2025 MSc Project Proposal

Cover page

1. Project Title

   Automating Galaxy Classification with Unsupervised Machine Learning

2. Research Area

   Science

3. Academic Level

   MSc

4. Primary Supervisor's Details

   a. **Full name:** Dr. Michelle Lochner
   b. **Email address:** mlochner@uwc.ac.za
   c. **University:** University of the Western Cape/ South African Radio Astronomy Observatory

5. Abstract

Upcoming telescopes such as the Square Kilometre Array (SKA) and the Vera C. Rubin Observatory, along with precursor instruments like MeerKAT and DECam, are revolutionising astronomical analysis and discovery. As data volumes surge, efficient methods for galaxy classification and rapid identification of rare sources are increasingly critical. Traditional approaches often rely on time-consuming human labelling, underscoring the need for automated solutions.

This project builds on Mohale & Lochner (2024) to develop and optimise the clustering algorithm Bayesian Gaussian Mixture Models (BGMM) for both optical and radio data. BGMMs can automatically group similar sources, accelerating training set creation, identifying morphologically complex groups, and revealing new patterns in large datasets. The student will create metrics to evaluate cluster stability, assess variability, and determine the optimal number of clusters.

A key innovation is preparing static clustering methods for integration into a cyber-machine interface, allowing humans (including citizen scientists) to focus on labelling ambiguous sources. The project will assess the reliability of BGMM probability estimates against known classes, paving the way for novel unsupervised active learning with next-generation telescopes.

Details of research project

1. Scientific Merit

This project builds on a robust foundation of research into unsupervised machine learning (ML) methods in astronomy, with a particular emphasis on harnessing the cyber-human interface to improve algorithmic performance (e.g., Lochner & Bassett 2021; Etsebeth et al. 2024; Mohale & Lochner 2024; Lochner & Rudnick 2025). Unsupervised ML is increasingly vital in modern astronomy, particularly in the radio domain, where advances in telescope technology, such as MeerKAT and the upcoming SKA, are providing unprecedented sensitivity and revealing a novel view of the universe.

The transformative nature of these new instruments introduces a critical challenge: existing labelled training sets are often inadequate for the fresh, high-dimensional datasets they produce. Traditional supervised learning methods struggle in this context, highlighting the need for automated, label-free approaches. Clustering algorithms, specifically Bayesian Gaussian Mixture Models (BGMM; Attias 1999), offer a promising solution by automatically grouping similar sources without requiring costly human annotation.

While BGMMs have seen some application in astronomy, their potential to generate scientifically useful catalogues remains largely untapped. This project aims to bridge this gap by developing and optimising BGMMs for practical use with both optical and radio data, increasing their utility for a broad range of scientific tasks. By incorporating selective human labelling through an innovative cyber-human interface, the project seeks to convert clustering outputs into training sets for downstream supervised learning or directly produce scientific-grade catalogues.

Beyond classification, BGMMs offer the potential to highlight groups of sources likely to contain morphologically complex or rare objects. This approach not only enhances the efficiency of anomaly detection algorithms but also opens avenues for discovering novel patterns or sub-classes of sources within large datasets. Ultimately, this project aims to set a new standard for unsupervised learning in astronomy, driving forward the field's capacity to handle the data-rich environments of next-generation telescopes.

2. Feasibility and Methods

The first objective of this project is to establish a robust method for understanding the impact of hyperparameter choices on the performance of Bayesian Gaussian Mixture Models (BGMM). This presents a unique challenge, as the unsupervised nature of BGMMs precludes straightforward evaluation using labelled data. Instead, we will develop and implement a suite of metrics to assess:

- **Cluster Stability:** Evaluating the consistency of cluster membership under perturbations to the model hyperparameters.
- **Inter- and Intra-Cluster Variability:** Quantifying the homogeneity within clusters and the distinctness between them.

- **Effective Distance Between Clusters:** Measuring the separability of clusters in high-dimensional feature space.

These metrics will provide a quantitative framework to assess the quality of clustering without requiring explicit labels.

