

AutoSourceID-Classifier

Star-galaxy classification using a convolutional neural network with spatial information

F. Stoppa^{1,6}, S. Bhattacharyya², R. Ruiz de Austri⁵, P. Vreeswijk¹, S. Caron^{3,4}, G. Zaharijas^{2,9}, S. Bloemen¹,
G. Principe^{12,13}, D. Malyshev¹⁴, V. Vodeb², P. J. Groot^{1,10,11}, E. Cator⁶, and G. Nelemans^{1,7,8}

¹ Department of Astrophysics/IMAPP, Radboud University, PO Box 9010, 6500 GL Nijmegen, The Netherlands
e-mail: f.stoppa@astro.ru.nl

² Center for Astrophysics and Cosmology, University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia

³ High Energy Physics/IMAPP, Radboud University, PO Box 9010, 6500 GL Nijmegen, The Netherlands

⁴ Nikhef, Science Park 105, 1098 XG Amsterdam, The Netherlands

⁵ Instituto de Física Corpuscular, IFIC-UV/CSIC, Valencia, Spain

⁶ Department of Mathematics/IMAPP, Radboud University, PO Box 9010, 6500 GL Nijmegen, The Netherlands

⁷ SRON, Netherlands Institute for Space Research, Sorbonnelaan 2, 3584 CA Utrecht, The Netherlands

⁸ Institute of Astronomy, KU Leuven, Celestijnenlaan 200D, 3001 Leuven, Belgium

⁹ Institute for Fundamental Physics of the Universe, Via Beirut 2, 34151 Trieste, Italy

¹⁰ Department of Astronomy and Inter-University Institute for Data Intensive Astronomy, University of Cape Town,
Private Bag X3, Rondebosch 7701, South Africa

¹¹ South African Astronomical Observatory, PO Box 9, Observatory, Cape Town 7935, South Africa

¹² Dipartimento di Fisica, Università di Trieste, 34127 Trieste, Italy

¹³ Istituto Nazionale di Fisica Nucleare, Sezione di Trieste, 34127 Trieste, Italy

¹⁴ Erlangen Centre for Astroparticle Physics, Nikolaus-Fiebiger-Str. 2, Erlangen 91058, Germany

Received 26 July 2023 / Accepted 17 October 2023

ABSTRACT

Aims. Traditional star-galaxy classification techniques often rely on feature estimation from catalogs, a process susceptible to introducing inaccuracies, thereby potentially jeopardizing the classification's reliability. Certain galaxies, especially those not manifesting as extended sources, can be misclassified when their shape parameters and flux solely drive the inference. We aim to create a robust and accurate classification network for identifying stars and galaxies directly from astronomical images.

Methods. The AutoSourceID-Classifier (ASID-C) algorithm developed for this work uses 32x32 pixel single filter band source cutouts generated by the previously developed AutoSourceID-Light (ASID-L) code. By leveraging convolutional neural networks (CNN) and additional information about the source position within the full-field image, ASID-C aims to accurately classify all stars and galaxies within a survey. Subsequently, we employed a modified Platt scaling calibration for the output of the CNN, ensuring that the derived probabilities were effectively calibrated, delivering precise and reliable results.

Results. We show that ASID-C, trained on MeerLICHT telescope images and using the Dark Energy Camera Legacy Survey (DECaLS) morphological classification, is a robust classifier and outperforms similar codes such as SourceExtractor. To facilitate a rigorous comparison, we also trained an eXtreme Gradient Boosting (XGBoost) model on tabular features extracted by SourceExtractor. While this XGBoost model approaches ASID-C in performance metrics, it does not offer the computational efficiency and reduced error propagation inherent in ASID-C's direct image-based classification approach. ASID-C excels in low signal-to-noise ratio and crowded scenarios, potentially aiding in transient host identification and advancing deep-sky astronomy.

Key words. methods: data analysis – techniques: image processing – astronomical databases: miscellaneous – stars: imaging – Galaxies: statistics

1. Introduction

Current and future large-scale astronomical photometric surveys are amassing, and will continue to amass, vast quantities of photometric data, including millions of images containing billions of stars and galaxies. This data influx necessitates a series of processing steps, such as source localization, anomaly detection, feature extraction, and star-galaxy classification, to be further optimized for efficiency in terms of both time and computational resources. With the advent of CMOS detectors capable of capturing large-format images of the night sky at cadences exceeding

1Hz, traditional methods and software for data processing will be inadequate to keep pace with the data acquisition rate.

To address this challenge, our research aims to enhance and streamline existing methods using machine learning tools. In AstroSourceID-Light (ASID-L; Stoppa et al. 2022), we demonstrated the potential for rapid and accurate source identification on images, achieving this in a fraction of the time required by the currently used methods. Building on this foundation, we now focus on star-galaxy classification, a fundamental data-processing task, and often the initial step for the scientific exploitation of survey data. Despite advances in source

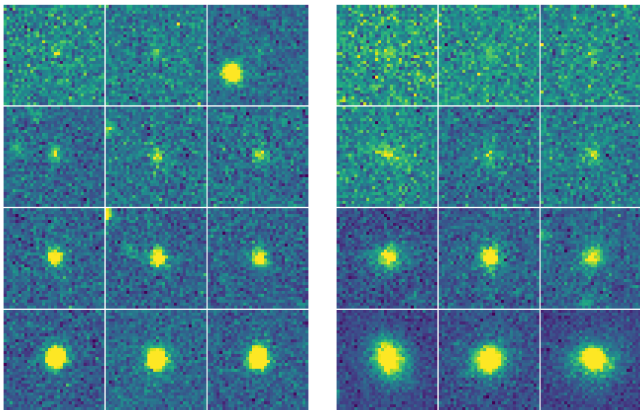


Fig. 1. Cutouts of stars (left) and galaxies (right) for $S/N = \{4, 10, 25, 65\}$ (top to bottom). The source is always at the center of the cutout. At a low S/N , sources are barely discernible, making their classification a complex task.

localization and feature extraction algorithms, there remains significant potential for improvement in star-galaxy classification methods, which are often applied from catalogs rather than directly from source images (Weir et al. 1995; Ball et al. 2006; Vasconcellos et al. 2011; Sevilla-Noarbe et al. 2018). Current classification methods, including classification tree methods based on the morphological features of the sources and Bayesian approaches that integrate available source information with prior knowledge about nearby star and galaxy populations, are prevalent but there is room for improvement (Henrion et al. 2011; López-Sanjuan et al. 2019).

Similar to many areas of astronomy, data analysis in star-galaxy classification has seen a surge in machine learning applications. The first machine learning application for star-galaxy classification was introduced in Odewahn et al. (1992), and it quickly became a core component of the astronomical image processing software SourceExtractor (Bertin & Arnouts 1996). More recently, a series of studies have employed various neural network architectures to address the star-galaxy classification problem (Odewahn et al. 2004; Fadely et al. 2012; Cabayol et al. 2018), with significant progress made in Kim & Brunner (2016). In this study, the authors demonstrated the effectiveness of convolutional neural networks (CNNs) in learning features directly from multiband (*ugriz*) optical images of different sources. When applied to the Canada–France–Hawaii Telescope Lensing Survey dataset (Heymans et al. 2012), their network was able to generate a classification score that matched the accuracy of a random forest-based algorithm (Breiman 2017), but with better-calibrated probability estimates, indicating a promising avenue for future research in this field.

Most of the methods mentioned above operate under the assumption that galaxies manifest as extended sources while stars appear as point sources. However, this distinction becomes less clear at fainter magnitudes and higher redshifts, where the morphological features of galaxies are less discernible. As a result, the effectiveness of these methods decreases under such conditions. Our study addresses this challenge by aiming to accurately classify all stars and galaxies within a survey, with a particular focus on those with a signal-to-noise ratio (S/N) near the detection limit, as illustrated in Fig. 1.

This task presents a unique set of challenges, primarily due to the need to identify a limited range of morphological characteristics. These challenges are further complicated by our deliberate

choice of single-band images. To address this, we have designed our method to be band-agnostic, allowing for flexibility in the choice of filters without compromising performance. While multiband sets of the same source would undoubtedly enrich the classification process, our approach is optimized to be effective even with a single image. This proves particularly advantageous for small- to medium-sized optical telescopes, where acquiring multiband data may not always be feasible or could be resource-intensive. By focusing on single-band images, we eliminated the need for repeated observations of the same source, thereby simplifying data acquisition. Furthermore, we introduced an ensemble technique, detailed later in this paper, that allows for a straightforward extension of our method to multiband images. This not only refines our classification methodology but also provides a clear path for adapting it to more complex datasets.

Given these challenges and the potential of single-band imagery, we have developed the AutoSourceID-Classifier (ASID-C). This tool is designed to take optical image cutouts of sources, which in our case are retrieved with ASID-L, and their positions in the full telescope image as input, and then output a probability for each source being either a star or a galaxy. To ensure the validity of our classification model’s predictions, we have employed an enhanced Platt scaling method (Platt 1999) to calibrate the network’s outputs.

The method presented here¹ is the fourth deep learning algorithm developed in the context of MeerLICHT/BlackGEM telescopes (Bloemen et al. 2016; Groot et al. 2022), following MeerCRAB, an algorithm for classifying real and bogus transients (Hosenie et al. 2021), ASID-L for source localization (Stoppa et al. 2022), and ASID-FE for feature extraction (Stoppa et al. 2023). Building on these deep learning algorithms, our ultimate goal is to establish the first fully automated machine learning detection pipeline for small- to medium-sized optical telescopes, thereby facilitating more efficient and accurate astronomical observations and analyses.

The rest of this paper is organized as follows: Sect. 2 provides details on the datasets used and the preprocessing steps undertaken. Section 3 describes the model, while Sect. 4 presents the results, including a comparison with SourceExtractor (Bertin & Arnouts 1996), a widely used tool for optical image analysis. Finally, in Sect. 5, we present our conclusions and discuss potential scientific applications of our tool.

2. Data

To develop our classifier, we constructed a dataset comprising source cutouts from images captured by the MeerLICHT telescope, paired with morphological classifications of sources from the Dark Energy Camera Legacy Survey (DECaLS; Blum et al. 2016). To the best of our knowledge, this dataset, containing approximately 12 million source cutouts, is the largest ever assembled for star-galaxy image classification in a machine learning context. In the following section, we outline the telescopes and catalogs used to create this dataset and provide a brief overview of the preprocessing steps involved in obtaining the image cutouts.

2.1. MeerLICHT images

The MeerLICHT telescope, which pairs a 65cm diameter primary mirror with a $10.5k \times 10.5k$ CCD detector, creates a 2.7 square degree monolithic field-of-view, sampled at

¹ <https://github.com/FiorenSt/AutoSourceID-Classifier>

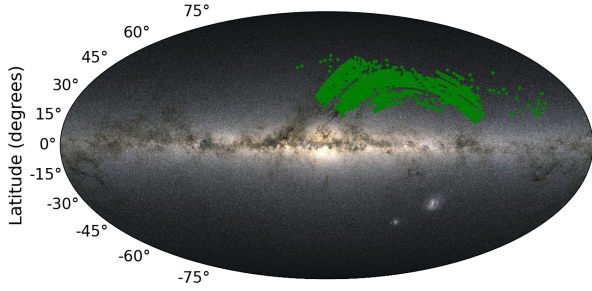


Fig. 2. Sky coordinates of the 718 MeerLICHT images (green dots) overlaid on an image of the Galactic Plane. This visualization demonstrates the diverse array of spatial densities covered by our dataset, facilitating a thorough evaluation of our model’s capacity to classify regions with varying degrees of crowdedness.

