Automatic Detection of Novelty Galaxies in Digital Sky Survey Data

Venkat Margapuri*, Basant Thapa*, and Lior Shamir*
Kansas State University, Manhattan, KS, USA

Abstract

Galaxy images of the order of multi-PB are collected as part of modern digital sky surveys using robotic telescopes. While there is a plethora of imaging data available, the majority of the images that are captured resemble galaxies that are "regular", i.e., galaxy types that are already known and probed. However, "novelty" galaxy types, i.e., little-known galaxy types are encountered on occasion. The astronomy community shows paramount interest in the novelty galaxy types since they contain the potential for scientific discovery. However, since these galaxies are rare, the identification of such novelty galaxies is not trivial and requires automation techniques. Since these novelty galaxies are by definition, not known, supervised machine learning models cannot be trained to detect them. In this paper, an unsupervised machine learning method for automatic detection of novelty galaxies in large databases is proposed. The method uses a large set of image features weighted by their entropy. To handle the impact of self-similar novelty galaxies, the most similar galaxies are ranked-ordered. In addition, Bag of Visual Words (BOVW) is assimilated to the problem of detecting novelty galaxies. Each image in the dataset is represented as a set of features made up of key-points and descriptors. A histogram of the features is constructed and is leveraged to identify the neighbors of each of the images. Experimental results using data from the Panoramic Survey Telescope and Rapid Response System (Pan-STARRS) show that the performance of the methods in detecting novelty galaxies is superior to other shallow learning methods such as one-class SVM, Local Outlier Factor, and K-Means, and also newer deep learning-based methods such as auto-encoders. The dataset used to evaluate the method is publicly available and can be used as a benchmark to test future algorithms for automatic
detection of peculiar galaxies.

Key Words: Entropy based algorithms, bag of visual words, Pan-STARRS, novelty detection, feature extraction.

1 Introduction

In the past two decades, Earth-based astronomical instruments have largely shifted from manually controlled telescopes to robotic telescopes that survey and image the entire sky [3], making their data available to the astronomy community through virtual observatories [6]. The astronomer community relies on the data from the observatories to aid in the furthering galactic scientific discovery. These powerful imaging instruments generate some of the world's largest databases, contain billions of astronomical objects, and lead to numerous scientific discoveries that were not possible in the pre-information era. Sloan Digital Sky Survey (SDSS) alone has produced data leading to more than $3 \cdot 10^4$ peer-reviewed papers, and it is very reasonable to assume that more discoveries of paramount scientific interest are hidden inside these databases. Any attempt to examine the abundance of information produced by the observatories is unrealistic and requires automation techniques to turn them into knowledge and scientific discoveries. One of the effective scientific tasks enabled by digital sky surveys is the identification of the databases. Most extra-galactic objects belong in the galaxy classification scheme, known as the "Hubble sequence" [13]. However, some galaxies do not fit any stage on the Hubble sequence and are considered "peculiar" galaxies [9]. Although these galaxies are rare, they are of high scientific interest as they carry important information about the past, present, or future universe. The Panoramic Survey Telescope and Rapid Response System (Pan-STARRS) is an array of two robotic telescopes synchronized to observe the same part of the sky simultaneously to increase the cost-effectiveness of its imaging power. Launched in 2008, Pan-STARRS used its wide 3^o field of view and 1.4 Gigapixel digital camera to image over $3.5 \cdot 10^9$ astronomical objects and generated the world's largest astronomical database of ~1.6PB.

In this paper the task of identifying novelty astronomical objects automatically is investigated. Deep-learning based auto-encoders technique is compared to statistical methods based on "shallow learning". The paper proposes two techniques for novelty detection - a detection algorithm that uses the concept of entropy of a set of pre-defined numerical image content descriptors and Bag of Visual Words technique that represents an image as a set features using Scale-Invariant Feature Transform (SIFT). The performance of the proposed techniques is compared against the performance of common "traditional" unsupervised machine learning algorithms such as One-Class Support Vector Machines (OCSVM), K-Means Clustering, Local Outlier Factor (LOF), and K-Nearest Neighbors algorithm which falls in the realm of supervised

* Department of Computer Science. Email: marven@ksuJ.edu, thapa@ksu.edu and lshamir@ksu.edu

learning. In addition, the deep learning technique of auto-encoders are applied and investigated.

2 Related Work

Relevant research in the area of study, while not abundant, is existent and studied to help pave a segue for the current work. The first attempt to identify peculiar galaxies on data from the Sloan Digital Sky Survey (SDSS), a sky survey with data analysis, faces challenges that largely overlap with the data analysis challenges of Pan-STARRS. It was done by using a large number of "citizen scientists" who observed the images manually over several years and determined whether the astronomical object is peculiar [4]. That initiative allowed the compilation of a large catalog of rare ring galaxies [18]. However, statistical analysis using ring galaxies detected automatically showed that many more ring galaxies were hidden inside [8]. Additionally, after several years of work involving over 10^5 volunteers, less than 10^6 objects were observed [23]. Applying the same method for the analysis of all objects imaged so far by Pan-STARRS will require over ~10^4 years to complete. The size of the data of digital sky surveys reinforces the use of automation.

