UC Irvine
UC Irvine Electronic Theses and Dissertations

Title
Bias Mitigation in Galaxy Zoo Using Machine Learning Techniques

Permalink
https://escholarship.org/uc/item/7241p065

Author
Silva do Nascimento Neto, Pedro

Publication Date
2019

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, availalbe at https://creativecommons.org/licenses/by/4.0/

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA,
IRVINE

Bias Mitigation in Galaxy Zoo Using Machine Learning Techniques

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Computer Science

by

Pedro Silva do Nascimento Neto

Dissertation Committee:
Professor Wayne Hayes, Chair
Professor Aaron Barth
Professor Eric Mjolsness

2019

© 2019 Pedro Silva do Nascimento Neto

DEDICATION

To my beloved wife, Elise.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

ACKNOWLEDGMENTS

I could not have achieved this without the incredible support of my wife, Elise Silva. Thank you for always being by my side, for your words of kindness, and for doing everything in your power so I could solely focus on this work. Thank you for believing in me and being there on all the good and bad days. Thank you for always jumping headfirst into the adventure with me. From the bottom of my heart, thank you.

I am also very thankful to my family. Even over 5000 miles away, you always provided me with the support and peace I needed to endure all this time away from you. I am especially thankful to my mother, Louracy Silva, I know how much you sacrificed so I could be here, and I will never forget. To my late grandmother, Josefa Ramos, your love was an inspiration, thank you for always believing in me.

I am very thankful to my advisor, Professor Wayne Hayes, for his guidance during my years at UCI. You were a great advisor, always present when I needed your help and understanding when I needed time to work on things at my own pace. It was refreshing knowing I could count on your support whenever necessary. Thanks for creating a work environment where I felt welcomed and happy to work on during all these years.

To my office mates for the past two years, Sridevi Maharaj and Matthew Portman, thanks for all the fun conversations and enriching discussions we had. You definitely made my time at UCI very memorable.

I am grateful for being part of the SpArcFiRe research group and for the many helpful weekly discussions we had. It has been a great pleasure working with every one of you over the years, and I'll miss you dearly. I am especially thankful to those who helped me immensely with debugging and improving SpArcFiRe's code these last months before my defense: William Schallock, Alan Barkley-Yeung, and Matthew Portman.

I would also like to thank all my co-authors for their help and partnership on the works we published together over the years I was a Ph.D. student: Wayne Hayes, Darren Davis, Rae Peng, John English, Sridevi Maharaj, Leon Cao, Sepehr Akhavan-Masouleh, and Li Li.

I have also learned a great deal from the mentors I had the opportunity to work with on the industry: Matt Wolff, Stephen Bailey, and Jaison George. Thank you for showing me how important and life-changing our work can be.

I am very fortunate to have made many friends during my years at UCI. You helped me immensely during these years by always being by my side when I needed it most. Whenever I look back to these years, I will be thinking of you.

Sections of this dissertation are reprinted material from publications as they appear on the *Monthly Notices of the Royal Astronomical* and the *MDPI Galaxies*. I thank the publishers and co-authors for authorizing me to re-use this material on my dissertation.

This work was partially supported by CAPES (Coordination for the Improvement of Higher Education Personnel - Brazil) through the Science Without Borders fellowship for Ph.D. Studies and by the Miguel Velez Scholarship for Graduate Studies.

This project benefited imensely from the data provided by the Galaxy Zoo (GZ) Project and the Sloan Digital Sky Survey (SDSS).

The GZ data are the result of the efforts of the Galaxy Zoo volunteers, without whom none of this work would be possible. Their efforts are individually acknowledged at authors.galaxyzoo.org.

Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web Site is http://www.sdss.org/.

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.

CURRICULUM VITAE

Pedro Silva do Nascimento Neto

EDUCATION

Doctor of Philosophy in Computer Science **2019**
University of California, Irvine *Irvine, California*

Master of Science in Computer Science **2017**
University of California, Irvine *Irvine, California*

Bachelor of Science in Computer Science **2014**
Federal University of Piauí *Teresina, Piauí, Brazil*

RESEARCH EXPERIENCE

Graduate Research Assistant **2014–2017**
University of California, Irvine *Irvine, California*

Undergraduate Research Assistant **2010–2012**
Federal University of Piauí *Teresina, Piauí, Brazil*

TEACHING EXPERIENCE

Teaching Assistant **2016–2019**
University of California, Irvine *Irvine, California*

PROFESSIONAL EXPERIENCE

Pinterest Labs Research Intern **2018**
Pinterest *San Francisco, California*

Machine Learning Software Engineering Intern **2017**
Syntiant Corp. *Irvine, California*

Data Science Intern **2015, 2016**
Cylance Inc. *Irvine, California*

REFEREED JOURNAL PUBLICATIONS

SpArcFiRe: Enhancing Spiral Galaxy Recognition Using Arm Analysis and Random Forests — **Sep 2018**
MDPI Galaxies

SpArcFiRe: Morphological Selection Effects due to Reduced Visibility of Tightly Winding Arms in Distant Spiral Galaxies — **Mar 2018**
Monthly Notices of the Royal Astronomical Society

On the Nature and Correction of the Spurious S-wise Spiral Galaxy Winding Bias in Galaxy Zoo 1 — **Jan 2017**
Monthly Notices of the Royal Astronomical Society

REFEREED CONFERENCE PUBLICATIONS

Improving Malware Detection Accuracy by Extracting Icon Information — **Jun 2018**
IEEE Conference on Multimedia Information Processing and Retrieval

REFEREED POSTERS

Modeling the Structure of BioGRID PPI Networks — **Dec 2018**
Rocky Mountain Bioinformatics Conference

On the Nature and Correction of the Spurious S-wise Spiral Galaxy Winding Bias in Galaxy Zoo 1 — **Mar 2017**
Brazilian Graduate Student Conference

SEMINAR PRESENTATIONS

Bias Mitigation in Galaxy Zoo using Machine Learning — **Oct 2019**
UCI Physical Sciences Machine Learning Nexus Bootcamp

PATENTS

Icon Based Malware Detection — **July 2019**
U.S. Patent No.: 10,354,173 B2

ABSTRACT OF THE DISSERTATION

Bias Mitigation in Galaxy Zoo Using Machine Learning Techniques

By

Pedro Silva do Nascimento Neto

Doctor of Philosophy in Computer Science

University of California, Irvine, 2019

Professor Wayne Hayes, Chair

Automated analysis of galaxy structure using machine learning had been attempted several times, but it never performed at a level where it could be reliably used, so in 2007, the Galaxy Zoo initiative, a project where anyone could assist in the morphological classification of galaxies, came to life. The project was a success, and since its inception, it inspired several other citizen projects, not only in Astronomy but also in different fields. It produced one of the most extensive labeled datasets of galaxy morphology, which is a prime in a field that has an abundance of data, but where most of it is unlabeled. This dataset of almost 1 Million Sloan Galaxies is still used today, especially for training machine learning classification models. A major concern is that this dataset contains known and measured human biases, some of which were never corrected.

In this dissertation, we explain how we trained a machine learning model that effectively removes some chirality biases using data from SpArcFiRe (a program designed to isolate and quantify arm structure in spiral galaxies), photometric data provided by Sloan Digital Sky Survey, and labels from Galaxy Zoo. We use it to detect and correct a chirality bias present when selecting a sample of spiral galaxies at any spirality threshold. We also employ a variant of this model to aid in detecting a selection bias that occurs due to reduced visibility of tightly winding arms in distant spiral galaxies.

Finally, we introduce a novel machine learning approach that combines feature vectors to perform classification by solving a linear system in the form of $A * x = b$. We achieve accuracies over 90% on both the Wine and Iris datasets. Currently, however, the method is too slow when compared with the scalability of other similar methods. We suggest possible directions for increasing its speed without compromising accuracy.

Chapter 1

Introduction

Today's Machine Learning (ML) algorithms are based on mathematical models that date back anywhere from decades (eg., artificial neural networks [McCulloch and Pitts, 1943, Hopfield, 1982]) to more than a century (eg., principal component analysis [Pearson, 1901]). However, their effectiveness usually scales with the amount of computational effort put towards their training. The result is that, due to Moore's Law [Moore, 1965], machine learning has become effective enough to be of practical use only in the past couple of decades. An average smartphone that most people carry on their pocket nowadays is orders of magnitude faster and has more memory capacity than what was available in the Apollo 11 mission that took the first humans to our moon [Kendall, 2019]. These two factors are also one of the main pillars in the ascension of Big Data. All of these circumstances combined created the perfect storm for the rise of Machine Learning.

Despite this, many fields still suffer from a lack of data. That is a big problem because most machine learning models are function approximators whose accuracy scales with the amount of training data. Fortunately, Astronomy is one of the fields that does not have such an issue. In fact, Astronomy has been plagued for years with the opposite issue: there is more

data available than people to evaluate it. In more concrete numbers, recent 3D modeling of the images from the Hubble telescope has put the number of observable galaxies in the visible universe in the order of 2 trillion [Conselice et al., 2016], which would require the lifetime of many graduate students and professional Astronomers to get a mere fraction of it properly analysed and classified. This data abundance makes it one of the best cases for the application of Machine Learning.

In order to train a robust supervised large scale ML model, a large labeled dataset is still necessary. This problem was one of the reasons that led to the creation of the Galaxy Zoo project in 2007. The idea was to have a public website where anyone could undergo training of about 5 to 10 minutes and start classifying galaxies. About 1 million galaxies from the Sloan Digital Sky Survey (SDSS) were used for this project. This initiative proved to be very successful. Within the first 24 hours, they were receiving an average of 70,000 classifications per hour, and at the end of the project, more than 150,000 people contributed over 50 million classifications. [Lintott et al., 2008, 2010]

A second problem arose with Galaxy Zoo, however. The dataset has been shown to contain several human biases, and while detecting and correcting them is not trivial, the Galaxy Zoo team has identified and described how to correct or lessen their effect of some of these biases. Lintott et al. [2008] described potential or observed biases in the human votes based on color, magnitude, resolution, surface brightness, and apparent size of the galaxy, as well as whether the image was displayed in color versus monochrome. Bamford et al. [2009] and Lintott et al. [2010] also discussed, quantified, and corrected various classification biases. Other biases include those concerning red spirals [Masters et al., 2010] and bar classification in GZ2 [Masters et al., 2011]. Many biases regarding detailed morphology, including arm winding, was done in relation to the much more detailed classifications in GZ2 [Willett et al., 2013, Hart et al., 2016] and GZ1 [Land et al., 2008]. We have a particular interest in spiral galaxy morphology and which biases affected their classification.

In this dissertation, we will present the work done on the detection, reduction, and correction of two such biases present on the Galaxy Zoo dataset. In chapter 2, we will describe in detail the ML model we used for the most significant portion of our work and the particular choices involved in model choice, training, and testing, along with a comparison with similar work. Chapter 3 describes how we went about understanding and correcting the chirality bias present on GZ1 with the aid of our ML model. In chapter 4, we will discuss how a selection bias manifested on the survey according to the pitch angle on spiral arms.

During the development of the work mentioned above, we saw time and again how some ML models are easy to be misinterpreted, especially on classification scenarios. Some of the more traditional ML models fail to identify cases where a new sample does not belong to any of the classes present on the training data or samples that lie in between classes. To that end, we developed a novel classification approach based on a linear combination of feature vectors described in chapter 5.

Chapter 2

Spiral Galaxy Recognition Using Arm Analysis and Random Forests

Automated quantification of galaxy morphology is necessary because the size of upcoming sky surveys will overwhelm human volunteers. Existing classification schemes are inadequate because (a) their uncertainty increases near the boundary of classes and astronomers need more control over these uncertainties, (b) galaxy morphology is continuous rather than discrete, and (c) sometimes we need to know not only the type of an object, but whether a particular *image* of the object exhibits visible structure. In this chapter we propose that regression is better suited to these tasks than classification, and focus specifically on determining the extent to which an image of a spiral galaxy exhibits *visible* spiral structure. We use the human vote distributions from Galaxy Zoo 1 (GZ1) to train a Random Forest to reproduce the fraction of GZ1 humans who vote for the "Spiral" class. We prefer the random forest model over other black box models because it allows us to trace *post hoc* the precise reasoning behind the regression of each image. Finally, we demonstrate that using features

The contents of this chapter are based on the paper *SpArcFiRe: Enhancing Spiral Galaxy Recognition Using Arm Analysis and Random Forests* by P. Silva, L. Cao, and W. B. Hayes, published in MDPI Galaxies 2018, 6(3), 95.

from SpArcFiRe – a code designed to isolate and quantify arm structure in spiral galaxies – improves regression results over and above using traditional features alone, across a sample of 470,000 galaxies from the Sloan Digital Sky Survey.

2.1 Introduction

Galaxies play a crucial role in our understanding of the cosmos at large. On at least three occasions in the past century, they have caused tectonic shifts in our understanding of the Universe. First, until the early 1920s, the Milky Way Galaxy was assumed to constitute the entire known Universe, with the so-called "spiral nebulae" being merely young solar systems—within the Milky Way—in the early stages of their formation. When the "Great Debate" [Smith, 1982] established spiral nebulae as separate "island universes" on par with the Milky Way, the known Universe suddenly expanded in size by many orders of magnitude, relegating humanity to live in just one among millions (now billions) of external galaxies. Several decades later, the high rotation speed of spiral galaxies revealed that they must contain far more mass than is visible in the stars [Oort et al., 1952, De Vaucouleurs, 1959], leading to the discovery of dark matter—suddenly quadrupling the amount of matter in the known Universe. Finally, in the late 20th century, the expansion rate of the universe, as measured by the motion of distant galaxies, revealed that there was even *more* "stuff" in the Universe than light and dark matter—which is now called "dark energy" [Perlmutter et al., 1999]. Today, linking the large-scale structure and evolution of the universe as a whole to the formation and evolution of individual galaxies is one of the Grand Challenges of the 21st century, as illustrated, for example, in the Illustris simulation of a large chunk of the known Universe [Nelson et al., 2015].

Galaxies typically contain billions or trillions of stars, all bound together by their mutual gravitation and that of the dark matter that dominates the mass of all galaxies [Binney

and Tremaine, 2011]. More than half of all galaxies in the local universe are *spirals*, with the remainder being either *elliptical* or *irregular* [Mihalas and Binney, 1981]. Spiral galaxies are characterized by a central bulge and two or more "arms" that emanate from and wind around the bulge; some galaxies also contain a central bar (Figure 2.1). Spiral galaxies are also sometimes called "disk" galaxies since most of the visible stars, and the arms themselves, are confined to a rotating disk.

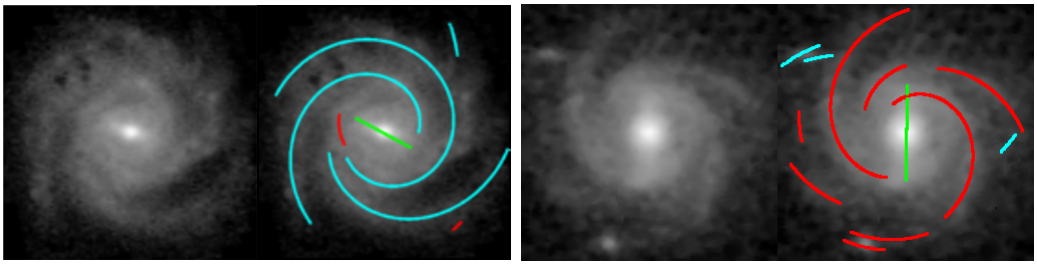


Figure 2.1: Two typical spiral galaxies along with SpArcFiRe's fit to each. These two galaxies have arms that wind in opposite directions as viewed from Earth. The directions are conventionally called S-wise (**left**) and Z-wise (**right**), since the arms wind in the same direction as those letters [Lintott et al., 2008]. SpArcFiRe [Davis and Hayes, 2012] depicts S-wise arm segments as cyan and Z-wise arm segments as red, while bars are depicted in green.

Despite decades of research, the formation of the arms of spiral galaxies is still not well understood [Binney and Tremaine, 2011, Sellwood, 2011]. At least part of the reason spiral arms defy theoretical description is that comparing *any* theory to observation requires objective *quantification*, and until recently there has been no reliable method of objectively quantifying spiral structure in images of spiral galaxies across large databases of galaxy images such as that contained in the Sloan Digital Sky Survey (SDSS, [York et al., 2000]). This changed in 2008 with the introduction of the citizen science project *Galaxy Zoo* [Lintott et al., 2008, 2010], which leveraged the Web to co-ordinate the volunteer efforts of hundreds of thousands of people to manually classify almost a million galaxy images from SDSS. While this was an amazing accomplishment, future sky surveys will overwhelm even the plethora of Galaxy Zoo volunteers: the Large Synoptic Survey Telescope (LSST) [Ivezić et al., 2019] and James Webb Space Telescope [Kalirai, 2018] will each produce far more data than the SDSS.

As a rough estimate of the amount of upcoming data, the Hubble Ultra Deep Field (HUDF) represents about 1/13,000,000 of the celestial sphere and contains about 10,000 galaxies at least 100 of which have visible structure (by our own estimate), suggesting that the entire sky contains upwards of 10^9 galaxies with visible structure at the resolution and depth of the HUDF. To quantify the morphology of this number of galaxies will require automated methods.

SpArcFiRe—the **SP**iral **ARC FI**nder and **RE**porter—is an algorithm designed to automatically extract structural information from the images of spiral galaxies [Davis and Hayes, 2012, 2014, Davis, 2014]. It was tested around a sample of 29,250 spiral galaxies from the Sloan Digital Sky Survey (SDSS), as selected by one of the PIs of the Galaxy Zoo project[1]. In the Galaxy Zoo project, arms are said to wind either S-wise, or Z-wise (cf. Figure 2.1). The fraction of humans who voted that a galaxy's arms wind S- or Z-wise are, respectively, P_S and P_Z; note that they do not need to sum to 1 since there were six choices for the humans to select from for each galaxy. We define the *spirality* of a galaxy as the sum $P_{SP} = P_S + P_Z$. The selection criterion for our 29,250 test spiral galaxies was: $(GZ1_{P_S} + GZ1_{P_Z}) > 0.8$ OR ($GZ2_{FeaturesOrDisk} > 0.7$ AND $GZ2_{NotEdgeOn} > 0.7$ AND $GZ2_{spiral} > 0.8$)[2]. Our sample also used the same magnitude limit as GZ1—17.7 in the red band.

Even though some galaxy images (e.g., elliptical galaxies or low-resolution spirals) do not have visible arms, we do not know in advance which images exhibit arms. For this reason, we run SpArcFiRe on *every* galaxy image, and our goal is to figure out when SpArcFiRe's output is meaningful, preferably using the output of SpArcFiRe itself. SpArcFiRe's job is to find spiral arms in spiral galaxies; often, it also marks noise as spiral structure. Thus, we wish to recognize when a galaxy image has visible spiral structure. Although ultimately

[1]Stephen Bamford, Personal Communication.

[2]$GZ1_{P_S}$ indicates the percent of Galaxy Zoo 1 votes for S-wise spiral. $GZ1_{P_Z}$ means the percent of Galaxy Zoo 1 votes for Z-wise spiral. $GZ2_{FeaturesOrDisk}$ means the fraction of Galaxy Zoo 2 votes that agree that there is a disk or feature in the object. $GZ2_{NotEdgeOn}$ indicates the fraction of Galaxy Zoo 2 votes that agree the disc is not face on. $GZ2_{spiral}$ indicates the percent of Galaxy Zoo 2 votes that agree that the object has spiral arms.

we hope to develop an objective, quantitative, continuous measure of galaxy morphology, for now, we focus on the simple task of reproducing what we call the *spirality* of a galaxy image: from the GZ1 catalog [Lintott et al., 2008, 2010], we define the *spirality* to be $P_{SP} = (GZ1_{P_S} + GZ1_{P_Z})$, representing the extent to which there is visible spiral structure in the image of the object. We emphasize that spirality is a measure of the *image*, not the object. We are not trying to classify galaxies; we are trying to discern if a particular image exhibits spiral structure that is unlikely to be caused by noise. For example, although elliptical galaxies should be assigned a spirality of zero, an edge-on disk *also* should be assigned a spirality of zero because spiral structure is not visible; thus, we wish to detect in both cases that SpArcFiRe's output should not be interpreted as representing spiral structure.

Since humans introduce certain types of biases into the classification scheme (for example, the chirality bias [Land et al., 2008, Lintott et al., 2010, Hayes et al., 2016]), we also wish to "dilute" such biases even though we train our method on human classifications. We do this by carefully choosing which inputs we allow our code to use. For example, we allow SpArcFiRe's measured pitch angle of spiral arms to be used as input to our regressor, but not the sign of the pitch angle [Hayes et al., 2016], thus reducing chirality discrepancies to about 1 part in 5000 [Hayes et al., 2016, Davis, 2014]. Our work follows up on existing work published in the astronomical literature [Shamir, 2009, Huertas-Company et al., 2010, Banerji et al., 2010, Kuminski et al., 2014, Dieleman et al., 2015, Ferrari et al., 2015], and extends it by either (a) using regression over classification; (b) using SpArcFiRe-derived features; or both.

2.1.1 Related Work

Arguably the most impactful and successful machine learning research published in the astronomical literature for similar tasks is Banerji et al. [2010] and Dieleman et al. [2015]. The former was one of the first to apply Machine Learning to try and reproduce the human

classifications of the GZ1 catalog [Lintott et al., 2008] and the latter focuses specifically on reproducing the vote distribution of the GZ2 catalog [Willett et al., 2013], a regression problem, exactly like the approach we explore on this chapter. The main difference here is that they are not concerned with the bias present in the dataset, so the smaller their Root Mean Squared Error (RMSE) is, "the better" their results are. In the recent Galaxy Zoo dataset releases, there has been an increased effort to eliminate human biases, but Hayes et al. [2016] have proven that these datasets, in particular, GZ1, still contain biases so there is a trade-off between lowering the RMSE of a model and avoiding the introduction of such biases on model predictions.

Banerji et al. [2010] present good results using neural networks. They classified Sloan galaxies in one of three categories: spiral, elliptical, and point sources/artifacts, using a neural network with inputs listed in Table 2.1. They found that on the entire sample of about 900,000 Sloan galaxies, they could reproduce the human GZ1 classifications in 92% of cases. Across a sample of brighter galaxies ($r < 17$), they correctly classify about 94% of galaxies. They do even better for a sample called the "Gold sample", in which galaxies are only included if the humans are themselves more than 80% confident in the classification. We do not believe the Gold sample comparison is meaningful, however, because it is crucial to know how good the machine learning model is when it *thinks* it is confident but is, in fact, mistaken, and the Gold sample completely disregards this aspect.[3]

Other interesting works published recently include Abd Elfattah et al. [2014], which uses Neural Networks and Empirical Mode Decomposition to perform galaxy classification but uses a very small test set of 108 objects so it is hard to predict how their models would fare when trying to classify a much larger set of objects, like our test set. Kuminski et al. [2014] makes a case for using "high-quality data", but we believe this will have the same

[3]In essence, comparing against the Gold sample says "look how well we do when the humans pre-select the easy ones for us!" More formally, it disregards false positives—galaxies which the prediction is confident but is actually way off.

Name	Description
C&P set	**colors and profile fitting**
$dered_g - dered_r$	$(g-r)$ color, deredenned
$dered_r - dered_i$	$(r-i)$ color, deredenned
$deVAB_i$	de Vaucouleurs fit axial ratio
$expAB_i$	exponential fit axial ratio
$lnLexp_i$	exponential disk fit log likelihood
$lnLdeV_i$	de Vaucouleurs fit log likelihood
$lnLstar_i$	Star log likelihood
$absMag\_X$	Absolute magnitudes in the 5 color bands
AM set	**adaptive moments**
$petroR90_i/petroR50_i$	concentration
$mRrCc_i$	adaptive (+) shape measure
aE_i	adaptive ellipticity
$mCr4_i$	adaptive 4th moment
$texture_i$	texture parameter

Table 2.1: Non-SpArcFiRe input parameters we used, identical to those used in Banerji et al. [2010], except for the absolute Magnitudes that also come from Sloan Digital Sky Survey (SDSS).

issues as Banerji et al. [2010]'s use of a "Gold Sample". Applebaum and Zhang [2015] uses an ensemble of Support Vector Machines to classify GZ2 galaxies achieving good results, and Ferrari et al. [2015] uses Linear Discriminant Analysis to classify galaxies from a couple different surveys—but classification is not our goal.