$0.56'' \text{ pix}^{-1}$ (Bloemen et al. 2016). With an average image quality of $2\text{--}3''$, point sources are sampled at $4\text{--}6$ pixels per FWHM. The telescope’s primary function is to enable simultaneous transient detection at radio and optical wavelengths, working in conjunction with the MeerKAT radio telescope (Jonas & MeerKAT Team 2016). The available filter set includes the SDSS *ugriz* filters and a broader *q*-band filter ($440\text{--}720$ nm), which roughly combines SDSS *g + r*. Images captured by the telescope undergo processing at the IDIA/ilifu facility through the BlackBOX software² (Vreeswijk et al., in prep.), which handles standard image processing tasks such as source detection, astrometric and photometric calibration, creation of the position-dependent point spread function (PSF), image subtraction, photometry, and transient detection.

To construct a robust and diverse training set, we used 12 million source cutouts derived from 718 distinct, nonoverlapping full-field images captured by the MeerLICHT telescope. These images span multiple filters, including *u*, *g*, *q*, *i*, and *r* bands, but it is important to note that each source appears only once in the dataset due to the nonoverlapping nature of the images. The selected images were marked “green” in the pipeline, indicating they were free from obvious anomalies. However, we did not impose any additional restrictions based on visibility parameters. As depicted in Fig. 2, the full-field images cover a wide range of spatial densities, from densely populated regions along the Galactic Plane to areas characterized by a sparse distribution of sources. This comprehensive coverage allows us to train and test the neural network’s ability to classify regions of varying crowdedness.

For each source in the images, a cutout of 34×34 pixels is created with the source, previously localized with ASID-L, at its center. This size was chosen to ensure that the vast majority of sources are comfortably enclosed within the cutout boundaries. For galaxies larger than 19 arcsec, the cutout will be smaller than the source itself; however, this will not prevent our network from learning that such images are likely to be galaxies. In addition, we match these localized sources with their counterparts in MeerLICHT’s catalogs. These catalogs, produced using SourceExtractor (Bertin & Arnouts 1996), provide approximate estimates for flux, location, and a stellarity parameter, which are useful for additional testing on the results of ASID-C. Finally, out of these, we retain the 12 million cutouts that have a matching source in the DECaLS dataset, which we discuss in more detail in the next section.

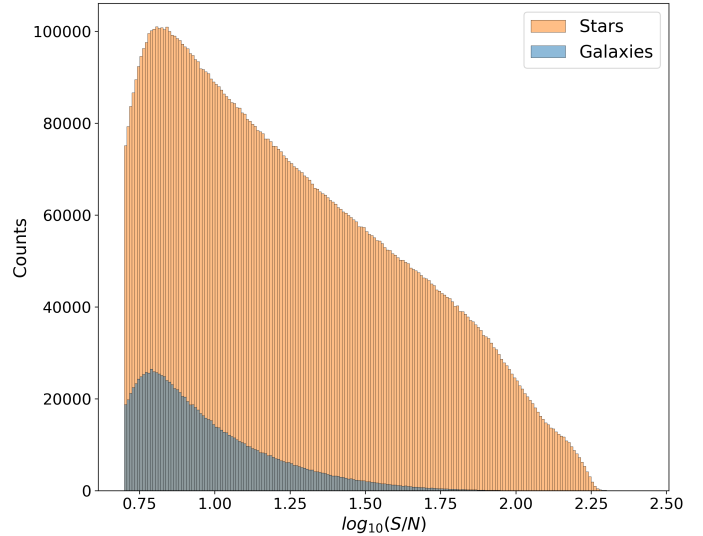


Fig. 3. Source counts by morphology as a function of S/N for our dataset. It reveals a 10:1 disproportion favoring stars and a distinct difference in S/N, where stars consistently display higher values.

2.2. Legacy survey

To complement the MeerLICHT images and provide a robust basis for star-galaxy classification, we used the catalogs from the Dark Energy Camera Legacy Survey (DECaLS), part of the DESI Legacy Imaging Surveys.

The DESI Legacy Imaging Surveys aim to map 14 000 square degrees of extragalactic sky in three optical bands (*g*, *r*, *z*) and combine it with four mid-infrared bands from the Wide-field Infrared Survey Explorer (WISE; Dey et al. 2019; Schlegel et al. 2021). This ambitious project is accomplished through three imaging projects that comprise the Legacy Surveys: DECaLS, the Beijing-Arizona Sky Survey (BASS; Zou et al. 2019), and the Mayall *z*-band Legacy Survey (MzLS; Silva et al. 2016). The current data release, Data Release 10 (DR10), is the tenth public data release from these surveys. Source detection in DR10 is performed using a filter matched to the PSF and spectral energy distribution (SED) of the sources on the stacked images, featuring a 6σ detection limit. Each image is processed using its PSF model to detect sources, and these processed images are then combined in a weighted manner to improve the detection of point sources. DR10 provides a classification of sources into five types based on their shape and structure, determined using a multiband, multi-epoch photometric model, including one for point sources and four for galaxies: round exponential galaxies (“REX”), deVaucouleurs (“DEV”) profiles, exponential (“EXP”) profiles, and Sersic (“SER”) profiles.

DECaLS, in particular, offers significant overlap with the fields observed by MeerLICHT, making it an ideal choice for our study. The depth and quality of the imaging data provided by DECaLS allow for reliable and detailed morphological information, which is crucial for differentiating between stars and galaxies, especially at faint magnitudes where the morphological features of galaxies may be less discernible. For our dataset, we only use DR10 sources detected by both ASID-L and the MeerLICHT pipeline in the 718 MeerLICHT images, as this allows us to use additional features estimated by the MeerLICHT pipeline, such as flux and S/N.

As depicted in Fig. 3, the resulting dataset from the match of MeerLICHT and DECaLS sources is characterized by an

² <https://github.com/pmvreeswijk/BlackBOX>

imbalance, with a 10:1 ratio favoring stars over galaxies and a noticeable difference in the S/N range. This imbalance underscores the challenge of our task; while the source population is predominantly stars at higher S/N values, the task of distinguishing between stars and galaxies becomes increasingly complex at lower S/N values.

2.3. Crossmatching with additional galaxies and quasar catalogs

To evaluate the accuracy of the galaxy labels within our dataset, we performed a crossmatch operation with external catalogs. For this crossmatch, we incorporated galaxies from the Two Micron All Sky Survey (2MASS; [Skrutskie et al. 2006](#)), the Lyon-Meudon Extragalactic Database (LEDA; [Paturel et al. 1995](#)), and the Kilo-Degree Survey Data Release 4 (KiDS; [Bilicki et al. 2021](#)). These galaxies were then cross-referenced with all the 12 million sources in our dataset, preserving those within a proximity of less than 0.56 arcsec (approximately 1 pixel). Of all the matching sources in the refined subset, approximately 12 640 galaxies, 99.3% were accurately identified as galaxies within the Legacy survey dataset. The remaining 0.7% misclassified sources pose a minor concern for our analysis, likely attributable to chance alignment and the fact that the sources we use are in a sky section first incorporated in the Legacy Survey in the latest DR10 release.

We further cross-referenced our catalog with a known list of quasars retrieved from the SIMBAD astronomical database ([Wenger et al. 2000](#)). Quasars, or quasi-stellar objects, are exceptionally luminous active galactic nuclei fueled by the accretion of material onto supermassive black holes at the heart of distant galaxies. Within our star-galaxy classification dataset, a minuscule fraction of the images, approximately 0.006%, are quasars. Due to their stellar-like appearance and intense luminosity, quasars are frequently misclassified as stars, a trend evident in our dataset where 95% of these quasars are erroneously labeled stars.

2.4. Data split, augmentation, and calibration

In this section, we discuss the process of data splitting, calibration, and augmentation, which are crucial steps in preparing our dataset for effective machine learning model training and evaluation.

The image cutouts dataset is divided into four subsets: 50% for training, 20% for validation, 20% for testing, and 10% for calibration. The training, validation, and test sets are used to train the model and evaluate its performance, while the calibration set is used to fine-tune the classifier's probability predictions. It is important to note that the split is made considering the disproportion of stars with respect to galaxies in the dataset, which is approximately 10:1. This ensures that the distribution of stars and galaxies in each subset reflects the actual distribution in the dataset, which is crucial for training a model that can accurately classify these celestial objects.

The dataset consists of 34×34 pixel cutouts, specifically designed to enable an augmentation step during the training stage aimed at improving the prediction of the network. We enhance the model's robustness and generalizability by applying a random one-pixel shift in any cardinal direction (up, down, left, or right) to each image. This random one-pixel shift reduces the cutout's dimensions to 32×32 pixels, contributing to a more diverse training dataset. Since the shift is only one pixel, it does not substantially impact the source's information; however, it

introduces variability in the training set that can be leveraged to identify sources in different positions. Additionally, this shifting helps mimic the potential misalignment of the sources in the cutouts in real scenarios.

Incorporating the calibration set ensures that ASID-C produces well-calibrated probability predictions, ultimately improving classification performance. The calibration process, described in Sect. 3.3, adjusts the classifier's output probabilities to better align with the true class probabilities, reducing the potential for over or underestimating probabilities. This step enhances the reliability and usefulness of the probability predictions, particularly in applications where accurate probability estimates are essential for downstream analysis or decision-making.

3. Method

In the field of image classification, machine learning techniques have been instrumental in transforming the way we analyze and interpret data. Among these techniques, CNNs have emerged as a powerful tool due to their ability to process image data in a robust and flexible manner. Introduced by [Fukushima & Miyake \(1982\)](#) and [LeCun et al. \(1995\)](#), CNNs have found widespread use in the computer vision community, providing a robust foundation for our work.

In this section, we outline our specific implementation of a machine learning model for star-galaxy classification. This model is designed to tackle the intricacies of our dataset, providing an optimized and custom-tailored approach to meet this classification challenge. Although the model is specifically tailored for our dataset and the star-galaxy classification task, its general structure and approach can be adapted for other astronomical datasets, telescopes, and purposes involving image data. This flexibility makes our model a versatile tool that can contribute to a wide range of astronomical research tasks.

3.1. Model

Our model uses a CNN, a type of neural network particularly effective for image analysis. In a CNN, an image is processed through multiple layers to generate feature maps. These maps are created by convolving each input feature map with a set of weights known as filters. Each feature map uses a distinct set of filters, allowing for a rich representation of the input data. This mechanism is at the core of our model's ability to analyze and classify the celestial images in our dataset.

While previous studies such as [Kim & Brunner \(2016\)](#) have used CNN architectures for star-galaxy classification, they often rely on multiple filter band images to enhance the classification results. In contrast, our approach simplifies this requirement by using single-band images. This makes our model versatile and independent of the specific filter applied.

Our deep neural network is designed with a dual-branch structure, specifically crafted to process image data and spatial information as inputs. This architectural choice is informed by the unique challenges of classifying celestial objects in optical telescope images, specifically, in our case, those from the MeerLICHT telescope. Spatial information serves a crucial role in our model, aiding the classifier in handling variations in the appearance of sources based on their location within the full-field MeerLICHT image. The PSF of sources, which influences their appearance, varies depending on their position within the image.

