

Metallicity Effects on Machine Learning Classification of Dusty Stellar Sources in the Magellanic Clouds

Sepideh Ghaziasgar ^{*1}, Mahdi Abdollahi ^{†1}, Atefeh Javadi ^{‡1}, Jacco Th. van Loon ^{§2}, Iain McDonald ^{¶3}, Joana Oliveira ^{||2}, and Habib G. Khosroshahi ^{**1,4}

¹School of Astronomy, Institute for Research in Fundamental Sciences (IPM), P.O. Box 19568-36613, Tehran, Iran

²Lennard-Jones Laboratories, Keele University, ST5 5BG, UK

³Jodrell Bank Centre for Astrophysics, Alan Turing Building, University of Manchester, M13 9PL, UK

⁴Iranian National Observatory, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

Abstract

Differences in metallicity between the Large Magellanic Cloud (LMC) and the Small Magellanic Cloud (SMC) offer an opportunity to examine whether environmental metallicity affects the performance of machine learning models in classifying dusty stellar sources. The five stellar classes studied include young stellar objects (YSOs), red supergiants (RSGs), post-asymptotic giant branch stars (PAGBs), and oxygen- and carbon-rich asymptotic giant branch stars (OAGBs and CAGBs), which are key phases of stellar evolution involved in dust production. Using spectroscopically labeled data from the Surveying the Agents of Galaxy Evolution (SAGE) project, we trained and evaluated a probabilistic random forest (PRF) classifier with four approaches: (1) separate training on LMC and SMC, including all five classes, (2) excluding the underpopulated PAGB class, (3) combined LMC and SMC datasets, and (4) cross-galaxy training and testing. The model achieved 93% accuracy on the SMC and 88% on the LMC across all five classes. In the SMC, PAGB sources were misclassified as YSOs, mainly because of their small sample size (4 objects). When PAGB was excluded, both the LMC and the SMC reached 92% accuracy. A combined dataset produced the same accuracy, and cross-galaxy training yielded similar results, indicating that metallicity does not significantly impact model performance. A comparison of absolute CMDs for the LMC and SMC confirms their similarity in stellar populations. These findings suggest that environmental metallicity has little effect on ML-based classification of dusty stellar sources, supporting the use of combined datasets and cross-galaxy models in low-metallicity environments.

Keywords: *stars: classification - stars: AGB, RSG, and post-AGB - stars: YSOs - galaxies: metallicity - galaxies: spectral catalog - galaxies: Local Group - methods: machine learning*

1. Introduction

The Large Magellanic Cloud (LMC) and Small Magellanic Cloud (SMC) are nearby dwarf galaxies at distances of approximately 50 and 60 kpc, with metallicities of about $0.5 Z_{\odot}$ and $0.2 Z_{\odot}$, respectively (Pietrzyński et al., 2013, Russell & Dopita, 1992, Subramanian & Subramaniam, 2009). Their proximity and contrasting chemical compositions make them excellent laboratories for investigating the relationship between stellar evolution, dust formation, and environmental metallicity (Ruffle et al., 2015).

Dusty stellar sources, including young stellar objects (YSOs) and evolved stars—oxygen- and carbon-rich asymptotic giant branch stars (OAGBs, CAGBs), red supergiants (RSGs), and post-asymptotic giant branch stars (PAGBs)—represent key phases in the stellar life cycle. These objects return heavy elements to the interstellar medium, shaping the dust content and chemical enrichment of galaxies (Boyer et al., 2011,

*sepideh.ghaziasgar@ipm.ir, Corresponding author

†m.abdollahi@ipm.ir

‡atefeh@ipm.ir

§j.t.van.loon@keele.ac.uk

¶Iain.Mcdonald-2@manchester.ac.uk

||j.oliveira@keele.ac.uk

**habib@ipm.ir

Höfner & Olofsson, 2018). Because of their brightness and strong infrared excesses, these populations are particularly well traced in the Magellanic Clouds.

Traditional classification of dusty stellar populations has relied primarily on photometric methods, such as color–magnitude diagrams (CMDs) and spectral energy distributions (SEDs). While such techniques are efficient for large samples, they can be ambiguous due to overlapping photometric signatures among different stellar classes. Spectroscopic observations, on the other hand, provide a more reliable basis for classification but remain limited in number. The Surveying the Agents of Galaxy Evolution (SAGE) and its spectroscopic follow-up SAGE-Spec surveys (Jones et al., 2017, Kemper et al., 2010, Meixner et al., 2006, Ruffle et al., 2015, Woods et al., 2011) have provided infrared photometry and spectroscopically confirmed dusty stellar sources in the Magellanic Clouds, enabling direct machine-learning studies of evolved stars and YSOs in both galaxies.