Finally, Kaggle.com—a website devoted to machine learning competitions—offered £10,000 (GBP) to the algorithm which best minimized the RMSE between the automatic regression scheme and the human vote distribution for *Galaxy Zoo 2*. The winning entry was a Deep Learning algorithm using convolutional Neural Networks [Dieleman et al., 2015]. It had an RMSE of about 0.07 relative to the human GZ2 vote distribution. Although this result is closer to the human votes than the result presented on this chapter, we are concerned about the professional use of deep learning techniques for several reasons:

- We would prefer a system with parameters that are understood and can be modified

by professional astronomers, and decision trees seem better suited to this task.

- Decision trees can be used to measure the quality of features[4] used to make a decision and thus are more suitable for our goals in this work. This is not the case for Deep Neural Networks, which do not yet easily provide a similar measure of feature quality.

- The most damning criticism against neural networks is that *they cannot explain their reasoning to us*. In particular, the way they make their decisions is not well understood, and research to better understand this issue is still in its infancy [Ribeiro et al., 2016, Nguyen et al., 2015]. In short, *we cannot learn from what they have learned*, or learn from how they make their decisions, because a neural net is a near-complete "black box". This is an *absolutely critical* disadvantage to a scientist who wants to know "why?" The goal of science is to *understand*, and a black box that cannot explain its rationale cannot provide us with new understanding.

For these reasons, we opt to use Decision Trees, which can be understood, dissected, and whose individual decisions can also be understood and dissected, if necessary. Understanding these decisions can teach us about galaxy characteristics and morphology in ways that "black box" machine learning algorithms cannot.

2.1.2 Regression, Not Classification, Because Galaxy Morphology Is Continuous, Not Discrete

In the real Universe, galaxy morphology is more continuous than discrete. Adjectives describing galaxies in the literature include: elliptical, spiral, dwarf, regular, irregular, giant,

[4]It is important to clarify that throughout this dissertation we will be using the term feature(s) to describe an individual measurable property or characteristic of a phenomenon being observed [Bishop, 2006] as it is commonly done in the machine learning literature as opposed to features as seen in an image-like globular clusters.

merger, peculiar, and so forth. Even among spiral galaxies, one sees adjectives such as flocculent, grand design, and barred. The range of spiral arm morphologies is enormous: one can quantify the shape, length, width, brightness profiles, color and color gradient, and contrast of individual spiral arms compared to the background disk in which they reside; one can quantify the "forking" structure of arms, how tightly they wind, and how many there are of various shapes and sizes. In addition, from the standpoint of theoretical astrophysics, there are at least three hypothesized mechanisms behind how spiral structure may form [Binney and Tremaine, 2011, Mihalas and Binney, 1981], and these mechanisms may not have clearly distinguishable visible signatures. Mergers or near impacts between galaxies can also form spiral-looking structures, providing still a 4th mechanism behind visible spiral structure. In short, although classification is a good "baby step" towards quantifying galaxy morphology, in reality, it is far too simplistic a view of the essentially continuous distribution of morphologies. As such, regression is the obvious next step in quantifying galaxy morphology.

As we will outline below, our goal is to isolate spiral galaxies and study their morphology. As such, when presented with the image of a random galaxy, our first question—the answer to which we are attempting to quantify in this chapter—is simply, "to what extent does this galaxy exhibit spiral structure?" If it displays significant spiral structure, then we are interested in knowing more. If not, we can move on to the next image... **but** we are also interested in allowing the user to choose a threshold defining the "extent" to which spiral structure is visible. This extent is the spirality, or P_{SP} measure, we are attempting to quantify.

All of the works presented in Section 2.1.1 provide good accuracies ($\geq$than 90%) but, except for Dieleman et al. [2015], they are performing classification rather than regression. There have been tremendous advances in Machine Learning towards improving classifiers, and most of these papers make use of those techniques, but that is not the goal of our work. Whereas in classification one is concerned in finding a line that best separates two or more classes (in this

scenario, spiral and non-spiral galaxies), in regression, we seek to learn about the underlying distribution, in this case, how to quantify the extent to which a galaxy image displays spiral structure[5], and at present the GZ votes are the best way to do that. Usually more information is gleaned from a continuous distribution than a discrete classification—in particular a user of the output can choose a confidence threshold themselves for classification that is more suitable for a certain task rather than relying on the table creator's subjective determination of where that threshold should lie. Peng et al. [2018], for example, used regression for a task where they needed to analyze how spirality prediction degraded as a function of image quality, a task for which classification gives limited information.

2.2 Methods

We are mostly concerned with correctly predicting spirality (the extent to which an image of a galaxy displays visible spiral structure) for images of galaxies, in which spiral structure is visible, that have a reasonably high resolution. In particular, since SpArcFiRe is designed to discern spiral structure in disk galaxies, we are most interested in isolating disk galaxies in which spiral structure is visible. By a judicious eyeball study of images at the low end of resolution, we have subjectively determined that spiral structure is invisible in Sloan galaxies if the full major axis of the observable image is less than about 13 pixels, so we ignore any galaxy smaller than this. This is similar to the cutoff of 4.5 arcseconds Petrosian radius used by the GZ1 team for galaxies with visible structure [Lintott et al., 2008]. Also following GZ1, we cut off galaxies dimmer than magnitude 17.7 in the R band. This leaves about 470,000 Sloan galaxies.

We created models using Weka [Hall et al., 2009] which provides many machine learning algo-

[5]High spirality is a strong indicator of a galaxy being spiral, but it's not a *necessary* condition. Galaxies with low spirality may be edge-on spirals, ellipticals, low-resolution spirals, or even disk galaxies without spiral structure, such as the Sombrero Galaxy.

rithms, an easy-to-use interface, and the ability to create sophisticated standalone command-line classifiers and regressors once the model has been trained. Weka provides, among many algorithms, a Neural Network algorithm, and a Random Forest (RF) algorithm. Neural Networks have been used with success in similar tasks like the convolutional model used by Dieleman et al. [2015]. These models excel in tasks where the input is spatially or temporal correlated like images or audio, so we briefly used the Neural Network algorithm to roughly reproduce the results of Banerji et al. [2010], having downloaded the same data they used from the Galaxy Zoo 1 survey [Lintott et al., 2008, 2010], which was a treated sample of the Sloan Digital Sky Survey Data Release 6 (SDSS DR6) [Adelman-McCarthy et al., 2008]. Since our machine learning algorithm uses the data only after SpArcFiRe has processed it, we found that Weka's Random Forest model had a lower RMSE, and as described in the previous section, a Random Forest model (described below) makes decisions that are easier for us to dissect and learn from. For the most advanced tasks, we recreated the same random forest models using Julia [Bezanson et al., 2017]; the results using Weka and Julia are virtually identical since the underlying mechanisms are the same.

To provide context, we explain the general idea of random forests. The "forest" part refers to a set of decision trees. Each decision tree has a set of input parameters. At each level of the tree, one asks if a particular parameter is in a specific range. For example, one level of the decision tree may ask if the galaxy has an absolute magnitude brighter than 18; another level may ask if it has a color redder than 0. The tree can be very deep, and once we arrive at a leaf node, we have a set of galaxies that satisfy an exact set of characteristics across the parameters that lie along the decision path to that node. The process of optimizing the decision tree is beyond the scope of this dissertation, but the *goal* is to optimize the leaf nodes to precisely define whatever output characteristic we are trying to reproduce. In our case, we are trying to reproduce the GZ1 human vote distribution. For example, one leaf may represent all galaxies where the human votes for (elliptical, spiral, other) are close to (0.80, 0.19, 0.01). This helps us determine what characteristics lead a decision tree to its

final regression values for each class.

The "random" part of a random forest refers to the fact that each decision tree's input parameters are chosen *randomly* from a larger set of input parameters provided by the user. The number of parameters to use for each tree is itself an integer parameter (fixed, in our case), as is the number of trees to use. Each tree effectively constitutes an "expert" in morphology quantification using its chosen set of parameters, and the forest is then a "mixture of experts", in which a voting mechanism is used to come up with the final output. A mixture of experts generally results in a much better prediction than a single tree trained on all parameters because the signal of each expert reinforces all the others, while the noise of the experts tends to cancel each other out (Sibley [2012] provides an excellent introduction to this idea).

Figure 2.2 is a simple example of a two-parameter decision tree. In this example, we will apply it only to galaxies that are clearly either spiral or elliptical. However, rather than a discrete classification, our goal is to provide just one number for each galaxy image: the extent to which it exhibits spiral structure. We use two familiar parameters: color and absolute magnitude. It is well known that elliptical galaxies tend to be both brighter and redder than spirals. Given a training set of galaxies that are truly either spiral or elliptical and given the colors and magnitudes of each, we perform the following set of operations to generate a two-parameter decision tree:

- Compute the mean magnitudes M_s, M_e for spirals and ellipticals, respectively.
- Compute the mean colors C_s, C_e for spirals and ellipticals, respectively.
- Compute a threshold color T_C intended to separate spirals from ellipticals; we will simply use the midpoint $T_C = (C_s + C_e)/2$.
- Similarly, compute a threshold magnitude $T_M = (M_s + M_e)/2$.

- Now, for each galaxy, first ask which side of the threshold its color is on, and then ask which side of the threshold its magnitude is on.

- This bins each galaxy into one of four leaf nodes, as in Figure 2.2.

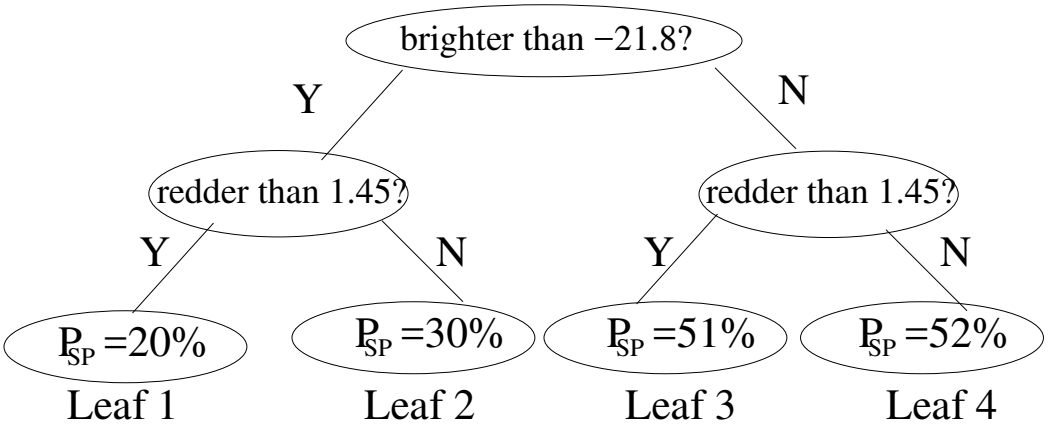


Figure 2.2: A simple 2-level, 2-parameter decision tree. The value of P_{SP} in each leaf node is the average extent to which training-set galaxies in that node display spiral structure, as measured by the GZ1 vote fraction for $P_S + P_Z$. That value is then used as the predicted P_{SP} for galaxies in the non-training set that land in that node. With proper optimization (beyond the scope of this dissertation), larger trees with more features can produce more accurate predictions of P_{SP}, P_{EL}, etc.

As we can see, the results are correlated with the correct answers but not strongly so: dim, blue-ish galaxies only have a slightly greater than 50% chance of being spiral, although it is true that bright, reddish galaxies are correctly measured as unlikely to be spiral.

The advantage of this method over other more opaque methods such as Neural Networks, or SVM, is that once we get to a leaf node of the decision tree, we know *exactly* why each galaxy is in that node—we can follow the decisions down the tree and build a boolean expression that describes all the galaxies at that node. If we wish, we can then ask ourselves if the decision path makes sense; we can look at the galaxies at that node, and ask if they form an interesting set. This kind of detailed, explicit decision-making analysis is (currently) absent in other machine learning methods although very recent work has begun to study this

question [Ribeiro et al., 2016, Nguyen et al., 2015], and is what allows us to be more confident that biases are unlikely to creep into the regression scheme.

2.3 Results

2.3.1 Features, Trees, and Forests

Referring again to Figure 2.2, we see that a 2-feature, 2-level decision tree based only on color and magnitude does a reasonable, though not great, job at separating spiral from non-spiral galaxies. Table 2.2 quantifies this in more detail, and provides a pair of features that gives better performance, although still only about 75% "correct" in total. In addition to the features listed in Table 2.1, Table 2.3 introduces quantities that are output by SpArcFiRe and used as additional input features for our model. Using these features, and in addition allowing the number of trees in the forest to increase, gives better results as summarized in Table 2.4 and described in more detail below.

We now explore in depth how many total trees should be in the forest, and how many randomly chosen features should be in each tree. Recall that the *total* number of features is fixed (and is 101 in our case), but that each decision tree chooses some random set of features. We will look at how both of these parameters change the results.

Presumably, the more features a particular tree uses, the better that tree will be, although more care needs to be put into training these models to avoid overfitting. Figure 2.3 plots the Pearson correlation between the GZ1 human vote proportion for P_{SP}, and our reproduction of that proportion, as a function of how many features are used by each tree. As can be seen, increasing the number of features used by each tree generally results in improvement. However, since each tree chooses a *random* subset of features, there is a bit of noise in the curve. It becomes less obvious that there is an improvement beyond about 35 features per

pair	p1	p2	p3	p4	correctAll	SPcapture	SPcontam
$rest_{ug}, MI$	0.20	0.30	0.52	0.51	65.7%	42.0%	48.7%
$deVAB_i, MRrCc_i$	0.12	0.65	0.38	0.85	74.3%	32.4%	14.8%

Table 2.2: Classification results for two-level, two-feature trees like that in Figure 2.2. Columns p_i represent the average fraction P_{SP}, across galaxies in leaf node i, of GZ1 humans who voted that object to be a spiral galaxy, across the training set. This value is then the assigned P_{SP} for any non-training-set galaxy placed in this leaf node. *correctAll*: assuming $P_{SP} > 0.5$ represent a positive spiral classification, the percentage across all galaxies of correct classifications; *SPcapture*: the fraction of true spirals that are captured by this classification scheme. *SPcontam*: the fraction of galaxies classified as spiral that are incorrectly classified. **Top row**: exactly the tree of Figure 2.2. **Second Row**: a pair that arguably performs better because it has a higher total correct classification, primarily because it has far less contamination of non-spirals, even though it has a smaller capture fraction. It demonstrates that we can have 75% correct classifications even with just a two-parameter, two-level tree. See Table 2.1 for the meaning of the input variables.

tree, so we use 35 in our final results below. We also see that the entire curve moves up as the number of trees in the forest increases.

Similarly, we would expect that as the number of trees in the forest is increased, the result would get better. Essentially, as more "experts" weigh into the decision, the better the results should be. Figure 2.4 demonstrates that this is indeed the case. Furthermore, unlike the case of choosing features, the curve is pretty much monotonically increasing: it seems that more trees are always better [Sibley, 2012]. In our results below, we use 150 total trees, each using 35 features out of our total set of 101 combined features from SpArcFiRe and SDSS.

2.3.2 Adding SpArcFiRe Features

As stated before, our goal is to test if adding SpArcFiRe's features (cf. Table 2.3) to the set of input features will improve our ability to reproduce the vote distribution of GZ1 for spiral galaxies, so instead of classification, we are using regression to achieve our results. This means that, rather than having a galaxy falling under a class (spiral, elliptical, and other),

Feature	Description
bar_scores	SpArcFiRe's various bar detection scores
avg(abs(pa))-abs(avg(pa))	pitch angle-weighted chirality consistency across arms
numArcs > L	SpArcFiRe's count of arms of various lengths
numDcoArcs > L	SpArcFiRe's count of dominant-chirality-only arms of various lengths
totalNumArcs	total number of arcs found by SpArcFiRe
totalArcLen	total length of all arcs found by SpArcFiRe
avgArcLen	average arc length across arcs found by SpArcFiRe
arcLenAtXX%	length of arc at XX = 25%, 50%, and 75% of total length of arcs
rankAtXX%	arc rank at XX = 25%, 50%, and 75% of total length of arcs (see text)
bulgeAxisRatio	axis ratio of bulge, if present
diskAxisRatio	axis ratio of entire galaxy image; values $\lesssim 0.2$ suggest an edge-on spiral rather than elliptical
disk/bulgeRatio	disk to bulge ratio
diskBulgeAxisRatio	ratio of (diskAxisRatio) / (bulgeAxisRatio)
gaussLogLik	Gauss Log Likelihood of ellipse fit
likelihoodCtr	likelihood of the center of the ellipse fit
abs(pa_alen_avg)	average pitch angle of arms, length-weighted
abs(pa_alen_avg_DCO)	average pitch angle only of arms of dominant chirality
twoLongestAgree	chirality agreement of two longest arcs (Boolean)

Table 2.3: Outputs from SpArcFiRe that are used as input features for our model, in addition to those from Table 2.1. See Davis and Hayes [2014] for full descriptions of these parameters. Parameters labeled "DCO" are measured only across arcs of "dominant chirality only"—that is, arcs of the "wrong" chirality, which are likely to be noise, are not included. The parameter "arcLenAt50%" means: lay arcs end-to-end sorted longest to shortest, resulting in a line of total length L, and measure the length of the arc that lies at the point L/2 along the line. If the arms are short at L/2, then short arcs tend to suggest the galaxy is either flocculent or non-spiral, whereas a long arc at this point suggests a more grand-design spiral. The "rankAt50%" feature is similar, except this is the integer rank of the arc touching the L/2 point. If the ratio ((diskAxisRatio) / (bulgeAxisRatio)) is close to 1, it is suggestive of an elliptical galaxy, whereas if this ratio is significantly less than one it suggests a spiral galaxy (since the bulge axis ratio tends to be 1 from any vantage point, but not so for the disk.)

our output is the extent to which an image of a galaxy displays spiral structure. This value, between 0 and 1, is represented by the percentage of humans that agree that a certain galaxy has visible spiral structure. We represent this idea by making the sum of GZ1 values $P_{SP} = P_S + P_Z$ as our *target* variable, and this is what we train our machine to reproduce—while simultaneously striving to eliminate the known P_S bias [Land et al., 2008, Hayes et al., 2016].

We built three different random forest models using the same hyperparameters (150 total

Total #Features	Features/Tree	# of Trees	Feature(s)	Correct
1	1	1	Color only	65%
1	1	1	Magnitude only	65%
2	2	1	color + mag	75%
3	3	1	col + mag + arcs	85%
7	7	1	various	∼90%
35	7	10	various	∼95%
35	7	50	various	∼97%
101	7	100	various	99.9%

Table 2.4: Illustration of how the results of the classification improve as we allow more complex trees, and larger forests. The lines in which the feature(s) are listed as "various" means that, choosing k features (k from the first column), it is not difficult to find k features that provide a correctness similar to the last column. Of course not *every* selection of k features will result in that correctness; correctness was enhanced when the feature set included at least some high-quality features (cf. Section 2.3.3).

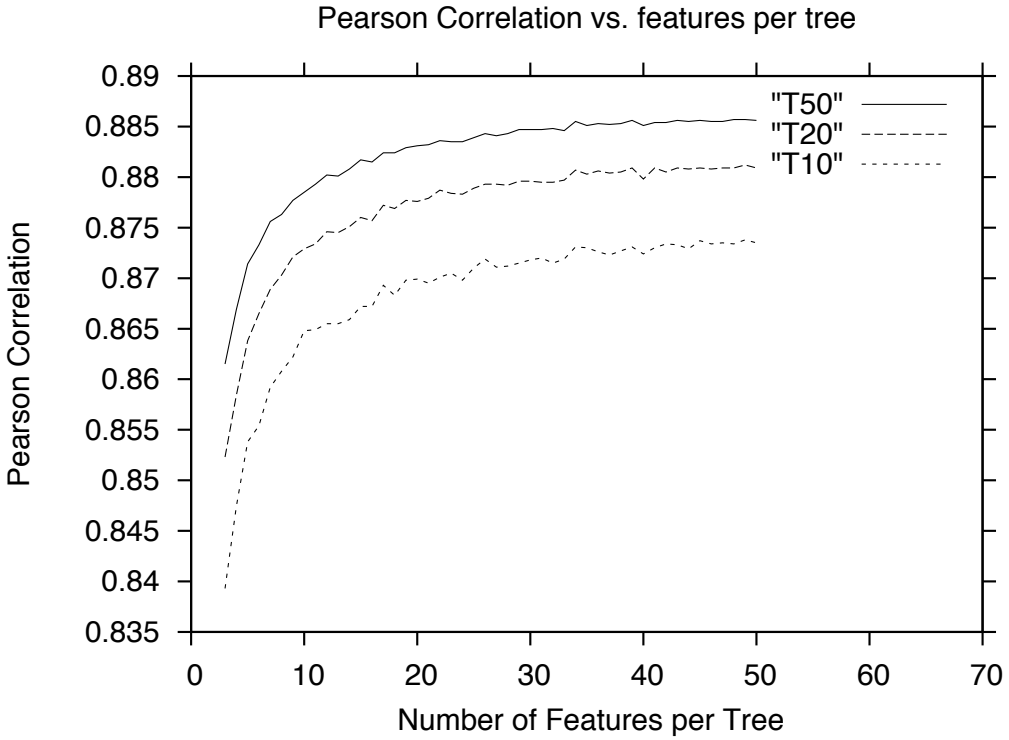


Figure 2.3: The Pearson correlation between the fraction of GZ1 humans voting for spiral, and our reproduction of that vote fraction, as a function of the number of features per tree that are chosen at random from the entire feature set. The three curves correspond to the cases where the total number of trees is 10, 20, or 50.

trees, each using 35 features) but with different feature sets. Model 1 uses only SDSS features, Model 2 uses only SpArcFiRe features, and Model 3 used both sets of features (this is the

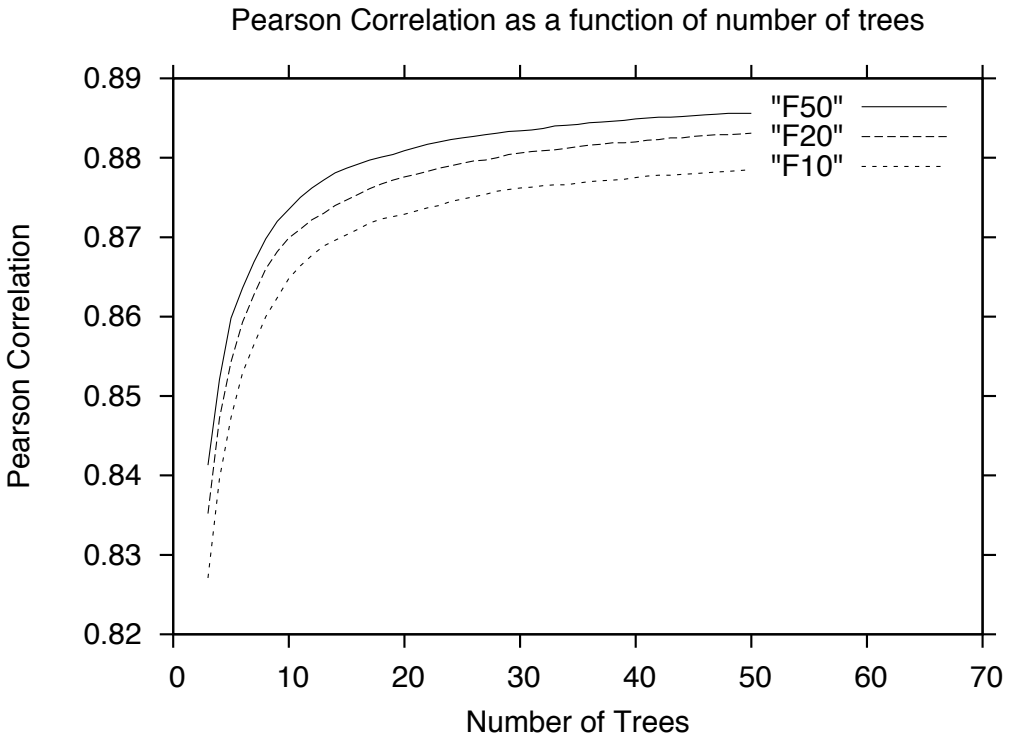


Figure 2.4: Similar to Figure 2.3, the Pearson correlation between the fraction of GZ1 humans voting for spiral, and our reproduction of that fraction, as a function of the number of trees in the forest. The three curves also show how the results change when the number of features per tree is 10, 20, or 50.

model we discuss throughout the chapter). We ran a 10-fold cross validation[6] [Refaeilzadeh et al., 2016, James et al., 2014, Kuhn and Johnson, 2013] in each one of those to get a more accurate measure of how those sets performed individually. Model 1 had a mean RMSE of 0.1518, while Model 2 had a mean RMSE of 0.1522, and both had Pearson correlations of about 0.85. Comparing these Pearson correlations using a Student's t-distribution test [Rahman, 1968], we find that they do not differ significantly (both are approximately 170 (sic) standard deviations away from a random distribution)—not surprising since the RMSE's differ only in the 4th digit. However, by combining the two feature sets, we get Model 3, which had a mean RMSE of 0.1404 and Pearson correlation of about 0.88. This RMSE dif-

[6]K-Fold cross validation is a method for measuring the quality of a learning algorithm by splitting the data into K buckets, training the algorithm in $K-1$ of these buckets and testing in the holdout bucket. We do this K times, each time holding out a different bucket and we report the average accuracy as the final accuracy of a model in that dataset. Finally, we choose the best out of the K models as our final model, having demonstrated that all the other K-1 models also perform reasonably well.

fers from the other two in the second digit, and, according to the Student's t-distribution, it differs from a random distribution by approximately 220 standard deviations—50 standard deviations further from random than both Models 1 and 2, meaning the p-value of Model 3's Pearson correlation is *many orders of magnitude* more statistically significant than the Pearson correlation of Models 1 and 2. These statistical tests demonstrate that SpArcFiRe features alone are about as good as SDSS features alone at predicting spirality, while the two combined significantly decrease the model error. This is already an indication that there is valuable information in both feature sets. We will explore feature quality in Section 2.3.3.