An example of automatic outlier detection applied to datasets of astronomical objects is the application of outlier detection to SDSS galaxy data to identify galaxies with unusual spectroscopic profile [2]. The method is based on unsupervised Random Forest [24], and was applied on the spectroscopic data of the galaxies rather than their images.

Substantial research has been done for general outlier detection. Among numerous approaches, the concept of entropy of features was used to mine outliers in databases [21]. Among more recent approaches, deep neural networks were used for automatic detection of outliers in data, including image data [7, 15]. While deep artificial neural networks, and in particular deep convolutional neural networks, have shown excellent performance in supervised learning of image data, the use of auto-encoders [7, 15] allows using the power of deep neural networks also for unsupervised machine learning.

3 Data

In the absence of a benchmark with ground truth for novelty galaxy detection, a controlled benchmark dataset of galaxy images from the Pan-STARRS sky survey is compiled. Each image is a 120 x 120 image in the JPG image format. The benchmark includes three datasets, such that each dataset contains 200 celestial objects. The first contains spiral galaxies, the second contains lenticular galaxies, and the third contains stars. The reason for using stars is that data analysis pipelines of digital sky surveys such as Pan-STARRS often struggle to classify between stars and galaxies, and therefore more objects identified as galaxies are in fact stars. Therefore, a practical algorithm for novelty galaxy detection needs to handle the existence of stars identified incorrectly as galaxies. The datasets are used such that in each run 200 galaxies from one dataset are combined with 10 galaxies from another dataset to create a dataset in which the majority of the galaxies are "regular" galaxies, but a small number of galaxies which are different from the majority of the galaxies are also included. That allows to develop and test methods for identifying galaxies that are different from most other galaxies. For instance, in a late-type universe that contains only spiral galaxies, a lenticular galaxy would be considered a rare novelty galaxy. Similarly, in a universe of just stars, a lenticular galaxy is considered peculiar. Therefore, it can be reasonably assumed that an unsupervised machine learning algorithm that is not trained on spiral galaxies yet automatically detects a small number of spiral galaxies among a large number of lenticular galaxies, is an algorithm that will also be able to identify other novelty galaxies without training. Figure 1 shows examples of the celestial objects as imaged by Pan-STARRS.

Figure 1: Example image of star (left), lenticular galaxy (center) and spiral galaxy (right) imaged by Pan-STARRS

The dataset is freely available at PanSTARRSData and can be used as a benchmark dataset for developing future algorithms for automatic detection of novelty galaxies.

4 Method

4.1 Entropy Based Algorithm

According to shallow supervised learning of image data, each image in the dataset is first converted to a set of numerical image content descriptors that reflect its visual content through numerical values. The set of numerical image content descriptors used in this study is WND-CHARM [19], that was proven effective to machine analysis of galaxy images [14, 17, 20, 22]. In summary, the WND-CHARM library computes a comprehensive set of 2883 numerical image content descriptors that reflect numerous aspects of the visual content such as the shape, color, edges, textures (e.g., Gabor, Haralick, Tamura), fractals, polynomial decomposition of the image (e.g., Chebyshev polynomials, Zernike), and statistics of the distribution of the pixel values (e.g., Radon features, multi-scale histograms, first four moments). That feature set is described in detail in [16, 19, 26], and is applied successfully to the task of galaxy image analysis [11, 25].

The feature extraction process computes 2883 numerical image content descriptors for each galaxy image. That large set is sufficiently comprehensive to reflect numerous aspects of the galaxy morphology [14, 17, 20, 22]. However, it can also be assumed that many of these descriptors are not informative for unsupervised detection of novelty galaxies, and possibly add noise to the system. In order to select the most informative features and avoid noise to better detect novel objects in the

dataset, a process of feature selection is required. Since the learning is unsupervised, many "traditional" feature selection algorithms are not suitable. Therefore, in this study, the concept of Entropy is used as a technique to perform unsupervised feature selection on datasets with a large number of features. The entropy of a system S with N possible outcomes is computed as $-\sum_{i=1}^{N} p_i \bullet log(p_i)$, where p_i is the frequency of outcome i in S. To compute entropy on the numerical image content descriptors, the value of each numerical content descriptor is convolved into a histogram of N bins, and p_i is the frequency of the values in the histogram bin i, such that $i \in \{1..N\}$. The intuition behind this method of feature selection is that informative features tend to have their values distributed in some non-random clusters of values, while non-informative features have their values randomly distributed.

Identification of novelty galaxies is unique in the sense that due to the enormous size of the datasets of galaxy images, a single one-of-a-kind peculiar galaxy is unlikely to exist. For instance, the future Vera Rubin observatory is expected to collect ~10^{10} galaxies, and therefore even an extremely rare one-in-a-million object is expected to appear in the dataset about 10^4 times. Therefore, an effective novelty galaxy detection algorithm is required to be sensitive to the number of galaxies in the dataset, and assume that many of the novelty galaxies are self-similar to each other.

