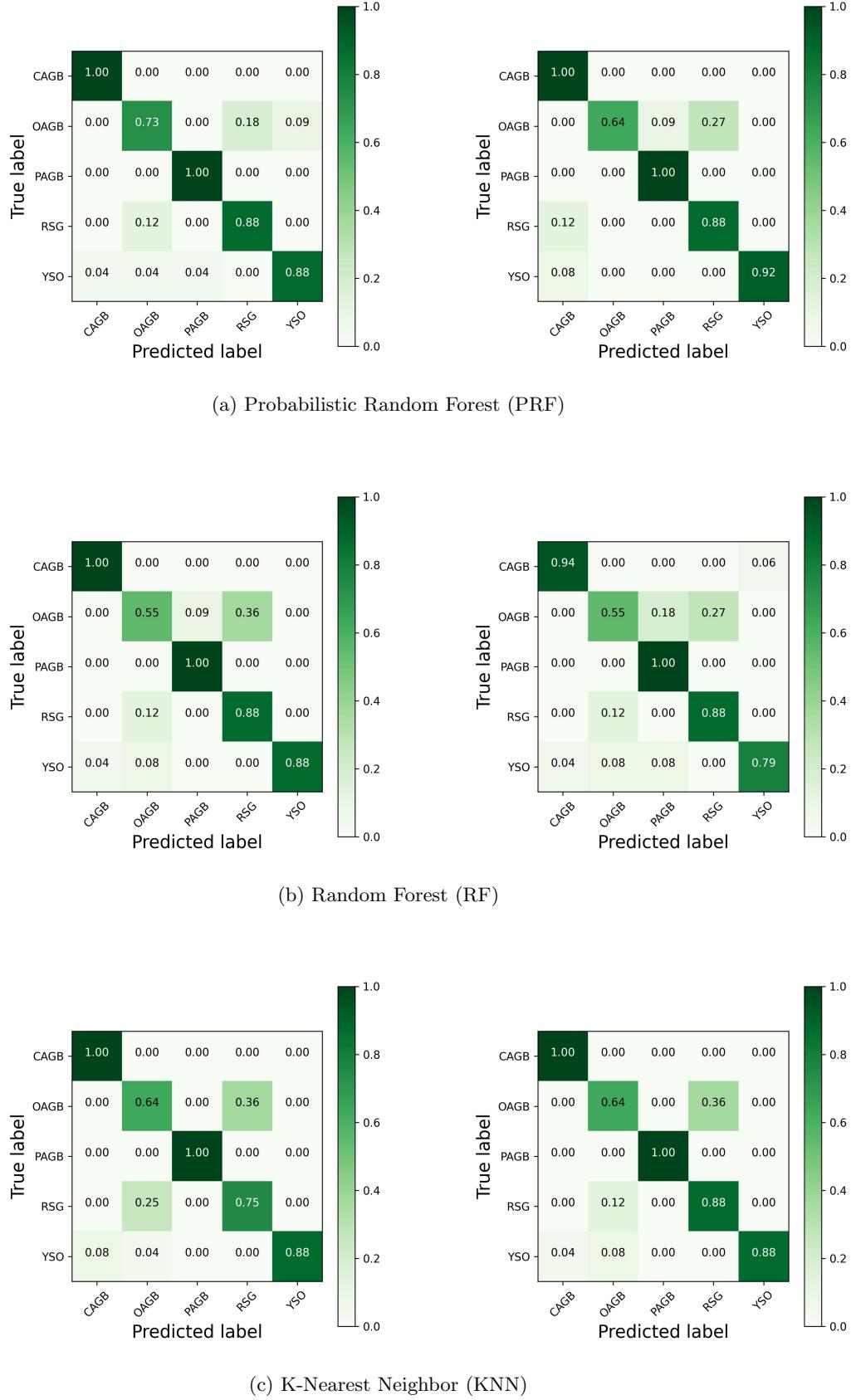
F. SUPPLEMENTAL MATERIALS

The results of confusion matrices for training different classifiers before and after augmenting the data using the SMOTE method are presented in Fig. F.1 and Fig. F.2. The classification reports under different settings, as discussed in Section 6, which examines the effect of metallicity, are presented in Table F.1, Table F.2, Table F.3, and Table F.4. The corresponding confusion matrices are shown in Fig.F.3 and Fig.F.4. In addition, the comparison