The 10-fold cross validation RMSE of Model 3 was 0.1404 but the the best model had an RMSE of 0.1374; this best one is the model we use for the remainder of the chapter (note that neither Model 1 nor 2 ever achieved such a low RMSE in any model we tested.). Table 2.5 shows our results, using both SDSS and SpArcFiRe's features, for the test set in a 10×10 confusion matrix. Each row represents one of 10 bins holding galaxies in which a certain fraction of humans voted for that value of spirality; each column represents one of 10 identical bins containing the predicted spirality from our method. Thus, "correct" predictions (within 10% of the human vote) appear along the diagonal of the matrix. The first off-diagonal elements represent where our prediction was 10%–20% off, etc.; the far corners represent our worst predictions.

Notice that our model has high sensitivity and specificity rates, which means that when it predicts that an object is spiral or non-spiral with high confidence, the prediction is very likely correct. For example, let's look at the case where our model predicts that an object is spiral with more than 90% of confidence, the penultimate column of the Table 2.5. If we consider a decision for spiral or non-spiral object being made above or below the 0.5 threshold, this gives us a sensitivity rate of more than 98%. The similar case happens for non-spiral predictions with more than 90% confidence (where $P_{SP} \leq 0.1$), the second column of the same table, in which, also considering a 0.5 threshold for a decision, our model gets

$P_{SP} \backslash RF_{SP}$	0.0–0.1	0.1–0.2	0.2–0.3	0.3–0.4	0.4–0.5	0.5–0.6	0.6–0.7	0.7–0.8	0.8–0.9	0.9–1.0	TOTAL
0.0–0.1	**21,238**	5042	1512	522	169	45	17	1	1	1	**28,548**
0.1–0.2	2471	**1803**	1145	594	254	111	28	5	0	0	**6411**
0.2–0.3	509	761	**668**	522	233	96	42	14	7	0	**2852**
0.3–0.4	184	345	424	**348**	209	120	58	23	5	2	**1718**
0.4–0.5	62	187	243	268	**191**	93	56	33	8	1	**1142**
0.5–0.6	47	128	215	200	188	**145**	76	38	11	3	**1051**
0.6–0.7	30	99	149	175	138	113	**76**	53	19	6	**858**
0.7–0.8	23	70	91	123	120	121	108	**80**	42	7	**785**
0.8–0.9	21	38	89	107	144	136	148	134	**80**	24	**921**
0.9–1.0	4	32	53	87	129	151	211	257	289	**303**	**1516**
TOTAL	**24,589**	**8505**	**4589**	**2946**	**1775**	**1131**	**820**	**638**	**462**	**347**	**45,802**

Table 2.5: Confusion Matrix of the best of our 10-fold cross-validated models. The rows represent the number of objects that have a GZ1 spirality between a specific interval. The columns represent how many of those our Random Forest predicted in the same and different intervals. Notice that these numbers are only for the test set, thus a total of 45802 objects, which represent a more accurate measure of how our Random Forest would perform in real-world situations. The same data are depicted pictorially in Figure 2.5.

more than 99% specificity rate.

Since we are doing regression, a more global way to visualize our results is to look at the correlation between our results and the GZ1 votes. Figure 2.5 shows a scatter plot where, for each galaxy, the x-axis represents the human vote fraction and the y-axis is our algorithm's prediction of the same value, across all 470,000 Sloan galaxies. Each red point represents one galaxy, and its distance from the $y = x$ line depicts our level of agreement. The clustering around the line $y = x$ suggests good agreement with GZ1. It is also notable that more than 98% of the galaxies have $|x - y| \leq 0.3$ and approximately 95% of the objects fall under $|x - y| \leq 0.2$.

For the sake of comparison with existing classifiers, we can turn our regressor into a classifier by choosing a boundary for the decision. If we choose that boundary to be 0.5, we will make our decision based on the majority vote, which mimics the choice of the Galaxy Zoo researchers in some releases [Lintott et al., 2008]. That would give our regressor an accuracy of approximately 93% based on the test set presented in Table 2.5, comparable to existing

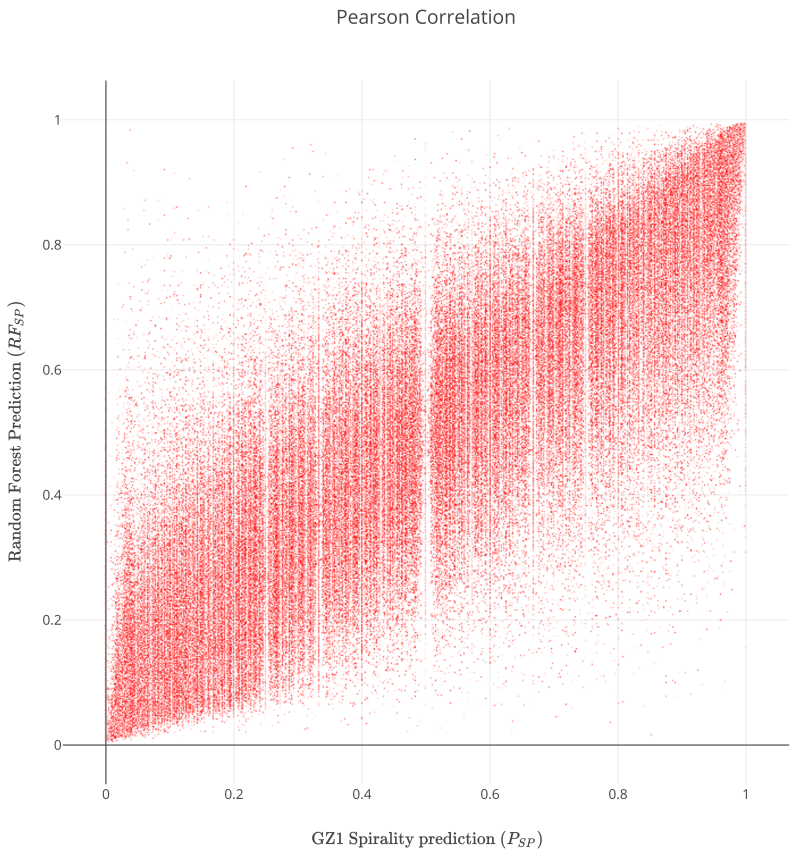


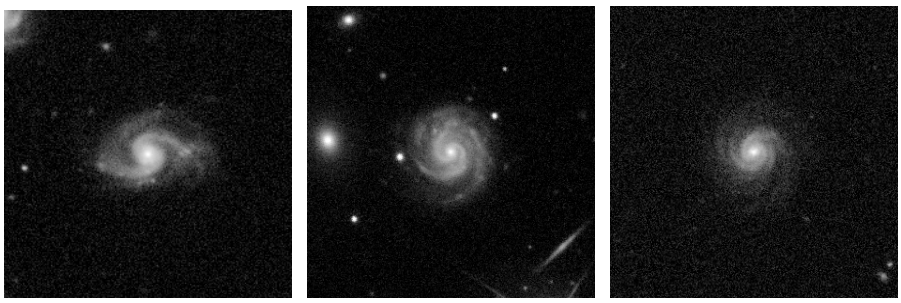
Figure 2.5: Scatter plot of both training data and test data depicted in Table 2.5: our predicted spirality (vertical) vs. the fraction of GZ1 humans voting for spiral (horizontal). The points cluster around the line $y = x$, depicting good agreement. Additionally, more than 98% of the galaxies have $|x - y| \leq 0.3$ and approximately 95% of the objects fall under $|x - y| \leq 0.2$. The vertical white lines appear because the fraction of human voters is a ratio of discrete integers.

methods.[7]

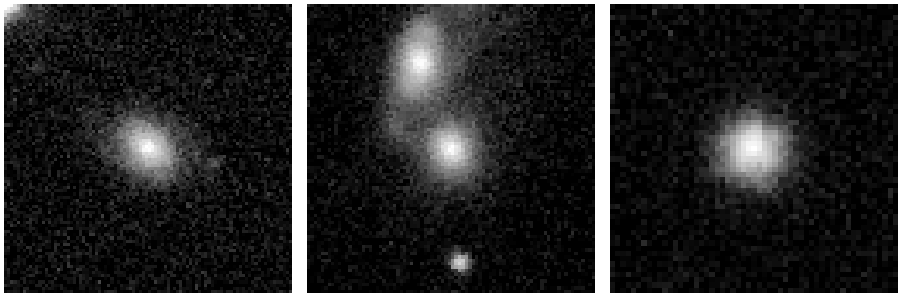
In Figure 2.6, we show some of our correctly classified objects. Those objects were cases where our model had a high agreement with the classifications provided by GZ1, and looking at the images we understand why. In Figure 2.6a, we display some of the spiral objects detected, while in Figure 2.6b we show the non-spiral objects detected, which belong to the

[7]A higher accuracy can be achieved if we use a boundary below 0.5. Note that, since there were six choices in GZ1, any vote receiving more than 1/6 of the votes can be a winning vote; for example, a vote of 40% could be considered a classification if all the other choices had less than 40% of votes. It is also possible to get better accuracy, using the same features, if we build a classifier rather than a regressor, but that is outside the scope of this work.

other classes of objects in GZ1: Elliptical, Merger, and Artefact, respectively.



(**a**) Spirals correctly detected by our model.



(**b**) Non-spirals correctly detected by our model.

Figure 2.6: Examples of images that had a high agreement of classification by both, our Random Forest Model and the GZ1 humans. (**a**) shows images of spirals where $P_{SP} \geq 0.90$ AND $|P_{SP} - F_{SP}| \leq 0.02$. Their SDSS IDs are, respectively, 1237654030325973054, 1237662306733916433, and 1237668298219847857; (**b**) shows images of non-spirals where $P_{SP} \leq 0.10$ AND $|P_{SP} - F_{SP}| \leq 0.02$. Their SDSS IDs are, respectively, 1237663529721397476, 1237661465447497940, and 1237661949201154382.

It is important to understand what is going on in the small fraction of objects for which our method performs poorly. These objects are in the opposite corners of the off-diagonal in Table 2.5: four objects from the bottom left corner and 1 from the top right corner. These are that objects with a high disagreement: $|x - y| \geq 0.9$. From our total of 45,802 galaxies in the test set, only five fall in this margin, and we show all of them in Figure 2.7. The top four rows depict the same problem: very faint arms that SpArcFiRe entirely failed to detect during the disk detection phase, so that it zoomed in past the arms, making it impossible for the arm detection code to find anything useful. This is a rare occurrence, and we are aware of this issue and are working on improving this specific step of SpArcFiRe [Davis and Hayes, 2012]. The object on the bottom row is clearly a merger, and arm-like features are

present, so our machine predicts a high spirality. One could argue that this is a *correct* prediction that the galaxy is not an elliptical galaxy, but the GZ1 humans correctly marked it as a merger and thus not a spiral at all. Since our machine has not been trained to detect mergers, it is unclear whether this should count as a misclassification.[8]

2.3.3 Feature Quality

Blindly adding features to a model does not guarantee that it will get better. Additional features might represent redundant information, which would not translate into more accurate classifiers for certain machine learning models, or worse, they would contribute to the curse of dimensionality [Jensen and Shen, 2009]. In order to make sure we are adding meaningful information, we further analyzed our feature set—again something that would be difficult to do with any model other than a random forest.

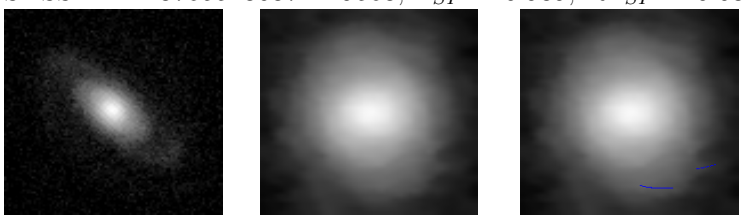
To check which features seem to be the most important overall, we created a feature ranking. As we have depicted in Figure 2.2, each node in a decision tree is a condition that splits the decision tree in two based upon a threshold in one variable. The measure used to make that decision is called impurity, and it is usually entropy for classification trees and variance for regression trees. It basically encodes how much information a particular feature, upon selection, adds to the decision process. The more outputs a feature can separate, the higher its entropy is going to be, thus decreasing the impurity of the decision tree. Thus, we compute how much each feature decreases the weighted impurity of a tree. In our case, since we are using random forests, the impurity decrease from each feature can be averaged, and the features are ranked according to this measure [Saabas, 2014].

Table 2.6 shows the top 10 features ranked by their importance along with the standard deviations of that score since this is an average over 150 decision trees. We can see that

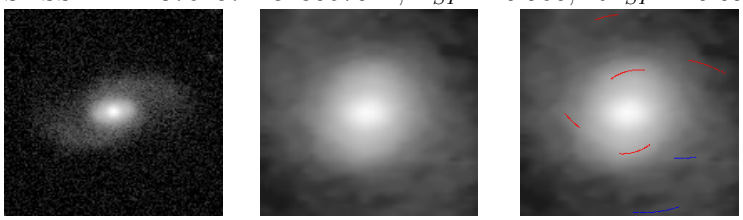
[8]One might argue that perhaps our "spirality" measure is more aptly called "non-ellipticity".

SDSS ID: 1237665330921275477; $P_{SP} = 0.905$; $RF_{SP} = 0.092$.

SDSS ID: 1237655130371129563; $P_{SP} = 0.939$; $RF_{SP} = 0.035$.



SDSS ID: 1237648722828067024; $P_{SP} = 0.960$; $RF_{SP} = 0.098$.



SDSS ID: 1237667255086219394; $P_{SP} = 0.906$; $RF_{SP} = 0.099$.

SDSS ID: 1237654948376608866; $P_{SP} = 0.0$; $RF_{SP} = 0.970$.

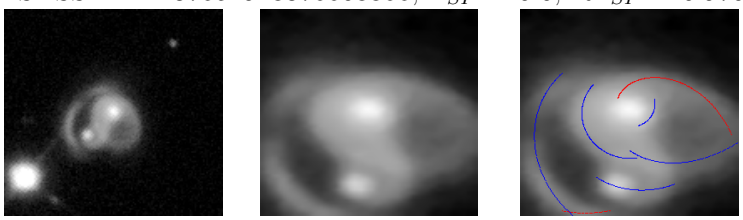


Figure 2.7: Grossly Misclassified Objects. In sets of 3, from left to right column, the images show the Original Input Image, the same image automatically cropped by SpArcFiRe, and the spiral Arcs detected on the image (if any). The SDSS Object IDs, the GZ1 Spirality prediction (P_{SP}), and our Random Forest Prediction (RF_{SP}) are shown above each trio of images. In all but the last, the problem is low-surface-brightness arms, which we know about and are working on this issue. Despite the disagreement in the 5th object, a merger, spiral structure is indeed present.

from the top 10 features, five come from SDSS and five from SpArcFiRe, suggesting again that the two feature sets contribute roughly equally to the quality of the results. The five best SpArcFiRe features are all related to the number of arcs greater or equal to a certain number of pixels, which is, not surprisingly, a strong indicator of the presence of spiral structure. Interestingly, in SpArcFiRe's favor, the best feature overall is the number of dominant-chirality-only arms equal or longer than 120, which is 30% more relevant than the most relevant feature from SDSS—far and away the most relevant feature among all features, *way* in front of the pack of other features in terms of importance.

Feature	**Score**	**Standard Deviation**
Number of dominant-chirality-only arms equal or longer than 120	0.039	0.080
Absolute Magnitude in the Z band	0.031	0.020
De-reddened magnitude in the R band	0.029	0.019
De Vaucoleurs fit axial ratio i band	0.028	0.013
Number of dominant-chirality-only arms equal or longer than 85	0.022	0.061
Number of dominant-chirality-only arms equal or longer than 100	0.022	0.057
Number of arcs equal or longer than 120	0.022	0.057
Exponential fit axial ratio i band	0.021	0.009
De-reddened magnitude in the G band	0.021	0.015
Number of dominant-chirality-only arms equal or longer than 80	0.021	0.060

Table 2.6: Top 10 best features for spirality prediction in decreasing order of importance. The standard deviation is measured across the 150 decision trees.

2.3.4 Comparison with Other Regression Methods

In Section 2.2, we argued extensively about why we prefer to use Random Forests over other methods because we want understandable models that perform regression rather than classification. Although we have explained why Random Forests are better than Neural Networks for understandability, here we test whether Random Forests compare favorably

Measure\Model	Random Forest	Ridge Regression	K-Nearest Neighbors
Pearson Correlation (PC)	0.8631	0.6753	0.7729
Root Mean Squared Error (RMSE)	0.1381	0.2495	0.1426
Mean Absolute Error (MAE)	0.0713	0.2011	0.0872
Mean Error (ME)	−0.00007	0.1789	−0.0025

Table 2.7: Measures of error and quality from models trained with both SDSS features and SpArcFiRe features. Note that these three models were each simultaneously trained from scratch on exactly the same data for this comparison, and thus the RMSE and Pearson correlations—which depend upon stochastic parameters—of this particular random forest (RF) model differs from our RF model discussed in the rest of the chapter.

against other regression methods. To that end, we trained a Ridge Linear Regressor and a K-Nearest Neighbors (KNN) Regressor using both SDSS and SpArcFiRe features set. Table 2.7 has a comparison of the results for all the models. As we can see, the Random Forest had by far the best performance out of the three models in all the measurements we used. Although the KNN model presents a comparable RMSE, its Mean Absolute Error (MAE) is over 20% worse than the random forest's MAE. The random forest also has a Pearson correlation almost 10% better than the next best model.

To make sure we aren't overfitting or have introduced different biases to the models, we also show the residual plots for all the models in Figure 2.8. For visualization purposes, we are using a random sample of 10,000 points from the test set for these plots. A model with no errors would have all the points with a residual value of zero. In a more realistic scenario, the residuals should not be either systematically high or low, meaning that they should be centered on zero throughout the range of predicted values [Bishop, 2006].

There is no apparent shape in any of the plots, other than a higher concentration of points in the $0 \leq P_{SP} \leq 0.5$ range, which is expected since the number of "ellpitical" classifications far outnumbers the spiral classifications in the GZ1 sample. A closer examination of the plots corroborates with the measures of error on Table 2.7 though. The ridge regression has higher residuals in the $0.4 \leq P_{SP} \leq 0.6$ range and the KNN model has the same issue in the $0.2 \leq P_{SP} \leq 0.5$ interval. The random forest plot, on the other hand, has a distribution more

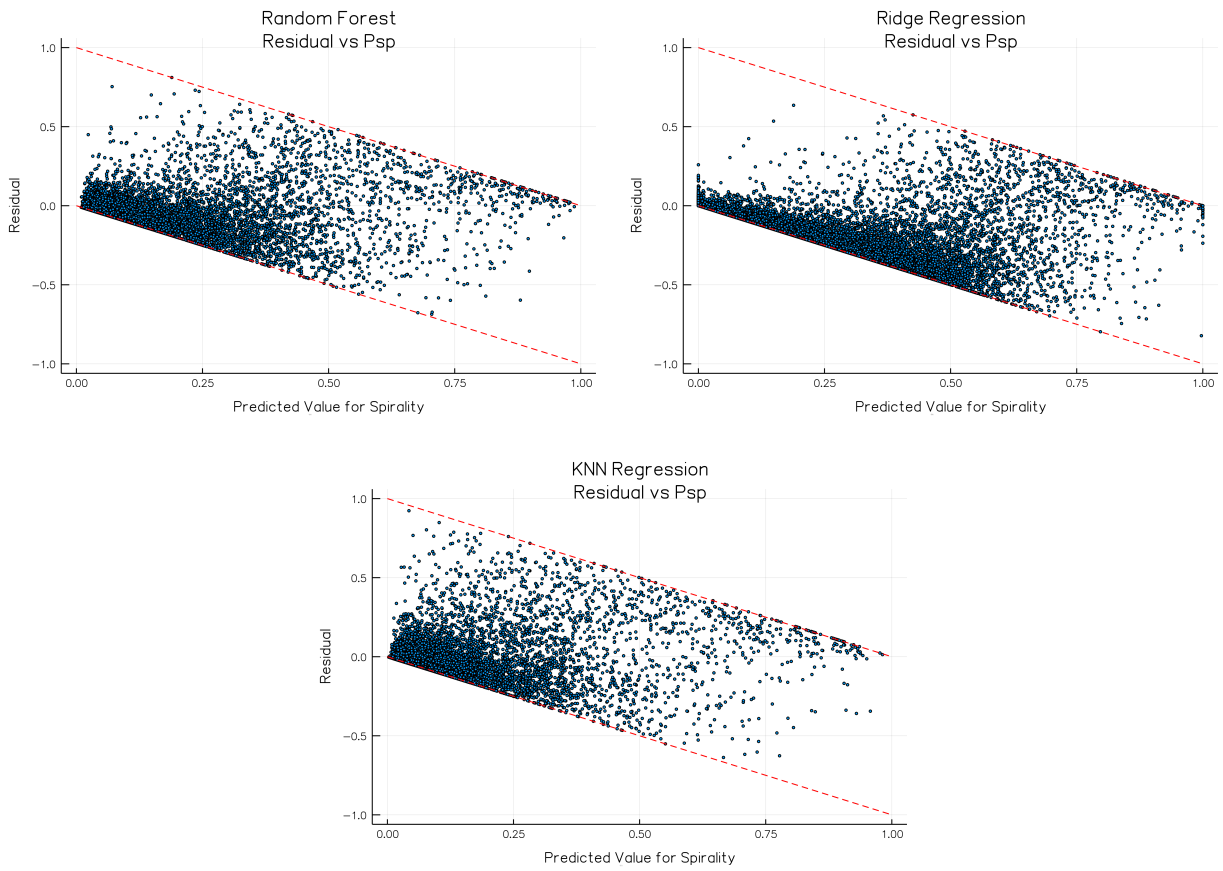


Figure 2.8: Residual plots for models trained with both SDSS features and SpArcFiRe features. For visualization purposes we used only 10000 points from the test set for these plots. The red dashed lines indicate the possible bounds that the values can fall on. Since both the inputs and outputs are constrained to be in the interval [0,1] (cf. Figure 2.5), the lower bound is determined by $f(x) = -x$ and the upper bound is determined by $f(x) = -x + 1$, for $0 \leq x \leq 1$.

symmetrically distributed about the x-axis, especially near the $P_{SP} = 0.5$ region, indicating that our method produces an unbiased, roughly uniform distribution of predicted spirality when the true spirality is close to 0.5.

2.4 Conclusions

Our results show that it is possible to have a model that performs well, is in agreement with human votes above 90% of the time, and also deal with the winding bias problem which was

addressed in more detail in Hayes et al. [2016]. In this sense, we "filter" the errors made by humans while still retaining the useful knowledge provided by the Galaxy Zoo. However, it is possible that Random Forests present a compromise at the intersection of understandability and precision, with RFs exceeding in the former while neural networks possibly excelling in the latter—possibly at the expense of reproducing human biases [Dieleman et al., 2015] or perhaps even introducing new ones.

What differentiates this from previous work is the addition of SpArcFiRe's output, which adds more information to the objects we are discriminating and helps to decrease the amount of bias present in the classifications provided in GZ1. These results demonstrate that SpArcFiRe adds valuable (rather than redundant) information; in turn, these new features can be used by other automatic machine learning classifiers and regressors to improve results. We provided some insights on what these models find more descriptive for spiral galaxies demonstrating the most important parameters used by random forests in Table 2.6.