Machine learning (ML) methods have become essential tools for identifying and classifying stellar sources from multiwavelength data (Baron, 2019). Among these, the Probabilistic Random Forest (PRF) classifier (Ghaziasgar et al., 2024, Kinson et al., 2022, Pennock et al., 2025, Reis et al., 2019) has proven particularly effective for handling uncertainties and imbalanced datasets. Previous studies have shown that supervised ML models trained on spectroscopic labels can achieve accuracies of around 90% in categorizing dusty stellar sources (Ghaziasgar et al., 2024).

However, differences in metallicity between the LMC and SMC may influence the photometric and spectroscopic properties of dusty stellar sources. In this study, we investigate whether the metallicity contrast between the two galaxies affects the performance of machine-learning classification of these objects, using spectroscopically labeled datasets from the Magellanic Clouds.

2. Data and Model

The dataset is obtained from the SAGE and SAGE-Spec surveys (Jones et al., 2017, Kemper et al., 2010, Meixner et al., 2006, Ruffle et al., 2015, Woods et al., 2011), which provide multiwavelength photometry and spectroscopically classified dusty stellar sources in the Magellanic Clouds. The combined spectroscopic catalog includes about 618 sources classified into five stellar categories: young stellar objects (YSOs), oxygen- and carbon-rich asymptotic giant branch stars (OAGBs and CAGBs), red supergiants (RSGs), and post-asymptotic giant branch stars (PAGBs).

The same 12 near- and mid-infrared filters used in these studies (Ghaziasgar et al., 2024, 2025a,b,c) were adopted here, as they effectively trace both photospheric and circumstellar dust emission. These features are particularly sensitive to dust and molecular absorption bands, making them suitable for identifying dusty stellar classes.

The full sample was naturally divided into two subsets corresponding to sources located in the Large and Small Magellanic Clouds, allowing a direct comparison between environments with different metallicities. The LMC subset contains 486 sources, while the SMC contains 132, reflecting the intrinsic population imbalance between the two galaxies. Due to the small number of PAGB stars in the SMC (only four objects), the Synthetic Minority Oversampling Technique (SMOTE; Chawla et al. 2011) could not be applied, since it requires at least six samples per class to generate synthetic data points.

The PRF classifier, previously identified as the most accurate model in Ghaziasgar et al. (2024), was adopted to evaluate classification performance. To assess the potential impact of metallicity in a systematic manner, we designed four experimental configurations: (1) training the model separately on the LMC and SMC datasets, (2) repeating the analysis after removing the sparsely populated PAGB class, (3) training and testing on the combined LMC and SMC catalog, and (4) cross-galaxy training and testing. Model performance was evaluated using precision, recall, F1-score, and accuracy, with confusion matrices used to visualize class-level outcomes.

3. Results

To test whether metallicity influences classification performance, we first trained the PRF model separately on the LMC and SMC catalogs while keeping all five stellar classes. The SMC experiment achieved an overall accuracy of 93%, correctly identifying four classes; however, PAGB sources were not recovered and were assigned mainly to the YSO category due to their extremely small sample size (four objects). The

limited size of the SMC set (132 objects) also prevented the use of SMOTE. These results are presented in the classification reports (Tables 1, 2) and the corresponding confusion matrices (Fig. 1).

Table 1: Classification report for the SMC catalog (five classes).

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 2: Classification report for the LMC catalog (five classes).