matrices of selected models for comparison to photometrically labeled data are shown in Fig. F.5.
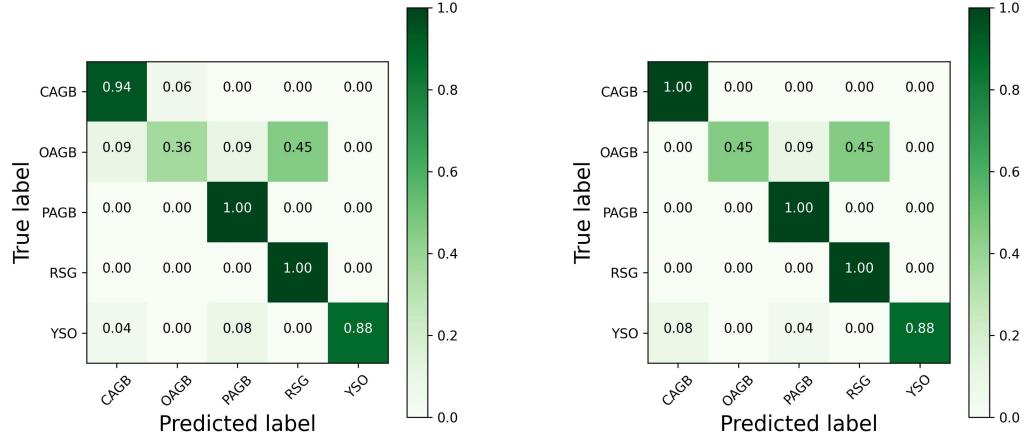
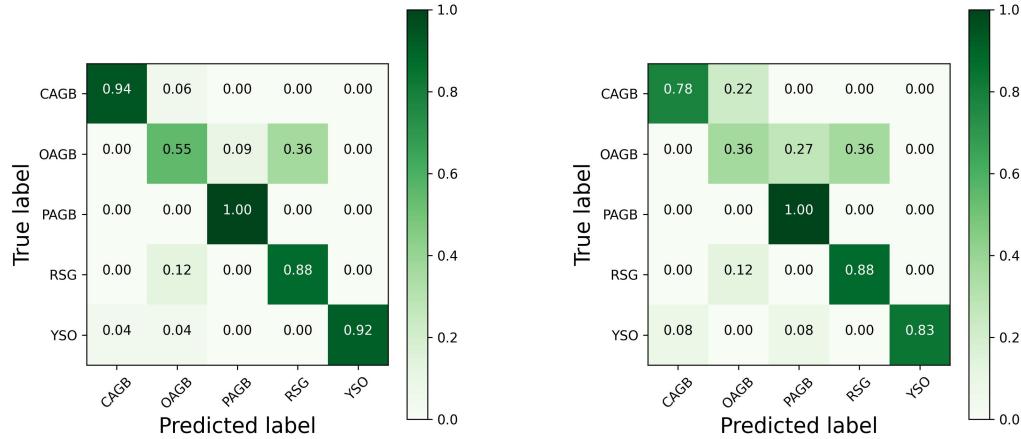**Figure D.1.** Feature correlation through different filters from the SAGE spectral catalogs.

Figure E.1. Feature correlation. This pairplot illustrates the pairwise relationships between each feature. The plot shows the correlation between 12 features for five dusty stellar classes that have been included.

(a) Probabilistic Random Forest (PRF)