Chapter 3

The Chirality Bias in Galaxy Zoo 1

The Galaxy Zoo 1 catalog displays a bias towards the S-wise winding direction in spiral galaxies which has yet to be explained. The lack of an explanation confounds our attempts to verify the Cosmological Principle, and has spurred some debate as to whether a bias exists in the real universe. The bias manifests not only in the obvious case of trying to decide if the universe as a whole has a winding bias, but also in the more insidious case of selecting which galaxies to include in a winding direction survey. While the former bias has been accounted for in a previous image mirroring study, the latter has not. Furthermore, the bias has never been *corrected* in the GZ1 catalog, as only a small sample of the GZ1 catalog was re-examined during the mirror study. We show that the existing bias is a human *selection* effect rather than a human chirality bias. In effect, the excess S-wise votes are spuriously "stolen" from the elliptical and edge-on-disk categories, not the Z-wise category. Thus, when selecting a set of spiral galaxies by imposing a threshold T so that $\max(P_S, P_Z) > T$ or $P_S + P_Z > T$, we spuriously select more S-wise than Z-wise galaxies. We show that when our provably unbiased random forest described in chapter 2 selects which galaxies are spirals independent

The contents of this chapter are based on the paper *On the nature and correction of the spurious S-wise spiral galaxy winding bias in Galaxy Zoo 1* by W. B. Hayes, D. R. Davis, and P. Silva, published in the Monthly Notices of the Royal Astronomical Society, Volume 466, Issue 4, May 2017, Pages 3928 - 3936.

of their chirality, the S-wise surplus vanishes, even if humans are still used to determine the chirality. Thus, when viewed across the entire GZ1 sample (and by implication, the Sloan catalog), the winding direction of arms in spiral galaxies as viewed from Earth is consistent with the flip of a fair coin.

3.1 Introduction

The *Cosmological Principle* is the assumption that at large scales the universe is homogeneous and isotropic. Homogeneity says that there is no special location in the Universe, and in particular that the Earth occupies no special location. Isotropy means that there is no preferred direction in the universe; for spiral galaxies, this means that the distribution of their spin axes should be spread uniformly at random on the celestial sphere. The two assumptions together imply that, as seen from the Earth, the distribution of observed arm winding directions of N spiral galaxies should be statistically consistent with N flips of a fair coin.

The Galaxy Zoo 1 (hereafter GZ1) project [Lintott et al., 2008, 2010] as described in the previous chapter was a website where humans were presented with random galaxy images from the Sloan Digital Sky Survey [York et al., 2000]. With each galaxy image they were given a choice of 6 "cartoon" galaxies and asked which cartoon most resembled the real galaxy. The GZ1 sample has almost 900,000 galaxies. After using SpArcFiRe [Davis and Hayes, 2014] to perform an ellipse fit of all GZ1 images, we concentrate on a subsample of 458,012 galaxies whose minor axis were larger than 14 pixels (semi-minor axis of 7 pixels), which we subjectively determined was the smallest sized disk on which spiral structure could be observed. For each galaxy, the number of votes for each of the 6 categories was converted into a fraction (Table 3.1).

Category	EL	EDGE	S-wise	Z-wise	MG	DK	Total
Winner by 50% majority	261700	53873	25102	23807	4431	755	369668
Percentage	57.14%	11.76%	5.48%	5.20%	0.97%	0.16%	80.71%
Winner by max vote	309591	73009	33007	31340	8406	2659	458012
Percentage	67.59%	15.94%	7.21%	6.84%	1.84%	0.58%	100%

Table 3.1: The 6 types of votes in Galaxy Zoo 1 across our sample of 458,012 GZ1 galaxies, along with the fraction of galaxies in each category as voted by the GZ1 humans. Note that not all galaxies have a winning vote that is a 50% majority, although every galaxy has a maximum vote (we ignore ties, which are rare).

As can be seen in Table 3.1, there is a significant excess of S-wise spiral galaxies, using either a majority-vote winner, or a less stringent "max vote" winner; similar surplusses of S-wise spirals are seen using other, more stringent criteria [Lintott et al., 2008, Land et al., 2008]. In our case, using the 50% majority-wins criterion, there are $25102 + 23807 = 48909$ galaxies with visible spiral structure, but there is an S-wise excess of 5.86σ (see Table 3.2) compared to 48909 coin flips; the "max vote" criterion shows an even stronger excess, with a statistical significance of 6.57σ. As we will see from Table 3.2 below, the effect gets smaller as we insist on higher human classification confidence, but never goes away even when 100% of humans agree on the chirality of a small set of galaxies. The statistical significance of this bias is detailed as a function of human confidence in the first quarter of Table 3.2, in which both the selection of galaxies, and their chirality, are chosen by GZ1 humans. As can be seen, the bias is detected at a level of somewhere between 3σ and 6σ, depending upon the human confidence level.

Whether this excess is real or not has been a matter of some debate. Lintott et al. [2008] and Land et al. [2008] show that the bias seems to disappear if galaxy images are flipped with 50% probability before being shown to humans, suggesting that somehow the humans are biased towards choosing S-wise galaxies. Whether the bias is a human cognitive bias, or perhaps due to website design or positioning of the buttons is unclear, and of little astronomical interest in any case. However, other studies [Longo, 2011, Shamir, 2012] have suggested that the bias is real rather than artifactual.

Spirality selector	Chirality determination	Spirality cutoff	|S-wise|	|Z-wise|	Sigma value	p-value
GZ1 humans	GZ1 humans	0.4	**32016**	30619	5.58σ	10^{-8}
		0.5	**25625**	24572	4.70σ	10^{-6}
		0.6	**20952**	20093	4.24σ	10^{-5}
		0.7	**16631**	16004	3.47σ	0.0002
		0.8	**12444**	11932	3.28σ	0.0004
		0.9	**7774**	7435	2.75σ	0.0030
GZ1 humans	SpArcFiRe	0.4	**31633**	31002	2.52σ	0.006
		0.5	**25417**	24780	2.84σ	0.002
		0.6	**20774**	20271	2.48σ	0.007
		0.7	**16533**	16102	2.39σ	0.010
		0.8	**12339**	12037	1.93σ	0.030
		0.9	**7708**	7501	1.68σ	0.050
Unbiased machine	GZ1 humans	0.4	29979	**30184**	0.84σ	0.250
		0.5	**19829**	19743	0.43σ	0.130
		0.6	**13130**	13093	0.23σ	0.400
		0.7	**8510**	8371	1.07σ	0.150
		0.8	**5028**	4895	1.34σ	0.100
		0.9	**2231**	2119	1.70σ	0.040
Unbiased machine	SpArcFiRe	0.4	**30103**	30060	0.18σ	0.40
		0.5	**19800**	19772	0.14σ	0.45
		0.6	13063	**13160**	0.60σ	0.30
		0.7	8371	**8510**	1.07σ	0.15
		0.8	4895	**5028**	1.34σ	0.09
		0.9	2121	**2229**	1.64σ	0.05

Table 3.2: Comparing the statistical significance of the chirality bias. **Selector**: who selects the sample (GZ1 humans or an unbiased machine learning algorithm); **Chirality determination**: who performs the chirality determination (GZ1 humans or unbiased SpArcFiRe algorithm); **Spirality cutoff**: include only galaxies for which $P_{SP} = P_S + P_Z >$ cutoff; **S-wise and Z-wise**: number of S-wise and Z-wise galaxies in above defined sample; the over-represented chirality is highlighted in bold; **Sigma and p-value**: standard deviation and p-value of difference between S- and Z-wise count compared to same number of coin flips.

In this chapter we put the problem to rest. We show below that the bias is almost certainly a human bias, and not a property of the actual GZ1 galaxies.

3.2 Nature of the bias

3.2.1 More S-wise than Z-wise spins for all values of "spirality"

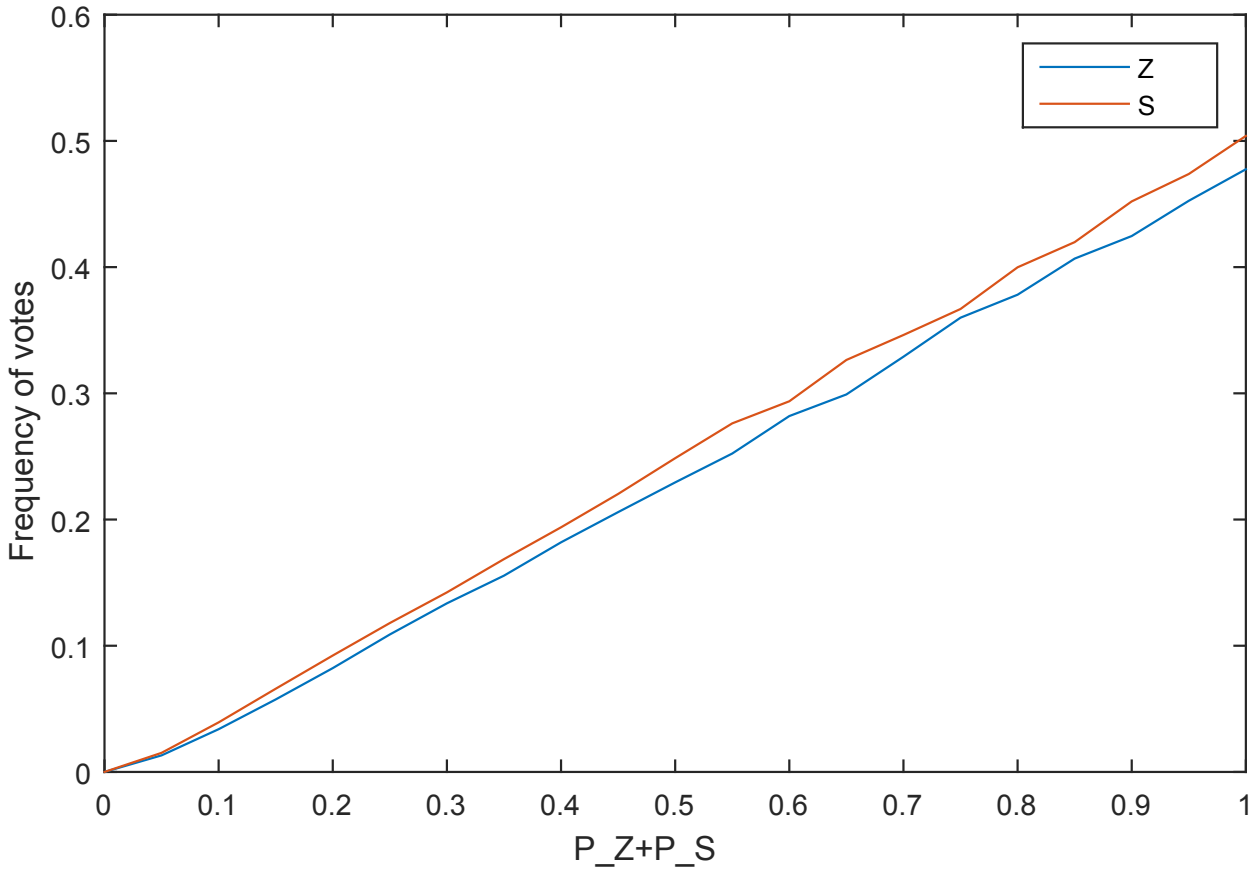


Figure 3.1: Lines joining the frequency histograms (vertical axis) of S-wise and Z-wise galaxies, according to GZ1 humans, having $x = P_S + P_Z$ (horizontal axis) among 20 equally-spaced bins in [0,1]. Note that *all* galaxies in the entire GZ1 sample are represented in this plot; galaxies to the left end tend to be elliptical or edge-on, while galaxies to the right end have clearly visible spiral structure. We see that the S-wise bias manifests across the entire spectrum, so for example near $x = 1$, we see that among all galaxies for which $P_S + P_Z \geq 0.95$, slightly more than half have $P_S \geq 0.95$, while slightly less than half have $P_Z \geq 0.95$. The selection effect manifests because any cutoff in $P_S + P_Z$ that is intended as a threshold above which a galaxy is considered to have visible spiral structure will automatically include more S-wise than Z-wise galaxies.

Figure 3.1 shows the frequencies of galaxies with the two winning chiralities, as voted by GZ1 humans, as a function of their sum $P_S + P_Z$. We refer to this sum as the *spirality* of a galaxy, and its value is meant to represent the probability that there exists any observable

spiral structure.[1] As can be seen, the S-wise bias manifests across all values of spirality even down close to zero, where the galaxies are unlikely to be spiral at all. Furthermore, we note that if one chooses any cutoff in spirality (or similarly $\max(P_S, P_Z)$) meant to isolate galaxies with visible spiral structure, then any such sample will automatically include more S-wise than Z-wise galaxies, because the S-wise curve is uniformly above the Z-wise one for all values of spirality. We shall demonstrate, as did Land et al. [2008] and Lintott et al. [2008], that this bias is spurious and not reprentative of the true chirality distribution.

3.2.2 Do humans actually disagree on chirality?

Land et al. [2008] briefly mentioned that there did not appear to be significant disagreement between humans about chirality. This statement seems at odds with Figure 3.1. Here we study that statement in detail, because understanding it may prove crucial to understanding where the bias comes from. To test the hypothesis that humans can disagree on the chirality of a galaxy, we introduce the idea of the *opposing vote*, which we define for any galaxy as *the most popular vote other than the most popular chirality*. Note that this is not quite the same as the second most popular vote, because if the most popular *chirality* is not the most popular vote *overall*, then the opposing vote is actually the winning vote overall. In other words,

1) When one of the two chiralities is the winning vote, then the opposing vote is the second most popular vote.

2) When neither of the two chiralities is the winning vote, the "opposing vote" is the winning vote for that galaxy.

[1] As distinct from GZ1's $P_{CS} = P_S + P_Z + P_{EDGE}$, which includes edge-on disks and represents if the galaxy, as seen from any direction, is a disk galaxy.

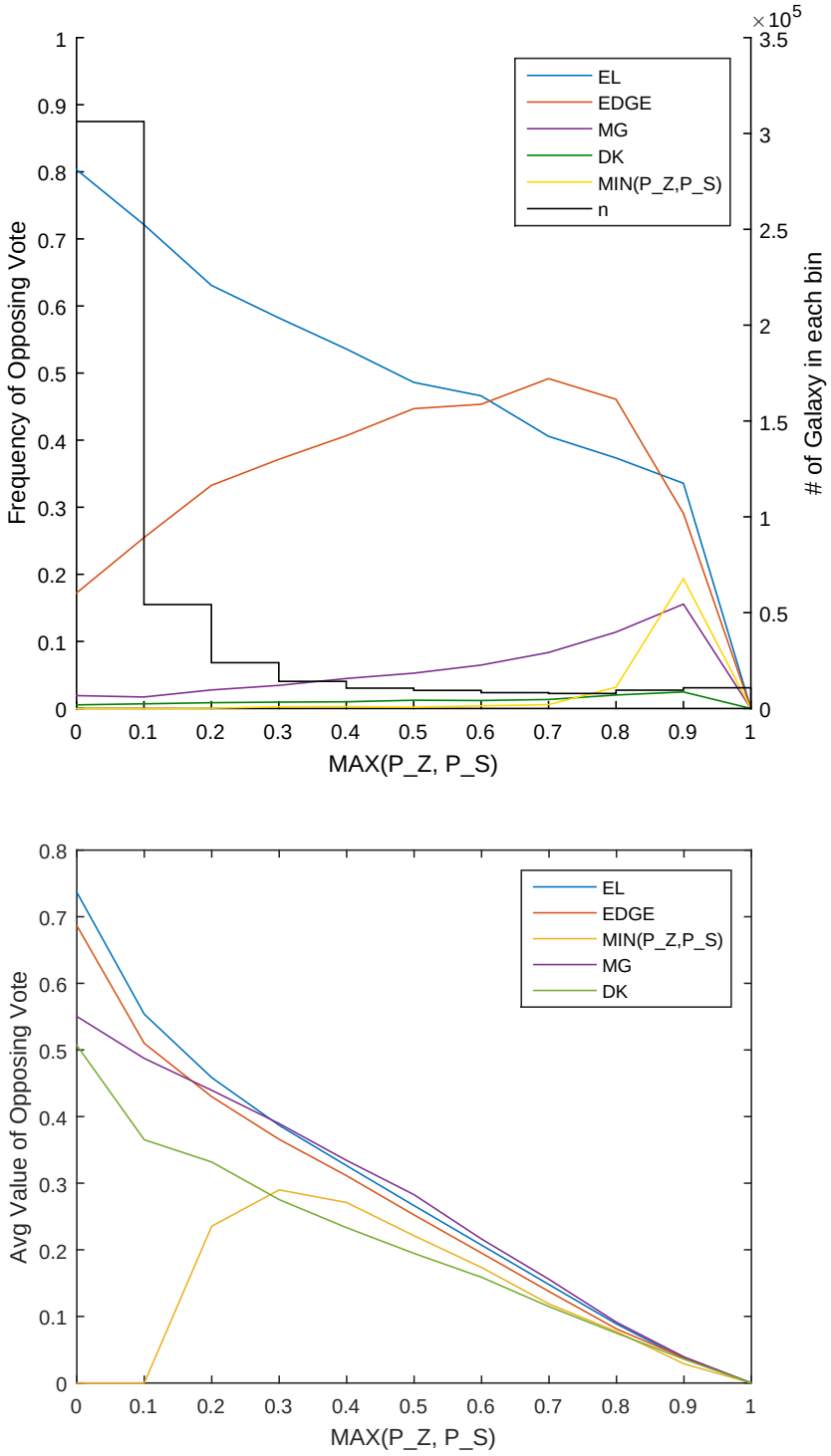


Figure 3.2: **Top**: Histogram of galaxy count (black bars) and frequency of the opposing vote (colored curves), as a function of the most popular chirality vote. As can be seen, the losing chirality is rarely the opposing vote. **Bottom**: Average value of the opposing vote, with the same horizontal axis.

Figure 3.2 describes the distribution and structure of the opposing votes, as a function of the most popular chirality, which is just $\max(P_S, P_Z)$, even if that winning chirality is not the winning vote across all 6 votes. The top half of Figure 3.2 shows the frequency that each of the other 5 votes (which are the losing chirality plus EL, EDGE, MG, DK) occur as the opposing vote. The first observation is that the losing chirality, $\min(P_S, P_Z)$, is almost never the opposing vote, even for very small values of the winning vote. That is to say, humans virtually never disagree on the chirality of a galaxy; even when only a small percentage of people actually choose a chirality, they still agree on that chirality.

Instead, the top half of Figure 3.2 demonstrates that the opposing vote is almost always either EDGE or EL. This tells us that the selection effect in Figure 3.1 occurs when people are uncertain whether they see spiral structure at all; the galaxy may be an edge-on disk galaxy with indistinct spiral structure, or appear to be elliptical that has faint spiral structure, but those that choose a chirality in that case tend, for whatever reason, to be slightly more inclined to choose S-wise over Z-wise, but even in those cases the humans tend to agree with each other on the chirality chosen. In other words, to arrive at the S-wise bias, humans are "stealing" votes from EDGE and EL, not from Z-wise. This allows the bias to exist even though humans virtually never disagree on chirality. Another interesting observation of the top half of Figure 3.2 is that near the origin, the EL and EDGE curves correctly show that, among galaxies that have no visible spiral structure, about 80% are elliptical and just under 20% are edge-on—in rough agreement with Table 3.1.

The bottom half of Figure 3.2 shows the average *value* of the opposing vote—i.e., the fraction of people, per-galaxy, who cast the opposing vote. As expected, as the winning chirality approaches 1, the average value of the opposing vote approaches zero. Also of interest is the fact that the sum of the winning chirality and the opposing vote value tends, on average, to be above 70%, so that the top two votes take the lion's share of the votes. We do see, however, that even though the losing chirality is rarely the opposing vote, it tends to be

a strong second when it does occur; these are probably galaxies that have strong spiral structure but are somehow disrupted so as to make the chirality unclear; one may hazard a guess that they may in fact be advanced mergers.

Finally, again referring to Figure 3.2, the sharp peak of the losing chirality (yellow line) at $x = 0.9$ in the top figure is not a problem because, as the lower figure shows, the *value* of that vote is tiny, as is the value of all the other non-winning votes (as they must be, since the winning chirality is taking 90% of the votes).

These graphs show that humnans do not significantly disagree with each other when determining chirality, which is an observation that is not at all obvious from any of the studies that have occurred to date. In fact it would be surprising if humans disagreed to any significant extent on winding direction, because in all but a very small number of cases, our own intuitive observation is that if there is a winding direction at all, it should be fairly obvious. These graphs strongly confirm this intuition. Together, Figures 3.1 and 3.2 demonstrate that the bias has nothing to do with humans disagreeing on winding direction. Instead, what is happening is that whenever there is uncertainty about whether or not there *exists* spiral structure *of any chirality*—that is, when a significant proportion of humans vote either edge-on or elliptical—then those that *do* vote for a chirality tend to vote for S-wise. The reason for this is still unknown, but the three obvious choices are (a) a human visual cortex bias; (b) something about the design of the web page for the GZ1 survey induces people to preferentially choose the S-wise button; or (c) there is a real chirality bias in the SDSS sample of galaxies.

3.3 Unbiased machine determination of winding direction

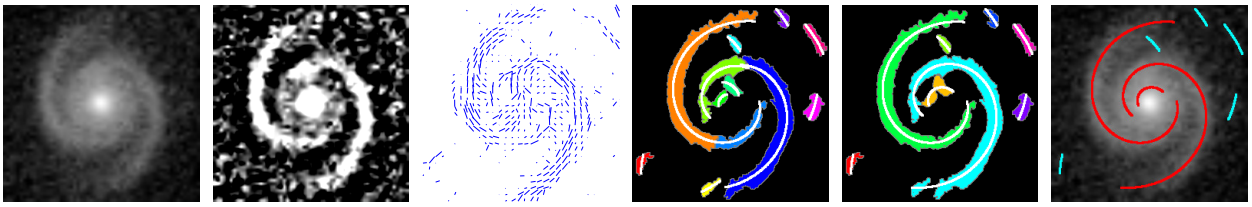
At this point we have two relevant observations: (1) galaxies voted S-wise significantly outnumber galaxies voted Z-wise, and (2) humans do not significantly disagree on chirality. In the absense of evidence for a human bias, this would directly imply that there is a real chirality bias in the universe. Land et al. [2008] convincingly demonstrated that this is not the case by having the GZ1 humans re-classify a subset of spiral galaxies while randomly left-right flipping each image with 50% probability. However, beyond demonstrating that the bias was human, they did not attempt to correct for it on a galaxy-by-galaxy basis.

SpArcFiRe[2] Davis and Hayes [2014] is an automated method that decomposes a spiral galaxy into its constituent arms. A very brief summary of how SpArcFiRe works is depicted in Figure 3.3. As described in Davis [2014], Davis and Hayes [2014], it was tested on a sample of 29,250 GZ galaxy images chosen by the leader of the Galaxy Zoo Project[3]. Among many other things, one of SpArcFiRe's outputs is a determination of the galaxy's winding direction. Given a list of found spiral arcs in a galaxy image, there are many ways to determine a global winding direction for the entire galaxy. Some arcs are longer than others, some may wind in the opposite direction to the majority, and some "arcs" are actually just noise mistaken for an arc. We found that the most reliable measure of the winding direction of the galaxy was a length-weighted vote of the winding direction of all the discovered spiral arcs.

To demonstrate that this measure is unbiased with regard to spin direction, we refer to Figure 3.4, which shows a scatter plot of the left-to-right mirrored vs. unmirrored value of the galaxy's pitch angle as measured by SpArcFiRe; in a perfect reversal, the two should be negatives of each other. We find that in 29,094 out of 29,250 cases (99.47% of cases), the

[2] *SP*iral *ARC FI*nder and *RE*porter

[3] Stephen Bamford, Personal Communication. The selection criteria were: $(GZ1_{P_S} + GZ1_{P_Z}) > 0.8$ OR $(GZ2_{FeaturesOrDisk} > 0.7$ AND $GZ2_{NotEdgeOn} > 0.7$ AND $GZ2_{spiral} > 0.8)$.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**)

Figure 3.3: Steps SpArcFiRe [Davis and Hayes, 2014] takes in describing a spiral galaxy image. **a)** The centered and de-projected image. **b)** Contrast-enhanced image. **c)** Orientation field (at reduced resolution for display purposes). **d)** Initial arm segments found via Hierarchical Agglomerative Clustering (HAC) of nearby pixels with similar orientations and consistent logarithmic spiral shape, overlaid with the associated logarithmic spiral arcs fitted to these clusters. **e)** Final pixel clusters (and associated arcs) found by merging compatible arcs. **f)** Final arcs superimposed on image (a). Red arcs wind S-wise, cyan arcs wind Z-wise.

two are negatives of each other to within 10^{-4} degrees. Even more relevant to this work, we find that in all but 5 cases (99.983% of cases), the chirality of the mirrored image is correctly flipped compared to the unmirrored case. Thus, the chirality determination of SpArcFiRe is unbiased, with respect to flipped images, to better than 2 parts it 10^4.