To handle the self-similarity of novelty galaxies, the intuition of the algorithm is that, given a set of galaxies, the farthest K^{th} neighbor amongst the K^{th} nearest neighbors of all the galaxies is a novelty galaxy. This allows the user of the algorithm to specify a minimum number of self-similar novelty galaxies. For example, consider a dataset of 100 galaxies with a K value of 10. The distance of each galaxy in the dataset is determined by its 10^{th} nearest neighbor. Therefore, if a galaxy has nine similar neighbors but is different from the remaining 90 galaxies, it will be assigned a high distance that reflects its dissimilarity from most of the galaxies. This simple mechanism might be inferior to other algorithms for the general case of novelty detection, but it is suitable for the detection of novelty galaxies as it provides the user with clear control over the number of self-similar novelty galaxies. This number changes with the type of galaxies considered, and therefore, the user is required to adjust the number based on the size of the dataset and the estimated frequency of different types of novelty galaxies.

The algorithm is described as follows:

1. Normalize the values in the dataset using Min-Max normalization.
2. Compute the entropy of each of the features of the dataset.
3. Choose a value between 0 and the greatest entropy of the features as the entropy threshold.
4. Apply the entropy threshold to the entropies of the features and set all entropies greater than the threshold to 0.
5. Pick a K, the order of the neighbor to be considered as the nearest neighbor. For instance, if the value of K is set to 5, the distance to the 5^{th} closest neighbor of each of the galaxies is used as the dissimilarity measure of that galaxy.
6. Compute the distance to the K^{th} neighbor of each of the galaxies using Minkowski distance i.e., weighted Euclidean Distance where the weights of the features are the entropy values obtained in Step 4.
7. Sort the galaxies by their distance to their K^{th} neighbor. Greater the distance, higher the likelihood that the galaxy is a novelty.

The algorithm depends on two parameters that control its performance:

1. **The order of the closest neighbor (K)**: If the value of K is lower than the number of novelty galaxies of a specific type, it is possible that the distance between a certain galaxy and its K^{th} neighbor is not larger than other non-novelty galaxies. Therefore, the user is required to select a value that is higher than the number of novelty galaxies of a certain type that are expected to exist in the dataset. The number depends on the size of the entire dataset and also not necessarily known to the user. In that case the user will need to attempt several K values and inspect the results to see if the detected novelty galaxies are indeed not "regular" galaxies.
2. **The value of the entropy threshold (Step 3 in the algorithm above)**: A high entropy threshold might lead to the rejection of features that carry information about the morphology of the galaxy. On the other hand, a low threshold might lead to the inclusion of noisy features.

The source code of the algorithm can be found at PanSTARRSNoveltyDetectionAlgorithm.

4.2 Bag of Visual Words

Bag of Visual Words (BOVW) [1, 10, 12] is a technique assimilated to image classification from the popular Bag of Words (BOW) technique used in information retrieval and natural language processing. The idea is to represent an image as a set of features. Each feature consists of keypoints and descriptors. Keypoints refer to the important defining points in an image that remain unaltered even upon the application of operations such as rotation, compression and expansion. Descriptors are the entities that describe the keypoints. The combination of keypoints and descriptors are used to construct vocabularies. Each image is represented as a frequency histogram of features present in the image. The histogram is leveraged to identify the similarity of one image to another.

The detection of keypoints on the images is made using Scale-Invariant Feature Transform (SIFT) [5, 10, 12]. The procedure is defined as follows:

1. **Construction of Scale Space**: The idea behind the construction of a scale space is to ensure that the detected features are not scale dependent. In some cases, an image can appear differently at different scales. However, the detection of similarity between images is required to be performed agnostic of the scale of the images. Gaussian blur is applied on the image to reduce the noise on the

image. The original is reduced in half resulting in a scaled image. Varying degrees of Gaussian blur are applied to original and scaled images resulting in images of varying scale space.

2. **Difference of Gaussian**: The procedure is to subtract one blurred version of an original image from another, less blurred version of the original image. The intuition is that the features of the images are enhanced providing images of better quality since they are put through a blurring effect in Step 1.
3. **Keypoint Localization**: The feature selection aspect of the algorithm lies in this step. Initially, the local minima and maxima of the images are identified by comparing each pixel in the image with every other pixel in its neighborhood. Later, the keypoints that provide the most information are kept and the low contrast keypoints are discarded. The prominence of the keypoints is identified using the second-degree Taylor expansion. Only the keypoints that result in a magnitude of 0.03 are kept and the others are discarded.
4. **Orientation Assignment**: This stage of the process assigns an orientation to each of the keypoints identified in Step 3 to make them invariant to rotation. Firstly, the magnitude and orientation for each of the pixels is computed where the former represents the intensity and latter represents the orientation of the pixel. The computation of magnitude and orientation warrants that the gradients in X and Y directions be computed. Assuming that the gradient in the X direction in G_x and Y direction is G_y, the magnitude is given by $\sqrt{G_x^2 + G_y^2}$ and orientation by atan(G_y/ G_x). The obtained magnitude and orientation are plotted as a histogram with orientation on the X axis and magnitude on the Y axis, where each bin represents a 10^o orientation yielding in 36 bins. The peak of the histogram is considered the orientation for the keypoint.
5. **Generation of Keypoint Descriptors**: The final step is obtaining the keypoint descriptors for each of the keypoints obtained in Step 4. The descriptors for a keypoint are identified by taking a 16 x 16 neighborhood around the keypoint. The neighborhood is then split into four 4 x 4-pixel neighborhoods. Similar to Step 4, a histogram is plotted between magnitude and orientation. However, the histogram is made up of only eight bins with each bin representing a 45^o orientation. Overall, 128 bins indicating magnitude and orientation for each keypoint are obtained.