(b) Random Forest (RF)
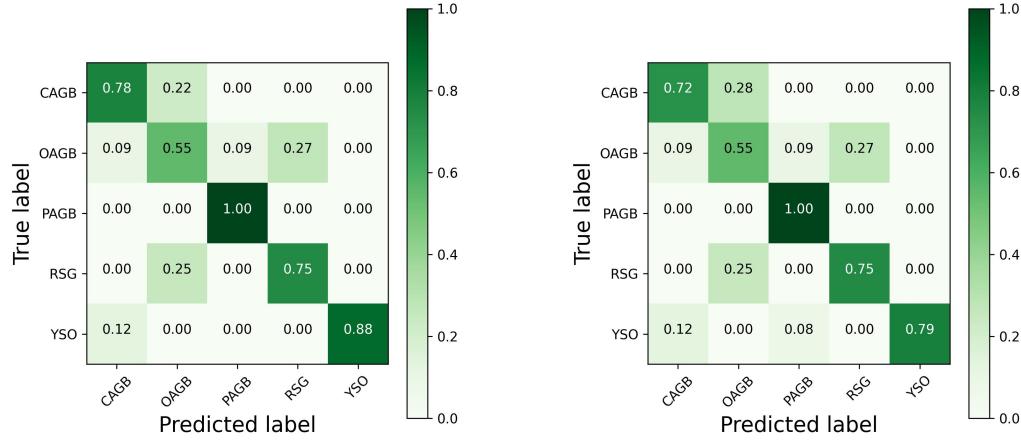

(c) K-Nearest Neighbor (KNN)

Figure F.1. Confusion matrix for different classifiers before (left panels) and after data augmentation (right panels).

(a) Support Vector Classification (SVC), with "poly" kernel.

(b) Support Vector Classification (SVC), with "rbf" kernel.

(c) Gaussian Naive Bayes (GNB).

Figure F.2. Confusion matrix for different classifiers before (left panels) and after data augmentation (right panels).

Table F.1. Classification report for the SMC catalog, including four classes of dusty stellar objects (see the left panel of Fig. F.4). The classification was performed using the following settings: number of estimators = 10, Keep probability = 0.8, and Simple PRF.

Class	Precision	Recall	F1-Score
CAGB	0.75	1.00	0.86
OAGB	0.86	0.78	0.88
RSG	0.67	1.00	0.80
YSO	1.00	1.00	1.00
accuracy			0.92
macro avg	0.85	0.94	0.88
weighted avg	0.95	0.92	0.92

Table F.2. Classification report for the LMC catalog, including four classes of dusty stellar objects (see the middle panel of Fig. F.4). The classification was performed using the following settings: number of estimators = 10, Keep probability = 0.8, and Simple PRF.

Class	Precision	Recall	F1-Score
CAGB	0.88	1.00	0.94
OAGB	0.90	0.83	0.86
RSG	0.93	0.93	0.93
YSO	0.97	0.93	0.95
accuracy			0.92
macro avg	0.92	0.92	0.92
weighted avg	0.92	0.92	0.92