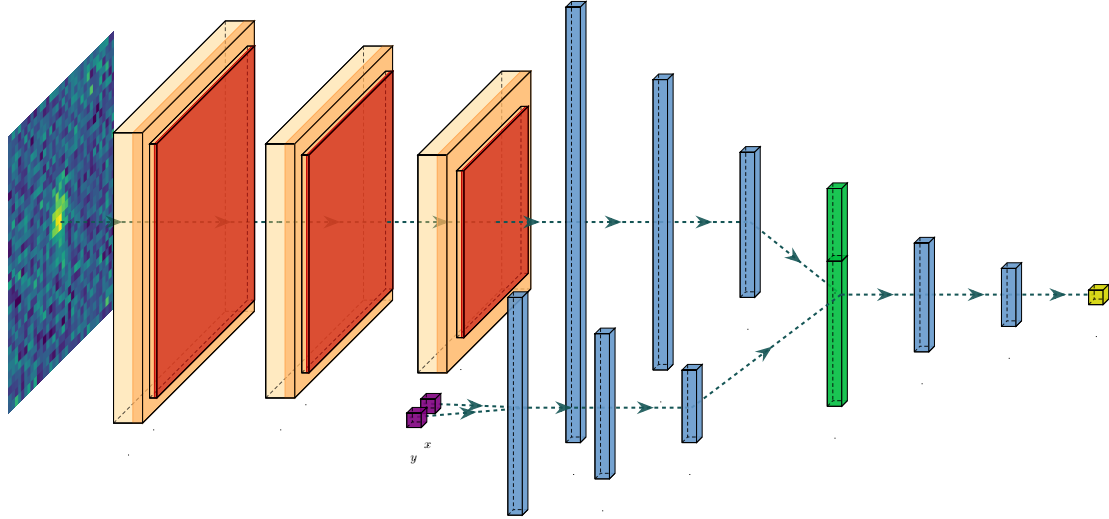


Fig. 4. Dual-branch network architecture used in this study. The primary branch processes the images through a series of convolutional layers (yellow), max-pooling layers (red), and dense layers (blue) to extract key features. The secondary branch processes the spatial information (pixel coordinates $[x, y]$) through multiple dense layers. Both branches merge through a concatenation layer (green), followed by additional dense layers and a final softmax layer for classification. This architecture allows the model to leverage image features and spatial information for more accurate star-galaxy classification.

For instance, sources at the center of the image are symmetrical, while those near the edges exhibit asymmetry. Furthermore, sources farther from the image’s center display an elongated PSF, causing stars to resemble galaxies. By integrating spatial information into our model, we can effectively account for these variations, thereby enhancing the accuracy of our model in classifying stars and galaxies. We validated the effectiveness of this approach by comparing the model’s performance with and without the spatial information branch. The results are shown in Appendix A.

As shown in Fig. 4, the model is a dual-branch neural network that accepts image data and spatial information as inputs. The imaging branch, a CNN, processes 32×32 pixel images of celestial objects, while the spatial branch handles the 2-dimensional location data. The CNN comprises three convolutional layers with 32 filters each, followed by max-pooling layers. The spatial branch, a fully connected network, includes two dense layers with 64 and 32 neurons, respectively. Outputs from both branches are concatenated and passed through two additional dense layers before reaching the final sigmoid activation function, which generates a value between 0 (galaxy) and 1 (star).

We train the model using the Adam optimizer (Kingma & Ba 2015). We implement several strategies to achieve better accuracy and convergence and prevent overfitting. As the network gets trained over multiple iterations and approaches the minima of the loss landscape, it is typically suggested to have a lower learning rate (η_t) for improving convergence (Wu et al. 2019). In our approach, we implement exponential decay with an initial learning rate (η_0) of 0.001 that decays exponentially after eight epochs (iterations) with a decay rate $\gamma = 0.99$.

To prevent overfitting, we implement early stopping, which ensures the training stops if the loss for the validation set does not improve over ten epochs. The model’s performance is assessed using precision, recall, and the area under the receiver operating characteristic (ROC) and precision-recall (PR) curves. Our approach allows for a low number of parameters, approximately 72k, making the algorithm fast and suitable for the needs of an automatic detection pipeline.

We use TensorFlow³ (Abadi et al. 2016) to implement the model and evaluate all possible hyperparameters with Weights & Biases⁴, a machine learning platform for developers to track, version control, and visualize results, which is especially helpful in our case for fixing hyperparameters. The results presented below were computed using an NVIDIA GeForce RTX 2080 GPU.

3.2. Loss function

In this work, we employ the binary focal Loss as our loss function, initially introduced by Lin et al. (2017) to tackle the class imbalance issue in object detection tasks. The focal loss is designed to prioritize hard-to-classify examples while reducing the weight of easy examples. It is defined as:

$$FL(y, p) = -\alpha_t (1 - p_t)^\gamma \log(p_t), \quad (1)$$

where

$$\alpha_t, p_t = \begin{cases} \alpha, & p; & y = 1 \\ 1 - \alpha, & 1 - p; & y = 0. \end{cases} \quad (2)$$

Here, y is the true label, p is the predicted probability, α is a weighting factor to balance the two classes, and γ is a focusing parameter that adjusts the rate at which easy examples are down-weighted. When $\alpha = 1$ and $\gamma = 0$, the binary focal loss simplifies to the well-known binary cross-entropy loss (Cox 1958). The binary cross-entropy loss is a common function for binary classification tasks, but it can be less effective when there is a class imbalance.

While the focal loss was initially designed for one-stage detectors in computer vision tasks with a significant class imbalance between foreground and background (on the order of 1000:1), we adapt this loss function to address the class imbalance between stars and galaxies (approximately 10:1) in our dataset. However, as shown in Fig. 5, we found no substantial

³ <https://www.tensorflow.org/>

⁴ <https://wandb.ai/site>

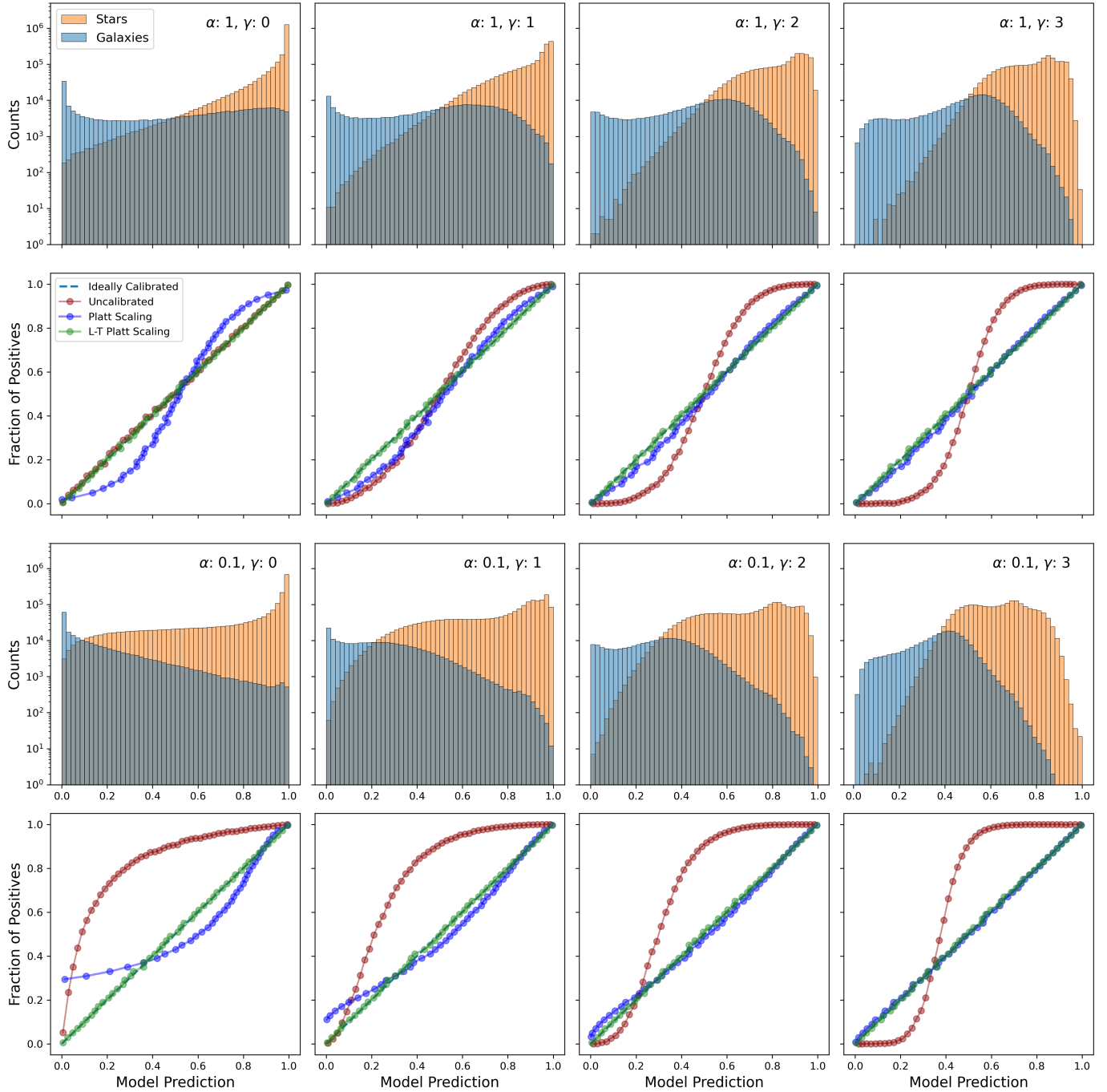


Fig. 5. Calibration analysis of the output of models with varying α and γ focal loss parameters. The first and third rows display the model predictions for the test set, with the color indicating the true class. The second and fourth rows showcase the calibration results for each corresponding model, presenting the uncalibrated, Platt-scaled, and Logit-Transformed Platt scaling results. This side-by-side comparison illustrates the effectiveness of the calibration techniques in enhancing the alignment of predicted probabilities with the observed frequency of positive class occurrences.

difference between the results of any of the focal loss models, especially after the calibration method introduced in the next section. Therefore, a focal loss with parameters $\alpha = 1$ and $\gamma = 0$, which equates to a binary cross-entropy, emerges as the optimal choice.

The lack of improvement from the focal loss, despite the 10:1 imbalance, can be attributed to the concept of the effective sample size (ESS; Bartoszek 2016; Cui et al. 2019). The ESS of our dataset, which is a measure of the number of independent observations that a given dataset is equivalent to, is large enough to

allow for accurate learning despite the imbalance. This suggests that when the dataset is sufficiently large, the use of the focal loss may not bring any additional benefit.

Although the focal loss does not yield a substantial improvement over the standard cross-entropy loss in our specific case, it offers an opportunity to investigate the impact of varying the focusing parameter (γ) and the weighting parameter (α) on probability calibration, which we explore in the subsequent section. The flexibility of the focal loss function could prove advantageous in future studies, especially when incorporating additional

features or diverse types of input data into the model, which may result in more complex classification scenarios.

3.3. Improving probability calibration with Logit-transformed Platt scaling

A crucial aspect of a well-performing classifier is its ability to provide reliable posterior class probability estimates that align with the actual likelihood of a positive class occurrence. This characteristic, known as calibration, is often visualized using probability calibration curves or reliability curves (Degroot & Fienberg 1983).

To construct a probability calibration curve, we apply the model to the calibration set, approximately one million cutouts, and divide the probability estimates into discrete bins, each representing a range of predicted probabilities. We then calculate the fraction of positive examples for each bin and plot these fractions against the predicted probabilities. A well-calibrated classifier will have a calibration curve closely following the diagonal line, indicating that the predicted probabilities accurately reflect the observed frequency of positive class occurrences.

