A catalog of broad morphology of Pan-STARRS galaxies based on deep learning

by

Hunter Goddard

B.S., Kansas State University, 2017

A THESIS

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Computer Science
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2021

Approved by:

Major Professor
Lior Shamir

Copyright

Abstract

Autonomous digital sky surveys such as Pan-STARRS have the ability to image a very large number of galactic and extra-galactic objects, and the large and complex nature of the image data reinforces the use of automation. Here we describe the design and implementation of a data analysis process for automatic broad morphology annotation of galaxies, and applied it to the data of Pan-STARRS DR1. The process is based on filters followed by a two-step convolutional neural network (CNN) classification. Training samples are generated by using an augmented and balanced set of manually classified galaxies. Results are evaluated for accuracy by comparison to the annotation of Pan-STARRS included in a previous broad morphology catalog of SDSS galaxies. Our analysis shows that a CNN combined with several filters is an effective approach for annotating the galaxies and removing unclean images. The catalog contains morphology labels for 1,662,190 galaxies with 95% accuracy. The accuracy can be further improved by selecting labels above certain confidence thresholds. The catalog is publicly available.

Table of Contents

List of Figures v

List of Tables vii

Acknowledgements viii

1 Introduction 1

2 Data 3

3 Image analysis method 5

3.1 Primary classification 6

3.2 Secondary Classification 11

4 Results 13

4.1 Comparison to an existing SDSS catalog 14

5 Conclusions 24

Bibliography 26

List of Figures

2.1 *Distribution of the r Kron magnitude of the galaxies in the dataset.* 4

3.1 *Distribution of the r exponential magnitude of the galaxies in the Galaxy Zoo dataset from which the annotations were taken.* 9

3.2 *Distribution of the redshift of the galaxies in the Pan-STARRS dataset. The redshift values were taken from SDSS.* 9

3.3 *Distribution of the redshift of the galaxies in the Galaxy Zoo dataset from which the annotations were taken.* 10

3.4 Confusion matrix and ROC curve of the classification of the 30% test samples. 11

3.5 Confusion matrix and ROC curve of the classification of the 1,200 ghosts and spiral galaxies. 12

4.1 Number of spiral and elliptical galaxies remaining when keeping only those at or above a certain model confidence. 14

4.2 The proportion of predicted labels that, when restricted to a minimum confidence threshold, agree with the annotations in[1]. For example, restricting the catalog to labels with 90% confidence or higher will have approximately 98% agreement with the annotations in[1] 16

4.3 Galaxies imaged by Pan-STARRS that were classified incorrectly as elliptical in[1] and as spiral in this catalog. 17

4.4 Galaxies imaged by Pan-STARRS that were classified as spiral in[1] but as elliptical in this catalog. 18

4.5 Galaxies imaged by SDSS and were classified incorrectly as elliptical in[1]. . . 19

4.6 Galaxies imaged by SDSS and were classified as spiral in[1] but as elliptical in this catalog. 20

List of Tables

3.1 Examples of images filtered for having too few foreground pixels or having too many saturated pixels. 7

4.1 Galaxies imaged by Pan-STARRS, SDSS, and HST. While the Pan-STARRS and SDSS images do not show clear spiral arms of the galaxies, HST shows that these galaxies are clearly spiral, and the arms can be identified. 15

4.2 Examples of images that were misclassified by the model. 22

4.3 Examples of images that were classified correctly by the model. 23

Acknowledgments

This research was funded by NSF grant AST-1903823. This publication uses data generated via the Zooniverse.org platform, development of which is funded by generous support, including a Global Impact Award from Google, and by a grant from the Alfred P. Sloan Foundation. The Pan-STARRS1 Surveys (PS1) and the PS1 public science archive have been made possible through contributions by the Institute for Astronomy, the University of Hawaii, the Pan-STARRS Project Office, the Max-Planck Society and its participating institutes, the Max Planck Institute for Astronomy, Heidelberg and the Max Planck Institute for Extraterrestrial Physics, Garching, The Johns Hopkins University, Durham University, the University of Edinburgh, the Queen's University Belfast, the Harvard-Smithsonian Center for Astrophysics, the Las Cumbres Observatory Global Telescope Network Incorporated, the National Central University of Taiwan, the Space Telescope Science Institute, the National Aeronautics and Space Administration under Grant No. NNX08AR22G issued through the Planetary Science Division of the NASA Science Mission Directorate, the National Science Foundation Grant No. AST-1238877, the University of Maryland, Eotvos Lorand University (ELTE), the Los Alamos National Laboratory, and the Gordon and Betty Moore Foundation.

Chapter 1

Introduction

With their ability to generate very large databases, autonomous digital sky surveys have been enabling research tasks that were not possible in the pre-information era, and have been becoming increasingly pivotal in astronomy. The ability of digital sky surveys to image large parts of the sky, combined with the concept of virtual observatories that make these data publicly accessible[2], has been introducing a new form of astronomy research, and that trend is bound to continue[3;4].

The Panoramic Survey Telescope and Rapid Response System (Pan-STARRS)[5;6] is a comprehensive digital sky survey covering $\sim 10^3$ degree^2 per night using an array of two 1.8m telescopes. Among other celestial objects, Pan-STARRS images a very large number of galaxies. Due to the complexity of galaxy morphology, the ability of current photometric pipelines to analyze these galaxy images is limited, and substantial information that is visible to the humans eye is practically unavailable to users of digital sky surveys data.

To automate the analysis of galaxy images, several methods have been proposed, including GALFIT[7], GIM2D[8], CAS[9], the Gini coefficient of the light distribution[10], Ganalyzer[11], and SpArcFiRe[12]. However, the ability of these methods to analyze a large number of real-world galaxy images and produce clean data products is limited, and catalogs of galaxy morphology were prepared manually by professional astronomers[13;14].

Due to the high volumes of data, the available pool of professional astronomers is not able

to provide the sufficient labor to analyze databases generated by modern digital sky surveys, leading to the use of crowdsourcing for that task[15–17]. The main crowdsourcing campaign for analysis of galaxy morphology was Galaxy Zoo[16], providing annotations of the broad morphology of galaxies imaged by Sloan Digital Sky Survey (SDSS), as well as other surveys such as the Cosmic Assembly Near-infrared Deep Extragalactic Legacy (CANDELS). However, analyzing the broad morphology of SDSS galaxies required ∼3 years of work performed by over 10^5 volunteers, and led to $\sim 7 \cdot 10^4$ galaxies considered "superclean". Given the huge databases of current and future sky surveys, it is clear that even when using crowdsourcing, the throughout of manual classification might not be sufficient for an exhaustive analysis of such databases.

The use of machine learning provided more effective methods for the purpose of galaxy image classification[18–27], and the use of such methods also provided computer-generated catalogs of galaxy morphology[1;28–35]. Automatic annotation methods were also tested on Pan-STARRS data by using the photometric measurements of colors and moments, classified by a Random Forest classifier to achieve a considerable accuracy of $\sim 89\%$[36].

Here we use automatic image analysis to prepare a catalog of the broad morphology of $\sim 1.7 \cdot 10^6$ Pan-STARRS DR1 galaxies. The catalog was generated by using a data analysis process that involves several steps and two convolutional neural networks (CNN) that automated the annotation process to handle the high volume of data.

Chapter 2

Data

The galaxy image data is sourced from the first data release (DR1) of Pan-STARRS[6;37;38]. First, all objects with Kron r magnitude of less than than 19 and identified by Pan-STARRS photometric pipeline as extended in all bands were selected.

To filter objects that are too small to identify morphology, objects that have Petrosian radius smaller than 5.5" were removed. To remove stars, objects that their PSF i magnitude subtracted by their Kron i magnitude was greater than 0.05 were also removed. That led to a dataset of 2,394,452 objects[33]. Objects that were flagged by Pan-STARRS photometric pipeline as artifacts, had a brighter neighbor, defect, double PSF, or a blend in any of the bands were excluded from the dataset. That led to a dataset of 2,131,371 objects assumed to be sufficiently large and clean to allow morphological analysis. Figure 2.1 shows the distribution of the r Kron magnitude of the galaxies in the dataset.

The galaxy images were then downloaded using Pan-STARRS *cutout* service. The images are in the JPG format and have a dimensionality of 120×120 pixels as in[1]. Pan-STARRS *cutout* provides JPG images for each of the bands. Here we use the images of the g band, as the color images using the y, i, and g bands are in many cases noisy, and do not allow effective analysis of the morphology. The process of downloading the data was completed in 62 days.