Next, we will focus on evaluating the reliability of the BGMM's probabilistic assignment of sources to clusters. This involves comparing the calculated probabilities with the stability of cluster membership when varying the number of clusters. This analysis is critical for identifying ambiguous sources—those with low assignment confidence—which is essential when generating clean training sets and scientific catalogues. Such prioritisation will enable targeted human labelling efforts through the proposed cyber-human interface.

The project will apply these methodologies to two diverse datasets, enhancing the generalisability of our findings:

1. **DECaLS Galaxy Zoo Optical Dataset (Walmsley et al. 2023)**: Previously utilised in Mohale & Lochner (2024), this dataset offers a valuable benchmark due to the availability of existing human labels. These labels allow us to directly evaluate the clustering algorithm's performance and validate our proposed metrics.
2. **MeerKAT Galaxy Cluster Legacy Survey (Knowles et al. 2022)**: This dataset, rich in morphologically complex and rare sources, provides a contrasting challenge due to its largely unlabelled nature. The metrics developed in this project will be instrumental in evaluating clustering performance on this dataset, demonstrating the efficacy of our approach in real-world, label-scarce scenarios.

By systematically applying BGMM to both optical and radio data, this project aims to establish pioneering guidelines for the effective use of unsupervised clustering in astronomical data analysis. The outcomes will contribute to both practical methodologies and broader scientific understanding, enhancing the utility of BGMMs for the next generation of astronomical surveys.

3. Availability of data and resources

This project makes use of publicly available data, namely optical Galaxy Zoo data from the DECaLS survey (Walmsley et al. (2023)) and radio data from the MeerKAT Galaxy Cluster Legacy Survey (Knowles et al. (2022)). Both datasets have been subjected to self-supervised learning methods to extract features from them, in Mohale & Lochner (2024) and Lochner & Rudnick (2025) so the student will be able to start immediately with this step already completed. In addition, the student will have access to the Ilifu supercomputing facility, should it be required.

4. Milestones and Timeline

Reproduce Mohale & Lochner (2024) Results (Months 1-3)
The initial phase involves reproducing the clustering algorithm and results from Mohale & Lochner (2024) using the DECaLS Galaxy Zoo optical dataset. This step will ensure a solid understanding of Bayesian Gaussian Mixture Models (BGMM) within an astronomical context and provide familiarity with the dataset, tools, and methodologies.

Develop Metrics and Apply to Optical Data (Months 4-12)
Following the reproduction of previous results, the focus will shift to developing new metrics for evaluating cluster stability, inter- and intra-cluster variability, and effective cluster separation. These metrics will be applied to the DECaLS Galaxy Zoo dataset to refine the BGMM algorithm and assess the reliability of its probability estimates against known labels.

Apply Methodology to Radio Data (Months 13-18)
The next phase extends the developed metrics and BGMM algorithm to the MeerKAT Galaxy Cluster Legacy Survey dataset. The approach will be adapted to manage the unlabelled nature of the radio data, showcasing the generalisability of the methods across different data modalities (optical and radio). This will provide insights into the performance of BGMMs with novel, high-dimensional datasets.

Thesis Writing and Submission (Months 19-24)
The final phase will involve compiling research findings into a cohesive Master's thesis and completing any outstanding research tasks.

5. Student Requirements

Good programming skills are critical for this project, primarily in python. Experience with machine learning is advantageous, but not essential.

References

- Lochner & Bassett (2021) - https://arxiv.org/abs/2010.11202
- Etsebeth et al. (2024) - https://arxiv.org/abs/2309.08660
- Mohale & Lochner (2024) - https://arxiv.org/abs/2311.14157
- Lochner & Rudnick (2025) - https://arxiv.org/abs/2411.04188
- Attias H., (1999) - https://proceedings.neurips.cc/paper_files/paper/1999/hash/74563ba21a90da13dacf2a73e3ddefa7-Abstract.html
- Walmsley et al. (2023) - https://arxiv.org/abs/2309.11425
- Knowles et al. (2022) - https://arxiv.org/abs/2111.05673