Table F.3. Classification report for the master dataset (LMC & SMC), including four classes of dusty stellar objects (see the right panel of Fig. F.4). The classification was performed using the following settings: number of estimators = 10, Keep probability = 0.8, and Simple PRF
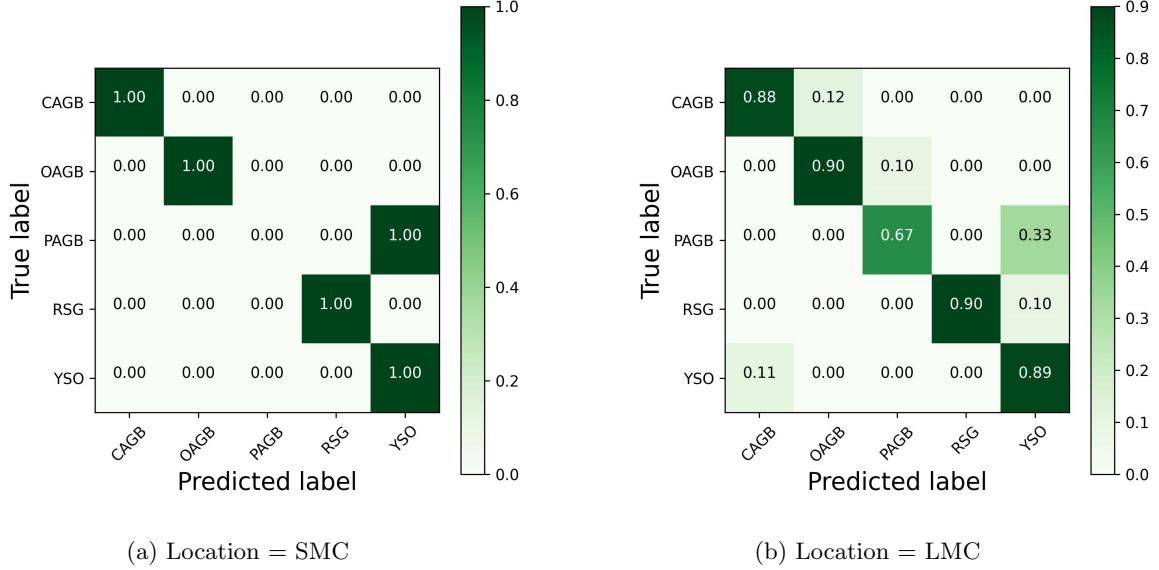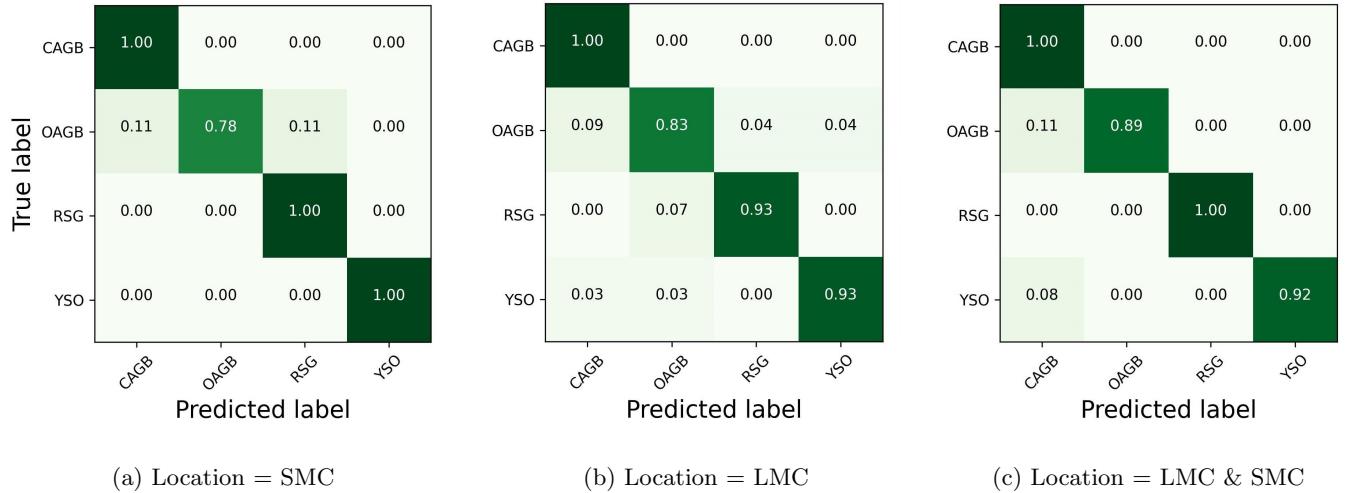
Class	Precision	Recall	F1-Score
CAGB	0.60	1.00	0.75
OAGB	1.00	0.89	0.94
RSG	1.00	1.00	1.00
YSO	1.00	0.92	0.96
accuracy			0.92
macro avg	0.90	0.95	0.91
weighted avg	0.95	0.92	0.93

Table F.4. Classification report for the case where the model was trained on LMC data and tested on SMC data, including four classes of dusty stellar objects (see Fig. 6). The classification was performed using the following settings: number of estimators = 10, Keep probability = 0.8, and Simple PRF.