The initial scale of the *cutout* was set to 0.25" per pixel. For each image that was

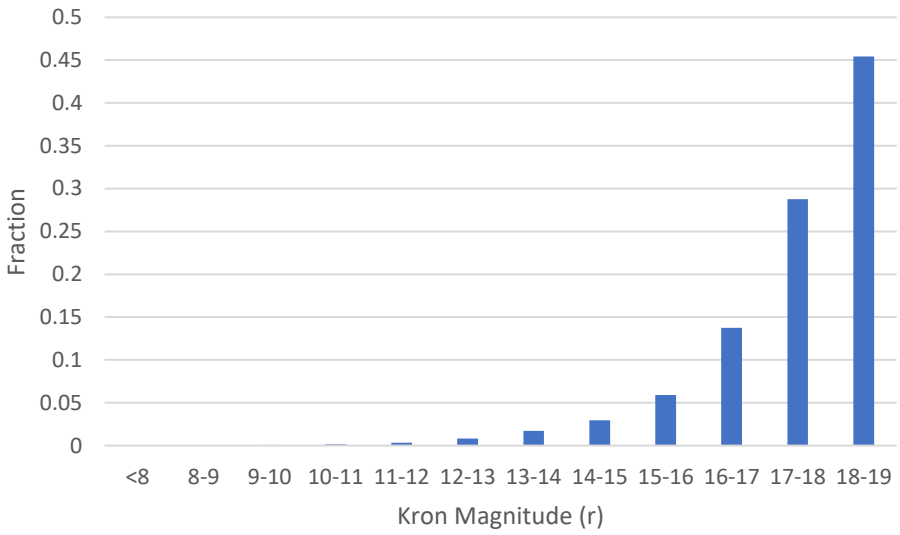


Figure 2.1: *Distribution of the r Kron magnitude of the galaxies in the dataset.*

downloaded, all bright pixels (grayscale value higher than 125) located on the edge of the frame were counted. If more than 25% of the pixels on the edge of the frame were bright, it is an indication that the object does not fully fit inside the frame. In that case, the scale was increased by 0.05", and the image was downloaded again. That was repeated until the number of bright pixels on the edge was less than 25% of the total edge pixels, meaning that the object is inside the frame. The JPG images are far smaller than the FITS images. A 120×120 JPG image retrieved through Pan-STARRS *cutout* service is normally of size of ∼3KB, while an image of the same dimensions in the FITS format will be ∼76KB. Although the FITS files provide more information, downloading the files in FITS format is substantially slower. While downloading the JPG images lasted 62 days, downloading the same number of images in the much larger FITS format will require a far longer period of time. The JPG images do not allow photometry, but they are smaller than the FITS files and provide visual information about the shape of the galaxy, which is the information required for the morphological classification of the galaxies. As explained in Section 3.1, the training of the neural network was done with images retrieved from Pan-STARRS, with the exact same size and format as the images that were annotated by the neural network after it was trained.

Chapter 3

Image analysis method

The filtering of the data described in Section 2 aims at removing objects that are not clean galaxy images. That allows to reduce the number of images downloaded and classified in the next step with the deep neural network. The removal of objects that are not galaxy images also makes the neural network more accurate due to the higher consistency of the data it is trained with.

To remove saturated images and images that have too few features to allow morphological classification, two additional filters are used. The first filter finds the ratio of fully saturated pixels (a grayscale value of 255 in the JPG image) to the total number of pixels and discards the image if this ratio is higher than 15:1000. Since a high number of saturated pixels is not expected in a clean galaxy image, the simple threshold of 1.5% is sufficient to identify and reject saturated images that are not galaxy images. This step rejected 30,220 objects that were identified as saturated.

The second filter uses the Otsu global threshold method[39] to separate the image into foreground and background pixels. If the number of foreground pixels is less than 1.8% of the total image, the image is marked as having too few distinguishable features. This filter rejected 375,107 galaxies that were identified as having too little foreground to allow identification. Together, these filters removed 405,327 images (∼19%) from the data set. The thresholds were determined experimentally by observing galaxy image samples. Table 3.1

shows examples of several objects that were filtered based on too few foreground pixels or too many saturated pixels.

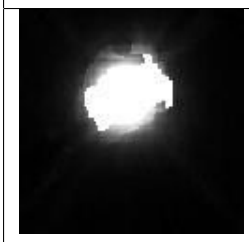
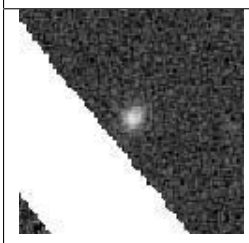
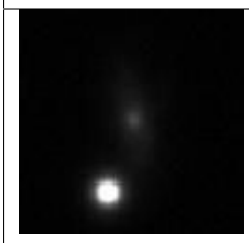
3.1 Primary classification

The classifier used for the purpose of annotating the galaxy images is a deep convolutional neural network (DCNN) based on the LeNet-5 architecture[40]. To adjust the model for input images of size 120×120 instead of 32×32, the kernel in the first convolutional layer was changed from 5×5 with stride 1 to 10×10 with stride 2, and the filter in the first pooling layer was similarly changed from 2×2 with stride 2 to 4×4 with stride 4. Each of the following layers has identical hyperparameters except for the output layer, where the number of classes is reduced from 10 to 2. The SoftMax output layer of the model provides a degree of certainty for the annotations that allows controlling the size/accuracy trade-off of the catalog, as will be discussed in Section 4.

Training samples were obtained using the debiased "superclean" Galaxy Zoo annotations. "Superclean" objects are objects on which 95% or more of the annotators agreed on their morphology with correction for the redshift bias[16]. That selection leads to a subset of very consistent annotations[16], but it also filters the vast majority of Galaxy Zoo annotations that do not satisfy these requirements. The Galaxy Zoo crowdsourcing campaign annotated galaxies imaged by SDSS, which is a different instrument with a different image processing pipeline. Although it has been shown in the past that neural networks trained with data from one telescope can be used to classify data acquired by other telescopes[41], it has also been shown that the accuracy of such networks is inferior to the accuracy of a neural network trained and tested with data from the same instrument[41]. Since a large number of galaxies annotated by Galaxy Zoo were also imaged by Pan-STARRS, the Pan-STARRS images of these galaxies can be fetched and be used as the training data, so that the images used to train the neural network are imaged by the same instrument that imaged the galaxies annotated by that network.

Due to the substantial overlap between the footprint of Pan-STARRS and SDSS, the

Table 3.1: *Examples of images filtered for having too few foreground pixels or having too many saturated pixels.*

Image	Saturated pixels (%)	Foreground pixels (%)
	6.1	10.7
	13.5	21.7
	30.9	34.9
	0.06	1.4
	0.16	1.1

idea of using SDSS data as labels to train machine learning systems with Pan-STARRS data has been used in the past. For instance,[42] used spectroscopic data from SDSS as labels for training a machine learning system that can determine the photometric redshift of Pan-STARRS galaxies.

In order to train the neural network with images from the same instrument that it is expected to annotate, the images of the galaxies annotated by Galaxy Zoo were retrieved from Pan-STARRS. Pan-STARRS has a different footprint than SDSS, so not all galaxies annotated by Galaxy Zoo are also imaged by Pan-STARRS. However, 22,456 Galaxy Zoo galaxies with "superclean" annotations were matched with galaxies in Pan-STARRS DR1 based on their right ascension and declination (within difference of 0.0001 degrees). These images were fetched from Pan-STARRS and were used for training the neural network.

Figure 3.1 shows the distribution of the r exponential magnitude of the Galaxy Zoo galaxies that their annotations were used for the compilation of the training set. The magnitude distribution is somewhat different from the magnitude distribution of the Pan-STARRS galaxies shown in Figure 2.1, which can be explained by the 17.77 limiting r Petrosian magnitude applied to the initial Galaxy Zoo sample[15]. As mentioned above, the SDSS images themselves were not used for the training.

Figures 3.2 and 3.3 show the histograms of the redshift distribution of the galaxies in Pan-STARRS and SDSS, respectively. The number of Pan-STARRS galaxies with redshift is small since Pan-STARRS does not collect spectra, and the spectra was only taken from SDSS galaxies that overlapped with Pan-STARRS galaxies. The two graphs show that the distribution of the redshift is similar in both datasets.