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

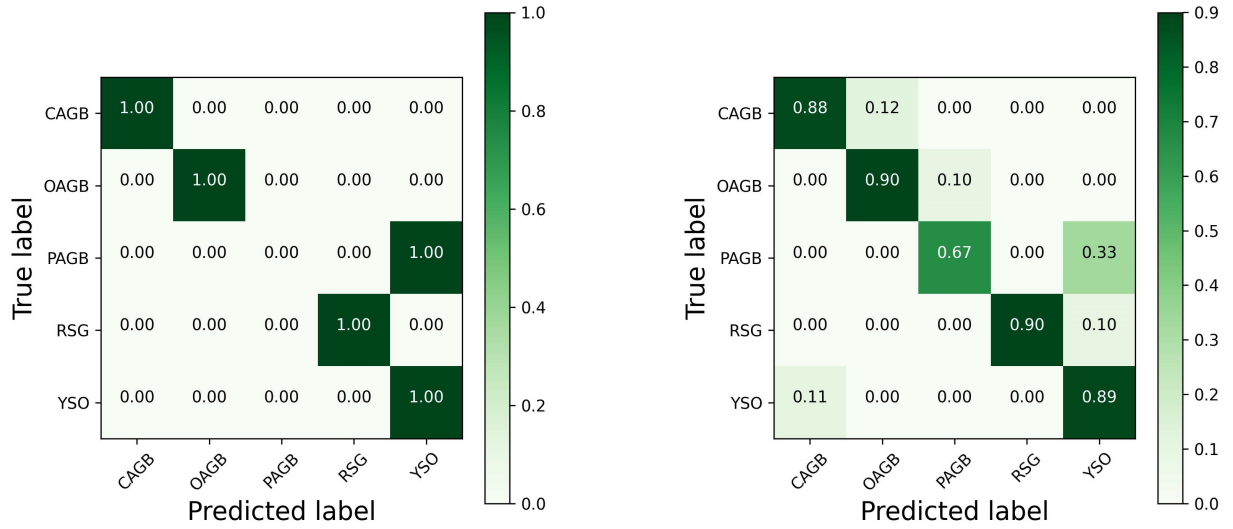


Figure 1: Confusion matrices for PRF classifier applied separately to the SMC (left) and LMC (right) using five stellar classes.

To remove the effect of the sparsely populated PAGB class, we repeated the analysis using only the four well-represented classes. In this configuration, both the LMC and SMC achieved an accuracy of 92%, confirming that PAGB imbalance was the primary limiting factor. The reports and confusion matrices are shown in Tables 3, 4 and Fig. 2. The similar diagonal patterns demonstrate that the main stellar groups are consistently learned, indicating that metallicity alone does not introduce measurable differences.

Table 3: Classification report for the SMC catalog (four classes).

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 4: Classification report for the LMC catalog (four classes).

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

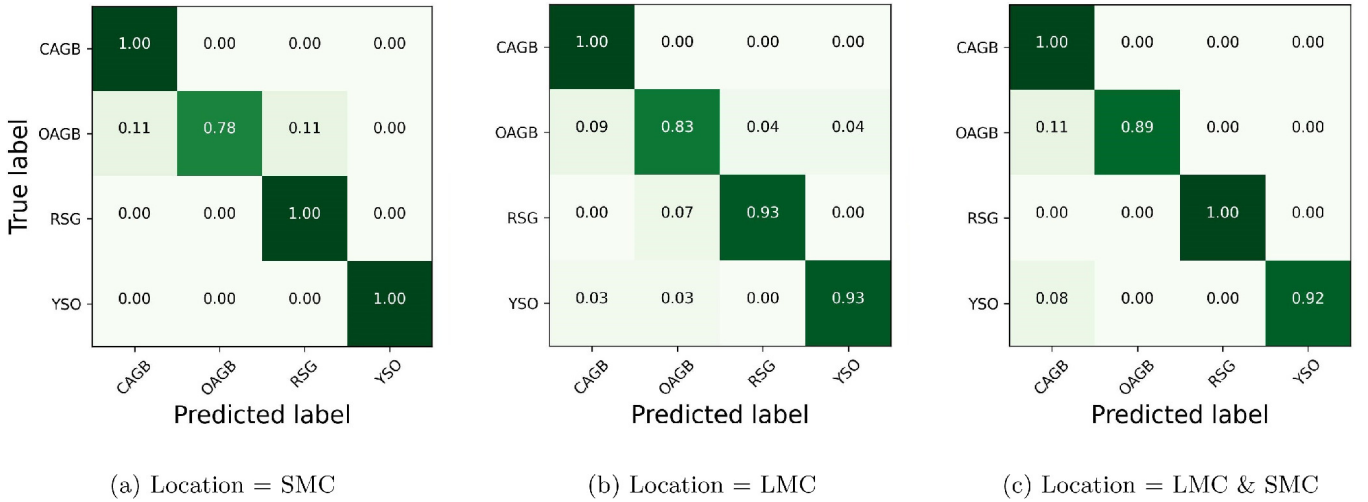


Figure 2: Confusion matrices for four-class PRF experiments: SMC-only (left), LMC-only (middle), and combined LMC&SMC (right).

Finally, we trained a model on the LMC sources and applied it to the SMC data. The accuracy again reached 92%, demonstrating that the model generalizes well across different metallicity environments. The classification report and confusion matrix are shown in Table 5 and Fig. 3.

Table 5: Classification report for the model trained on LMC and tested on SMC (four classes).