However, it is often the case that the raw output scores of a classifier do not perfectly align with these observed frequencies. This discrepancy can be due to various factors, including the complexity of the data, the model's assumptions, and the training process. To address this issue, a variety of post-calibration methods have been developed, aiming to adjust the output scores and improve their alignment with the true probabilities.

One such method is Platt scaling, a widely used technique for calibrating the outputs of a model. Platt scaling transforms the raw output values to better align with the true probabilities of the predicted classes (Platt 1999). The standard Platt scaling formula is:

$$P(y = 1|f) = \frac{1}{1 + \exp(Af + B)}, \quad (3)$$

where A and B are parameters estimated from the data using the Maximum Likelihood method. This method has been shown to improve the calibration and performance of machine learning models (Niculescu-Mizil & Caruana 2005; Kull et al. 2019). However, Platt scaling can struggle when the output of a classifier is not sigmoid-shaped (Kull et al. 2017). To address this, we use a simple modification of the traditional Platt scaling method: applying a logit transformation to the model's output scores prior to Platt scaling. While this transformation is recognized in the calibration literature (Filho et al. 2023), existing studies have not thoroughly investigated its advantages and disadvantages. However, in our specific case, the Logit-Transformed (L-T) Platt scaling shows superior performance, providing empirical evidence of its effectiveness over simple Platt scaling.

The L-T Platt scaling formula is given by:

$$P(y = 1|x) = \frac{1}{1 + \exp(A \logit(f) + B)}. \quad (4)$$

In the formula above, the term $\logit(f)$ in the equation represents the logit transformed output score of the model. The parameters A and B , estimated from the data, are similar to those in the traditional Platt scaling method. These parameters are optimized to achieve the best possible calibration for the specific dataset and model at hand.

We applied the L-T Platt scaling to a range of models, each trained with different focal loss parameters. The resulting models' predictions and their calibration, illustrated in Fig. 5,

demonstrate that this method can recover almost perfectly calibrated probabilities for each model.

The calibration process is a critical step in our methodology, enabling the direct comparison of different models. As discussed in Sect. 3.2, post-calibration, the performance of the models converges, indicating the robustness of our dataset. This robustness obviates the need for correction factors from the focal loss, suggesting that a model trained with a simple cross-entropy loss function can directly yield optimal results.

Interestingly, our analysis also suggests that the calibration technique itself may not be necessary for our specific case. However, the process provides valuable insights into the performance of different models and serves as a useful tool for model comparison and evaluation. Therefore, while not essential for our specific task, the calibration process contributes to a more comprehensive understanding of our models' performance and the impact of different loss functions. This knowledge can serve as a valuable guide for future research and model development in star-galaxy classification and beyond. The insights gained can potentially influence a wider range of applications, fostering advancements in the broader field of astronomical image analysis.

4. Results

This section presents a thorough analysis of our network's performance using a test set of over 2 million cutouts, focusing on ASID-C's robustness and versatility across various challenging astronomical scenarios.

We first employ a Uniform Manifold Approximation and Projection (UMAP⁵; McInnes et al. 2018) as a dimensionality reduction technique to visualize the model's decision-making process. We then explore the model's performance across different S/N bins, MeerLICHT bands, and high stellar density regions, particularly those near the Galactic Plane. We also evaluate ASID-C computational efficiency as a function of increasing dataset sizes.

We use a set of relevant metrics to gauge the effectiveness of our results. Unlike deterministic classifiers that assign discrete labels to each source, our probabilistic classifier provides a probability that determines whether each source is a star or a galaxy. The performance evaluation of probabilistic classifiers often involves transforming probability estimates into class labels by setting a specific probability threshold. For example, a source is classified as a star if $p > 0.5$ and a galaxy if $p \leq 0.5$. However, this approach can be misleading if the predictions are not well-calibrated. Therefore, we employ several performance metrics that are suitable for probabilistic classifiers: the area under the curve (AUC) for the Receiver Operating Characteristic (ROC; Swets 1996) curve, the area under the precision-recall curve (AUPRC; Davis & Goadrich 2006) and Brier score (Brier 1950).

Finally, we compare ASID-C's performance with that of SourceExtractor, a widely used software in astronomy, and eXtreme Gradient Boosting (XGBoost), a state-of-the-art machine learning decision tree algorithm for tabular data. This comparative analysis serves to highlight the relative strengths and limitations of each approach in the context of star-galaxy classification.

⁵ <https://umap-learn.readthedocs.io/en/latest/index.html>

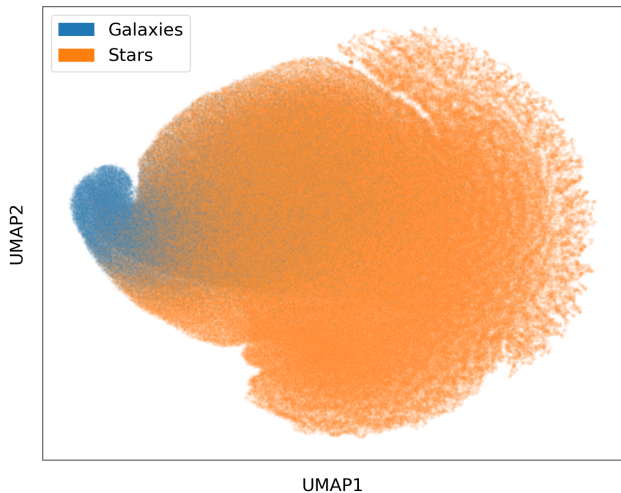


Fig. 6. UMAP latent space visualization. The axes represent the UMAP embeddings, with each point colored by their true label, star (orange) or galaxy (blue).

4.1. ASID-C performance and misclassifications

This section delivers an in-depth assessment of ASID-C, focusing on both its performance and instances of misclassification. The goal is to provide a nuanced perspective on the model’s capabilities and constraints.

To gain a visual understanding of how ASID-C differentiates between stars and galaxies, we employ a visualization and dimensionality reduction technique called UMAP. The latter, proposed as an alternative to T-distributed stochastic neighbor embedding (t-SNE; Maaten & Hinton 2008), is renowned for its ability to preserve the global structure of data and its efficient implementation.

We first adapted the trained model to integrate UMAP into our deep neural network by removing the final classification layer (of the model depicted in Fig. 4). This adjustment enables us to concentrate on the output from the preceding 32-dimensional dense layer, which is abundant with insights into the features learned by the network. Subsequently, we employed this adapted model to generate predictions on the test set and apply UMAP to the latter. UMAP provides a condensed summary of the decision-making process, unveiling the intricate patterns the model uses to make accurate classifications and offering a more nuanced understanding of the network’s performance. This method highlights the complex interplay of features contributing to the successful differentiation between stars and galaxies.

As depicted in Fig. 6, the UMAP visualization reveals a distinct separation between stars and galaxies, with a clear clump corresponding to galaxies. It is important to note that the axes in Fig. 6 represent the UMAP embeddings and do not lend themselves to direct interpretation as with techniques like Principal Component Analysis (PCA). The construction of a UMAP visualization requires selecting specific parameters that will influence the final output. In our case, we chose a Correlation metric, 30 nearest neighbors, and a minimum distance of 1.

Figure 7 presents the UMAP embeddings colored by their full model prediction. This visualization further emphasizes the separation between galaxies, predominantly predicted with a 0, and stars, predominantly predicted with a 1. The intermediate region of predictions, ranging between 0.4 and 0.6, corresponds to the area where distinguishing between the two classes is most challenging.

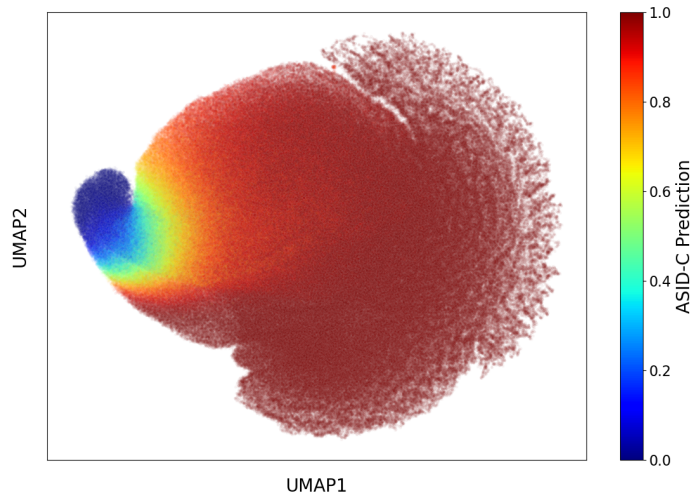


Fig. 7. UMAP latent space visualization. The axes represent the UMAP embeddings, with each point colored by their full-model predicted value.

The visual insights obtained from the UMAP analysis have been instrumental in the model’s iterative development process. They not only facilitated targeted refinements in the model’s architecture but also enriched our conceptual understanding of the classification challenge at hand. This visualization serves as a diagnostic tool, allowing us to pinpoint both the strengths and areas for improvement in ASID-C’s decision-making algorithm. While the model has demonstrated good performances, as outlined in earlier sections, it is worth noting that the classification is not entirely dichotomous, and there is a discernible overlap between the two classes, indicating complexities in the task that warrant further investigation. Subsequent sections will delve into the key factors contributing to this complexity.

4.1.1. Performance across signal-to-noise ratio

To further evaluate the robustness of ASID-C, we analyzed its performance across five distinct bins of S/N. This analysis aims to shed light on how well the classifier performs under varying levels of noise, a crucial aspect of its applicability in real-world astronomical observations.

To provide a comprehensive view, we present both the receiver operating characteristic (ROC) and precision–recall (PR) curves. The ROC curve plots the true positive rate against the false positive rate at various threshold levels, providing a holistic view of the classifier’s discriminative power. A curve that closely follows the top-left border of the plot indicates a well-performing classifier. The area under the ROC curve (ROC-AUC) serves as a single scalar metric that quantifies the classifier’s ability to distinguish between the two classes across all thresholds.

Conversely, the PR curve shows the trade-off between precision, True Positives / (True Positives + False Positives), and recall, True Positives / (True Positives + False Negatives) at various thresholds. A perfect PR curve would reach the top-right corner of the plot, indicating both high precision and high recall. This curve is particularly useful for evaluating performance in imbalanced datasets, and it allows the assessment of the classifier’s performance separately for stars and galaxies. As for the ROC-AUC, the area under the PR curve (PR-AUC) provides a summary measure of the classifier’s performance across all thresholds.

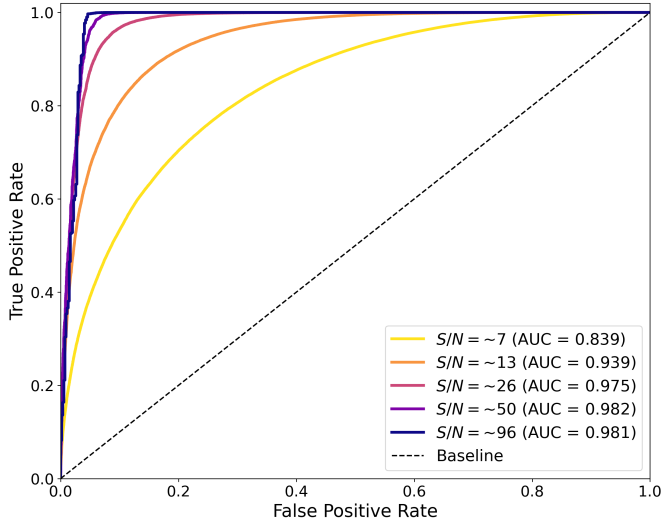


Fig. 8. ROC curve of ASID-C evaluated across five distinct S/N bins, ranging from low to high signal-to-noise ratios. The plot reveals a general trend of increasing ROC-AUC values with higher S/N, highlighting the model’s robustness to varying noise conditions. Notably, even at the lowest S/N bin, ASID-C demonstrates resilience with a ROC-AUC value of 0.84, underscoring its applicability in challenging observational scenarios.