Galaxy Zoo manual annotations have been shown in the past to be sensitive to the spin direction of the galaxies[43]. To eliminate the possible effect of spin patterns, the training set was augmented such that all galaxies were mirrored, and both the original and mirrored image of each galaxy were used in the training set. That resulted in a training set of 31,564 spiral images and 13,348 elliptical images. Mirroring the spiral galaxies ensures a symmetric dataset that is not biased by certain preferences of the human volunteers who annotated the galaxies. That is, while mirroring the images in the training set is often used

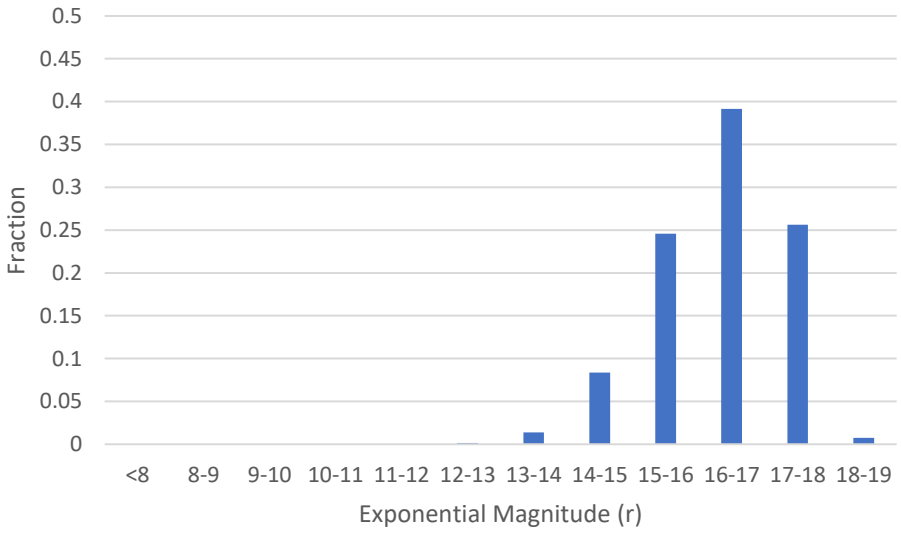


Figure 3.1: *Distribution of the r exponential magnitude of the galaxies in the Galaxy Zoo dataset from which the annotations were taken.*

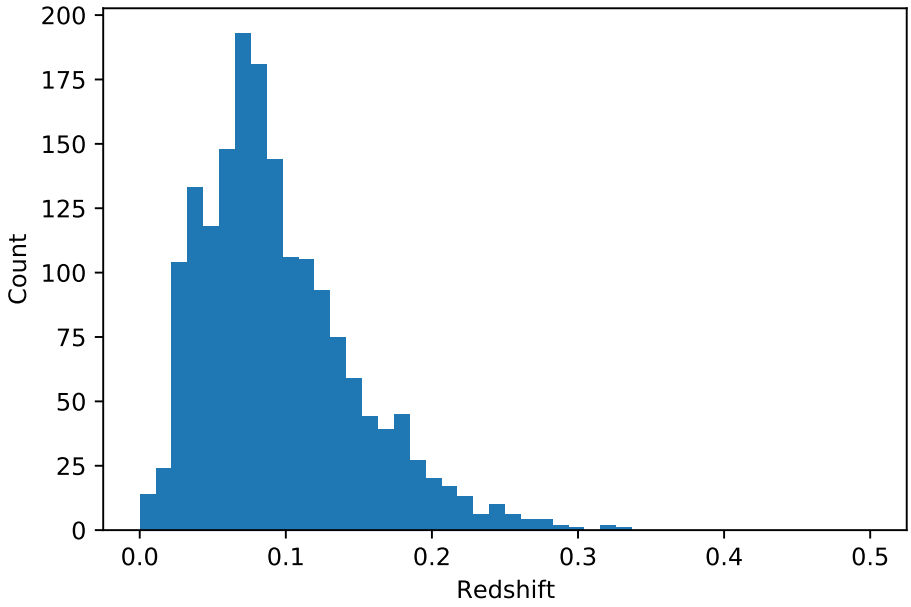


Figure 3.2: *Distribution of the redshift of the galaxies in the Pan-STARRS dataset. The redshift values were taken from SDSS.*

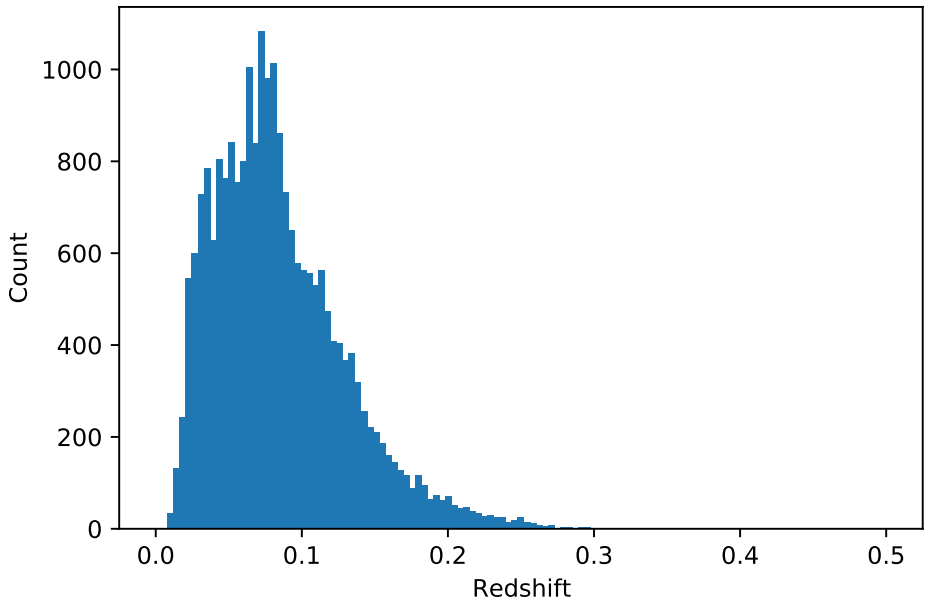


Figure 3.3: *Distribution of the redshift of the galaxies in the Galaxy Zoo dataset from which the annotations were taken.*

when training deep neural networks for augmenting the data and increasing the number of training samples, in this case it was also used to produce a symmetric unbiased dataset. Mirroring of the elliptical galaxies was done to ensure consistency in the manner training data are handled, and avoid a situation in which different classes are handled differently.

The classifier is implemented in Python 3 using TensorFlow[44] and Keras[45]. The model was trained for 250 epochs on a 70% training subset and ended with 98.7% accuracy when evaluated against the remaining 30% testing subset. Figure 3.4 shows the confusion matrix and receiver operating characteristic (ROC) curve of the classification. The high accuracy shows that although the galaxy images were labeled with annotations made with SDSS galaxies, the annotations were still consistent in Pan-STARRS images. That consistency indicates that the two sky surveys are roughly equivalent in the information they provide about the morphology of the galaxies.

Loss was computed using categorical cross entropy, and stochastic gradient descent (SGD) was used as the optimizer. Various activation functions including ReLU were tested, and gave comparable classification accuracy. The tanh activation used by LeNet-5 gave the best

performance and therefore was used for the model. Classification on the total data set (excluding those removed by the filtering step) labeled 904,550 images as elliptical galaxies and 821,494 images as spiral galaxies.

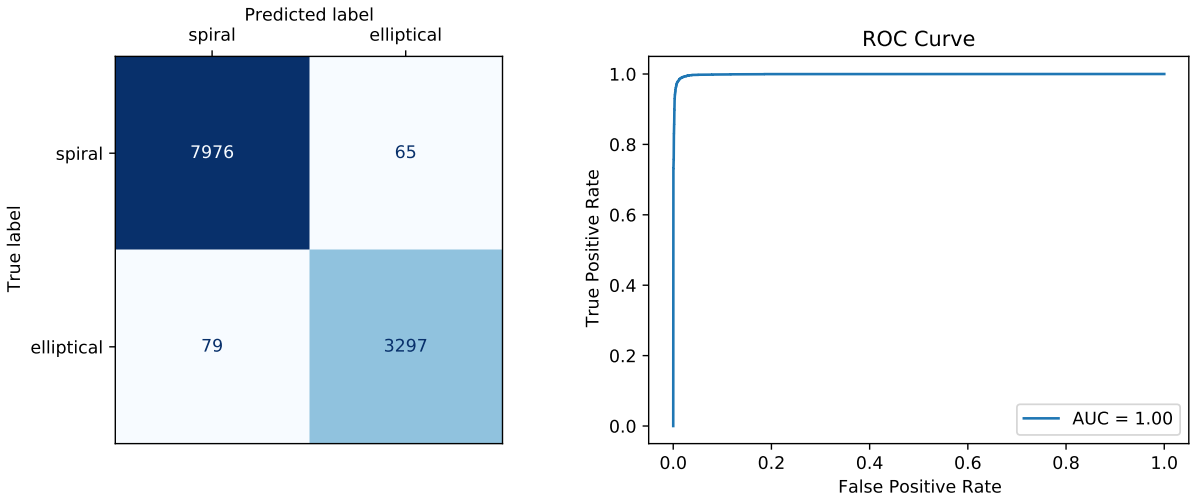


Figure 3.4: *Confusion matrix and ROC curve of the classification of the 30% test samples.*

3.2 Secondary Classification

Following the classification described in Section 3.1, the set of images predicted as spiral was shown to contain a significant number of "ghosts", or unclean images. The CNN classifier interpreted the unclean images as patterns of spiral features, leaving the elliptical predictions relatively clean.

To remove these ghosts, we constructed a second deep CNN to separate them from the true spirals. The architecture of this model is simpler than the first, using three convolutional layers with filter sizes of 7×7×8, 5×5×32, and 3×3×64, ReLU activation function, and a single SoftMax output layer. Between the convolutional layers there are max pooling layers that each reduce the input dimensions by half. The model uses the Adam optimizer and categorical cross entropy for loss.