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

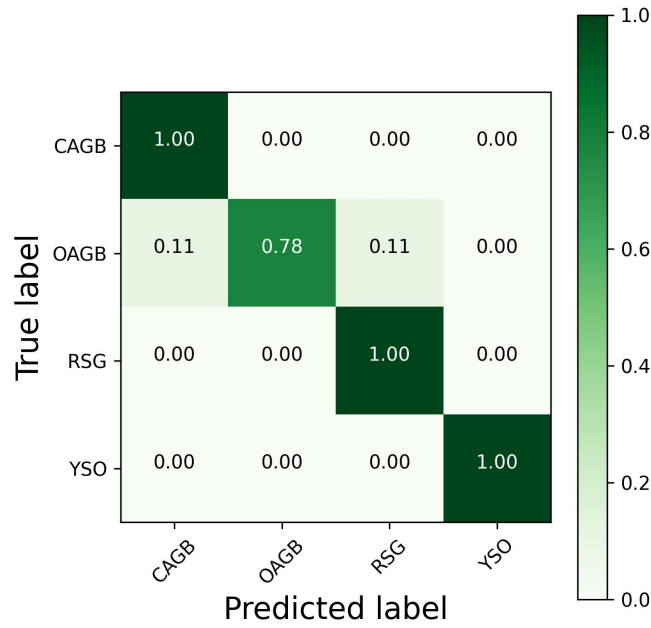


Figure 3: Confusion matrix for the model trained on LMC and tested on SMC (four classes).

In summary, these results show that the metallicity contrast between the LMC and SMC has no measurable effect on the PRF classifier’s performance. Across all four experimental configurations, including separate training, the runs without the PAGB class, the combined datasets, and the cross-galaxy tests, the model reaches nearly identical accuracies and produces similarly structured confusion matrices. This demonstrates that a classifier trained in one low-metallicity environment remains reliable when applied to the other. In practice, this allows the use of unified training sets, reduces dependence on limited spectroscopic samples, and supports extending machine-learning classification to other low-metallicity dwarf galaxies in the Local Group.

Acknowledgements

The authors thank the School of Astronomy at the Institute for Research in Fundamental Sciences (IPM) and the Iranian National Observatory (INO) for supporting this research. S. Ghaziasgar is grateful for the support of the Byurakan Astrophysical Observatory (BAO).

References

- Baron D., 2019, [arXiv e-prints](#), p. [arXiv:1904.07248](#)
- Boyer M. L., et al., 2011, [Astron. J.](#) , **142**, 103
- Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P., 2011, [arXiv e-prints](#), p. [arXiv:1106.1813](#)
- Ghaziasgar S., Abdollahi M., Javadi A., van Loo J. T., McDonald I., Oliveira J., Khosroshahi H. G., 2024, [Communications of the Byurakan Astrophysical Observatory](#), **71**, 377
- Ghaziasgar S., Abdollahi M., Javadi A., van Loon J. T., McDonald I., Oliveira J., Khosroshahi H. G., 2025a, [arXiv e-prints](#), p. [arXiv:2509.05531](#)
- Ghaziasgar S., Masoudnezhad A., Javadi A., van Loon J. T., Khosroshahi H. G., Khosravaninezhad N., 2025b, in McIver J., Mahabal A., Fluke C., eds, IAU Symposium Vol. 19, IAU Symposium. pp 67–72 ([arXiv:2211.03403](#)), [doi:10.1017/S1743921323001308](#)
- Ghaziasgar S., et al., 2025c, [Astrophys. J.](#) , **986**, 168
- Höfner S., Olofsson H., 2018, [Astron. Astrophys. Rev.](#) , **26**, 1
- Jones O. C., et al., 2017, [Mon. Not. R. Astron. Soc.](#) , **470**, 3250
- Kemper F., et al., 2010, [Publ. Astron. Soc. Pac.](#) , **122**, 683
- Kinson D. A., Oliveira J. M., van Loon J. T., 2021, [Mon. Not. R. Astron. Soc.](#) , **507**, 5106
- Kinson D. A., Oliveira J. M., van Loon J. T., 2022, [Mon. Not. R. Astron. Soc.](#) , **517**, 140
- Meixner M., et al., 2006, [Astron. J.](#) , **132**, 2268
- Pennock C. M., et al., 2025, [Mon. Not. R. Astron. Soc.](#) , **537**, 1028
- Pietrzyński G., et al., 2013, [Nature.](#) , **495**, 76
- Reis I., Baron D., Shahaf S., 2019, [Astron. J.](#) , **157**, 16
- Ruffle P. M. E., et al., 2015, [Mon. Not. R. Astron. Soc.](#) , **451**, 3504
- Russell S. C., Dopita M. A., 1992, [Astrophys. J.](#) , **384**, 508
- Subramanian S., Subramaniam A., 2009, [Astron. Astrophys.](#) , **496**, 399
- Subramanian S., Subramaniam A., 2011, in Astronomical Society of India Conference Series. p. 144
- Woods P. M., et al., 2011, [Mon. Not. R. Astron. Soc.](#) , **411**, 1597