As illustrated in Fig. 8, the ROC-AUC values improve with increasing S/N, although the rate of improvement varies between bins. While the classifier exhibits excellent performance in higher S/N bins, it also maintains a commendable ROC-AUC of 0.84 in the lowest S/N bin. This demonstrates ASID-C’s resilience to noise and its capability to perform adequately even when the signal is weak.

Having examined the ROC curves, we now turn our attention to the PR curves. It is worth noting that in the context of astronomy, the metrics used in Precision-Recall curves are often referred to as purity and completeness, respectively. While mathematically identical, the choice of terminology often depends on the specific field and application.

As depicted in Fig. 9, the PR-AUC values for both stars and galaxies improve as the S/N increases. It is important to note that while the PR-AUC values for stars consistently outperform those for galaxies, this discrepancy is largely due to the inherent class imbalance in the dataset, where stars are more numerous than galaxies. This imbalance naturally affects the precision and recall metrics, making it more challenging to achieve high PR-AUC values for galaxies. Nonetheless, even in the lowest S/N bin, ASID-C maintains a respectable PR-AUC, underscoring its resilience to noise and its ability to classify both stars and galaxies effectively under less-than-ideal observational conditions.

This analysis confirms ASID-C’s robustness across different S/N conditions and its adaptability to class imbalances, making it a valuable tool for a range of astronomical applications, especially those requiring performance in low S/N settings.

4.1.2. Performance across MeerLICHT bands

ASID-C is engineered to be band-agnostic, a feature that amplifies its versatility across the diverse MeerLICHT bands. While the network is designed to operate effectively on single-band images, the reality of modern astronomy often involves multi-band observations. To gain insights into any band-specific

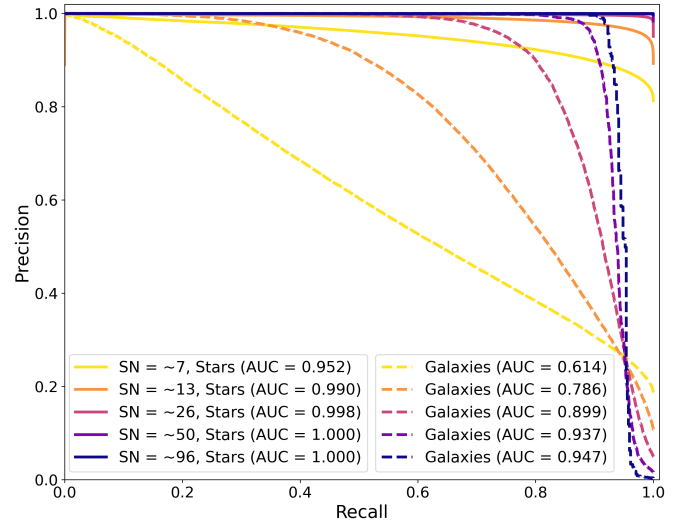


Fig. 9. Precision–recall curves of ASID-C evaluated across five distinct S/N bins. The PR-AUC values show an upward trend with increasing S/N, highlighting the model’s adaptability to different noise conditions. The PR-AUC values for stars are consistently higher than those for galaxies; this discrepancy is primarily attributable to the class imbalance in the dataset, where stars outnumber galaxies. Despite this challenge, ASID-C maintains commendable PR-AUC values even in the lowest S/N bin, emphasizing its resilience and effectiveness in classifying both celestial objects under varying observational conditions.

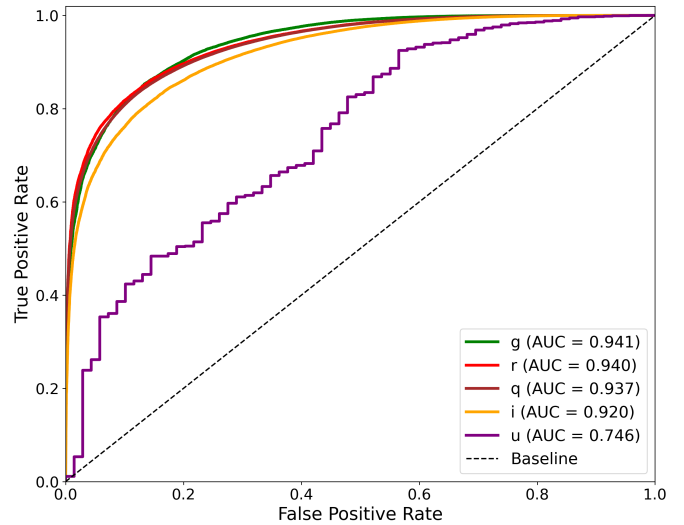


Fig. 10. ROC curve of ASID-C across five MeerLICHT bands. The figure shows consistently high ROC-AUC values across most bands. A noticeable performance dip in the *u* band is observed, likely due to the intrinsic color properties of galaxies. In this band, galaxies often lack features that distinguish them from stars, making classification more challenging.

nuances, we evaluated the network’s performance on the test set, where the band of the original full-field image for each source cutout is known.

Figure 10 shows the ROC curve of ASID-C for five MeerLICHT bands. The network maintains high accuracy scores for classifying stars across most bands, but a notable exception arises in the *u* band. The *u* band, sensitive to ultraviolet wavelengths, exhibits lower accuracy, likely attributable to the intrinsic color properties of galaxies. Galaxies are generally redder and emit less in the ultraviolet spectrum, making them appear

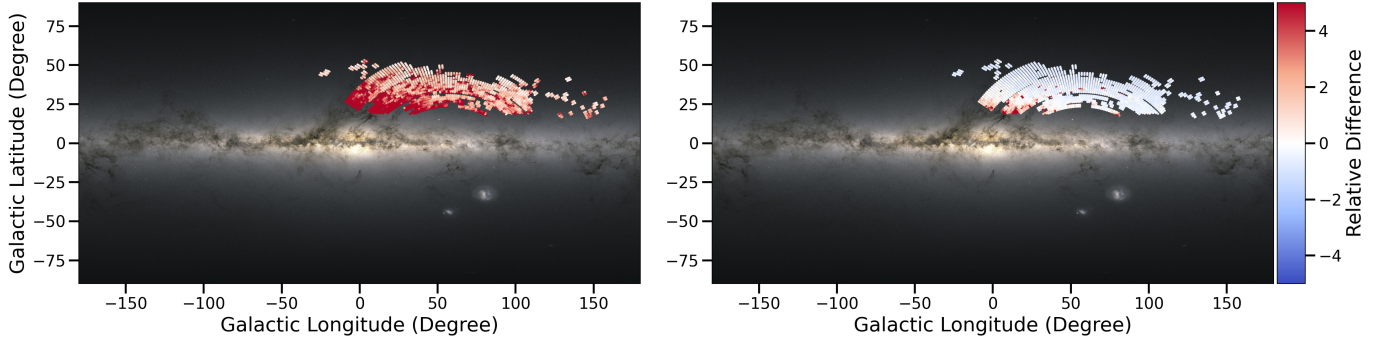


Fig. 11. Comparison of SourceExtractor (left) and ASID-C (right) performance across the Galactic Plane. The tiles are colored based on the relative difference between the actual and estimated numbers of galaxies in that region. The left panel highlights the tendency of SourceExtractor to overestimate galaxies in regions with high stellar density, indicating a bias in its classification approach. The right panel showcases the ability of ASID-C to classify galaxies accurately, even in regions with high stellar density and complex structures, thereby enabling reliable identification of celestial objects in such challenging environments.

more like stars and thus contributing to the classification challenges in this band (Strateva et al. 2001; Driver et al. 2006). Further complicating this issue, galaxies with a red peak, which are generally of an early morphological type, have been found to have higher surface brightness than those in the blue bands. This characteristic could also contribute to the difficulties in classifying galaxies in the u band (Wyder et al. 2007).

To tackle two intertwined challenges – namely, the combination of predictions from multiple observations of the same source and the variability in network performance across different bands – we introduce an ensemble methodology. This approach integrates observations of the same source across various bands into a unified, more reliable classification. In this approach, each band would be assigned a weight based on its performance metrics. Specifically, bands that yield more accurate classifications would be given higher weights, thereby influencing the final prediction more substantially and mitigating the impact of less reliable bands like the u band. The ensemble approach can be mathematically represented as:

$$\text{Weighted Prediction} = \frac{\sum_{i=1}^n w_i p_i}{\sum_{i=1}^n w_i}, \quad (5)$$

where w_i would denote the weight assigned to the i th band, and p_i would represent the prediction for the i th observation.

In summary, the ensemble methodology serves as a robust extension of ASID-C, optimizing its performance across the MeerLICHT bands and offering a balanced, reliable classification scheme that is well-suited for the complexities of modern astronomical observations.

4.1.3. ASID-C performance in high stellar density regions

In this section, we showcase the applicability of ASID-C on images oriented toward the Galactic Plane, a region teeming with stars and with a typical number of galaxies that are, however, difficult to differentiate. These areas, distinguished by a markedly high stellar density and complex background, traditionally present considerable challenges for classification techniques. Factors such as significant interstellar extinction, confusion with Galactic structures, and high source overlap further complicate the task. Despite the relative scarcity of galaxies compared to stars, their accurate identification and classification

are crucial for various astronomical studies, particularly for identifying transients’ host galaxies and understanding the structure of our own Galaxy.

For the analysis presented in this section, we applied a threshold to the predictions, considering sources with a predicted value less than 0.5 as galaxies. Although not optimal, this threshold was chosen based on the distribution of predictions and the known characteristics of the dataset. Furthermore, we visually compare ASID-C’s results with those of SourceExtractor’s stellerity parameter, but a more in-depth comparison between the two tools will be presented in a later section.

Figure 11 illustrates the performance of both SourceExtractor (left panel) and ASID-C (right panel) across the Galactic Plane. The tiles are colored based on the relative difference between the actual and estimated numbers of galaxies in that region. The left panel of Fig. 11 shows the performance of SourceExtractor, which tends to overestimate galaxies in regions with high stellar density, indicating a bias in its classification approach. In contrast, the right panel of Fig. 11 showcases the performance of ASID-C in the same regions. ASID-C accurately classifies galaxies almost independently of the number of sources in the region. Even amidst the high density of stars and complex structures close to the Galactic Plane, ASID-C enables reliable identification of galaxies.

The robust performance of ASID-C in high stellar density regions demonstrates its versatility and potential for broad applications in astronomy. Its ability to accurately classify galaxies amidst complex structures and dense star populations can significantly enhance our understanding of celestial phenomena.

4.1.4. Evaluation of prediction timings

The computational efficiency of a model is a critical factor, especially when dealing with large astronomical datasets. To evaluate the computational efficiency of ASID-C, we measured the time taken to predict the class of different sizes of datasets on a GPU-accelerated system, which significantly benefits deep learning models like ASID-C. The datasets used in this analysis ranged from 5000 to 100 000 images. We performed the prediction fifty times for each dataset size and calculated the average time taken. This approach mitigates potential variability in the timing results due to factors such as system load or GPU thermal throttling.