The implementation of BOVW technique for novelty detection is as follows:

1. Extract the set of features from each of the images in the data set using Scale-Invariant Feature Transform (SIFT).
2. Convert the extracted features into visual words by using the K-Means Clustering algorithm. The centers identified by the algorithm form the vocabulary of visual words.
3. Compare the features of each of the images against the vocabulary and create histograms for each of the images in both the training and testing data sets.
4. Select a K, the order of the neighbor to be considered as the nearest neighbor and compute the Euclidean distance from each galaxy to its K^{th} neighbor using the data from the histogram.
5. Sort the galaxies by the distance to their K^{th} neighbor. Greater the distance, higher the likelihood that the galaxy is a novelty.

The source code of the algorithm can be found at PanSTARRSVisualBOWAlgorithm.

5 Method

The concept of 'rank' is used to express the performance of the proposed techniques. Rank *r* is the number of query galaxies determined by the algorithm as the most likely to be novelty galaxies. If a novelty galaxy is among these *r* galaxies, the attempt is considered a hit, and otherwise a miss. Since candidates of novelty galaxies are inspected manually, a method that returns false positives is acceptable as long as the novelty galaxies are among a set that is small enough for manual analysis. Note that the problem of novelty galaxy detection does not require identifying all novelty galaxies, as novelty galaxies of the same type are expected to be present multiple times in galaxy datasets acquired by robotic telescopes.

5.1 Entropy Based Algorithm

Figure 2 shows the performance of the Entropy based algorithm stated in Section 4.1 when the K parameter is set to 5, 10, and 20. The results show that the performance of the algorithm when identifying spiral galaxies among lenticular galaxies is better than the performance of the algorithm when identifying stars among lenticular galaxies. This is partly explained by the fact that lenticular galaxies and stars are more similar in morphology to each other compared to lenticular and spiral galaxies.

5.2 Bag of Visual Words

The performance of the Bag of Visual Words technique described in Section 4.2 is shown in Figure 3 when the value of K is set to 5, 10 and 20. From the results, it is inferred that the performance of the technique while identifying stars is far superior compared to the performance of the technique while identifying spiral and lenticular galaxies. It is perhaps due to the similarities observed between the images of spiral and lenticular galaxies. While the galaxies are structurally different, both lenticular and spiral galaxies contain a sea of nebulous matter around them. The error rate for stars is significantly lower because the images of stars contain no nebulous matter around them and are structurally circular. This characteristic of stars aids the technique in being identified better.

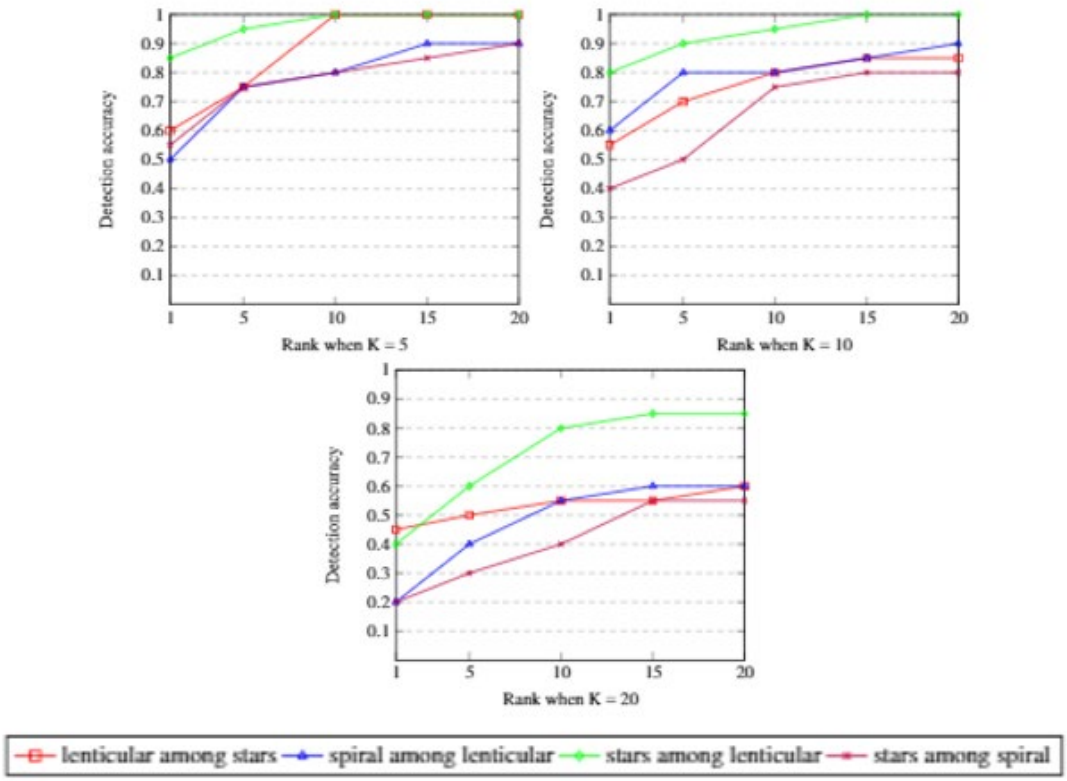

Figure 2: Detection accuracy when using different datasets and ranks using entropy-based algorithm