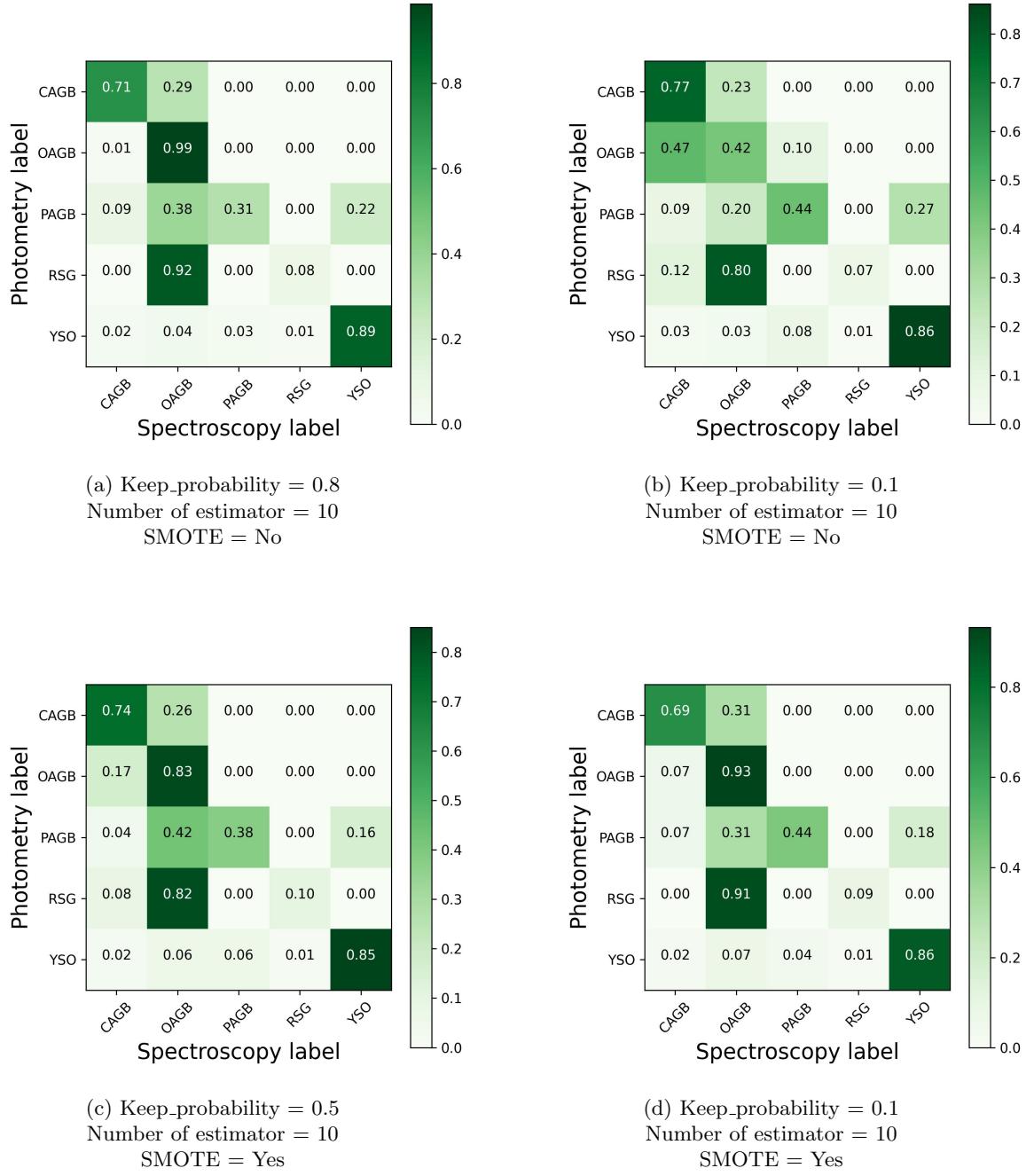| Class	Precision	Recall	F1-Score
CAGB	0.75	1.00	0.86
OAGB	0.86	0.78	0.88
RSG	0.67	1.00	0.80
YSO	1.00	1.00	1.00
accuracy			0.92
macro avg	0.85	0.94	0.88
weighted avg	0.95	0.92	0.92

(a) Location = SMC

(b) Location = LMC

Figure F.3. Confusion matrices for the best-performing classifier (PRF) applied separately to the SMC (left) and LMC (right) datasets, rather than using a combined dataset, to account for metallicity differences between the two galaxies. This classification was performed considering five stellar classes. The settings for this classification are: Number of Class = 5, Keep_Probability = 0.8, Number of Estimator = 10, and SMOTE = No.

(a) Location = SMC

(b) Location = LMC

(c) Location = LMC & SMC

Figure F.4. Confusion matrices for the best-performing classifier (PRF) applied separately to the SMC (left), LMC (middle), and the combined LMC & SMC dataset (right) to assess the impact of metallicity differences between the two galaxies. The model was trained with four stellar classes (see Table F.1, Table F.2, Table F.3 for more details. The settings for this classification are: Number of Class = 4, Keep_probability = 0.8, Number of estimator = 10, and SMOTE = No.

(a) Keep_probability = 0.8
Number of estimator = 10
SMOTE = No

(b) Keep_probability = 0.1
Number of estimator = 10
SMOTE = No

(c) Keep_probability = 0.5
Number of estimator = 10
SMOTE = Yes

(d) Keep_probability = 0.1
Number of estimator = 10
SMOTE = Yes

Figure F.5. Comparison matrices of selected models for comparison to photometrically labeled data.