The results, shown in Fig. 12, demonstrate that the time ASID-C took to predict the class of the images scales linearly

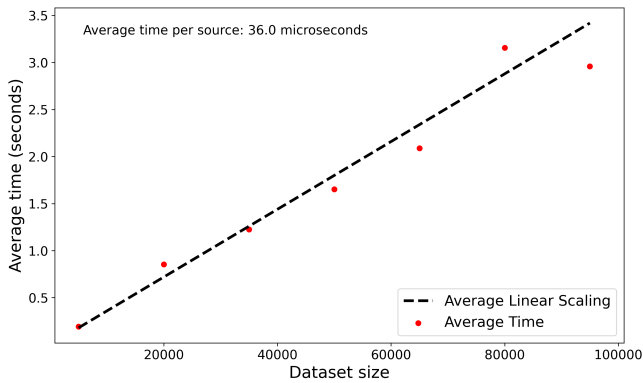


Fig. 12. Average time taken by ASID-C to predict on different sizes of datasets. The red dots represent the average time taken per dataset size, and the black dashed line represents the expected time based on linear scaling.

with the dataset size. This linear scaling is a desirable property, as it means that the time required to process a large dataset can be accurately estimated based on the time taken to process a smaller dataset sample. ASID-C demonstrates excellent computational efficiency, with linear scaling of prediction time with dataset size and a time per source of approximately $36 \mu\text{s}$. This efficiency, combined with its predictions' high accuracy and reliability, makes ASID-C a powerful tool for star-galaxy classification in large astronomical surveys.

4.2. Direct comparison with SourceExtractor

SourceExtractor (Bertin & Arnouts 1996), a widely used software in the field of astronomy for source extraction, feature estimation, and classification, serves as a benchmark in our study. Despite its extensive use in the astronomical community and its integration into the MeerLICHT official pipeline, SourceExtractor exhibits certain limitations in the context of star-galaxy classification.

Figure 13 illustrates the performance of SourceExtractor for star-galaxy classification. The left panel shows the distribution of SourceExtractor's predictions, which displays a bias toward a large number of objects with a probability value of ≈ 0.5 , close to the binary classifier threshold. This bias can increase misclassifications, particularly for galaxies, which are often assigned probabilities close to the 0.5 threshold. The right panel of Fig. 13 depicts the performance of SourceExtractor as a function of S/N. It becomes evident that the tool struggles with sources with a lower S/N, leading to decreased classification performance for these sources.

In contrast, our method, ASID-C, generates a more refined and well-calibrated prediction probability compared to SourceExtractor, as depicted in Fig. 14. The left panel shows the distribution of ASID-C's star-galaxy predictions, which are more evenly distributed and well-calibrated. The right panel of Fig. 14 demonstrates superior performance of ASID-C, particularly for sources with low to medium S/N. ASID-C is more robust and effective in classifying sources across a wider range of S/N values, thereby outperforming SourceExtractor.

The comparative analyses presented in this section underscore the effectiveness of ASID-C in star-galaxy classification, particularly in challenging scenarios involving sources with lower S/N. To further substantiate this, we evaluate the performance of the classifiers with the previously introduced Area

Under the Precision-Recall Curve (AUPRC) and the Brier Score (Brier 1950), both of which are computed as functions of source S/N. The latter is a measure of the reliability of a classifier's probabilistic predictions. It calculates the mean squared difference between the predicted probabilities for each class and the actual outcomes. A lower Brier Score is indicative of more accurate probabilistic forecasts, with a score of zero signifying a flawless classifier.

Table 1 provides a comprehensive view of the performance metrics across different S/N values for ASID-C, SourceExtractor (SE), and a baseline model (Base), where the labels are randomly assigned with a 10:1 proportion. The best performance for each metric and S/N value is highlighted in bold font.

From the table, it is evident that ASID-C consistently outperforms both SE and Base across all S/N values and metrics. ASID-C particularly excels in low S/N regions, where classifying stars and galaxies becomes more challenging. This superior performance is reflected in the higher AUPRC values and lower Brier scores for ASID-C compared to SE and Base. The significantly lower Brier scores for ASID-C suggest that it generates more reliable and well-calibrated probability predictions, especially for galaxies near the detectability threshold.

In the context of our study, the high AUPRC values observed for stars across all three methods – ASID-C, SE, and Base – raise an important point of consideration. These elevated values can be attributed to the imbalance in the dataset, which is heavily skewed toward stars. Such an imbalance can artificially inflate the AUPRC values, as the metric is sensitive to the number of true positives, which are abundant for stars in this dataset. Therefore, while these high AUPRC values may suggest excellent performance, they should be interpreted with caution.

Given the significant classification challenges posed by low S/N regions, the ability of ASID-C to handle them effectively underscores not only its robustness but also its potential as a highly effective tool for star-galaxy classification in future astronomical studies.

4.3. Comparison with retrained SourceExtractor

While SourceExtractor is a versatile tool capable of extracting a wide array of features across various telescopes, its general-purpose design may not yield optimal results for specialized tasks such as star-galaxy classification. To ensure a fairer comparison with our ASID-C model, we employed an XGBoost classifier (Chen & Guestrin 2016). This gradient-boosting decision tree is particularly proficient in handling tabular data and was trained on a feature set extracted by SourceExtractor. This feature set comprises magnitudes calculated at three different aperture sizes, an elongation parameter, which quantifies the degree to which a source is elongated, and (x, y) , the same location information used in ASID-C. For the XGBoost model training, we used a binary logistic objective and ran a total of 200 boosting rounds. It is worth noting that the training set for this XGBoost model is identical to that used for ASID-C, with the key distinction being the data format; the XGBoost model uses only tabular data. This approach allows us to directly assess the efficacy of a two-step process – initial feature extraction followed by machine learning on tabular data – against ASID-C's method of direct image-based classification.

One key attribute of tree-based machine learning methods like XGBoost is their ability to quantify feature importance. In our analysis, we found that the classifier's performance is predominantly influenced by two features: the smallest aperture and

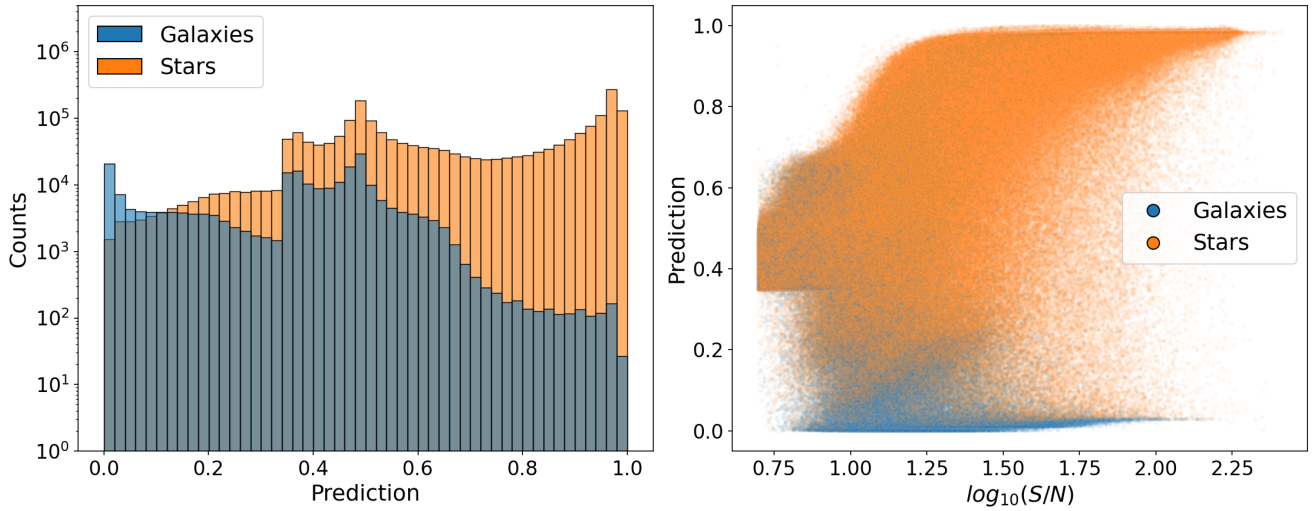


Fig. 13. Star-galaxy prediction for SourceExtractor. Left: the histogram of the predictions, colored by true class, shows a bias toward values of ≈ 0.5 , indicating a lack of calibration. Right: star-galaxy classification performance of SourceExtractor as a function of source S/N. The figure shows a decrease in classification performance for sources with lower S/N, indicating that SourceExtractor struggles to classify these sources accurately.

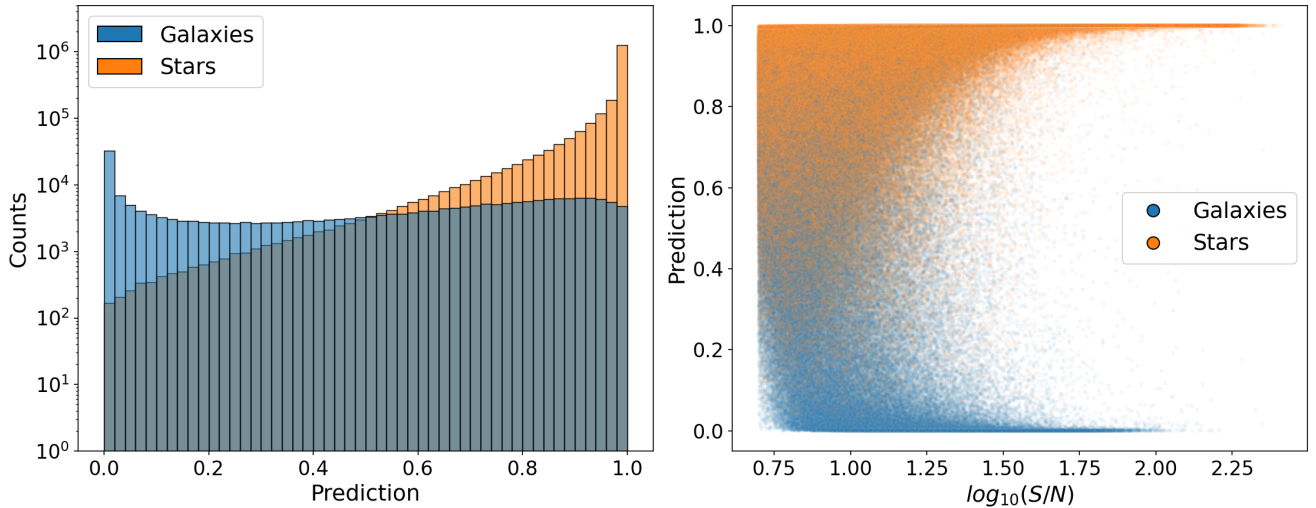


Fig. 14. Star-galaxy prediction for ASID-C. Left: the histogram of the predictions, colored by true class, shows a well-calibrated set of predictions. Right: star-galaxy classification performance of ASID-C as a function of source S/N. The figure shows that ASID-C maintains a high level of performance across a wider range of S/N values, thereby outperforming SourceExtractor, particularly for sources with low to medium S/N.

the elongation. These two parameters alone play a significant role in the decision-making process. Further investigation into SourceExtractor’s methodology for star-galaxy classification led us to an additional parameter known as “seeing,” which quantifies the atmospheric blurring of astronomical objects within the entire full-field image. This parameter can be estimated either through a PSF model of the image or manually provided. Currently, we do not have a machine learning-based method for rapid estimation of the seeing parameter, but we plan to explore this avenue in future work, along with a more comprehensive study on the influence of global parameters such as seeing on local source predictions.