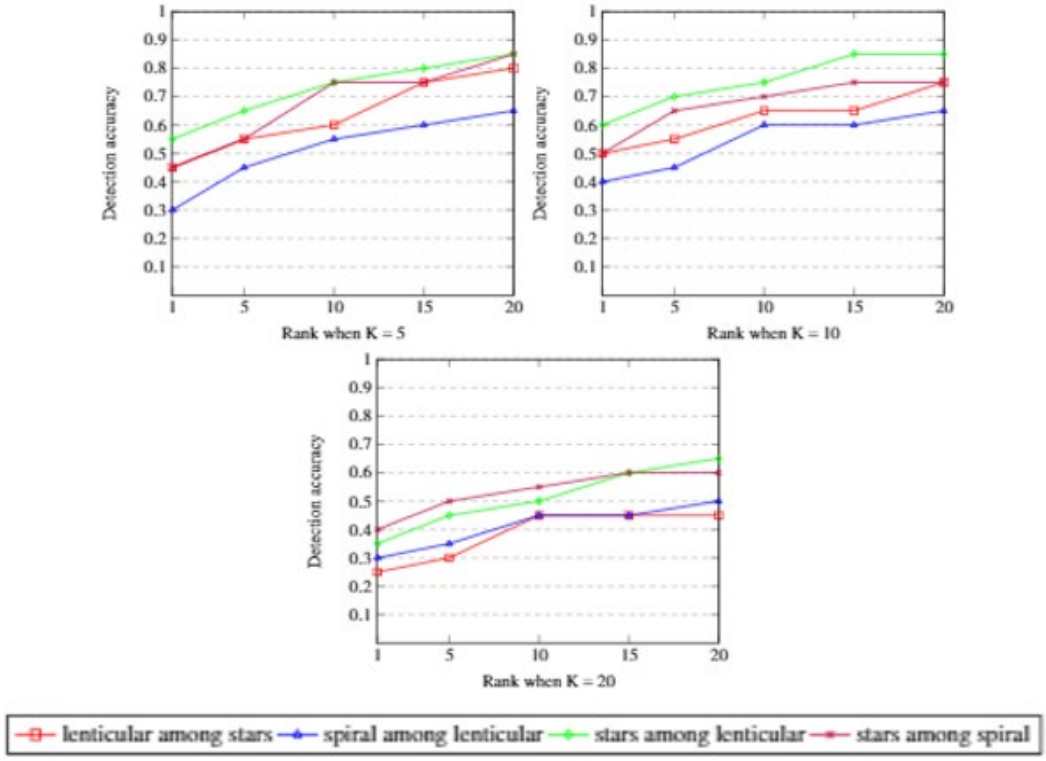

Figure 3: Detection accuracy when using different datasets and ranks using bag of visual words

6 Comparison to Novelty Detection Algorithms

Since the problem of automatic novelty galaxy detection is relatively new, not many proposed novelty detection algorithms for galaxies are available in the existing literature. Hence, the performance of the proposed algorithm is compared against "traditional" novelty detection algorithms such as one-class SVM, K-Means, and Local Outlier Factor (LOF), as well as the deep learning-based auto-encoders.

6.1 Comparison to Deep Learning with Auto-Encoders

Auto-encoders [7] are a class of unsupervised machine learning using artificial neural networks (ANN). A typical artificial neural network consists of an input layer, which inputs the data to the layers of the neural network, several hidden layers, and an output layer, which outputs the outcome. Each of the hidden layers in the network performs computations on the weighted inputs and transfers the computed result to the next layer. An auto-encoder can be conceptualized as a specific type of neural network that copies the input values to the output without requiring a target variable. Since target variables are not required, it is a good fit for unsupervised learning [15].

For this experiment, a deep auto-encoder is used. The auto-encoder architecture comprises of ReLU activation function in the encoding layers and sigmoid activation function in the decoding layers. The loss function used is binary cross-entropy and the optimizer used is RMSProp. The size of the input of 120 x 120. Auto-encoders with three different architectures are developed. Architecture#1, with hidden layers of sizes 128, 64, 32, 64, 128, architecture#2, with hidden layers of sizes 1024, 512, 256, 512, 1024 and architecture#3 with hidden layers of sizes 2048, 1024, 512, 1024, 2048. In each of the datasets, the "regular" galaxy images are split into two groups, one containing 180 images to train the auto-encoder, and another of 20 images to test on the auto-encoder to obtain the reconstruction losses. Then, the "novelty" galaxy images are tested on the auto-encoder, and the loss of the "novelty" galaxies is compared to the loss of the "regular" galaxies. For evaluation, the 30th to the 90th percentile of reconstruction loss values on "regular" galaxies are used as thresholds, and the percentage of "novelty" galaxies identified from amongst 200 images of "novelty" galaxies is computed as shown in Figure 4.