We then ran SpArcFiRe on the entire Galaxy Zoo sample of galaxies, in order to determine the chirality of galaxies in an unbiased manner. However, we still used the human GZ1 determination of $P_S + P_Z$ to select which galaxies actually display spiral structure. The second quarter of Table 3.2 details the statistical significance of the winding direction bias, as a function of the *human* confidence in observed spiral structure, but when the chirality is determined by the unbiased SpArcFiRe algorithm. The statistical signifance of the S-wise bias is weaker than in the first quarter of Table 3.2, but surprisingly, the bias is still significant to somewhere between 2σ and 3σ.

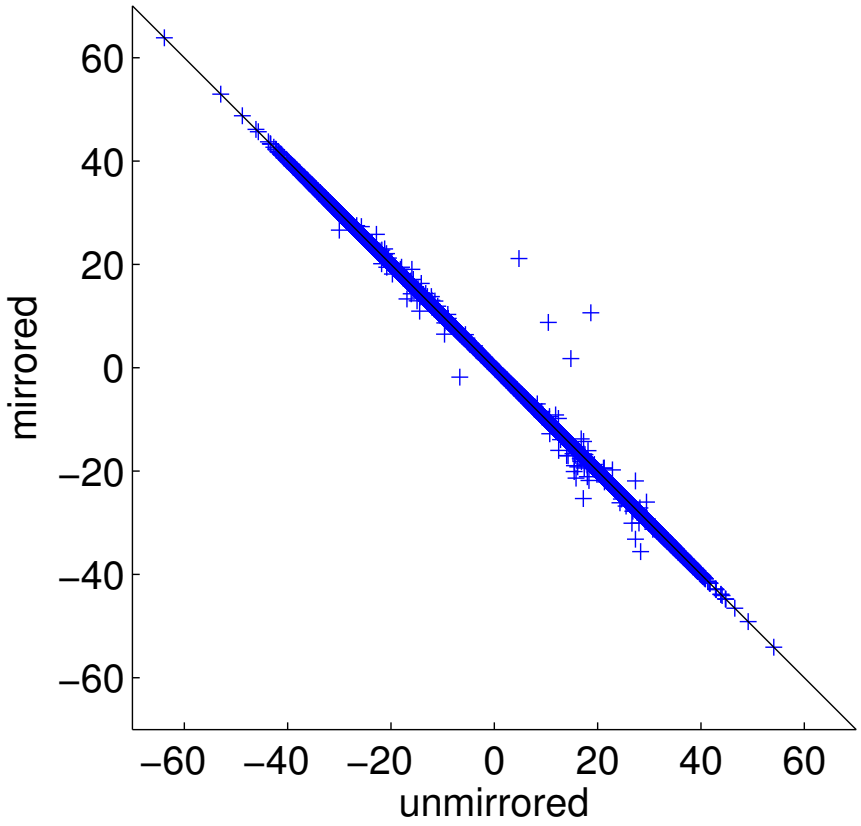


Figure 3.4: Galaxy-level pitch angles reported by SpArcFiRe using unmirrored and left-to-right mirrored input images across 29,250 "clear" spiral galaxies (see text for definition). These galaxy-level pitch angles are calculated as the arc-length-weighted average of all arcs agreeing with the dominant winding direction (as determined by an arc-length-weighted vote). We see that for almost all galaxies the measured pitch angle almost exactly negative, as it should be. The diagonal line gives $y = -x$; cases on this line are visually underrepresented due to overlap. More importantly, only 5 cases out of 29,250 disagree on chirality, showing that SpArcFiRe is chirality-unbiased to a level of almost 1 part in 10,000.

3.4 Unbiased machine determination of spirality

Our goal in this section is to explain how we created a random forest that was capable of reproducing the spirality $P_S + P_Z$, while being simultaneously unable to reproduce either P_S nor P_Z alone. That is, we want to create a spirality measure for a galaxy that is provably independent of chirality.

3.4.1 Building a selector that is unbiased to chirality

As alluded to earlier, the problem is not in the actual determination of chiraltiy. Humans do not disagree with each other on chirality, and in fact the human determination of chirality agrees with the SpArcFiRe determination of chirality in between 95% and 98% of cases on the GZ1 clean sample, depending upon SpArcFiRe's own determination of its certainty[Davis, 2014, tables 5.1 and 5.2, column "80"], and the cases of chirality disagreement between GZ1 humans and SpArcFiRe appear randomly distributed.

Figure 3.1 points to the problem: S-wise galaxies outnumber Z-wise ones for *any* set of galaxies selected using a criterion of either $\max(P_S, P_Z)$ or $P_S + P_Z$ greater than some threshold α, and P_S and P_Z are taken from the human GZ1 vote values. Thus, we must determine some method of determining if there is a selection bias and if so, try to eliminate it.

To do this, we need to create a sample of galaxies that have visible spiral structure ("spirality"), but selected in a way that is unbiased to winding direction. To do this we create a machine learning algorithm that is provided with attributes of the galaxy that are independent of winding direction, and tell it to attempt to reproduce $P_S + P_Z$. We then demonstrate that it can reproduce $P_S + P_Z$ with reasonable accuracy and then show that it is unable to simultaneously reproduce winding direction to any level better than chance.

In order to create such an algorithm, we need to ensure that the features it uses (that is, the measurements of the galaxy) are features that are independent of chirality. This may not be trivial, as recent work has suggested that even photometric data may be able to recover winding diretion to a significant degree [Shamir, 2016]. We choose our attributes to include some photometric attributes that were disjoint with those that Shamir [2016] found to be correlated with chirality, in addition to several SpArcFiRe outputs with all chirality information removed.

Our list of input attributes to our machine learning algorithm, assumed to be independent of chirality, are as follows. From the SDSS database, we allow parameters used by Banerji et al. [2010] (colors, de Vaucouleurs fit axial ratios, exponential fit axial ratios, exponential disk fit log likelihood, de Vaucouleurs fit log likelihood, star log likelihood, ratios of Petrosian radii, Adaptive shape measures, adaptive ellipticities, adaptive 4th moment, and a texture parameter), as well as absolute magnitudes and disk-to-bulge ratios. From SpArcFiRe [Davis and Hayes, 2014] we allow all numerical output parameters including pitch angles after having taken their absolute value. Such parameters include counts and lengths of spiral arcs, the absolute value of their pitch angles, and the number and length of arcs of agreeing and disagreeing chiralities (with the actual chirality replaced by "majority" and "minority").

Finally, since it is known that machine learning algorithms tend to reproduce the input distribution of target values we trained our machine on a set of galaxies that were 50-50 S-wise and Z-wise according to the GZ1 humans.

We applied two filters to the data to build a dataset of high-confidence spiral galaxies. For S-wise the rule was $((P_S + P_Z > 0.6) \cap (P_S > 0.5) \cap (P_S - P_Z > 0.3))$; which means that there are at least 60% of the votes for spirality, at least 50% of the final votes were for P_S, and there is at least 30% more votes for P_S than for P_Z. The first and second rules are effective in filtering out other types of objects and the third in making sure that the humans have a higher agreement not only in spirality but also in chirality. Similarly to build our Z-wise set the rule was $((P_S + P_Z > 0.6) \cap (P_Z > 0.5) \cap (P_Z - P_S > 0.3))$. After this we sampled 19500 objects from each class, to assure we had a balanced dataset, totaling 39000 objects.

We built a random forest model to predict chirality. As it is common with these models, we performed what is called a "hyperparameter" search across possible machine configurations to decide what was the optimal number of trees per forest and the optimal number of features per tree. We built 45 models with number of trees varying on the interval {10,15,20,25,30,35,40,45,50} and number of features per tree from {30,40,50,60,70}. We used

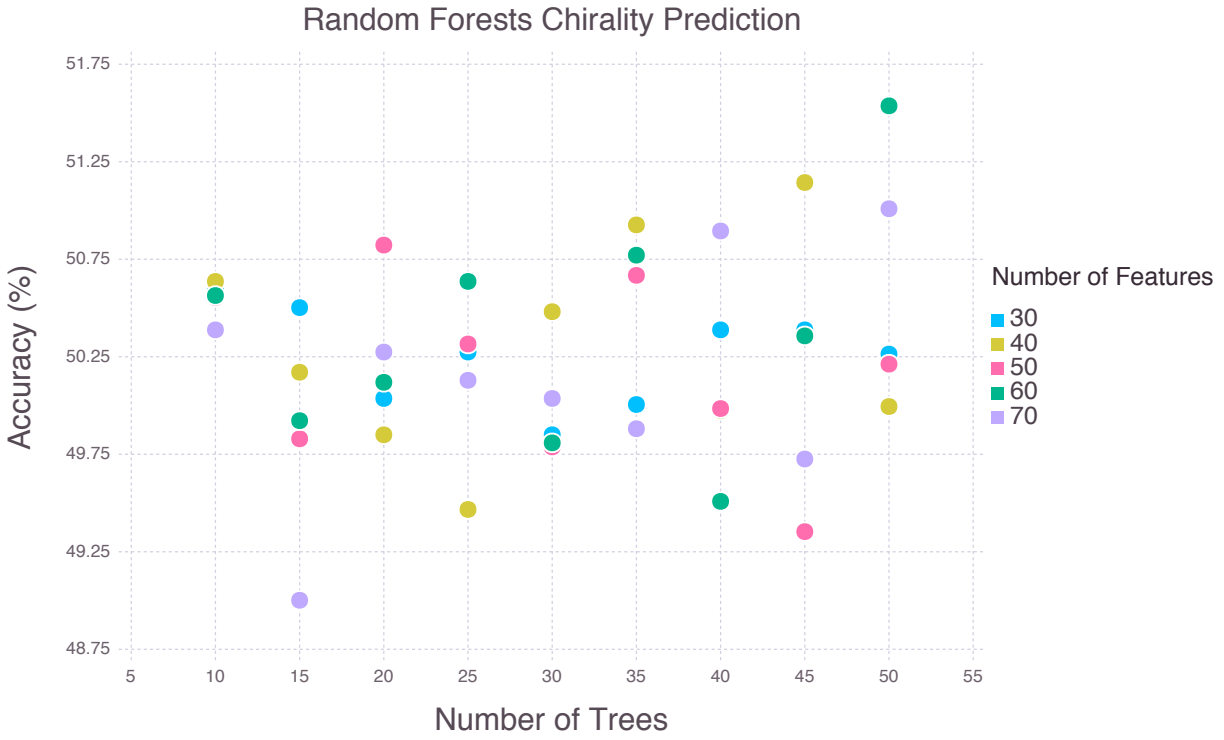


Figure 3.5: Chirality Prediction using Random Forests with 45 different architectures, based on the number of trees and the number of features for each forest.

a Bernoulli distribution to sample 75% of the data for training the forests and the remaining 25% to test the models.

As mentioned above we expect not to be able to predict chirality with any level of confidence. The accuracy of the models range from from 49% to 51.25%, heavily centered around 50, as it is portrayed in Figure 3.5. Due to chance in data sampling and the way that Random Forests are built we expected this variability to occur. Notice that 15, or 1/3 of all the models built, have an accuracy below 50%, i.e they are worse than a coin flip for tracking chirality. To make sure that our models are not able to indeed predict chirality we decided to further investigate the 3 models that had at least 51% accuracy.

When performing a classification task in machine learning one can have a notion of how confident a model is. For our case we set up the target to be either 0, to predict P_Z, or

1, to predict P_S. The model outputs a number P within this range: if $P < 0.5$ we say the predicted class for that object was P_Z, if $P > 0.5$ the predicted class is P_S. The more confident a model is that an object belongs to either class the closer the output value will be to those limits (0 and 1). That means that when a model is not very confident of its output, the values predicted will fall mostly in the middle value of the range. To better visualize this distribution Figure 3.6 shows histograms from the 3 models that had at least 51% accuracy. As we had supposed, the distributions are heavily centered, indicating that the model has a very low level of confidence on its predictions. Also, from the 3 models at least 77% of the predictions were between 0.4 and 0.6, i.e., the model had less than 20% of confidence on the output for at least 77% of the objects.

To ensure once and for all that these models performance was due to chance we rebuilt those three models using the same data,the same Bernoulli distribution and the same split for test and training and at the end we got different accuracy values for the 3 models that were previously at least 51%. The best out of the 3 now has an accuracy of 50.23%.

We conclude that with these models, given these attributes we are unable to retain any information that correlates in any way to chirality of spiral galaxies.

3.4.2 Using the same machine to predict spirality

Now that we have a list of attributes and a machine that is unable to predict chirality in the form of either P_S or P_Z alone, we use a machine with the same input attributes and hyperparameters to reproduce the sum $P_S + P_Z$, which we term the *spirality* of a galaxy.

Given a list of human spirality votes $P_S + P_Z$ for each galaxy, we train the machine on 75% of the galaxies and test it on the remaining 25%; we do this four times, for four non-overlapping

[4]This means that the model predicted a value p on the interval $0 < p < 0.5$ when the expected value was 0 and and $0.5 <= p <= 1$ when the expected value was 1.

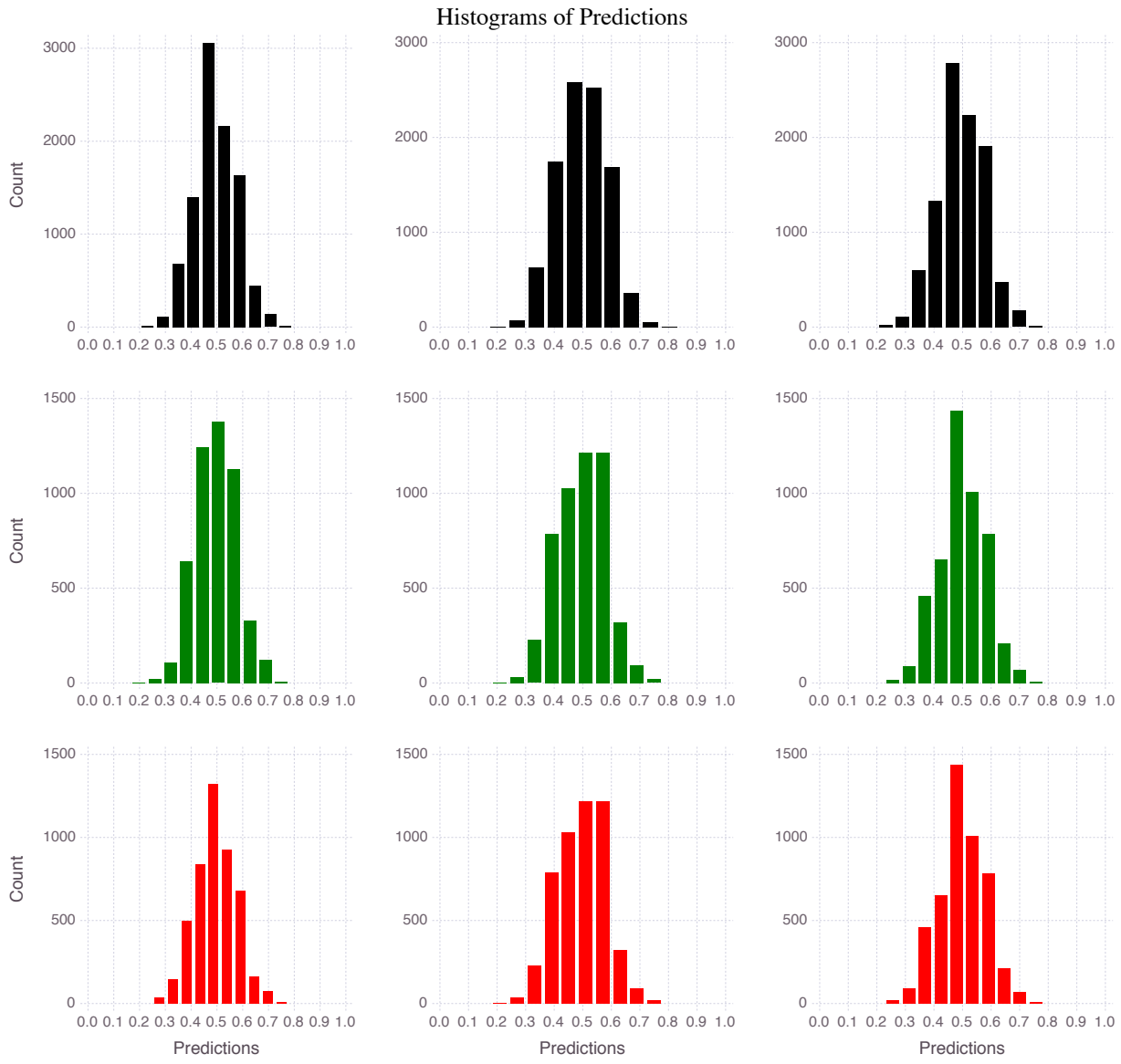


Figure 3.6: Histograms of the predictions for the 3 models that had the highest accuracy. Each column represents a model with the first being the best overall, using 50 trees and 60 features and an accuracy of 51.54%, the second, using 45 trees and 40 features and an accuracy of 51.14%, and the third using 50 trees and 70 features and an accuracy of 51.01%. The first row amounts for all the values predicted by the model, and the second and third rows shows the values that the model predicted correctly and incorrectly, respectively.[4]

25% subsets. The concatenation of these four 25% test subsets constitute our database of machine-determined spiralities that are independent of chirality. Given a particular galaxy, the difference between the human value $P_S + P_Z$ and our predicted spirality P_{SP} is the error

for that galaxy. The root mean squared error across all galaxies is a typical measure used to assess the accuracy of a predicted model. In our case, we were able to produce a machine with an RMSE of 0.137 across our sample of 450,012 galaxies. This is quite a bit larger than what other machines have done; for example the Kaggle winner was able to produce an RMSE of just 0.07 [Dieleman et al., 2015]. However, they made no effort to remove human bias, and thus it is not surprising that they are able to reproduce exactly how the humans voted better than we can.

3.5 Results

The bottom half of Table 3.2 shows the results of our chirality bias study when our unbiased machine from Section 3.4 selects galaxies based on predicted spirality. As can be seen, using this machine to perform selection virtually eliminates the chirality bias, *even if humans still choose the chirality*. This confirms our statement earlier that the GZ1 humans have a *selection* bias, not a chirality bias. In fact there is no significant difference between the two subtables in the lower half of Table 3.2: as long as our machine learning algorithm performs the selection based on unbiased spirality, it doesn't matter if the winding direction is determined by humans, or by SpArcFiRe. In either case, the S-wise bias is either vastly reduced, or reversed, apparently at random.

3.6 Discussion

Ideally we would like to integrate our new catalog into the GZ1 catalog so as to publish a "corrected" vote catalog in which the chirality bias has been removed. However, this is not as simple as rescaling the P_S and P_Z values to our values. Recall that the S-wise votes are "stolen" from the edge-on and elliptical categories. Thus, we would need to re-scale *all* the

vote values on a galaxy-by-galaxy basis, not just the two chirality votes, in order to fully correct the bias. Furthermore, we would like to do this in a way that only minimally changes the values of the human votes. Creating a machine algorithm that simultaneously removes the bias, and also minimizes the change in human vote values, is non-trivial, and left for future work.

Chapter 4

The Pitch Angle Selection Bias in Galaxy Zoo 1

The Galaxy Zoo project has provided a plethora of valuable morphological data on a large number of galaxies from various surveys, and their team members have identified and/or corrected for many biases. In this chapter we study a new bias related to spiral arm pitch angles, which first requires selecting a sample of spiral galaxies that show observable structure. One obvious way is to select galaxies using a threshold in spirality, which we define as the fraction of Galaxy Zoo humans who have reported seeing spiral structure. Using such a threshold, we use SpArcFiRe to measure spiral arm pitch angles. We observe that the mean pitch angle of spiral arms increases linearly with redshift for $0.05 < z < 0.085$. We hypothesize that this is a selection effect due to tightly wound arms becoming less visible as image quality degrades, leading to fewer such galaxies being above the spirality threshold as redshift increases. We corroborate this hypothesis by first artificially degrading images of

The contents of this chapter are based on the paper *SpArcFiRe: morphological selection effects due to reduced visibility of tightly winding arms in distant spiral galaxies* by T. R. Peng, J. E. English, P. Silva, D. R. Davis, and W. B. Hayes, published in the Monthly Notices of the Royal Astronomical Society, Volume 479, Issue 4, October 2018, Pages 5532 - 5543.

nearby galaxies, and then using a Random Forest trained on Galaxy Zoo data to provide a spirality for each artificially degraded image. It correctly predicts that the detected spirality of a fixed galaxy decreases as image quality degrades. We then use these spiralities to corroborate the hypothesis that the mean pitch angle of those galaxies remaining above a fixed spirality threshold is higher than those eliminated by the selection effect. We find that SpARcFiRe's ability to accurately measure pitch angles decreases as the image degrades, but that spirality decreases more quickly in galaxies with tightly wound arms, leading to the selection effect. This demonstrates that users who select samples of galaxies using a threshold of Galaxy Zoo votes must carefully consider the possibility of selection effects on morphological measures, even if the measure itself is believed to be objective and unbiased.

4.1 Introduction

The arms of spiral galaxies are still not fully understood [Binney and Tremaine, 2011], in part because there is no widely accepted method of quantifying their visible structure. While the light profiles of elliptical galaxies are fairly easy to model [Peng et al., 2010], no such easy quantification exists for spirals. The human-based classification scheme *Galaxy Zoo* provides an initial idea of the structure of galaxies [Lintott et al., 2008, 2010, Willett et al., 2013]. However, human classifications provide only a very rough quantification, and many biases have been noted, some of which may be directly attributable to humans [Lintott et al., 2008, Land et al., 2008, Hayes et al., 2016], some may be inescapable selection effects [Bamford et al., 2009,], while others may be caused by the finiteness of the number of human classifiers and the decrease of human classifiers further down the Galaxy Zoo 2 classification tree [Masters et al., 2011].

There has been extensive work studying biases in the Galaxy Zoo dataset, most prominently by the Galaxy Zoo team. Bamford et al. [2009] performed an extensive analysis of the

morphological biases in the Galaxy Zoo 1 catalog, distingushing between *selection* effects and *classification* bias. A selection effect occurs when objects disappear from the catalog due to being unobservable; the most common of this is the Malmquist effect that makes intrinsically dim objects invisible with increasing distance, so that the catalog is biased towards bright objects at high redshifts. A *classification* bias (again according to Bamford et al. [2009]) occurs when an object can be seen but is mis-classified due to image degradation (especially by untrained classifiers such as the vast majority of volunteers in the Galaxy Zoo project); the primary classification bias detailed by Bamford et al. [2009] was the effect of spiral galaxies being misclassified as elliptical ("early type") since the spiral arms became less visible as images became smaller and more noisy with increasing redshift.

Lintott et al. [2010, 2008] studied several different selection effects in the Galaxy Zoo 1 catalog, including winding direction, color, and the human biases that ironically arise when the human subjects are aware that their potential biases are under scrutiny. Redshift and magnitude biases in the selection of barred galaxies were studied in Masters et al. [2011] and Hoyle et al. [2011], while Darg et al. [2011] studied possible selection effects due to oversampling of ellipticals in multi-mergers. Willett et al. [2013] did extensive work on classification biases in the Galaxy Zoo 2 decision tree under the assumption that no true galaxy evolution takes place within the volume of the Sloan survey. Hart et al. [2016] demonstrated and showed how to correct a redshift bias in spiral arm count.

SpArcFiRe [Davis and Hayes, 2012, 2014, Davis, 2014] is the latest and most comprehensive among recent attempts [Odewahn et al., 2002, Au, 2006, Au et al., 2006, Shamir, 2011, Davis et al., 2012] to quantify spiral structure in an automated fashion. It seems to agree with humans as well as can be expected for all the measures given by the Galaxy Zoo 1 catalog, and has played a hand in analyzing exactly how the Galaxy Zoo 1 human classifications have a selection bias (but not a disagreement bias) when selecting which direction the arms wind in a spiral galaxy [Hayes et al., 2016]. SpArcFiRe's analysis of a spiral galaxy includes parameters

that describe each individual spiral arm segment found in an image: these parameters include the pixels corresponding to each arm segment, its average length, width, location, pitch angle, and bounds (error estimates) in each of these. (See Figure 3.3; Davis and Hayes [2014], Davis [2014] provide more detail.)