For training, several hundred ghost images were initially selected from the set of galaxy images that were mistakenly predicted as spirals, and an equal number of spiral galaxy

images were randomly selected from the original spiral training set. These images were divided into 70% training and 30% testing subsets as before. The model converged during training, and the images originally labeled as spirals were further classified into true spirals and ghosts. This process was repeated several times by selecting additional training images from those labeled as "ghosts" until the size of the training set reached 4,000 images. Testing the neural network shows that the network identifies "ghosts" with accuracy very close to 100%, and almost no false positives. Figure 3.5 shows the confusion and matrix and ROC curve when testing 1,200 images of spiral galaxies and ghosts. The final iteration of this classifier identified a total of 63,854 images as "ghosts" ($\sim 7.8\%$), removing them from the set of spiral galaxies.

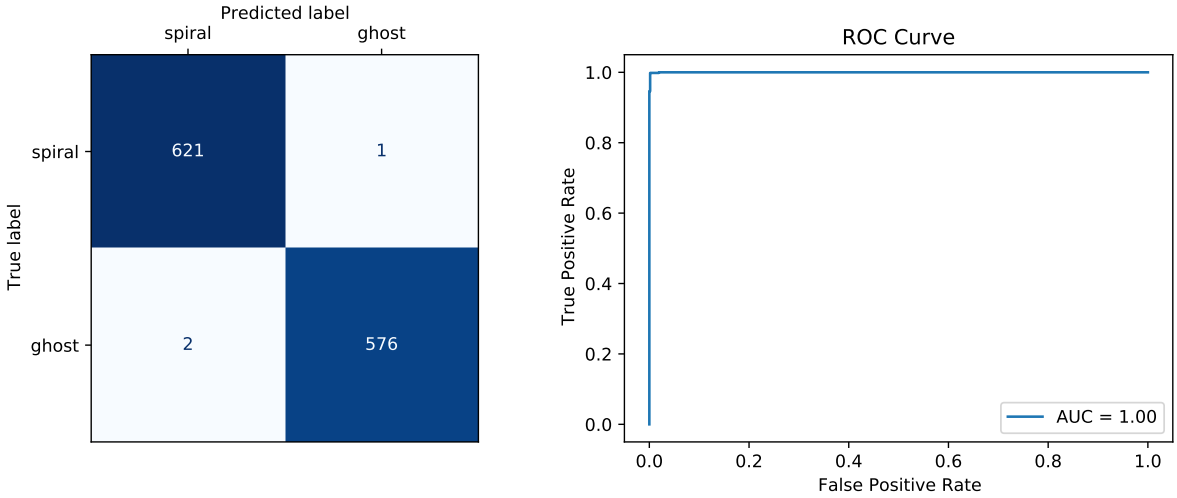


Figure 3.5: *Confusion matrix and ROC curve of the classification of the 1,200 ghosts and spiral galaxies.*

Chapter 4

Results

The application of the methods described in Section 3 to the Pan-STARRS images described in Section 2 provided a catalog of 1,662,190 galaxies. The catalog is accessible through a simple CSV file that can be downloaded at https://figshare.com/articles/PanSTARRS_DR1_Broad_Morphology_Catalog/12081144. Each row in the catalog is a galaxy, and includes the Pan-STARRS object ID of the galaxy, its right ascension, declination, and the probability of the galaxy to be spiral or elliptical as estimated by the SoftMax layer of the CNN as described in Section 3. Figure 4.1 shows the number of galaxies available after applying a threshold to the output of the SoftMax layer of the model.

The catalog includes 904,550 galaxies identified as elliptical and 757,640 identified as spiral. It should be noted that the annotation of a galaxy as an elliptical galaxy means that no spiral features were identified. However, the ability of an algorithm or a person to identify spiral features largely depends on the ability of the optics to provide a detailed image. Therefore, the identification of a galaxy as elliptical does not necessarily guarantee that the galaxy indeed does not have spiral features, but that the optics cannot identify such features[46]. For instance, Table 4.1 shows examples of galaxies imaged by Pan-STARRS and SDSS, and the same galaxies imaged by Hubble Space Telescope (HST). As the table shows, these galaxies do not have clear visible spiral arms in the Earth-based telescopes, while the arms are seen clearly in the HST images.

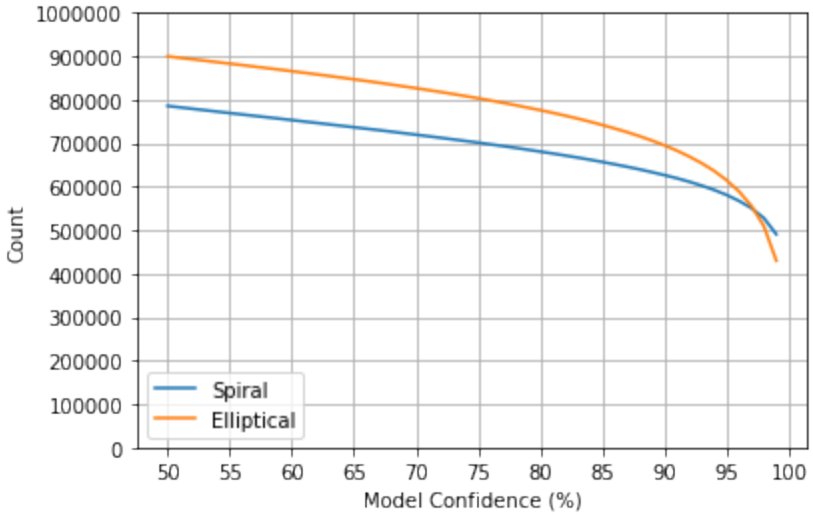


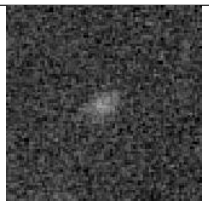
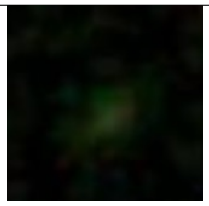
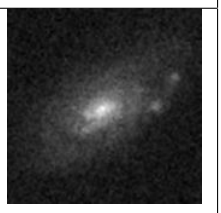
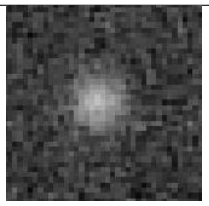
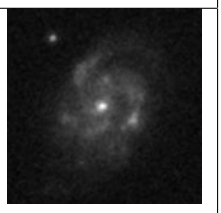
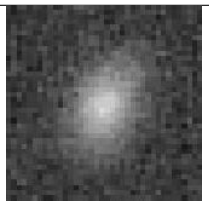
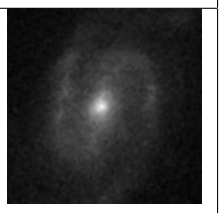
Figure 4.1: *Number of spiral and elliptical galaxies remaining when keeping only those at or above a certain model confidence.*

4.1 Comparison to an existing SDSS catalog

In the absence of a large manually annotated galaxy morphology catalog of Pan-STARRS galaxies, the evaluation of the consistency of the annotations was done using annotations of SDSS galaxies that were also imaged by Pan-STARRS. The largest catalog of broad morphology of SDSS galaxies is[1], with annotation of $\sim 3 \cdot 10^6$ galaxies. Although SDSS is a different sky survey, the footprint of SDSS overlaps with the footprint of Pan-STARRS. Since the[1] catalog is large, it is expected that some galaxies in[1] will also be included in the catalog of Pan-STARRS galaxies described in this paper.

To evaluate the catalog, the annotations were compared to the annotations of SDSS galaxies in[1] with high degree of confidence of the annotations. Since the images of[1] are collected and processed by the SDSS pipeline, their object identifiers naturally do not match the identifiers of Pan-STARRS objects. Therefore, the objects were matched by their coordinates, with tolerance of 0.0001^o to account for subtle differences in measurements between the two telescopes. This produced 13,186 total matches with 1,961 having 90% or higher confidence in the[1] catalog. Figure 4.2 shows the degree of agreement between the annotations of the galaxies in the catalog and the annotations of the galaxies in[1] with high confidence level.

Table 4.1: *Galaxies imaged by Pan-STARRS, SDSS, and HST. While the Pan-STARRS and SDSS images do not show clear spiral arms of the galaxies, HST shows that these galaxies are clearly spiral, and the arms can be identified.*

Coordinates	Pan-STARRS	SDSS	HST
(150.165^o,1.588^o)			
(150.329^o,1.603^o)			
(149.951^o,1.966^o)			

When comparing the accuracy of the catalog to the accuracy of[1], the algorithm was more accurate in identifying spiral galaxies, while the algorithm used in this catalog was more accurate in the identification of elliptical galaxies. The algorithm used in[1] is a "shallow learning" algorithm[47], which is a different paradigm of machine learning compared to the deep convolutional neural network used here. Shallow learning features such as textures and fractals might better reflect spiral arms, and therefore increase the ability of the algorithm to detect spiral galaxies. Elliptical galaxies are more consistent in shape than spiral galaxies, which can increase the performance of deep convolutional neural networks that their accuracy depend on the consistency of the images.