6.2 One-Class Support Vector Machines (OCSVM)

The OCSVM algorithm is applied to each of the four datasets using the scikit-learn library. The performance of the algorithm is measured as the number of actual "novelty" galaxies identified by the algorithm divided by the total number of "novelty" galaxies attempted. Ideally, only the ten "novelty" galaxies are identified as "novelty" galaxies by the algorithm, in which case the detection rate would be 100%. However, the observation on all four datasets is that the algorithm identifies a large portion of "regular" galaxies also identified as "novelty" galaxies while also misidentifying some "novelty" galaxies as regular galaxies. So, the performance of the algorithm is similar to that of novelty

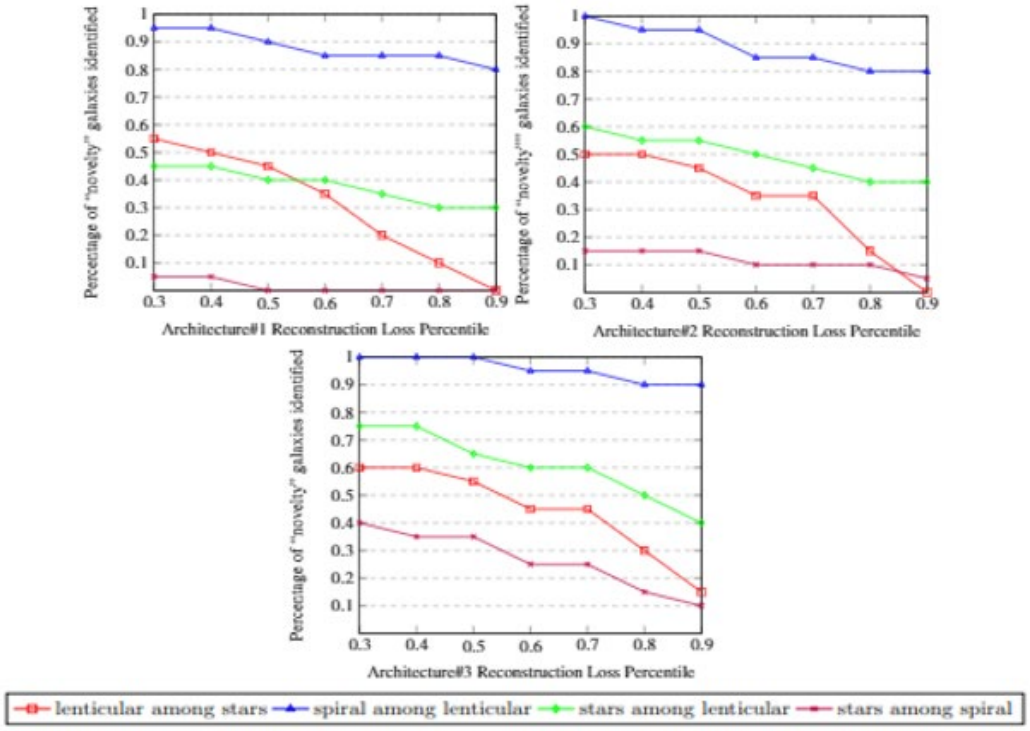

Figure 4: Detection accuracy when using different datasets and ranks using auto-encoders

galaxy detection by random chance. The outcomes of the algorithm are shown in Figure 5.

6.3 Local Outlier Factor (LOF) Algorithm

The Local Outlier Factor (LOF) algorithm produces a score that provides an insight into the likelihood of a data point being an outlier in a given dataset. The scikit-learn LOF library is used to apply the algorithm to each of the datasets. Since the algorithm is unsupervised, no alteration is made to the datasets. A score close to 1 means that the sample is an inlier, while outliers have a larger LOF score. The results show that for each of the datasets, all of the values obtained for the LOF scores are 1, indicating that the algorithm considers all of the images, including the outliers, as the same class as the inliers. As a result, the accuracy obtained using the algorithm is 0 % on all four of the datasets.

7 Comparison to Novelty Detection Algorithms

Automation techniques in the field of astronomical discovery and analysis are the need of the hour considering the enormous amount of information recorded by modern sky surveys using robotic telescopes. The infrequent occurrence of novelty galaxies makes the problem of novelty galaxy detection complex since conventional machine learning classifiers don't always perform well owing to lack of enough training data.

The proposed unsupervised novelty detection algorithm uses a comprehensive set of numerical image content descriptors, and therefore depends on feature selection. Entropy is shown as a useful way to select features for the galaxies are the most frequent. The results are as shown in Figure 6.