Incorporating the external seeing parameter output of the official MeerLICHT pipeline, BlackBOX, into our XGBoost training, we observed a substantial improvement in performance, highlighting the importance of a global parameter that characterizes the entire image. Both XGBoost and ASID-C were retrained

with this additional parameter, and their performance, along with that of the other methods, is presented in Fig. 15.

The analysis reveals that an XGBoost model trained on MeerLICHT-specific features outperforms the original SourceExtractor and is competitive with ASID-C in AUC. The incorporation of the seeing parameter into both the XGBoost and ASID-C models resulted in marked improvements, with the latter achieving the highest ROC-AUC overall. Furthermore, to quantify the impact of the seeing parameter in ASID-C, we calculated relative false positive rate (FPR) differences. These are expressed as percentages and are evaluated at the optimal threshold for each method, using ASID-C with seeing as the reference point. Specifically, ASID-C with seeing shows a substantial improvement of 46.67% FPR over SourceExtractor, 25.20% over SE XGBoost, and 15.54% over the version of ASID-C without seeing. When compared to SE XGBoost integrated with seeing, ASID-C with seeing registers a 9.46% FPR improvement.

Table 1. Performance metrics by S/N.

S/N	~7	~13	~26	~50	~96
AUPRC ASID (s)	0.952	0.990	0.998	1.000	1.000
AUPRC SE (s)	0.841	0.975	0.998	1.000	1.000
AUPRC Base (s)	0.813	0.893	0.950	0.984	0.997
AUPRC ASID (g)	0.614	0.786	0.899	0.937	0.947
AUPRC SE (g)	0.271	0.575	0.782	0.853	0.782
AUPRC Base (g)	0.187	0.107	0.050	0.016	0.003
Brier ASID	0.108	0.044	0.012	0.002	0.000
Brier SE	0.279	0.188	0.076	0.034	0.015
Brier Base	0.250	0.185	0.140	0.113	0.101

Notes. The AUPRC, due to its formulation, is calculated separately for both stars (s) and galaxies (g) as the positive class. The table shows that ASID-C consistently outperforms both SE and the baseline model across most S/N values and metrics.

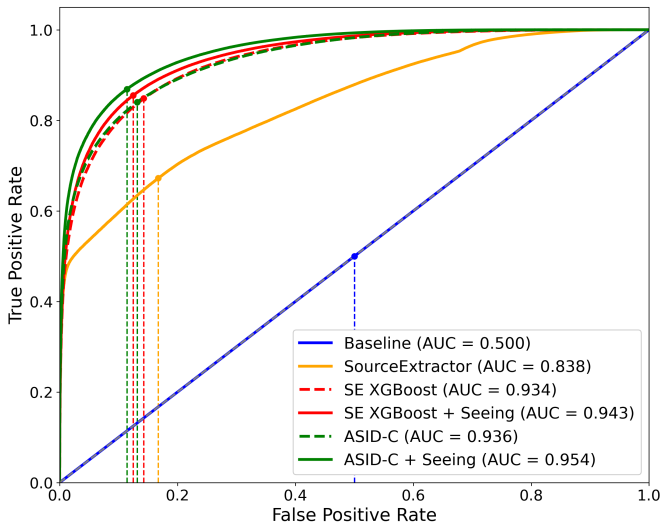


Fig. 15. ROC curves and area under the curve (AUC) values for multiple methods, including two variants of ASID-C, two XGBoost models, SourceExtractor, and a Baseline model. Optimal threshold points are indicated on each curve, and vertical dashed lines show the false positive rate at these optimal points. The ASID-C model, when trained with the additional seeing parameter, exhibits the highest overall AUC.

While XGBoost proves effective, especially when tailored to specific datasets, it is important to note its computational limitations. The two-step approach – feature extraction followed by classification – does introduce a margin of error, although this appears to be minimal, mitigating concerns about the classification process’s integrity. On the other hand, ASID-C operates directly on image data, thereby eliminating the risk of error propagation associated with a two-step process. This makes ASID-C more robust, particularly for large or diverse datasets or when computational resources are constrained.

In summary, while XGBoost presents a viable alternative, it comes with its own set of challenges, which are inherently mitigated by ASID-C’s direct approach to classification.

5. Conclusion and discussion

This study introduced ASID-C, a dual-branch CNN specifically designed for star-galaxy classification. ASID-C incorporates

both image data and spatial information, specifically the location of the source within the image, to provide a robust and effective solution for distinguishing between stars and galaxies in astronomical images from the MeerLICHT telescope.

A standout feature of ASID-C’s performance is the near-perfect calibration of its predictions, particularly for models trained with cross-entropy loss. While Platt scaling, especially when enhanced with a logit transformation, can improve the calibration of model predictions, we found that the cross-entropy models already yield almost perfectly calibrated predictions. This ensures that the predicted probabilities closely mirror the actual class proportions, a critical attribute for many applications. A well-calibrated model can significantly improve decision-making processes, particularly when an accurate estimate of class membership probability is paramount.

To rigorously assess the capabilities and limitations of ASID-C, we employed the UMAP technique to explore its latent space. This analysis shed light on the underlying data structures that aid in distinguishing between stars and galaxies, offering a deeper understanding of the model’s classification capabilities. We then conducted a targeted evaluation across multiple dimensions. In terms of the S/Ns, ASID-C proved resilient, maintaining commendable accuracy even under low S/N conditions. It also excelled in high stellar density regions, which are traditionally challenging for classification. However, the model’s performance was not uniformly strong across all MeerLICHT bands; it faltered in the u band. To address this and, at the same time, provide a way to combine predictions on multiband observations of the same source, we proposed an ensemble methodology that enhances the model’s multiband reliability. Lastly, we verified the model’s computational scalability, affirming its suitability for large-scale astronomical applications.

Our evaluation of ASID-C’s performance demonstrates its superiority over the widely used SourceExtractor tool, particularly regarding the reliability and calibration of its probability predictions. Notably, ASID-C shows exceptional advantage at low S/N sources, a challenging area where many existing tools struggle. We also compared ASID-C’s performance with a gradient-boosted decision tree model, XGBoost, trained on features extracted by SourceExtractor. While XGBoost easily outperforms SourceExtractor results, being trained specifically on the dataset at hand, our results indicate that ASID-C outperforms it in terms of both performance, but particularly in efficiency. Importantly, the inclusion of the global seeing parameter in ASID-C further amplifies its performance. This addition allows ASID-C to achieve significant improvements in the FPR compared to other methods, solidifying its position as a highly reliable and efficient tool for astronomical classification tasks.

One of the most promising applications of ASID-C lies in real-time star-galaxy classification. The rapid processing speed of CNNs, combined with the robustness of ASID-C, makes it an ideal tool for time-sensitive astronomical observations. For instance, transient events, such as supernovae or gamma-ray bursts, require immediate follow-up observations to capture their rapidly changing properties. By providing an accurate and immediate classification of sources, ASID-C can help astronomers quickly identify the nature of the transient’s host galaxy or rule out false positives, thereby streamlining the transient detection and follow-up process.

Beyond binary star-galaxy classification, ASID-C could potentially be extended to classify galaxies based on their morphology. Galaxy morphology, such as spiral, elliptical, or irregular, provides valuable insights into the formation and evolution of galaxies. By training ASID-C on a dataset labeled with

galaxy morphologies, we could leverage its robust feature extraction capabilities to identify the morphological characteristics of galaxies. This would expand the capabilities of ASID-C and contribute to our understanding of galaxy evolution.

While our current work focuses on optical data, ASID-C has the potential to handle data across multiple wavelengths, such as infrared, ultraviolet, or radio. This multiwavelength approach could enhance classification accuracy by providing a more comprehensive view of the sources. For instance, certain types of galaxies or stars may exhibit unique characteristics at specific wavelengths that are not apparent in optical data. By incorporating multiwavelength data into ASID-C, we can leverage these unique characteristics to improve classification performance.

In conclusion, ASID-C represents a significant advancement in the field of star-galaxy classification. Its strong performance and adaptability make it a valuable tool for the astronomical community, opening up new opportunities for research and study. By enhancing our ability to classify celestial objects accurately, ASID-C contributes to our understanding of the universe and accelerates the pace of discovery and exploration in astronomy. The insights gained from this study provide valuable guidance for future research and applications in star-galaxy classification.

Acknowledgements. F.S. and G.N. acknowledge support from the Dutch Science Foundation NWO. S.B. and G.Z. acknowledge the financial support from the Slovenian Research Agency (grants P1-0031, I0-0033 and J1-1700). R.R. acknowledges support from the Ministerio de Ciencia e Innovación (PID2020-113644GB-I00). G.P. acknowledges support by ICSC – Centro Nazionale di Ricerca in High Performance Computing, Big Data and Quantum Computing, funded by European Union – NextGenerationEU. PJG is partly supported by NRF SARChI Grant 111692. The MeerLICHT telescope is a collaboration between Radboud University, the University of Cape Town, the South African Astronomical Observatory, the University of Oxford, the University of Manchester and the University of Amsterdam, and supported by the NWO and NRF Funding agencies. The Legacy Surveys consist of three individual and complementary projects: the Dark Energy Camera Legacy Survey (DECaLS; Proposal ID #2014B-0404; PIs: David Schlegel and Arjun Dey), the Beijing-Arizona Sky Survey (BASS; NAOJ Prop. ID #2015A-0801; PIs: Zhou Xu and Xiaohui Fan), and the Mayall z -band Legacy Survey (MzLS; Prop. ID #2016A-0453; PI: Arjun Dey). DECaLS, BASS and MzLS together include data obtained, respectively, at the Blanco telescope, Cerro Tololo Inter-American Observatory, NSF’s NOIRLab; the Bok telescope, Steward Observatory, University of Arizona; and the Mayall telescope, Kitt Peak National Observatory, NOIRLab. Pipeline processing and analyses of the data were supported by NOIRLab and the Lawrence Berkeley National Laboratory (LBNL). The Legacy Surveys project is honored to be permitted to conduct astronomical research on Iolkam Du’ag (Kitt Peak), a mountain with particular significance to the Tohono O’odham Nation. NOIRLab is operated by the Association of Universities for Research in Astronomy (AURA) under a cooperative agreement with the National Science Foundation. LBNL is managed by the Regents of the University of California under contract to the U.S. Department of Energy. This project used data obtained with the Dark Energy Camera (DECam), which was constructed by the Dark Energy Survey (DES) collaboration. Funding for the DES Projects has been provided by the U.S. Department of Energy, the U.S. National Science Foundation, the Ministry of Science and Education of Spain, the Science and Technology Facilities Council of the United Kingdom, the Higher Education Funding Council for England, the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, the Kavli Institute of Cosmological Physics at the University of Chicago, Center for Cosmology and Astro-Particle Physics at the Ohio State University, the Mitchell Institute for Fundamental Physics and Astronomy at Texas A&M University, Financiadora de Estudos e Projetos, Fundação Carlos Chagas Filho de Amparo, Financiadora de Estudos e Projetos, Fundação Carlos Chagas Filho de Amparo a Pesquisa do Estado do Rio de Janeiro, Conselho Nacional de Desenvolvimento Científico e Tecnológico and the Ministerio da Ciencia, Tecnologia e Inovacao, the Deutsche Forschungsgemeinschaft and the Collaborating Institutions in the Dark Energy Survey. The Collaborating Institutions are Argonne National Laboratory, the University of California at Santa Cruz, the University of Cambridge, Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas-Madrid, the University of Chicago, University College London, the DES-Brazil Consortium, the University of Edinburgh, the Eidgenössische Technische Hochschule

(ETH) Zurich, Fermi National Accelerator Laboratory, the University of Illinois at Urbana-Champaign, the Institut de Ciències de l’Espai (IEEC/CSIC), the Institut de Física d’Altes Energies, Lawrence Berkeley National Laboratory, the Ludwig Maximilians Universität München and the associated Excellence Cluster Universe, the University of Michigan, NSF’s NOIRLab, the University of Nottingham, the Ohio State University, the University of Pennsylvania, the University of Portsmouth, SLAC National Accelerator Laboratory, Stanford University, the University of Sussex, and Texas A&M University. BASS is a key project of the Telescope Access Program (TAP), which has been funded by the National Astronomical Observatories of China, the Chinese Academy of Sciences (the Strategic Priority Research Program “The Emergence of Cosmological Structures” Grant # XDB09000000), and the Special Fund for Astronomy from the Ministry of Finance. The External Cooperation Program of Chinese Academy of Sciences (Grant # 114A11KYSB20160057) and Chinese National Natural Science Foundation (Grant # 12120101003, # 11433005) also support the BASS. The Legacy Survey team makes use of data products from the Near-Earth Object Wide-field Infrared Survey Explorer (NEOWISE), which is a project of the Jet Propulsion Laboratory/California Institute of Technology. NEOWISE is funded by the National Aeronautics and Space Administration. The Legacy Surveys imaging of the DESI footprint is supported by the Director, Office of Science, Office of High Energy Physics of the U.S. Department of Energy under Contract No. DE-AC02-05CH1123, by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility under the same contract; and by the U.S. National Science Foundation, Division of Astronomical Sciences under Contract No. AST-0950945 to NOAO.