4.1.1 The Observation

We have run SpArcFiRe on the same set of SDSS [York et al., 2000, Eisenstein et al., 2011] galaxies as did the Galaxy Zoo 1 survey, and intend to publish our output data soon. However, in our initial excitement, we looked for as many correlations as we could think of between our new analysis of SDSS galaxies, and existing data. One of the most interesting we found was an apparent positive correlation between average pitch angle of spiral galaxy arms and redshift, even after carefully selecting a uniform volume-limited sample of galaxies to account for the Malmquist bias (Figure 4.1). In our initial examination of the results, we realized that it might be a selection effect: tightly winding arms (ie., arms with a low pitch angle) may become less distinct and eventually disappear as a galaxy image degrades with distance. Thus, at large distances, galaxies with loosely winding (ie., high pitch angle) arms may be selected more easily than tightly-winding (low pitch angle) galaxies, all other properties being equal. At this point we realized that it would be wise to perform a detailed analysis of how SpArcFiRe responds to images of decreasing resolution and increasing noise. That is the primary purpose of this chapter.

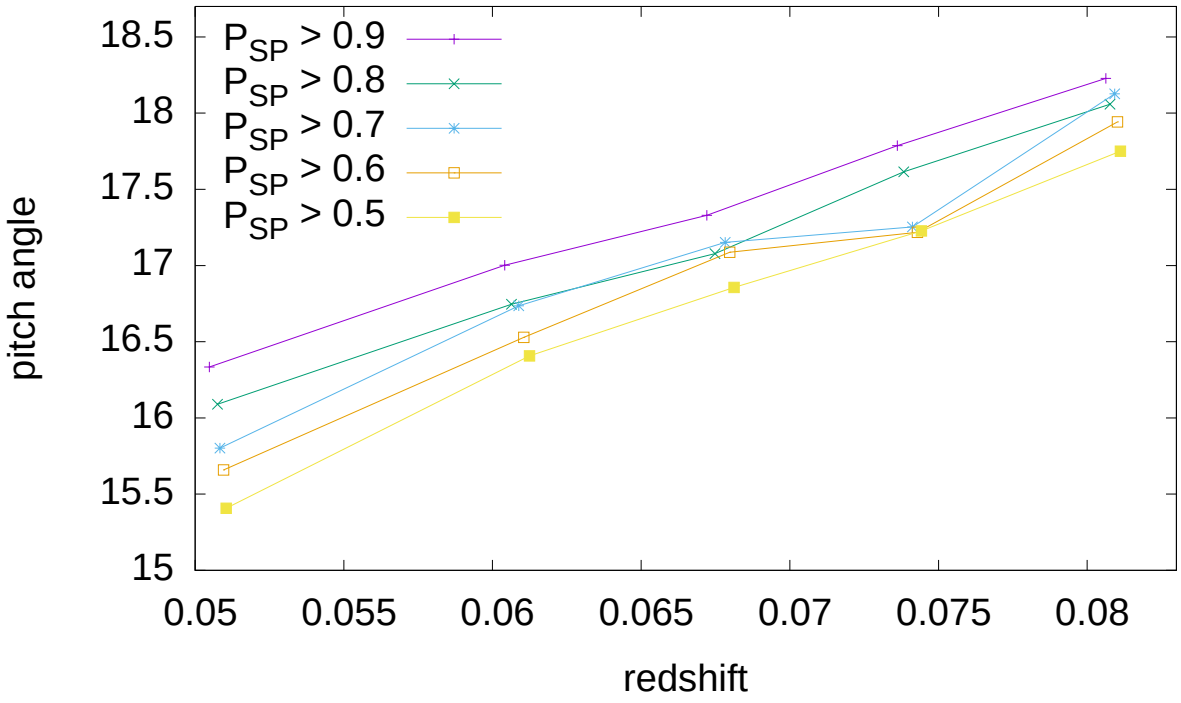


Figure 4.1: **The observation that spurred this study.** Within a volume-limited sample accounting for the Malmquist bias ($z < 0.085$, absolute magnitude brighter than -22.25 in the r band) of SDSS spiral galaxies (GZ1 spirality $P_{\mathrm{SP}} \geq x$ for the values of x displayed), SpArcFiRe observes their mean pitch angle to increase with redshift. Linear extrapolation predicts that this value would reach 90 degrees at about $z = 1.2$, suggesting the observed increase is either spurious, or must become sublinear with redshift. The lines, for $P_{\mathrm{SP}} \geq \{0.9, 0.8, 0.7, 0.6, 0.5\}$ have a total of 2896, 3639, 4106, 4473, and 4843 galaxies, respectively. (Galaxies accumulate as P_{SP} decreases.) For each line, the set of galaxies were divided equally into 5 bins sorted by redshift. The mark on the curve then represents the point (mean redshift, mean pitch angle) across the set of galaxies in that bin. Comparing the 5 curves, note that the mean pitch angle increases slightly with increasing spirality, suggesting a possible selection effect: loosely winding arms (corresponding to larger pitch angles) are more visible than tightly-wound arms, leading to more such galaxies being included in the sample. The same effect (loosely winding arms being more visible) may cause a selection bias that increases with redshift due to image degradation. The purpose of this chapter is to test that hypothesis.

4.2 Image degradation

4.2.1 "Clear" set of spiral galaxy images

We define the *spirality* of a galaxy image as the fraction of humans from GZ1 who voted that they saw either S-wise or Z-wise arms.[1] More precisely, if P_{S} and P_{Z} constitute the fraction of people who voted for each[2], the *spirality* $P_{\mathrm{SP}} = P_{\mathrm{S}} + P_{\mathrm{Z}}$.

We picked 7536 clear spiral galaxy images from SDSS for our experiment. The precise selection came from the union of the following two sets: (a) a complete volume-limited sample of galaxies out to $z = 0.085$ with magnitude in the SDSS r-band brighter than -22.25 and spirality greater than 0.7 and a Petrosian 90% radius greater than 6 arc-seconds (15 pixels), which we subjectively determined to be the smallest galaxy in which arms could be seen; this set comprises 4106 spirals, and is exactly the same set depicted on the $P_{\mathrm{SP}} > 0.7$ curve in Figure 4.1. The second set is (b) any galaxy with a GZ1 spirality greater than 0.9, which added another 3435 galaxies to our set.[3] The goal now is to degrade these images artificially and observe the effect.

4.2.2 Image degradation using Sunpy

Sunpy [SunPy Community et al., 2015] is a tool primarily used to generate artificial images from the Illustris simulation [Nelson et al., 2015]; since images from a simulation can be generated with almost arbitrary clarity, Sunpy was used to produce degraded images of

[1] We emphasize that spirality is technically associated with an *image*, not with an object. The spirality of an image of a spiral galaxy can decrease towards zero as the image becomes more degraded—which is precisely the effect we're trying to quantify in this chapter. Spirality can also become zero if a disk galaxy it tilted so far as to be edge-on, because then there is no *visible* spiral structure.

[2] Called P_{CW} and P_{ACW} in the dataset but since "clockwise" is ambiguous, the terms S-wise and Z-wise have since been adopted.

[3] In hindsight perhaps we should have chosen 0.7 for the lower spirality cutoff for both sets. However, considering our blurred set contains almost a million images, we think the sample size is big enough.

simulated galaxies by mimicking the degradation that occurs to real images of galaxies. We add a point spread function (full width half maximum or FWHM(PSF)), redefine the pixel size, and add noise to an image, in that order. In the latter case, SunPy assumes an input image with ostensibly zero noise. Then, to add noise to arrive at a S/N of K, it adds up all the signal S across the entire image, and then adds Gaussian pixel noise of total value KS. Here we have used Sunpy to artificially degrade a set of high quality SDSS images of real galaxies; these do not have zero noise to start with, so our degraded images will have S/N slightly lower than specified.

Starting with our clear set of 7536 SDSS galaxies, we degrade them with the following parameters: we vary the FWHM(PSF) from 4" (clear) to 128" (very blurry) in geometric steps of $2^{1/4}$ arc seconds; and we vary the S/N from 256 (clear) down to 8 (very noisy) in geometric steps of 2. We set the pixels-per-arcsecond ratio to a constant of 1/3.2 of the FWHM(PSF), corresponding roughly to SDSS that has an average FWHM(PSF) of about 1.3 arc seconds and a pixel size of 0.4 arc seconds. Throughout the chapter the term "PSF" refers to FWHM(PSF).

4.2.3 Blurring Pipeline

We ran our blurring program on all 7536 galaxy images across the above set of FWHM(PSF) and S/N values, generating 949,536 blurred images. Running SpArcFiRe on these images produced 622,585 galaxies with usable output from SpArcFiRe; the remaining 326,951 images were so badly degraded that SpArcFiRe failed to find anything in the image, either because it could not isolate the galaxy at all, or because it could not find any spiral arcs.

Figure 4.2 depicts an example galaxy (SDSS DR12 19-digit ID 1237648702972625038) and its degraded images. The top half of Figure 4.2 depicts the original unblurred image at the far top-left, along with several samples of the degraded image: moving to the right

increases the FWHM(PSF) (note the pixelation observable to the right since we set the pixels-per-FWHM(PSF) to the constant 3.2), and moving down decreases (degrades) the S/N ratio. The bottom half of Figure 4.2 is the corresponding chart of the spiral arcs found by SpArcFiRe. The bottom right has 2 dark squares because the images are so badly degraded that SpArcFiRe failed to produce any output (not surprising, considering the input images they came from).

4.3 Results

4.3.1 Pitch angle *vs.* image degradation: raw data

In this section we study the "raw" effect of image degradation on mean pitch angle. We first take the unblurred images of our 7536 "clear" spiral galaxies, and compute a galaxy-level pitch angle. As explained in Davis and Hayes [2014], these galaxy-level pitch angles are computed from an arc-length weighted mean pitch angle across the spiral arcs discovered by SpArcFiRe. We define two different types of means: one of them includes *all* arcs found, including sign, even though some of the arcs are almost certainly noise; we simply call this one the mean pitch angle. The second type includes only those arcs that agree with the *dominant chirality*, which is defined as the winding direction with the longest total length of all arcs of one sign. This dominant chirality has been shown to agree very well with the GZ1 human majority vote on which direction the whole galaxy "winds", to the point that SpArcFiRe is actually a more reliable indicator of winding direction than the average human.[4] Pitch angles computed using only arcs that agree with the dominant chirality are labelled **DCO** (Dominant Chirality Only), and we believe it is a more reliable indicator

[4]Galaxy Zoo ranks human voters based upon how often they agree with the majority; SpArcFiRe ranks in the upper quartile of this quality measure, making it more reliable than at least 75% of humans for determining winding direction.

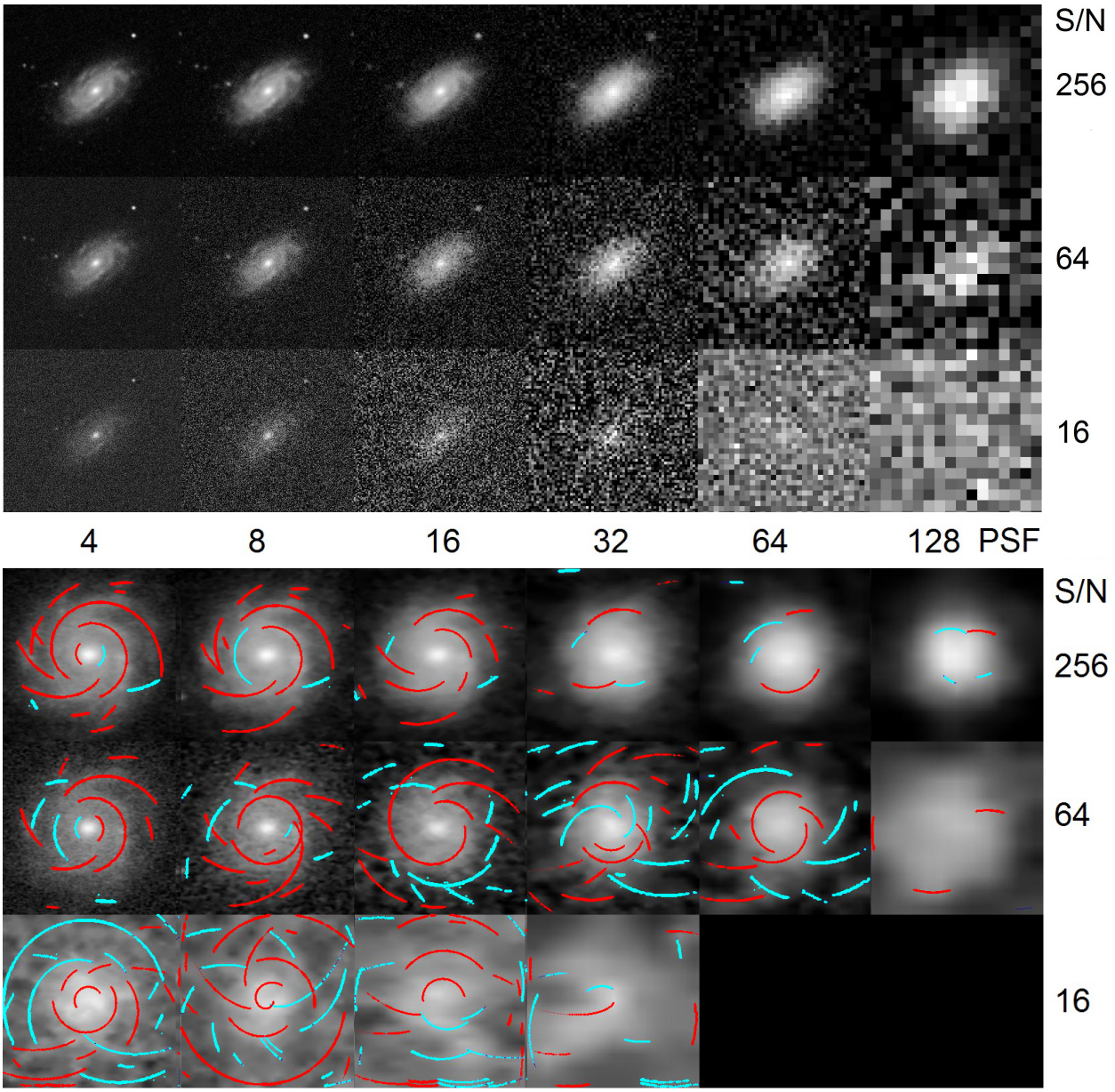


Figure 4.2: **Top**: FWHM(PSF) blurring and noise added to SDSS galaxy 1237648702972625038 using Sunpy. **Bottom**: The corresponding output images generated by SpArcFiRe, which have been cropped and de-projected so the disk appears face-on. The two black squares indicate that SpArcFiRe failed to find an object in the image. Observe that at least the global chirality is correct up to FWHM(PSF) 16 for S/N as low as 64, but by S/N 16 the output arcs are mostly noise—unsurprising because the arms seem to be invisible to the human eye in the input images as well.

of global pitch angle, since arcs whose sign disagree with the dominant chirality are much more likely to be noise arcs than real arms [Davis and Hayes, 2014]. Obviously, since the non-DCO mean pitch angle includes arcs of opposite sign, their mean will be closer to zero (ie., smaller in magnitude) than the DCO mean pitch angles. So for example in Figure 4.2, the red (S-wise) arcs are clearly the dominant chirality in the most clear images, and most of the cyan arcs are noise in those images; the absolute value of the mean pitch angle including sign will obviously be less than the absolute value of the DCO mean pitch angle, since in the former case there will be cancellation.

In order to determine the effect of image degradation on pitch angle, we first sort the clear (unblurred) images of our 7536 galaxies according to absolute value of their galaxy-level pitch angle. We refer to the upper quartile of this list as the "loosely winding" (high pitch angle) set, and the lower quartile as the "tightly winding" (low pitch angle) set. For all plots that we will show in Figures 4.3–4.8, we plotted three different values for both tightly and loosely winding galaxies. The first value is the absolute value of the mean pitch angle across the group. The second value is the absolute value of the difference between a galaxy's unblurred pitch angle and the blurred pitch angle. The third value is the *fractional* difference between the unblurred and blurred pitch angle. In each case we provide results both for the DCO and non-DCO cases.

Figures 4.3 and 4.4 show the above values as a function of FWHM(PSF), for DCO and non-DCO pitch angles, respectively. Figure 4.3 shows how the DCO pitch angle changes as we increase the FWHM(PSF) with a constant S/N of 256 (the best S/N value). The red and yellow lines depict absolute pitch angles of loosely wound and tightly wound galaxies, respectively; the pitch angles of these two groups of galaxies become indistinguishable at around FWHM(PSF)=40. The green and the blue lines depict the (absolute value of) difference between the blurred and unblurred pitch angles for loose and tight, respectively; the curve depicting loosely wound galaxies is always above that for tightly wound, meaning,

presumably in proportion to the fact that loosely wound pitch angles are larger than tightly wound ones. Finally, we also plot the *fractional* differences which are the black and the purple dashed lines, as the purple (tightly wound) and black (loosely wound) lines. We see that, as a fraction, the tightly wound galaxies are more strongly affected by blurring than the loosely wound ones.

Figure 4.4 is similar to Figure 4.3, but including all arcs (not just DCO ones). The qualtitative observations are similar to those of Figure 4.3, thought not as stark since there is more noise to start with in the non-DCO analysis of pitch angle.

Figures 4.5 and 4.6 show the average pitch angles (DCO and non-DCO, respectively) as a function of signal to noise ratio, with the FWHM(PSF) held constant at 4 (the best FWHM(PSF)). Again, most of the qualitative observations of 4.3 and 4.4 hold, and the difference in pitch angle between the tightly (red) and loosely (yellow) wound arms disappears at about a S/N of 16, just as it disappears at FWHM(PSF) 40 in the previous plots; and again, the tightly wound (purple) arms suffer a larger percentage perturbation than the loosely wound (black) arms. Also, the non-DCO pitch angles (Figure 4.6) are more adversely affected than the DCO (Figure 4.5) ones.

In Figures 4.7 and 4.8, we combine the effects of FWHM(PSF) and S/N together by setting the product of the two to 1024. Thus, as the FWHM(PSF) increases, the S/N decreases in tandem to keep the product at 1024. We can see that when both types of degradation are applied together, the resulting pitch angle measures degrade far more quickly, with the difference between the two sets vanishing at about the point where the FWHM(PSF) is about 16 (S/N of 64) for the DCO arcs, and even earlier for non-DCO arcs—about FWHM(PSF) 8, S/N 128.

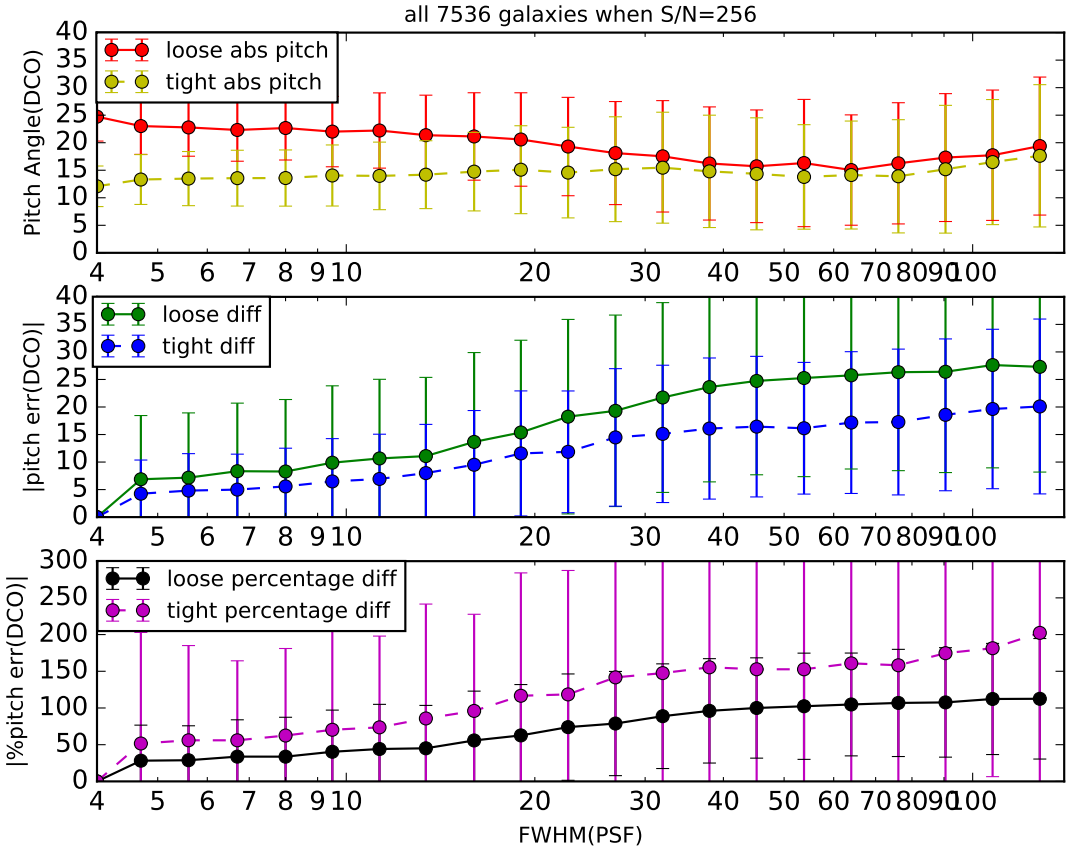


Figure 4.3: Various types of changes in the pitch angle as a function of FWHM(PSF), comparing galaxies in the upper ("loose") and lower ("tight") quartiles of pitch angle measured on the unblurred images. In the first vertical plot, the solid red (loose) and dashed yellow (tight) curves show the mean pitch angles of the two groups as a function of FWHM(PSF); clearly the red (loose) arms have a higher pitch angle at the low end of FWHM(PSF); interestingly, the two averages meet at around FWHM(PSF) 40, meaning that effectively the difference between the two has become invisible. In the second vertical plot, the solid green (loose) and dashed blue (tight) curves demonstrate how the error in absolute value of pitch angle increases with FWHM(PSF), as expected, and that the error in the loosely winding arms increases more rapidly. However, in the third vertical plot, the solid black and dashed purple lines demonstrate that *as a fraction*, the error of the tightly wound arms increases more rapidly. All of these measures are averages that discard arcs that wind in the non-dominant direction. ("DCO" on the vertical axis means "dominant chirality only".) The S/N ratio is held constant at 256 (the clearest S/N) throughout. Error bars in all cases are 1 sigma.

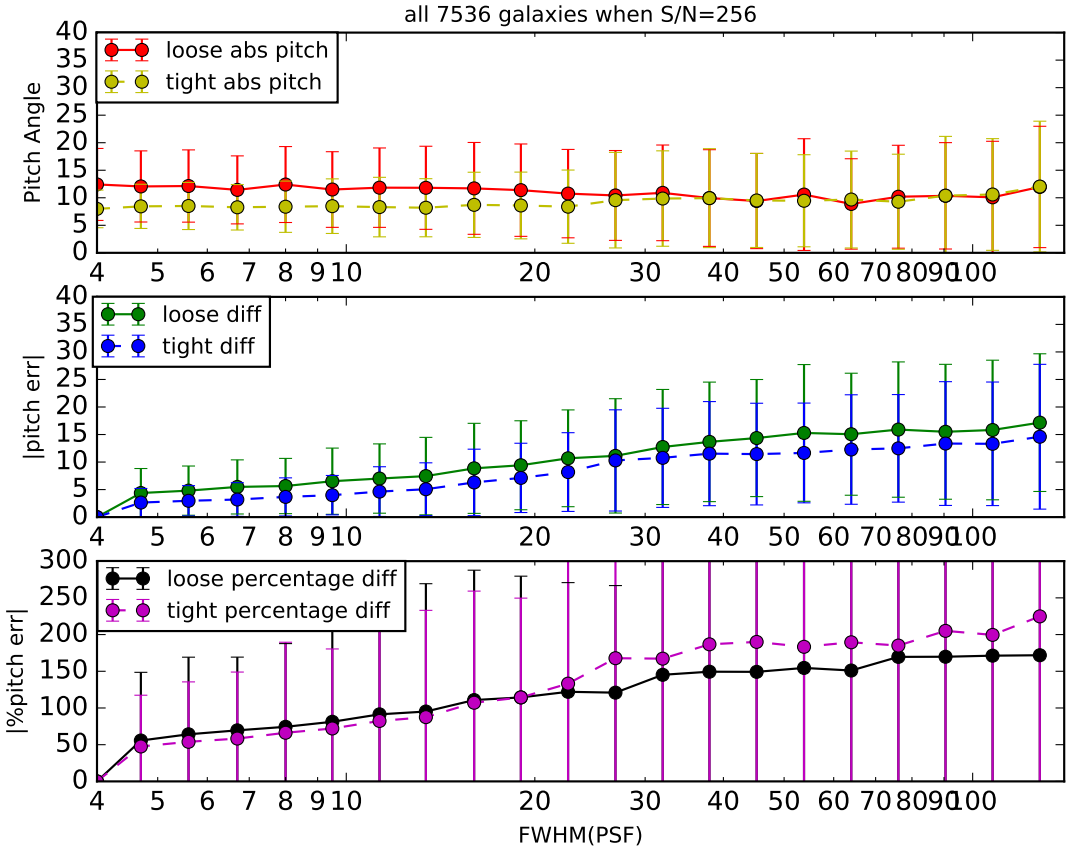


Figure 4.4: Exactly the same description as Figure 4.3, except the means now include arcs that wind in the non-dominant direction (which are often but not always noise). Most of the above observations still hold qualitatively.