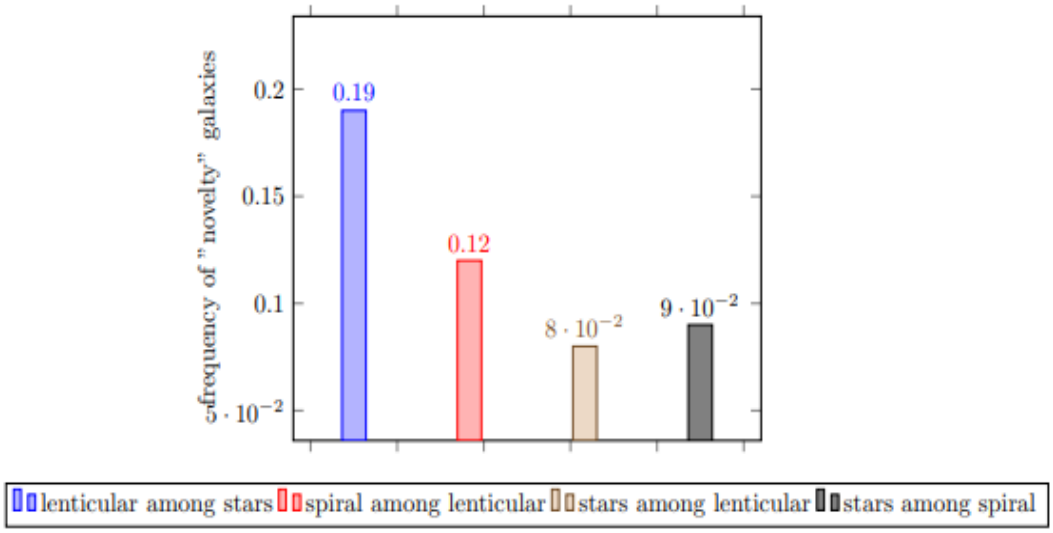

Figure 5: Detection accuracy when using different datasets and ranks using OCSVM

6.4 K-Means Clustering Algorithm

K-Means is a simple and established unsupervised learning algorithm which works by choosing a centroid value for each randomly chosen cluster, and iteratively assigning each data point to a cluster that best fits based on the Euclidean distance between the data point and the centroids of the clusters. K-Means is typically used for automatic clustering. However, in some cases it can be used for novelty detection by identifying small clusters. If a small cluster is identified, the cluster may contain a small number of self-similar samples that are different from the other samples in the dataset. Therefore, K-Means is an algorithm that could be possibly used for novelty detection in the current scenario. The algorithm is tested with two through 10 clusters. The performance is measured as the number of novelty galaxies among regular galaxies in the cluster in which novelty

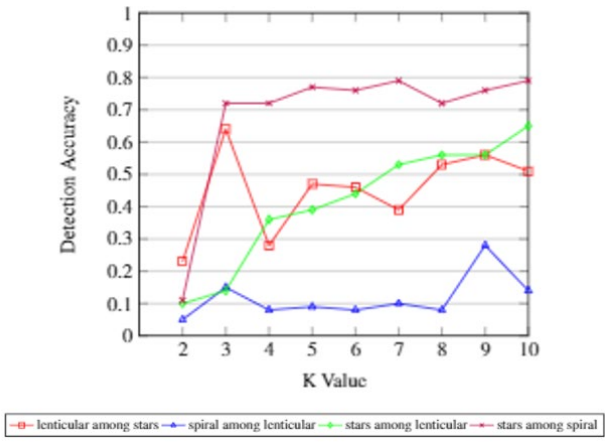

Figure 6: Detection accuracy using different datasets and using K-Means Clustering

problem of unsupervised detection of novelty galaxies.

The Visual Bag of Words technique assimilated to novelty detection identifies the distinctive features of galaxies to help identify novelty galaxies. The technique can scale to images of different dimensions and orientations since it is built to be scale and orientation invariant.

The methods proposed in the paper outperform "traditional" methods such as one-class SVM, K-Means, and newer methods based on deep neural networks such as auto-encoders. It should be noted, however, that the relatively low number of annotated samples does not allow efficient training of an autoencoder, that normally requires a high number of samples. The dataset used for the experiments is publicly available and can be used for the development and testing of new algorithms for novelty galaxy detection in large astronomical databases.

The downside of the evaluation is that it is performed on a relatively small and controlled dataset, far smaller than the huge datasets generated by modern digital sky surveys. The efficacy of the method will be tested in the future by applying it to extremely large image databases and evaluating its ability to identify real novelty galaxies hidden among millions of celestial objects that have not been inspected yet.

Acknowledgement

The research was funded in part by NSF grant number AST-1903823.

References

[1] Singh Aishwarya, "A Detailed Guide to the Powerful Sift Technique for Image Matching," *Medium*, 2019.

[2] T. Amarbayasgalan, B. Jargalsaikhan, and K. H. Ryu, "Unsupervised Novelty Detection using Deep Autoencoders with Density Based Clustering," *Applied Sciences*, 8(9):1468, 2018.

[3] K. Borne, "Virtual Observatories, Data Mining, and Astroinformatics," *Planets, Stars and Stellar Systems*, 2:403-443, 2013.

[4] R. J. Buta, "Galactic Rings Revisited–I. CVRHS Classifications of 3962 Ringed Galaxies from the Galaxy Zoo 2 Database," *Monthly Notices of the Royal Astronomical Society*, 471(4):4027-4046, 2017.

[5] Tyagi Deepanshu, "Introduction to SIFT," *Medium*, 2019.

[6] S. G. Djorgovski, A. A. Mahabal, A. J. Drake, M. J. Graham, and C. Donalek, "Sky Surveys," *Planets, Stars, and Stellar Systems*, 2:223-281, 2013.