References

- Abadi, M., Agarwal, A., Barham, P., et al. 2016, ArXiv e-prints [arXiv:1603.04467]
- Ball, N. M., Brunner, R. J., Myers, A. D., & Tchong, D. 2006, *ApJ*, **650**, 497
- Bartoszek, K. 2016, *J. Theor. Biol.*, **407**, 371
- Bertin, E., & Arnouts, S. 1996, *A&AS*, **117**, 393
- Bilicki, M., Dvornik, A., Hoekstra, H., et al. 2021, *A&A*, **653**, A82
- Bloemen, S., Groot, P., Woudt, P., et al. 2016, *Proc. SPIE*, **9906**, 990664
- Blum, R. D., Burleigh, K., Dey, A., et al. 2016, *Am. Astron. Soc. Meeting Abstracts*, **228**, 317.01
- Breiman, L. 2017, *Classification and Regression Trees* (Routledge)
- Brier, G. W. 1950, *Monthly Weather Rev.*, **78**, 1
- Cabayol, L., Sevilla-Noarbe, I., Fernández, E., et al. 2018, *MNRAS*, **483**, 529
- Chen, T., & Guestrin, C. 2016, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*
- Cox, D. R. 1958, *J. Roy. Stat. Soc. Ser. B (Methodological)*, **20**, 215
- Cui, Y., Jia, M., Lin, T.-Y., Song, Y., & Belongie, S. 2019, ArXiv e-prints [arXiv:1901.05555]
- Davis, J., & Goadrich, M. 2006, in *Proceedings of the 23rd International Conference on Machine Learning*, 233
- Degroot, M. H., & Fienberg, S. E. 1983, *The Statistician*, **32**, 12
- Dey, A., Schlegel, D. J., Lang, D., et al. 2019, *AJ*, **157**, 168
- Driver, S. P., Allen, P. D., Graham, A. W., et al. 2006, *MNRAS*, **368**, 414
- Fadely, R., Hogg, D. W., & Willman, B. 2012, *ApJ*, **760**, 15
- Filho, T. S., Song, H., Perello-Nieto, M., et al. 2023, *Mach. Learn.*, **112**, 3211
- Fukushima, K., & Miyake, S. 1982, *Pattern Recogn.*, **15**, 455
- Groot, P. J., Bloemen, S., Vreeswijk, P. M., et al. 2022, *SPIE Conf. Ser.*, **12182**, 121821V
- Henrion, M., Mortlock, D. J., Hand, D. J., & Gandy, A. 2011, *MNRAS*, **412**, 2286
- Heymans, C., Van Waerbeke, L., Miller, L., et al. 2012, *MNRAS*, **427**, 146
- Hosenie, Z., Bloemen, S., Groot, P. J., et al. 2021, *Exp. Astron.*, **51**, 319
- Jonas, J., & MeerKAT Team. 2016, in *MeerKAT Science: On the Pathway to the SKA*, 1
- Kim, E. J., & Brunner, R. J. 2016, *MNRAS*, **464**, 4463
- Kingma, D. P., & Ba, J. 2015, in *Adam: A Method for Stochastic Optimization. Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*
- Kull, M., Filho, T. M. S., & Flach, P. 2017, *Electron. J. Stat.*, **11**, 5052
- Kull, M., Perello-Nieto, M., Kängsepp, M., et al. 2019, ArXiv e-prints [arXiv:1910.12656]
- LeCun, Y., Bengio, Y., et al. 1995, *The Handbook of Brain Theory and Neural Networks*, 3361, 1995
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. 2017, in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2999
- López-Sanjuan, C., Vázquez Ramió, H., Varela, J., et al. 2019, *A&A*, **622**, A177
- Maaten, L., & Hinton, G. 2008, *J. Mach. Learn. Res.*, **9**, 2579

- McInnes, L., Healy, J., Saul, N., & Großberger, L. 2018, *J. Open Source Softw.*, **3**, 861
- Niculescu-Mizil, A., & Caruana, R. 2005, *Proceedings of the 22nd International Conference on Machine Learning*
- Odehahn, S. C., Stockwell, E. B., Pennington, R. L., Humphreys, R. M., & Zumach, W. A. 1992, *AJ*, **103**, 318
- Odehahn, S. C., de Carvalho, R. R., Gal, R. R., et al. 2004, *AJ*, **128**, 3092
- Paturel, G., Bottinelli, L., & Gouguenheim, L. 1995, *Astrophys. Lett. Commun.*, **31**, 13
- Platt, J. 1999, in *Advances in Large Margin Classifiers* (MIT Press)
- Schlegel, D., Dey, A., Herrera, D., et al. 2021, *Am. Astron. Soc. Meeting Abstracts*, **53**, 235.03
- Sevilla-Noarbe, I., Hoyle, B., Marchã, M. J., et al. 2018, *MNRAS*, **481**, 5451
- Silva, D. R., Blum, R. D., Allen, L., et al. 2016, *Am. Astron. Soc. Meeting Abstracts*, **228**, 317.02
- Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, *AJ*, **131**, 1163
- Stoppa, F., Vreeswijk, P., Bloemen, S., et al. 2022, *A&A*, **662**, A109
- Stoppa, F., Ruiz de Austri, R., Vreeswijk, P., et al. 2023, *A&A*, **680**, A108
- Strateva, I., Ivezić, Ž., Knapp, G. R., et al. 2001, *AJ*, **122**, 1861
- Swets, J. A. 1996, in *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers* (Psychology Press)
- Vasconcellos, E. C., de Carvalho, R. R., Gal, R. R., et al. 2011, *AJ*, **141**, 189
- Weir, N., Fayyad, U. M., & Djorgovski, S. 1995, *AJ*, **109**, 2401
- Wenger, M., Ochsenbein, F., Egret, D., et al. 2000, *A&AS*, **143**, 9
- Wu, Y., Liu, L., Bae, J., et al. 2019, ArXiv e-prints [arXiv:1908.06477]
- Wyder, T. K., Martin, D. C., Schiminovich, D., et al. 2007, *ApJS*, **173**, 293
- Zou, H., Zhou, X., Fan, X., et al. 2019, *ApJS*, **245**, 4

Appendix A: Incorporating spatial information

The dual-branch structure of our model, ASID-C, is designed to incorporate a variety of features. A key component we have integrated is spatial information, which plays a crucial role in astronomical image analysis. The appearance of celestial objects can vary depending on their location in the image, complicating the task of distinguishing between stars and galaxies. To address this, ASID-C considers the location of each celestial object within the image. This approach allows the model to account for variations in the PSF and other spatially dependent effects, enhancing its accuracy and reliability even under challenging imaging conditions.

To validate the effectiveness of this approach, we compared the performance of the model with and without the spatial information branch. The results showed a significant improvement when spatial information was included, underscoring its importance in star-galaxy classification. Figures A.1, A.2, and A.3 illustrate the comparison of loss, Area Under the Receiver Operating Characteristic Curve (AUC), and Area Under the Precision-Recall Curve (AUPRC) respectively, between the models with and without the spatial information branch.

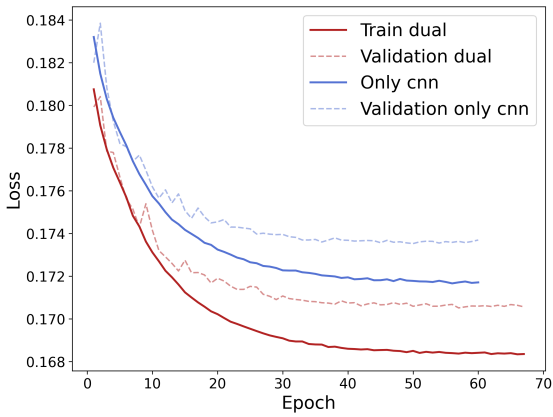


Fig. A.1. Comparison of loss between models with and without the spatial information branch. The model incorporating spatial information exhibits a lower loss, indicating improved performance.

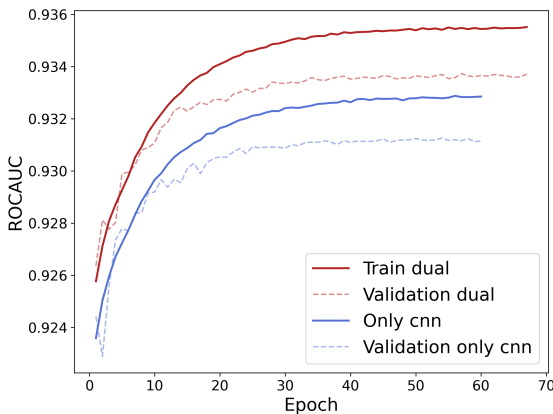


Fig. A.2. Comparison of AUC between models with and without the spatial information branch. The model incorporating spatial information achieves a higher AUC, demonstrating its superior classification performance.

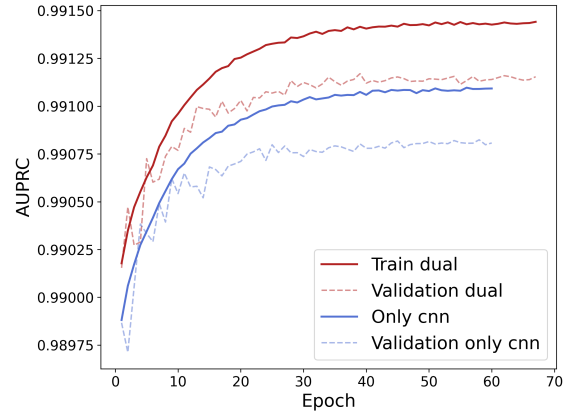


Fig. A.3. Comparison of AUPRC between models with and without the spatial information branch. The model incorporating spatial information shows a higher AUPRC, indicating a better balance between precision and recall.

data from multiband images, or metadata from external catalogs. This adaptability ensures our model can evolve to meet the changing needs and challenges of star-galaxy classification.

Moreover, the dual-branch structure of ASID-C allows for the inclusion of other features, such as PSF information, color