Figure 4.3 shows galaxies that were classified as elliptical galaxies by the[1] catalog, but as spiral in this catalog, and by visual inspection seem spiral galaxies. Figure 4.4 shows galaxies classified in[1] as spiral but in this catalog as elliptical. Careful manual inspection of the images show that the galaxies in Figure 4.3 are spiral galaxies, but in many of the cases

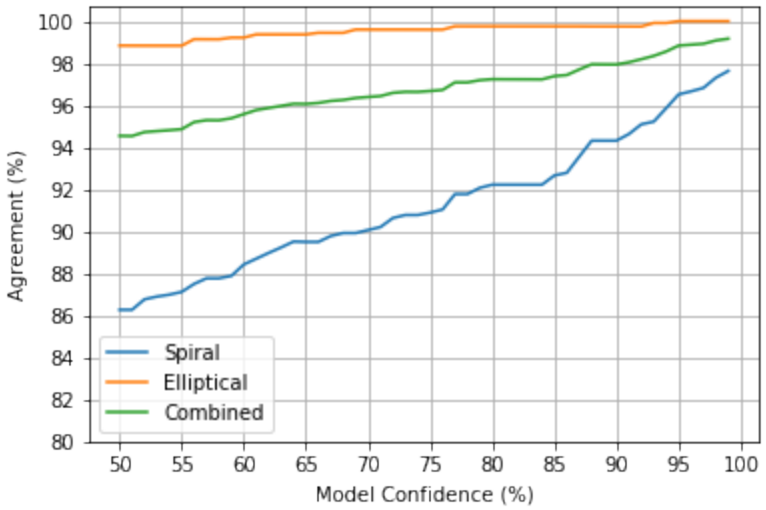


Figure 4.2: *The proportion of predicted labels that, when restricted to a minimum confidence threshold, agree with the annotations in*[1]*. For example, restricting the catalog to labels with 90% confidence or higher will have approximately 98% agreement with the annotations in*[1]

.

the arms are dim. It is therefore possible that the shallow learning algorithm used in[1] failed to detect these spiral galaxies due to the weak presence of the spiral arms.

Figure 4.4 shows that the galaxies classified as elliptical do not have clear spiral arms. However, given that the resolution of Pan-STARRS is limited, it is possible that these galaxies are spiral, as shown in Figure 4.1, where spiral arms not visible in Pan-STARRS become clearly visible using a space-based instrument with higher resolution.

Figures 4.5 and 4.6 show the same galaxies in Figures 4.3 and 4.4 imaged by SDSS, and classified as elliptical in[1]. Figure 4.5 shows that some of the galaxies are ring galaxies or interacting systems, while some of them are clear spiral galaxies that were misclassified by the algorithm. Figure 4.6 shows galaxies classified as spiral in the[1] catalog.

The comparison between the shallow learning and deep neural network shows that while the deep neural network leans towards elliptical galaxies, the shallow learning algorithm is more sensitive to spiral galaxies. The shallow learning algorithm used in[1] is described in detail in[47–49]. In summary, it computed 2883 numerical image content descriptors from each galaxy image. These image features include edges, textures, fractals, polynomial decomposition, statistical distribution of pixel intensities, and more to provide a comprehensive

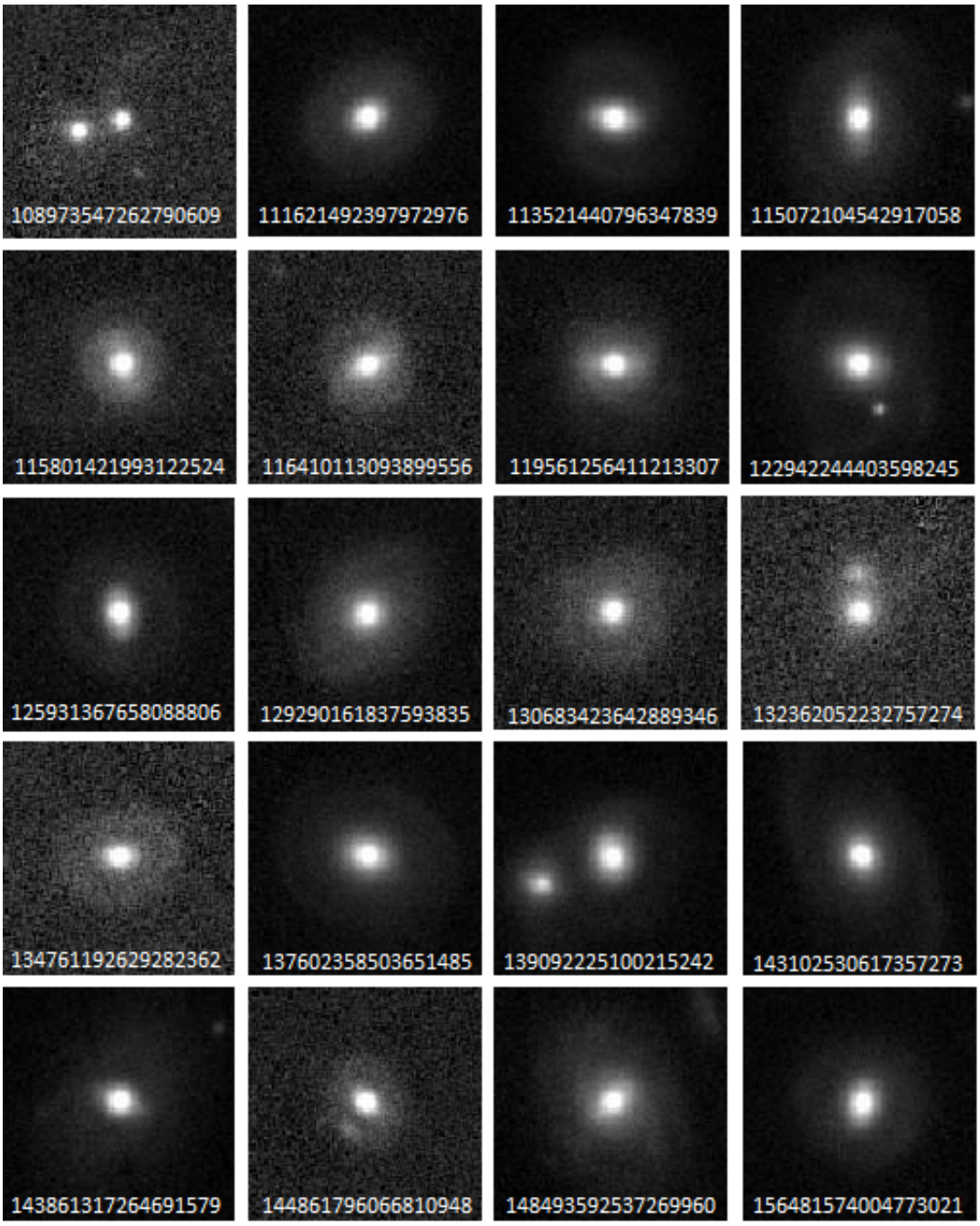


Figure 4.3: *Galaxies imaged by Pan-STARRS that were classified incorrectly as elliptical in*[1] *and as spiral in this catalog.*