4.3.2 The spirality selection effect

As we have seen, a galaxy's measured pitch angle will change as the image degrades, and the mean measured pitch angle in a group of selected galaxies will change correspondingly. Note, however, that it only makes sense to measure the pitch angle of a spiral galaxy; elliptical galaxies don't have arms. Recall that to make the observation depicted in Figure 4.1, we selected galaxies for which some fraction of Galaxy Zoo Humans said that they saw spiral structure; this fraction determines the *spirality* of the galaxy's image, and putting a threshold on the spirality gives us a selection of galaxies for which we have some level of certainty of them having visible spiral structure. In order to duplicate the observation of

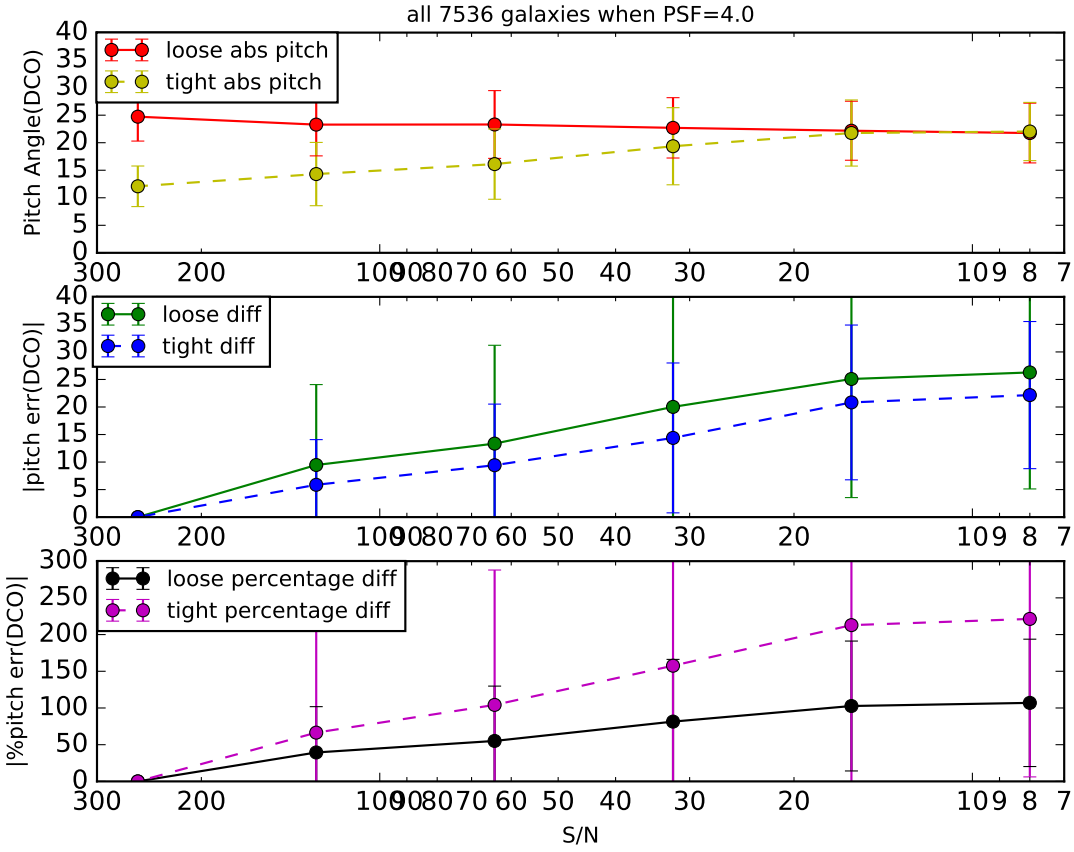


Figure 4.5: Similar curve descriptions as Figure 4.3 except now the FWHM(PSF) is held constant and the S/N is changed; highest S/N are on the left, with images degrading towards the right. The tight and loose pitch angles become indistinguishable at about S/N=16. Note the change in tight pitch angles is quite a bit more strongly affected by noise than it was in the blurring case; we are not sure why. As with Figure 4.3, arcs of the wrong winding direction are excluded.

Figure 4.1 on a set of artificially blurred galaxies, we need a way to estimate how humans would have voted on their spiralities. For this we turn to our machine learning expertise. We have previously described a machine learning algorithm that demonstrated how we could eliminate a human-created bias in a winding direction survey [Hayes et al., 2016]. Here we extend that algorithm and make it more accurate and robust. The main difference here is which attributes our random forest is allowed to use as inputs. Since we ran SpArcFiRe on all the blurred images, we are allowing the machine to use most of SpArcFiRe's outputs. We also allow it to use colors and magnitudes from SDSS, because we assume those would

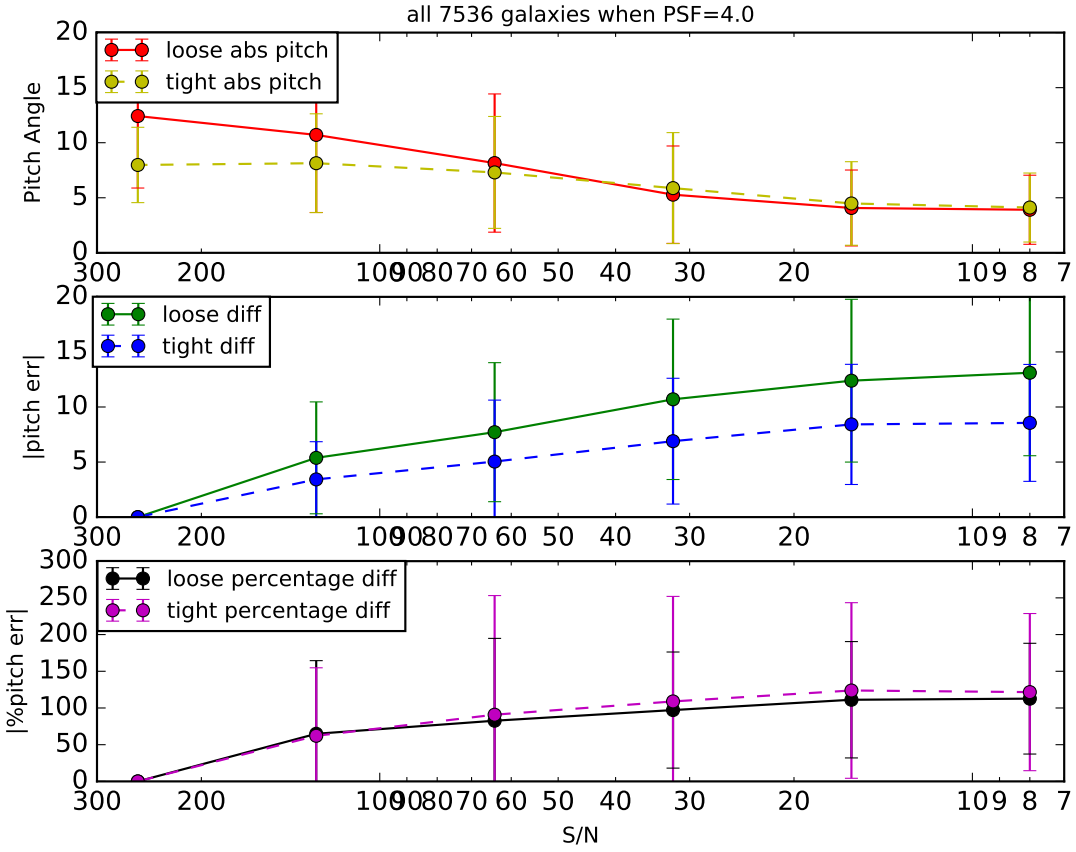


Figure 4.6: Similar to Figure 4.5 but including arcs of the wrong winding direction.

not change when the image becomes blurred or noisy. However, some parameters that come from SDSS may vary with the noise added by blurring the objects. A more sophisticated analysis would be required to decide if, for example, Petrosian radii vary with noise, but that is outside of the scope of this work. In order to avoid bias introduced by this potential issue we decided not to use such attributes. We ended up with 105 features per object, and we only used the following classes of variables (the ones that we believe are unaffected by blurriness) from SDSS: absolute colors, de-reddened magnitudes in the g, i and r bands, and k-correction for $z = 0$ in all bands.

We used a random forest with 35 trees, with a 95-5 split for training and testing. As a measure of how good our spirality prediction is, we compute the root mean squared error

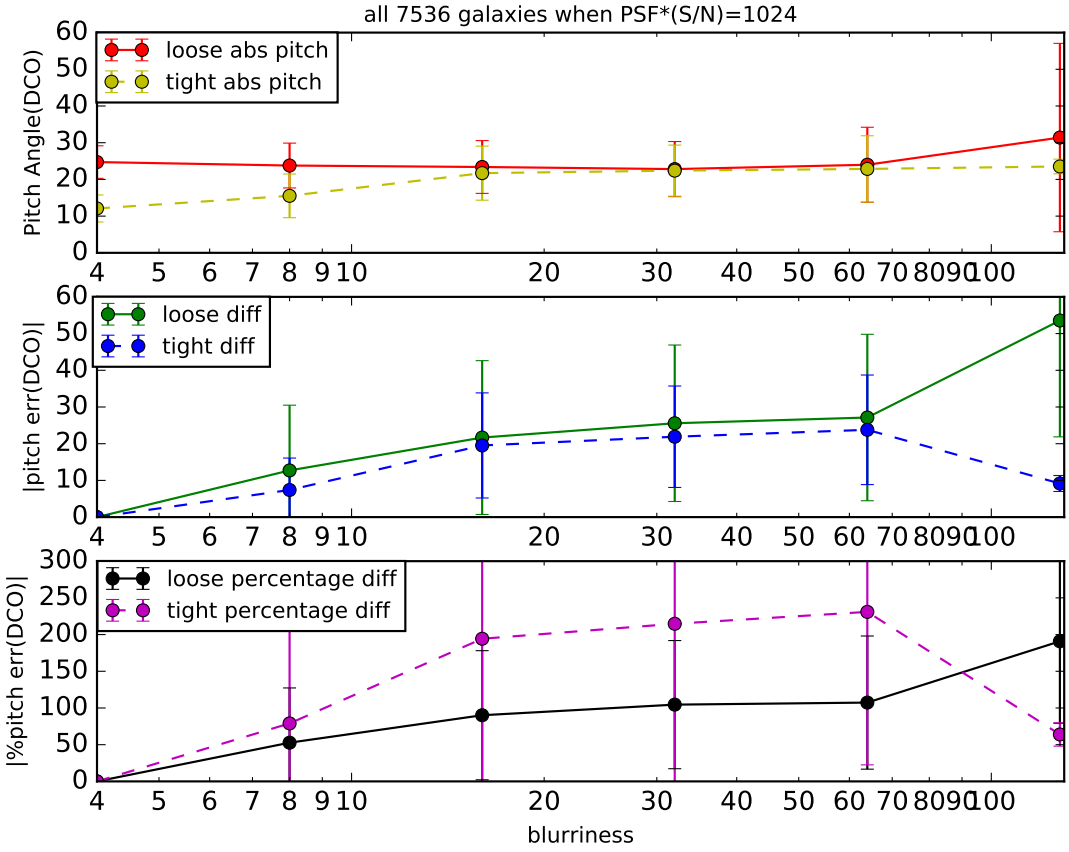


Figure 4.7: An attempt to account for degrading both S/N and blurring in one plot: on the left we have FWHM(PSF)=4 and S/N=256. As we move to the right, we increase FWHM(PSF) and decrease S/N simultaneously to maintain their product at 1024. As may be expected, the loose (solid red) and tight (dashed yellow) meet each other earlier, at a FWHM(PSF) of 16 (S/N=64), which is earlier in both measures than occurred in either individually. Excludes arcs of the wrong chirality.

(RMSE) between the predicted and real spirality for the 5% of the set called the test set, after training on 95% of the initial set. Our final RMSE was 0.14, slightly higher than the ones reported in Hayes et al. [2016], which is expected since we are using fewer features.

In Hayes et al. [2016], the RMSE in predicting spirality was about 0.137 across all spiralities, but more detailed analysis has shown that the RMSE in predicting spiralities was heavily skewed: near spirality zero (ie., for elliptical galaxies), our machine was extremely accurate, having RMSE in the 0.02 range; but at the high end of spirality (above about 0.7), the

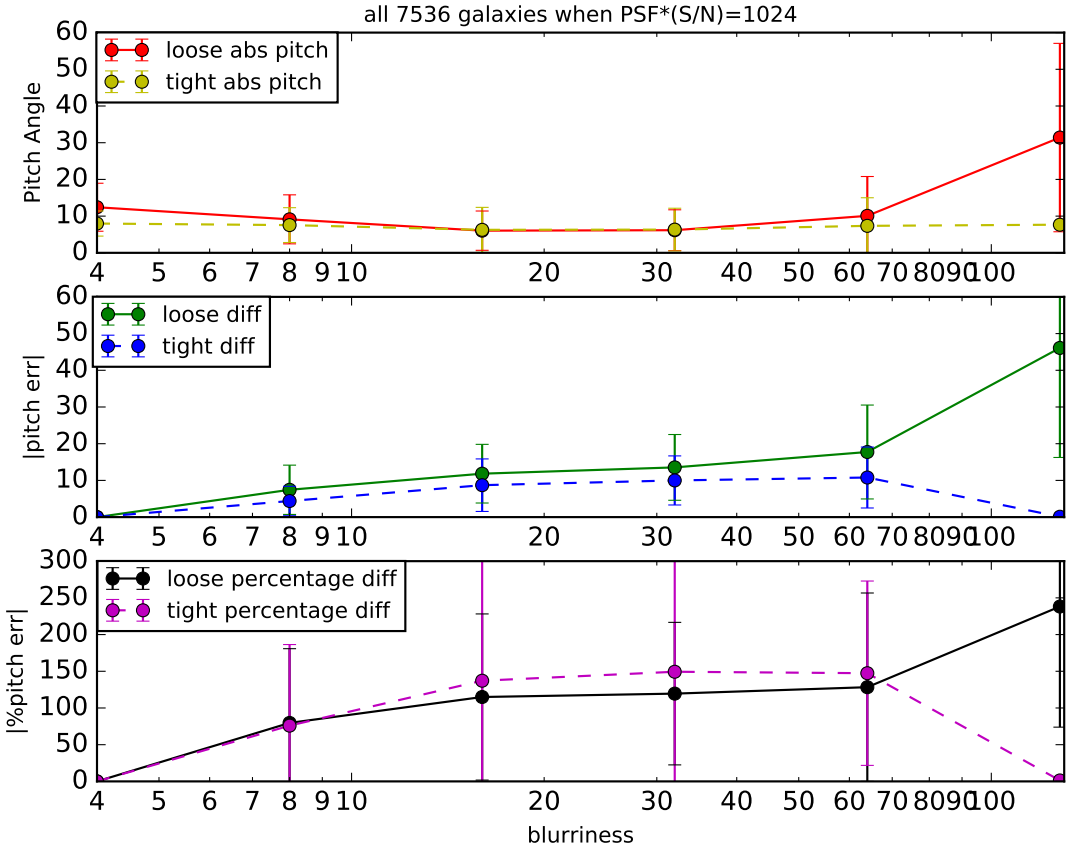


Figure 4.8: As Figure 4.7, but now including arcs of the wrong chirality. Now the loose and tight arcs become indistinguishable even earlier, around FWHM(PSF) 8 or perhaps slightly higher (S/N $\approx$ 128).

RMSE was much larger, in the 0.30 range. This is precisely the opposite of what we want, because we want to be able to precisely pinpoint which galaxy images indeed show spiral structure, not which ones do not. Further study into the issue revealed that the problem was that there are somewhere between 2x and 6x more elliptical galaxies than face-on spiral ones in the Galaxy Zoo sample[5], which simply made our machine train more intensively to reduce its error on that most voluminous sample of galaxies (the ellipticals), at the expense of making larger errors on the much smaller sample of spirals.[6] The solution is simple: we

[5]There are 2x as many ellipticals as spirals if you insist on 90% certainty in both; if you only insist on 50% majority then the number is closer to 6x.

[6]Note that the issue is *not* that there are not enough spirals to train on. Tens of thousands of spirals is plenty. The problem was just that the machine "tried harder" to reduce the error of prediction for the

reduced the number of ellipticals that we trained on by a factor of 16, thereby reversing the trend: there are now 2-6 times as many spirals (a few tens of thousands) and only a few thousand ellipticals to train on, so now the machine is very good at recognizing galaxies with high spirality (say, above 0.7) and less good at predicting the spirality of galaxies at the low-spirality end. This is fine, because we really don't care how bad the machine is at predicting spirality of any galaxy with spirality less than 0.5, so long as it doesn't say the galaxy *is* a spiral galaxy. This we have achieved: our RMSE is now about 0.15 across the entire spectrum of spiralities, which has reduced our RMSE at the high end by a factor of 2.[7]

With this new machine, we can now estimate spiralities for our set of 622,585 degraded galaxy images. The results we report below are of an analysis made on the blurred objects only, which were not part of the training nor test set.

Figure 4.9 shows spirality as a function of FWHM(PSF) and S/N. We can see that the spiralities of galaxies almost uniformly become smaller as the FWHM(PSF) increases, although the interaction with S/N is more complex. Note that in the high S/N case (256), the spirality drops very quickly with increasing FWHM(PSF): this is because, in the absence of also adding noise, the images become smooth blobs which effectively look elliptical, so the machine correctly labels these images as having low *observable* spiral structure. As the S/N decreases, the added noise is sometimes interpreted as arcs, and our machine (which looks for arcs) mistakes these for spiral structure; this is why, as we decrease the S/N, the spirality decreases more slowly with increasing FWHM(PSF). This effect is greatest around S/N 16 or 32, but once the S/N drops to 8, the spirality again drops sharply with increasing FWHM(PSF) as the machine starts to recognize the images as nothing but blurry noise.

Figure 4.10 depicts the number of artificially degraded galaxy images that have a spirality

largest sub-sample it could find, which was the ellipticals.

[7]As we have mentioned in Hayes et al. [2016], this RMSE is worse than the Kaggle winner's 0.07, but the Kaggle winner [Dieleman et al., 2015] made no attempt to detect or reduce human-induced biases.

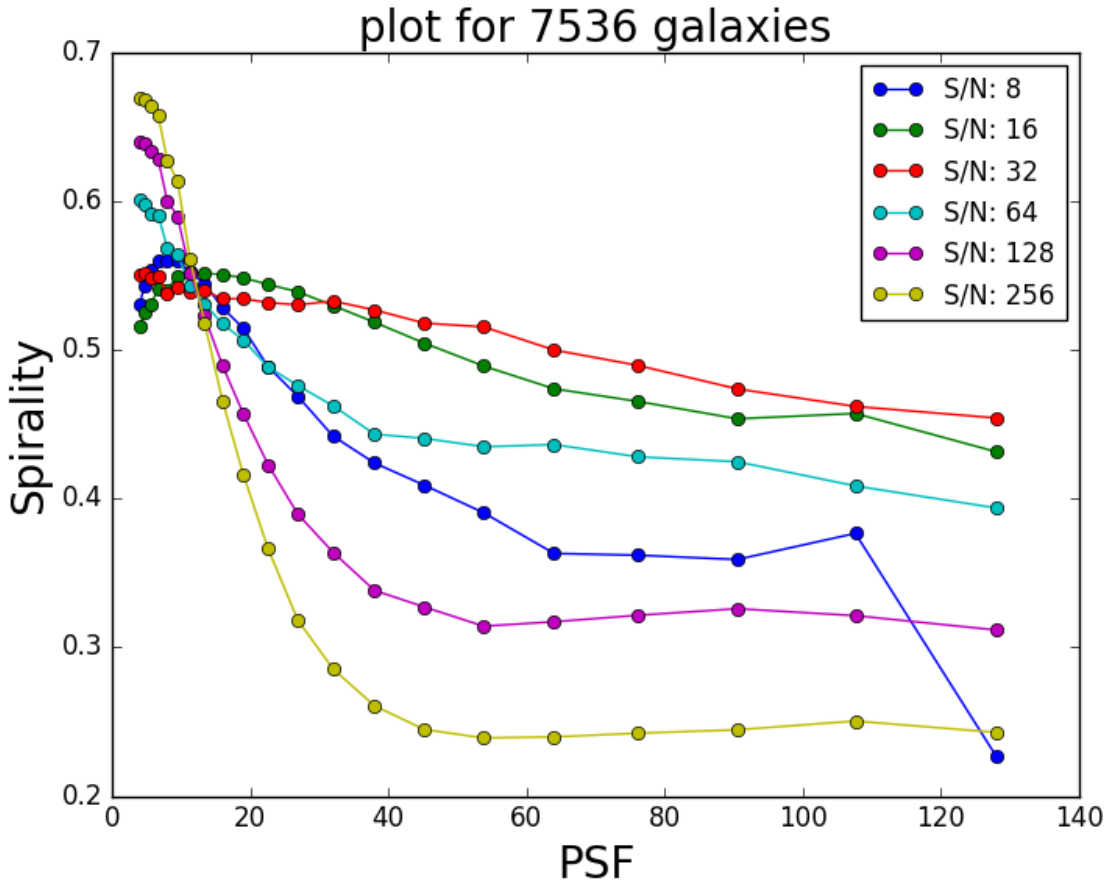


Figure 4.9: Mean spirality decreases with FWHM(PSF), although the interaction with S/N seems more complex.

greater than a threshold of 0.7, as a function of FWHM(PSF) and S/N. It shows that as the blurriness gets bigger, there are fewer images that meet the threshold of 0.7 in spirality.

Figure 4.11 depicts the mean pitch angle (DCO) of galaxies meeting the 0.7 threshold in spirality, as a function of FWHM(PSF) and S/N. It is fairly clear that the mean pitch angle decreases with decreasing S/N, but the relationship between pitch angle and FWHM(PSF) is less clear in this group; except for the highest S/N case, there is a sharp drop in mean pitch angle as the FWHM(PSF) grows at the small end, but then the pitch angle slowly increases again once the FWHM(PSF) is above about 20. We are not sure why this is; however, as the next section shows, all of this data comes together nicely when we set the same selection criteria for real and artificially blurred galaxies.