[7] E. Kuminski, J. George, J. Wallin, and L. Shamir, "Combining Human and Machine Learning for Morphological Analysis of Galaxy Images," *Publications of the Astronomical Society of the Pacific*, 126(944):959, 2014.

[8] C. Lintott, K. Schawinski, S. Bamford, A. Slosar, K. Land, D. Thomas, E. Edmondson, K. Masters, R C. Nichol, M. J. Raddick, and A. Szalay, "Galaxy Zoo 1: Data Release of Morphological Classifications for Nearly 900,000 Galaxies," *Monthly Notices of the Royal Astronomical Society*, 410(1):166-1768, 2011.

[9] C. J. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu, and P. Murray, "Galaxy Zoo: Morphologies Derived from Visual Inspection of Galaxies from the Sloan Digital Sky Survey," *Monthly Notices of the Royal Astronomical Society*, 389(3):1179-1189, 2008.

[10] G. Lowe, "Sift-the Scale Invariant Feature Transform," *Int. J*, 2(91-110):2, 2004.

[11] Z. Lu, L. Wang, and J. R. Wen, "Image Classification by Visual Bag-of-Words Refinement and Reduction," *Neurocomputing*, 173:373-384, 2016.

[12] Ning Minghao, "Sift (Scale-Invariant Feature Transform)," *Medium*, 2019.

[13] W. W. Morgan and N. U. Mayall, 1957. "A Spectral Classification of Galaxies," Arp and Halton, Atlas of Peculiar Galaxies, *Publications of the Astronomical Society of the Pacific*, 69(409):291-303,1966.

[14] L. Shamir, "Evaluation of Face Datasets as Tools for Assessing the Performance of Face Recognition Methods," *International Journal of Computer Vision*, 79(3):225-230, 2008.

[15] L. Shamir, "Automatic Morphological Classification of Galaxy Images," *Monthly Notices of the Royal Astronomical Society*, 399(3):1367-1372, 2009.

[16] L. Shamir, "Morphology-Based Query for Galaxy Image Databases," *Publications of the Astronomical Society of the Pacific*, 129(972):024003, 2016.

[17] A. Schutter and L. Shamir, "Galaxy Morphology—An Unsupervised Machine Learning Approach," *Astronomy and Computing*, 12:60-66, 2015.

[18] L. Shamir, "Automatic Detection of Full Ring Galaxy Candidates in SDSS," *Monthly Notices of the Royal Astronomical Society*, 491(3):3767-3777, 2020.

[19] L. Shamir, A. Holincheck, and J. Wallin, "Automatic Quantitative Morphological Analysis of Interacting Galaxies," *Astronomy and Computing*, 2:67-73, 2013.

[20] L. Shamir, T. Macura, N. Orlov, D. M. Eckley, and I. G. Goldberg, "Impressionism, Expressionism, Surrealism: Automated Recognition of Painters and Schools of Art," *ACM Transactions on Applied Perception (TAP)*, 7(2):1-17, 2010.

[21] L. Shamir, N. Orlov, D. M. Eckley, T. Macura, J. Johnston, and I. G. Goldberg, "Wndchrm–an Open-Source Utility for Biological Image Analysis," *Source Code for Biology and Medicine*, 3(1):1-13, 2008.

[22] L. Shamir and J. Wallin, "Automatic Detection and Quantitative Assessment of Peculiar Galaxy Pairs in Sloan Digital Sky Survey," *Monthly Notices of the Royal Astronomical Society*, 443(4):3528-3537, 2014.

[23] T. Shi and S. Horvath, "Unsupervised Learning with Random Forest Predictors," *Journal of Computational and Graphical Statistics*, 15(1):118-138, 2006. Ming-Jian Zhou and Jun-Cai Tao. "An Outlier Mining

Algorithm Based on Attribute Entropy," 2011.
[24] K. Sun, J. Zhang, C. Zhang, and J. Hu, "Generalized Extreme Learning Machine Autoencoder and a New Deep Neural Network," *Neurocomputing*, 230:374-381, 2017.
[25] J. Yang, Y. G. Jiang, A. G. Hauptmann, and C. W. Ngo, "Evaluating Bag-of-Visual-Words Representations in Scene Classification," *Proceedings of the International Workshop on Multimedia Information Retrieval*, pp. 197-206, September 2007.
[26] Y. Zhang, R. Jin, and Z. H. Zhou, "Understanding Bag-of-Words Model: A Statistical Framework," *International Journal of Machine Learning and Cybernetics*, 1(1-4):43-52, 2010.

Venkat Margapuri is a researcher working on his PhD. in Computer Science at Kansas State University. His interests are in the areas of machine learning, scientific computing, robotics and data science.

Basant Thapa is currently pursuing a Master's in Software Engineering from Kansas State University. He graduated from Wichita State University with a Bachelor's in Computer Science and a minor in Mathematics. He is interested in advancing his study in the field of big data solutions and machine learning.

Lior Shamir is an Associate Professor of computer science at Kansas State University. He received his Ph.D from Michigan Technological University in 2006, and postdoc at the National Institutes of Health (NIH).