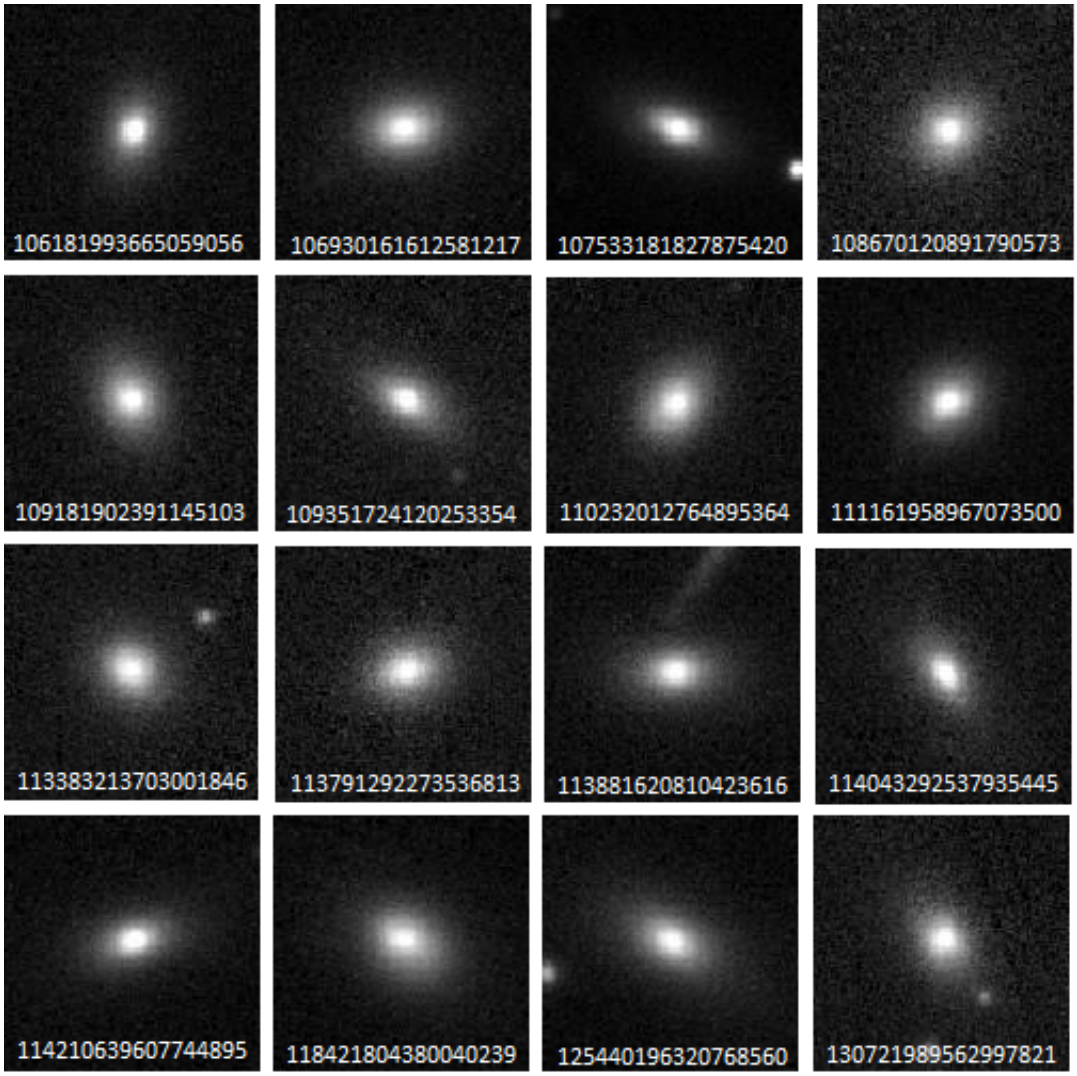


Figure 4.4: *Galaxies imaged by Pan-STARRS that were classified as spiral in*[1] *but as elliptical in this catalog.*

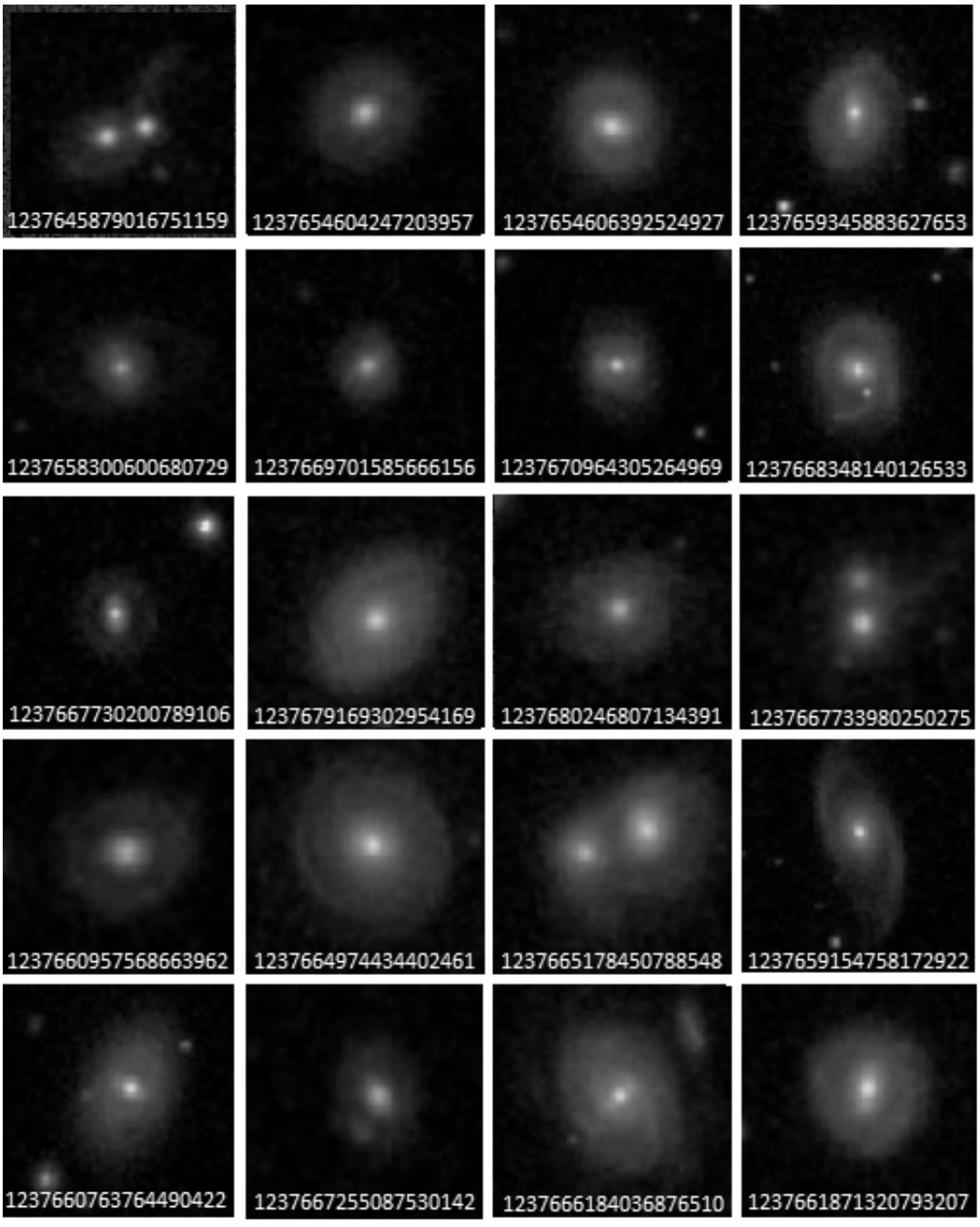


Figure 4.5: *Galaxies imaged by SDSS and were classified incorrectly as elliptical in*[1].

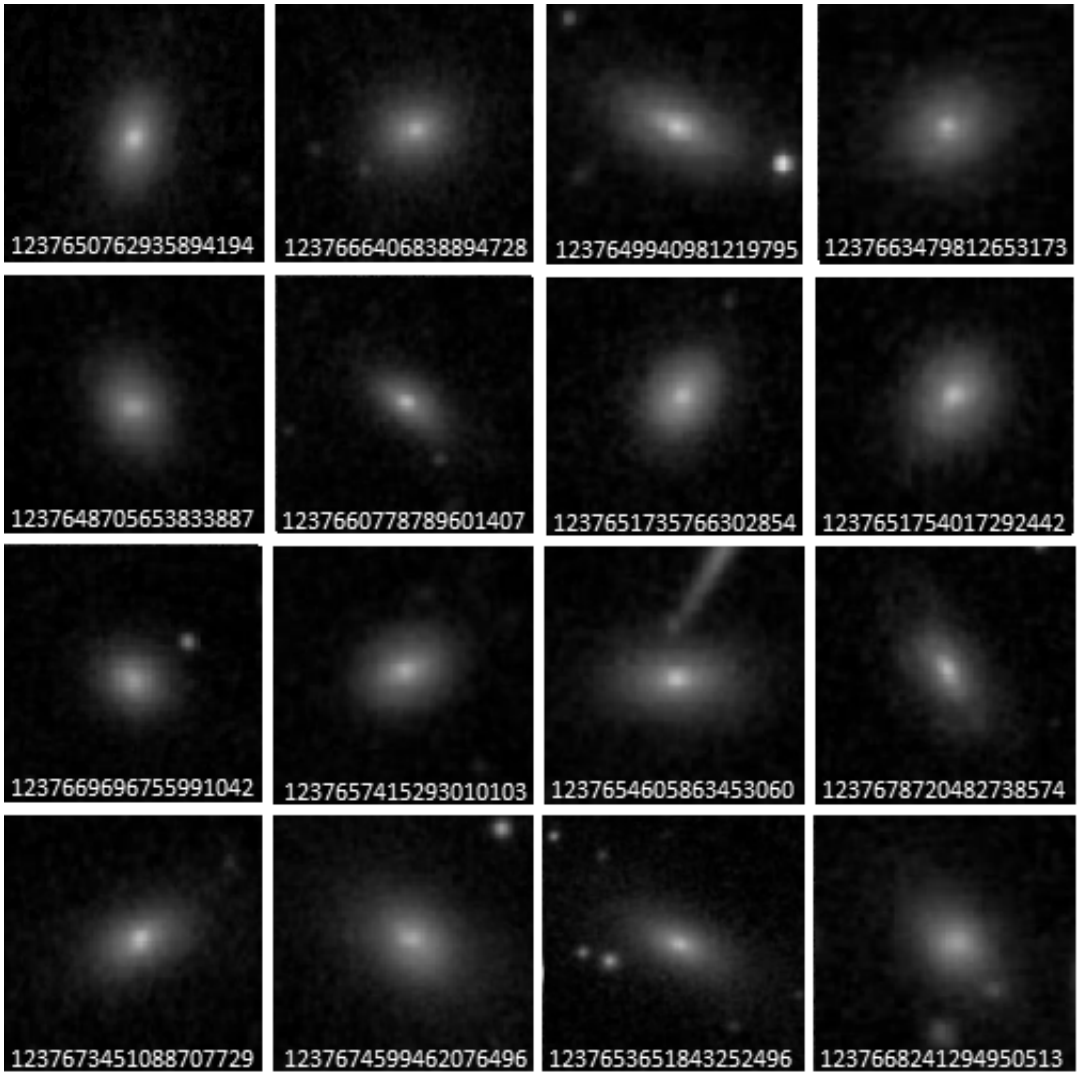


Figure 4.6: *Galaxies imaged by SDSS and were classified as spiral in*[1] *but as elliptical in this catalog.*

numerical representation of the image. These features are filtered for the most informative features, weighted for their informativeness, and then classified using an instance-based classifier. Instance-based classifiers have the advantage of handling effectively rare instances, imbalanced classes, and variations inside the classes[50–52]. Since the variability in the class of spiral galaxies is higher than the variability in the class elliptical galaxies, it is possible that the instance-based classifier used in[1] can be more accurate in the identification of spiral galaxies.

Experiments done by[53] compared the accuracy of deep convolutional neural networks to the shallow learning algorithm used in[1] for the purpose of automatic morphological classification of galaxies. The results showed that the CNN provided better accuracy compared to the older shallow learning algorithm, especially in cases of faint tidal features. That can explain some of the missclassified galaxies shown in Figure 4.3, in which the arms are visible, but are relatively dim. However, the experiments also showed that the shallow learning algorithm used in[1] was better able to handle the more complex cases, in which the CNNs struggled to make clear classification[53]. Since a collection of spiral galaxies is more likely to contain more rare objects, and since the variability among spiral galaxies is higher, an instance-based classifier such as the one used in[1] can be more effective in the identification of spiral galaxies compared to elliptical galaxies.

Table 4.2: *Examples of images that were misclassified by the model.*

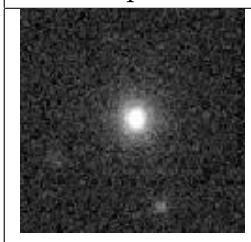
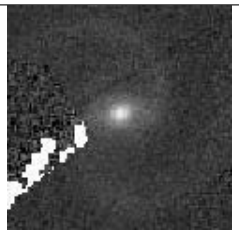
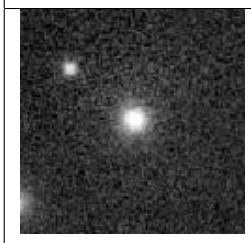
Misclassified as spiral	Confidence	Misclassified as Elliptical	Confidence
	0.5181		0.6017
	0.5614		0.6093
	0.5940		0.6455
	0.7677		0.7493
	0.9114		0.7647

Table 4.3: *Examples of images that were classified correctly by the model.*

Classified as spiral	Confidence	Classified Elliptical	Confidence
	0.9999		0.9999
	0.9988		0.8799
	0.9718		0.7568
	0.7446		0.6780
	0.5163		0.5637

Chapter 5

Conclusions

While digital sky surveys are capable of collecting and generating extremely large databases, one of the obstacles in fully utilizing these data is the automatic analysis. Image data, and in particular images of extended objects, are more challenging to analyze due to the complex nature of the image data. Here we created a catalog of Pan-STARRS galaxies classified by their broad morphology into elliptical and spiral galaxies. The likelihood of the annotations provided by the SoftMax layer allows the selection of the objects such that a more consistent catalog by sacrificing some of the galaxies that their classification is less certain. The catalog is available in the form of a CSV file at https://figshare.com/articles/PanSTARRS_DR1_Broad_Morphology_Catalog/12081144. The classification accuracy is favorably comparable to the ∼89% classification accuracy achieved when using the photometric features provided by the Pan-STARRS photometric pipeline[36].

As space-based missions such as Euclid and ground-based missions such as the Rubin Observatory are expected to generate high volumes of astronomical image data, computational methods that can label and organize real-world astronomical images are expected to become increasingly pivotal in astronomy research. Such methods can provide usable data products, and are expected to become important for the purpose of fully utilizing the power of these missions. While convolutional neural networks have demonstrated their ability to classify galaxies by their morphology, a practical solution needs to handle noise, bad data,

and inconsistencies that are typical to large real-world datasets. As shown in this paper, the deep neural network is not sufficient to provide clean data products. Instead, a combination of several algorithms that complete a full data analysis pipeline was needed. With the increasing robustness of such systems, it is also expected that protocols that combine multiple neural networks and filtering algorithms will be used to provide detailed morphological information. That information will become part of future data releases of digital sky survey.

The processing was done by first downloading the galaxy images to another server, and the analysis of the data was done on the remote server. The reason for using that practice is because the data analysis is based on solutions designed specifically for the task of galaxy annotation, and not on "standard" tasks provided by common services such as CasJobs[54]. Although the smaller JPG images were used, downloading all images still required a substantial amount of time. Using the more informative FITS images would have increased the required time to download the data by an order of magnitude, and analyzing data of much larger digital sky surveys such as the Rubin Observatory will become impractical using this practice. Therefore, future surveys might provide users not merely with certain specific predesigned tasks, but might also allow processing time for user-designed programs to access the raw data without the need to download it to third-party servers.

Bibliography

[1] Evan Kuminski and Lior Shamir. Computer-generated visual morphology catalog of ∼3,000,000 sdss galaxies. *The Astrophysical Journal Supplement Series*, 223(2):20, 2016.

[2] SG Djorgovski, RJ Brunner, AA Mahabal, SC Odewahn, RR de Carvalho, RR Gal, P Stolorz, R Granat, D Curkendall, J Jacob, et al. Exploration of large digital sky surveys. In *Mining the Sky*, pages 305–322. 2001.

[3] Kirk Borne. Virtual observatories, data mining, and astroinformatics. In *Planets, Stars and Stellar Systems*, pages 403–443. 2013.

[4] S George Djorgovski, Ashish Mahabal, Andrew Drake, Matthew Graham, and Ciro Donalek. Sky surveys. In *Planets, Stars and Stellar Systems*, pages 223–281. Springer, 2013.

[5] Nicholas Kaiser. Pan-starrs: a wide-field optical survey telescope array. In *Ground-based Telescopes*, volume 5489, pages 11–22. International Society for Optics and Photonics, 2004.

[6] HA Flewelling, EA Magnier, KC Chambers, JN Heasley, C Holmberg, ME Huber, W Sweeney, CZ Waters, T Chen, D Farrow, et al. The pan-starrs1 database and data products. *arXiv preprint arXiv:1612.05243*, 2016.

[7] Chien Y Peng, Luis C Ho, Chris D Impey, and Hans-Walter Rix. Detailed structural decomposition of galaxy images. *AJ*, 124(1):266, 2002.

[8] L Simard. Photometric redshifts and the luminosity-size relation of galaxies to z= 1. 1. In *Photometric Redshifts and the Detection of High Redshift Galaxies*, volume 191, page 325, 1999.

[9] Christopher J Conselice. The relationship between stellar light distributions of galaxies and their formation histories. *The Astrophysical Journal Supplement Series*, 147(1):1, 2003.

[10] Roberto G Abraham, Sidney Van Den Bergh, and Preethi Nair. A new approach to galaxy morphology. i. analysis of the sloan digital sky survey early data release. *ApJ*, 588(1):218, 2003.

[11] Lior Shamir. Ganalyzer: A tool for automatic galaxy image analysis. *ApJ*, 736(2):141, 2011.

[12] Darren R Davis and Wayne B Hayes. Sparcfire: Scalable automated detection of spiral galaxy arm segments. *ApJ*, 790(2):87, 2014.

[13] Preethi B Nair and Roberto G Abraham. A catalog of detailed visual morphological classifications for 14,034 galaxies in the sloan digital sky survey. *The Astrophysical Journal Supplement Series*, 186(2):427, 2010.

[14] Anthony Baillard, Emmanuel Bertin, Valérie De Lapparent, Pascal Fouqué, Stéphane Arnouts, Yannick Mellier, Roser Pelló, J-F Leborgne, Philippe Prugniel, Dmitry Makarov, et al. The efigi catalogue of 4458 nearby galaxies with detailed morphology. *A&A*, 532:A74, 2011.

[15] Chris J Lintott, Kevin Schawinski, Anže Slosar, Kate Land, Steven Bamford, Daniel Thomas, M Jordan Raddick, Robert C Nichol, Alex Szalay, Dan Andreescu, et al. 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.

[16] Chris Lintott, Kevin Schawinski, Steven Bamford, Anže Slosar, Kate Land, Daniel Thomas, Edd Edmondson, Karen Masters, Robert C Nichol, M Jordan Raddick, et al. Galaxy zoo 1: data release of morphological classifications for nearly 900 000 galaxies. *MNRAS*, 410(1):166–178, 2011.

[17] Kyle W Willett, Chris J Lintott, Steven P Bamford, Karen L Masters, Brooke D Simmons, Kevin RV Casteels, Edward M Edmondson, Lucy F Fortson, Sugata Kaviraj, William C Keel, et al. Galaxy zoo 2: detailed morphological classifications for 304 122 galaxies from the sloan digital sky survey. *MNRAS*, page stt1458, 2013.

[18] Lior Shamir. Automatic morphological classification of galaxy images. *Monthly Notices of the Royal Astronomical Society*, 399(3):1367–1372, 2009.

[19] M Huertas-Company, L Tasca, D Rouan, D Pelat, J Kneib, O Le Fevre, P Capak, J Kartaltepe, A Koekemoer, H Mccracken, et al. A robust morphological classification of high-redshift galaxies using support vector machines on seeing limited images. *A&A*, 497(3):743, 2009.

[20] Manda Banerji, Ofer Lahav, Chris J Lintott, Filipe B Abdalla, Kevin Schawinski, Steven P Bamford, Dan Andreescu, Phil Murray, M Jordan Raddick, Anze Slosar, et al. Galaxy zoo: reproducing galaxy morphologies via machine learning. *MNRAS*, 406(1):342–353, 2010.

[21] Lior Shamir, Anthony Holincheck, and John Wallin. Automatic quantitative morphological analysis of interacting galaxies. *Astronomy and Computing*, 2:67–73, 2013.

[22] Andrew Schutter and Lior Shamir. Galaxy morphology—an unsupervised machine learning approach. *Astronomy and Computing*, 12:60–66, 2015.

[23] Evan Kuminski, Joe George, John Wallin, and Lior Shamir. Combining human and machine learning for morphological analysis of galaxy images. *Publication of the Astronomical Society of the Pacific*, 126(944):959–967, 2014.

[24] Sander Dieleman, Kyle W Willett, and Joni Dambre. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society*, 450(2):1441–1459, 2015.

[25] Alex Hocking, James E Geach, Yi Sun, and Neil Davey. An automatic taxonomy of

galaxy morphology using unsupervised machine learning. *MNRAS*, 473(1):1108–1129, 2017.

[26] Evan Kuminski and Lior Shamir. A hybrid approach to machine learning annotation of large galaxy image databases. *Astronomy and Computing*, 25:257–269, 2018.

[27] Pedro Silva, Leon Cao, and Wayne Hayes. Sparcfire: Enhancing spiral galaxy recognition using arm analysis and random forests. *Galaxies*, 6(3):95, 2018.

[28] Marc Huertas-Company, JA Aguerri, M Bernardi, S Mei, and J Sánchez Almeida. Revisiting the hubble sequence in the sdss dr7 spectroscopic sample: a publicly available bayesian automated classification. *arXiv preprint arXiv:1010.3018*, 2010.

[29] Luc Simard, J Trevor Mendel, David R Patton, Sara L Ellison, and Alan W McConnachie. A catalog of bulge+ disk decompositions and updated photometry for 1.12 million galaxies in the sloan digital sky survey. *ApJS*, 196(1):11, 2011.

[30] Lior Shamir and John 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.

[31] M Huertas-Company, R Gravet, G Cabrera-Vives, PG Pérez-González, JS Kartaltepe, G Barro, M Bernardi, S Mei, F Shankar, P Dimauro, et al. A catalog of visual-like morphologies in the 5 candels fields using deep-learning. *arXiv preprint arXiv:1509.05429*, 2015.

[32] Marc Huertas-Company, Pablo G Pérez-González, Simona Mei, Francesco Shankar, Mariangela Bernardi, Emanuele Daddi, Guillermo Barro, Guillermo Cabrera-Vives, Andrea Cattaneo, Paola Dimauro, et al. The morphologies of massive galaxies from z˜ 3-witnessing the 2 channels of bulge growth. *arXiv preprint arXiv:1506.03084*, 2015.

[33] Ian Timmis and Lior Shamir. A catalog of automatically detected ring galaxy candidates in panstarss. *The Astrophysical Journal Supplement Series*, 231(1):2, 2017.

[34] Nicholas Paul, Nicholas Virag, and Lior Shamir. A catalog of photometric redshift and the distribution of broad galaxy morphologies. *Galaxies*, 6(2):64, 2018.

[35] Lior Shamir. Automatic detection of full ring galaxy candidates in sdss. *MNRAS*, 491 (3):3767–3777, 2019.

[36] A Baldeschi, A Miller, M Stroh, R Margutti, and DL Coppejans. Star formation and morphological properties of galaxies in the pan-starrs 3\pi survey-i. a machine learning approach to galaxy and supernova classification. *arXiv preprint arXiv:2005.00155*, 2020.

[37] KW Hodapp, N Kaiser, H Aussel, W Burgett, KC Chambers, M Chun, T Dombeck, A Douglas, D Hafner, J Heasley, et al. Design of the pan-starrs telescopes. *AN*, 325 (6-8):636–642, 2004.

[38] Kenneth C Chambers, EA Magnier, N Metcalfe, HA Flewelling, ME Huber, CZ Waters, L Denneau, PW Draper, D Farrow, DP Finkbeiner, et al. The pan-starrs1 surveys. *arXiv preprint arXiv:1612.05560*, 2016.

[39] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

[40] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[41] H Domínguez Sánchez, M Huertas-Company, M Bernardi, S Kaviraj, JL Fischer, TMC Abbott, FB Abdalla, J Annis, S Avila, D Brooks, et al. Transfer learning for galaxy morphology from one survey to another. *MNRAS*, 484(1):93–100, 2019.

[42] Paula Tarrío and Stefano Zarattini. Photometric redshifts for the pan-starrs1 survey. *arXiv preprint arXiv:2005.06489*, 2020.

[43] Kate Land, Anže Slosar, Chris Lintott, Dan Andreescu, Steven Bamford, Phil Murray, Robert Nichol, M Jordan Raddick, Kevin Schawinski, Alex Szalay, et al. Galaxy zoo:

the large-scale spin statistics of spiral galaxies in the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 388(4):1686–1692, 2008.

[44] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org.

[45] François Chollet et al. Keras. https://keras.io, 2015.

[46] Levente Dojcsak and Lior Shamir. Quantitative analysis of spirality in elliptical galaxies. *New Astronomy*, 28:1–8, 2014.

[47] Lior Shamir, Nikita Orlov, D Mark Eckley, Tomasz Macura, Josiah Johnston, and Ilya G Goldberg. Wndchrm–an open source utility for biological image analysis. *Source Code for Biology and Medicine*, 3(1):13, 2008.

[48] Nikita Orlov, Lior Shamir, Tomasz Macura, Josiah Johnston, D Mark Eckley, and Ilya G Goldberg. Wnd-charm: Multi-purpose image classification using compound image transforms. *Pattern Recognition Letters*, 29(11):1684–1693, 2008.

[49] Lior Shamir, Tomasz Macura, Nikita Orlov, D Mark Eckley, and Ilya G Goldberg. Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art. *ACM Transactions on Applied Perception*, 7(2):1–17, 2010.

[50] Xiuzhen Zhang, Yuxuan Li, Ramamohanarao Kotagiri, Lifang Wu, Zahir Tari, and Mo-

hamed Cheriet. Krnn: k rare-class nearest neighbour classification. *Pattern Recognition*, 62:33–44, 2017.

[51] Yuxuan Li and Xiuzhen Zhang. Improving k nearest neighbor with exemplar generalization for imbalanced classification. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 321–332, 2011.

[52] Sankha Subhra Mullick, Shounak Datta, and Swagatam Das. Adaptive learning-based k-nearest neighbor classifiers with resilience to class imbalance. *IEEE Transactions on Neural Networks and Learning Systems*, 29(11):5713–5725, 2018.

[53] Mike Walmsley, Annette MN Ferguson, Robert G Mann, and Chris J Lintott. Identification of low surface brightness tidal features in galaxies using convolutional neural networks. *MNRAS*, 483(3):2968–2982, 2019.

[54] Nolan Li and Ani R Thakar. Casjobs and mydb: A batch query workbench. *Computing in Science & Engineering*, 10(1):18–29, 2008.