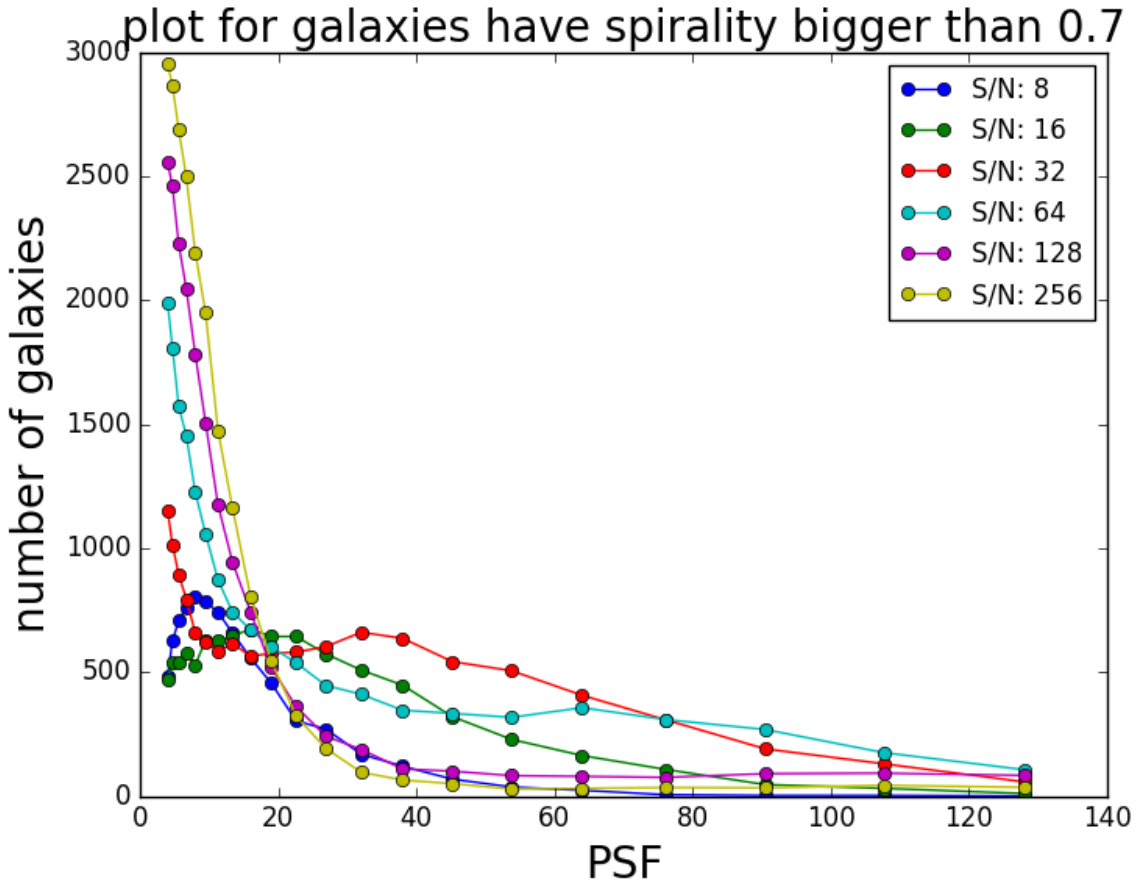


Figure 4.10: Heading towards explaining the selection effect in SDSS, we impose a cutoff of 0.7 in spirality, and count how many galaxies in our sample have spirality above 0.7 after being degraded in both FWHM(PSF) and S/N. We see the number of galaxies above the 0.7 spirality cutoff decreases rapidly with increasing blurring and decreased S/N.

4.3.3 Comparing real *vs.* blurred images

Despite the complex interplay between pitch angle, FWHM(PSF), S/N, and spirality threshold described above, we now show in Figures 4.12 and 4.13 that our blurred images successfully mimic how pitch angle is affected by image degradation in the real sky.

We hypothesize that the real issue is not with any one of galaxy angular size, FWHM(PSF), S/N, or spirality. Instead, we observe that on any given night, there could be different FWHM(PSF) values at the observing site, differing levels of noise, etc. Thus, we choose a relatively stringent threshold of 0.8 in spirality, and look at the pitch angle of galaxies

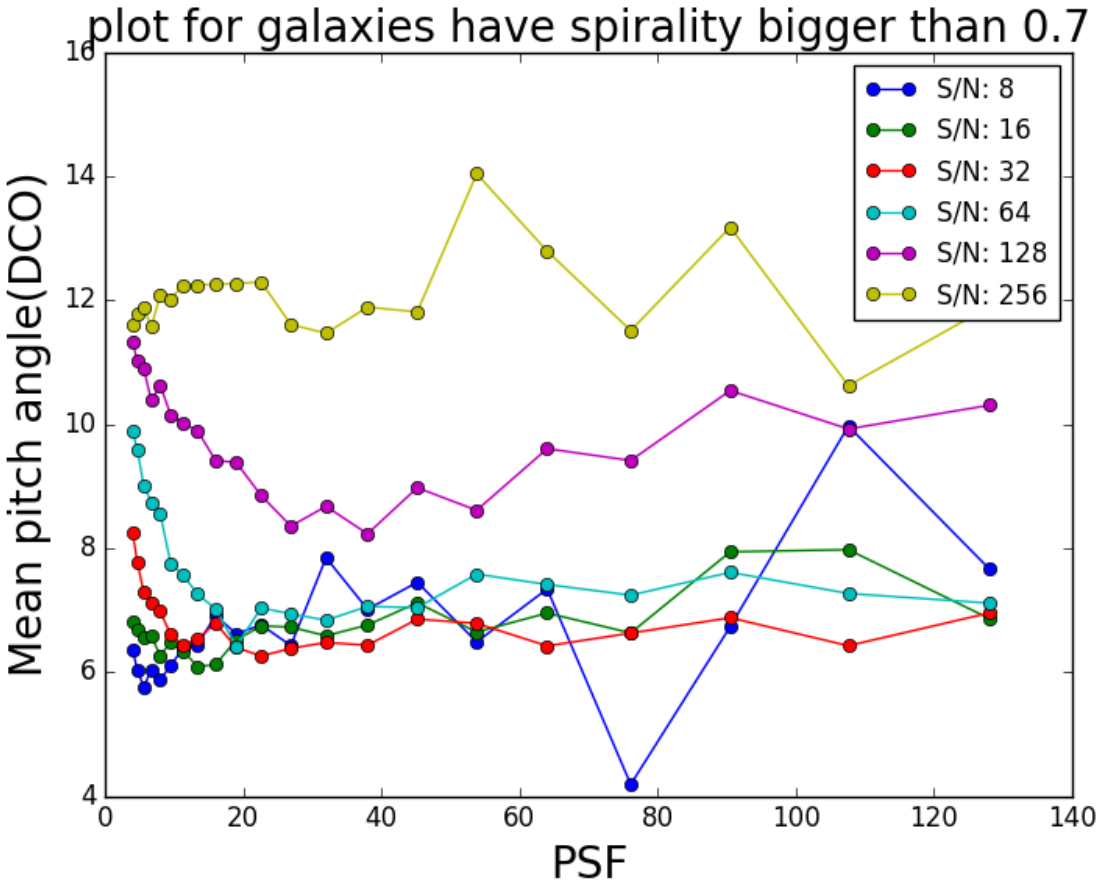


Figure 4.11: The vertical axis of this plot is the mean pitch angle that have bigger than a 0.7 spirality. And the horizontal axis of this plot is FWHM(PSF). we also used different colors to represent S/N values.

as a function of *radius in units of the FWHM(PSF)*, and allowing any S/N that gives a galaxy above the spirality threshold. Figure 4.12 demonstrates that, using these criteria, the pitch angle as a function of (radius/FWHM(PSF)) for real SDSS images ("SDSS-all"), vs. images from our artificially blurred set, give mean pitch angle curves that are virtually indistinguishable from each other. Similar plots occurs for other spirality thresholds.

At this point, it appears we have shown that there is a selection effect such that tightly wound spiral arms are harder to see than loosely wound ones. It stands to reason, then, that it should be possible to mimic the procedure used to produce complete volume-limited samples. In particular, if our hypothesis is correct, we should be able to create a lower threshold in

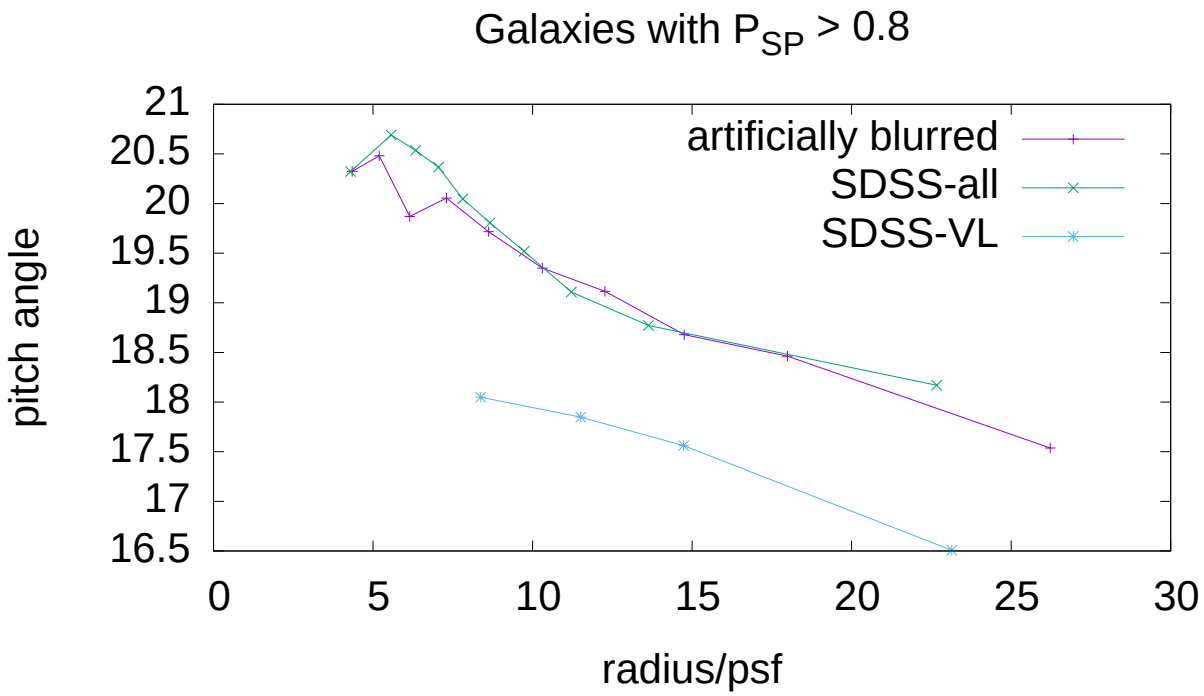


Figure 4.12: The line SDSS-VL is the exact same set of (3622 volume limited) galaxies with $P_{\rm SP} > 0.8$ as Figure 4.1, but plotted against radius/FWHM(PSF); the SDSS-all curve is all (36,384) SDSS galaxies for which PS¿0.8, which will include some dimmer, closer galaxies than the volume limited sample; and the purple curve are (24,347) images from our 7536 nearby galaxies that have been artificially blurred, chosen using similar criteria to the "SDSS-all" sample. As can be seen, the artificially blurred sample does a very good job of matching the "SDSS-all" sample, corroborating our selection effect hypothesis.

pitch angle, and eliminate all galaxies below that threshold. With a high enough threshold, we should be able to produce a sample of galaxies that avoid the selection effect because their pitch angles are all big enough to avoid the selection effect. We test this procedure in Figure 4.13. While we do see that the selection effect is reduced as the pitch angle threshold is increased, there is no obvious cutoff that seems capable of completely eliminating the effect of "lower pitch angle with increasing radius in units of the FWHM(PSF)". Clearly, there is more here than we have been able to discern; the details of how to account for or correct for this pitch angle selection effect will presumably depend on the details of the data being used and the measurements and analysis being done with the data. The results presented here clearly demonstrate that a selection bias exists in terms of detecting spirality; future

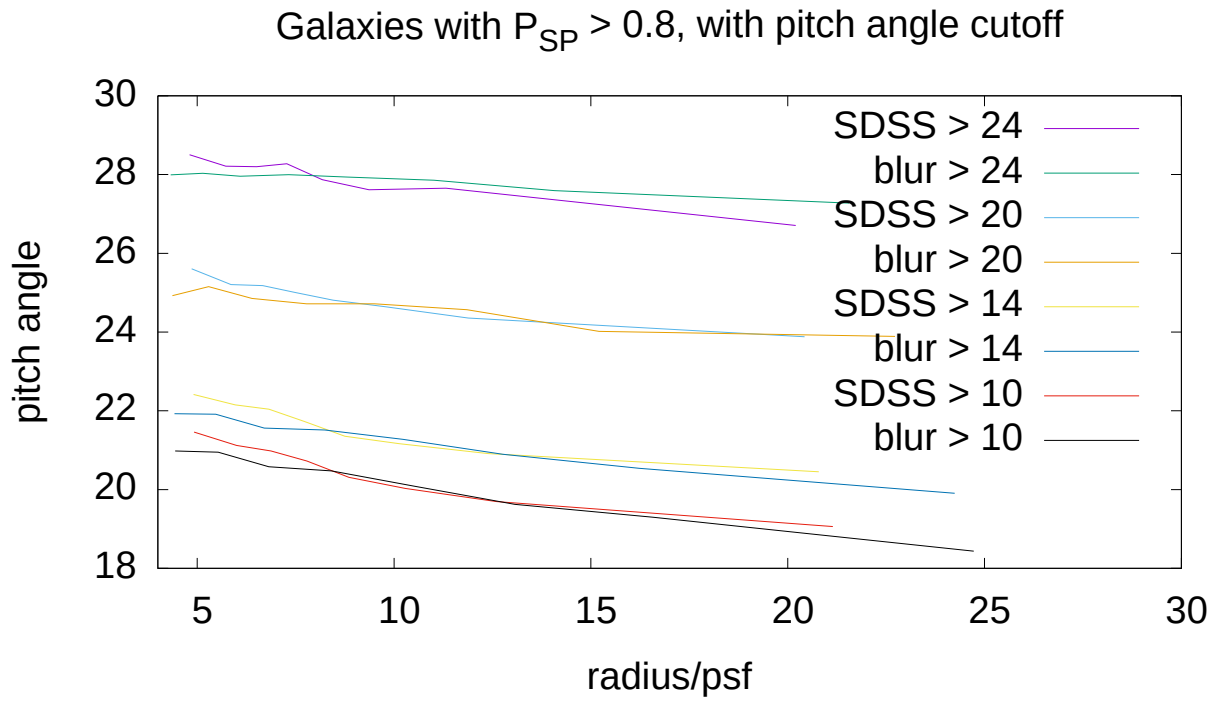


Figure 4.13: Another comparison of measured pitch angles between real SDSS galaxies vs. artificially blurred ones. Here, all galaxies have a spirality $P_{\rm SP} > 0.8$, and we have attempted to account for the selection effect by imposing a lower limit on pitch angle, to remove galaxies with too-low pitch angles from the sample. As we can see, with equal selection limits, the real and artificially blurred galaxies have a statistically identical mean pitch angles for the 4 values of pitch angle thresholds 10, 14, 20, and 24 degrees. This again suggests we are correctly modelling how observed pitch angle changes with image degradation. However, the slope of all the lines are still negative, suggesting that none of the thresholds we have chosen are strong enough, although the slopes become less negative with increasing threshold. Above a threshold of 24 degrees, the sample sizes become too small to be meaningful.

programs to quantify spiral structure (or its evolution) in survey data will need to take this effect into account, but this will probably require simulations tailored to the details of the data and methodology being used.

4.4 Conclusion

We have demonstrated that there exists a morphological selection effect when attempting to isolate "spiral" galaxies based upon a threshold in the human Galaxy Zoo 1 votes. In particular, it seems that tightly wound spiral arms, being less visible with increasing image degradation, are less prone to be included in a selected set of galaxies based on said spirality threshold. We were able to reproduce the effect to high precision by creating a set of artificially degraded images in tandem with a machine learning algorithm to predict the spiralities of said blurred images. However, we were not fully successful in attempting to account for the selection effect by adding a threshold to pitch angle in an effort to explicitly eliminate tightly-wound spiral galaxies. Further work will be required to determine if the pitch angle variations we observe are some deeper selection effect, or a physical effect in the real universe.

Our analysis was restricted only to the SDSS red band images, since that band provides the greatest number of clear images. However, studying these effects in other wavebands will be important as we attempt to measure pitch angle differences across wavebands [Martínez-García et al., 2014, Davis et al., 2015, Pour Imani et al., 2015, Pour-Imani et al., 2016].

Chapter 5

A Classifier based on a linear combination of feature vectors

5.1 Motivation

Our work on galaxy morphology classification has led to us questioning some of the current methods widely used to classify galaxies. Like we explained in chapter 2, although it is much easier for us to categorize some objects into classes, we believe they are not very representative of the problem. We believe regression is more suited to perform these tasks because the concept of spirality in and on itself is a continuous measure, but even just regression is insufficient. We would like to understand what does it mean when an object is 60% spiral, where do the remaining 40% lie?

Take, for example, figure 5.1, which shows the late states of a merger between two galaxies, but it still presents spirals structure. So is that 60% merger and 40% spiral? Ideally, we would like to have a method that is not only able to quantify how much a particular object belongs to a class but how it relates between classes. Another valuable desirable property

Figure 5.1: Hubble image of NGC 2623, a merger of two spiral galaxies.

is being able to identify when an object does not belong to any of the proposed classes. An excellent example of that is Hanny's Voorwerp [Lintott et al., 2009] shown in figure 5.2, a rare type of astronomical object called a quasar ionization echo, discovered during the Galaxy Zoo project. If we want to handle large surveys for ML systems to classify, we would like to be able to know about theses occurrences automatically. The problem is that most ML models tend to use either a softmax function for multi-class classification in order to get a probability distribution of the outputs that add up to one, or a variation of this technique. In this chapter, we describe a new method we developed with these goals in mind.

5.2 Algorithm Summary

The algorithm works by solving a linear system in the form of $A * x = b$. Given an input matrix *input* with pre-known, well-defined classes $C_{1..k}$, and features $F_{1..m}$ we create a $m \times k$ matrix A. A is built using the code in algorithm 1. To compute $A_{i \times j}$ we first select all the rows in the input data that have label C_i and then from that subset of A we select feature F_m

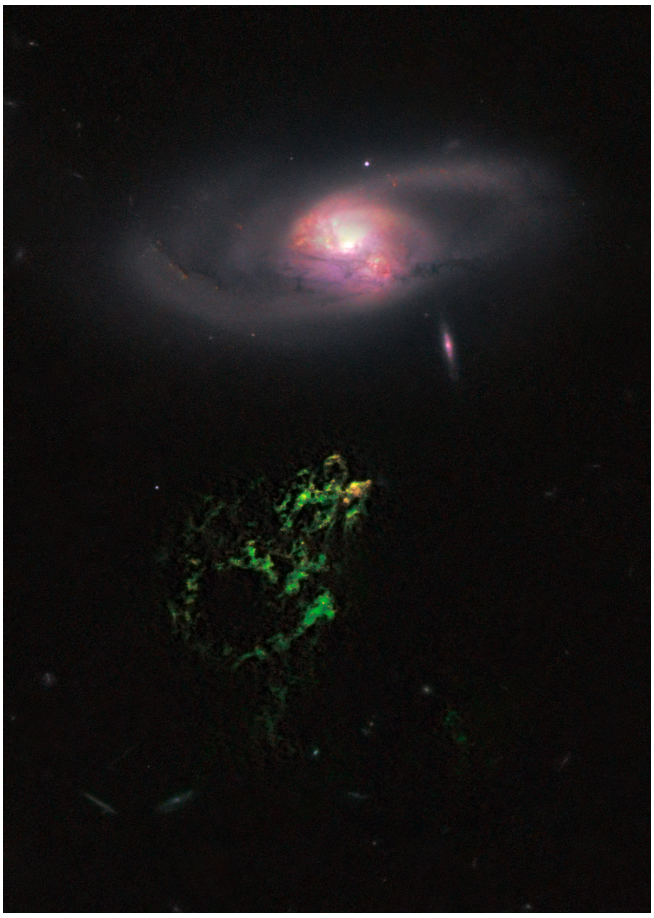


Figure 5.2: Hubble image of Hannys Voorwerp at the bottom, a rare object discovered during the Galaxy Zoo project, and the spiral galaxy IC 2497 at the top.

which yields a subset of the input matrix $input_{p\times 1}$, where p is the number of samples in the *input* labeled C_i. We then compute the probability density function using a kernel density estimation (KDE) on this subset. From there, we pick the value of x that corresponds to the peak on y in the KDE. That value, the mode of the distribution, is going to be the value for $A_{i\times j}$. We do this for all features and classes until we fill out all the values of A.

Given a new instance, b is simply the k-vector representation of its features. We then solve for the n-element vector x using the standard $A * x = b$ linear equation. This results in a x vector that has scalar elements $x_{1..k}$, where each element corresponds to a label, so the magnitude of each element will correspond to its similarity to the samples with that label.

Algorithm 1 Building the A matrix

```
Given an input matrix
for all Features of input do
    for all Classes of input do
        x_fit ← select all samples of input with label C_j at feature F_i
        kde ← compute kernel density estimation of x_fit
        A[i, j] ← mode of kde
    end for
end for
```

Since we observe the magnitudes of the outputs to perform our final classification, all the features must be on the same scale, otherwise the method will be biased towards the feature with a higher magnitude. The easiest way to have this condition satisfied is by having all the values in *input* standardized feature-wise. This way, all the features will have the same scale, with $\mu = 0$ and $\sigma = 1$, while conserving their shape. Notice that b must also go through the same transformations *input* went through.

5.2.1 Algorithm Complexity

As explained in 5.2, there are two main steps on this algorithm, building the A matrix and getting a prediction for a new sample. The first part is comparable to training against most ML methods. Our method takes $O(m \times k \times kdecost)$ where *kdecost* is the cost to compute the KDE for each element o A. This algorithm is heavily dependent on the quality of the KDE computed for each feature element of matrix A. The KDE, in turn, varies wildly with the bandwidth parameter, and choosing it is a non-trivial task. We performed a cross-validated grid-search over 30 possible values to pick this parameter. While this is very costly computationally, it guarantees that we have the most accurate KDE.

The cost of getting a prediction for a new sample is the cost of solving the linear system. That cost is generally between $O(n^2)$ and $O(n^3)$ depending on which algorithm is used to solve the system. We use the least-squares algorithm to solve our linear system, which solves

it by computing a vector x that minimizes the squared Euclidean 2-norm. The least-squares has a complexity of $O(k^2 \times m)$ [Tan et al., 2018] for m features, which determine the number of rows of the A matrix and k classes which determine the number of columns of the A matrix.

5.3 Similar Methods

In this section we discuss about some of the similarities and differences between our method and others that are closely related to it: Principal Component Analysis (PCA) [Pearson, 1901], Linear Discriminant Analysis (LDA) [Fisher, 1936, Rao, 1948], Neighborhood Components Analysis (NCA) [Goldberger et al., 2005], and Othorgonal least squares [Chen et al., 1991].

PCA is unsupervised, so it does not take label information into consideration, and is generally used as a pre-processing step on the ML pipeline for dimensionality reduction. It uses the eigenvectors of a co-variance matrix to find the directions that maximize the variance in the dataset. LDA has a similar approach, using eigenvectors, but it uses the labels of the dataset, so it is a supervised approach, like ours, and it tries to maximize the separation between multiple classes. It requires labelled data so that scatter matrix both within and between each class can be calculated and then used to maximize the distance between the eigenvectors that represent the data. NCA is geared towards classification like our approach and adjusts its linear transformation to maximize its softmax function over distances. It also is capable of reducing the dimensionality of the data using a rectangular matrix to project every example onto a smaller subspace.

Orthogonal least squares is a sequential selection process in which at each step the next data point to be chosen as a basis function center corresponds to the one that gives the

greatest reduction in the sum-of-squares error Bishop [2006]. This methods poses the task as an optimization problem and focuses on finding values for the expansion coefficients. Our method however, focuses on finding a feature vector distribution per class that is more representatives of the pair, so when a new sample b is classified the coefficients of x can be interpreted as the classification vector for the new sample. While we use the kernel density estimator as a step to achieve that end other functions and kernels could be used to the same end.

5.4 Experiments

We decided to test the efficacy of our algorithm on two well-known datasets available at the UCI Machine Learning Repository [Dua and Graff, 2019]. We tested our method on the Iris Dataset, a dataset that contains 150 samples and 4 measurements of data to quantify the morphological variation of Iris flowers of three related species [Anderson, 1935, Fisher, 1936]. We also tested it on the Wine Dataset, a dataset with 178 samples with 13 continuous features that are the results of a chemical analysis of wines grown in the same region in Italy but derived from 3 different cultivars [Forina et al., 1998].

We performed a stratified sampling of 80% of the data to be our training set and used the remainder 20% as our test set. For the wine dataset, we achieved an accuracy of 94.4%, correctly predicting 34 out of 36 samples on the test set. Figure 5.3 shows the KDE for the features on the training set for the Wine dataset, used to build its matrix A.

We did the same procedure on the iris dataset and achieved an accuracy of 66.6% correctly predicting 20 out of 30 samples on the test set. This performance was below our expectations, so we set out to understand why, and the first half of figure 5.4 provides us with some clarification. The first 3 features of classes *versicolor* and *virginica* have a very similar

distribution. That means those classes are harder to separate, and our algorithm makes more errors in that scenario. To help disentangle those classes, we decided to insert some new features on the dataset, which are just a non-linear transformation of the original features. For this example, we added the cosine of the original features. Their distribution can be seen in the second half of figure 5.4. This approach not only doubles the number of features we have, but they seem to add enough variability to the data that leads to an improvement in performance. Using the original features and the cosine features, we achieve an accuracy of 90% correctly predicting 27 out of the 30 samples of the test set.

5.5 Discussion

5.5.1 Understanding x

The way we determine which class a test sample belongs to is by examining the values of the x vector after solving our linear equation. x is a k-dimensional vector and each element corresponds to the similarity of b to its respective class. The values from x are not bounded, and our understanding is that the higher the value of x is, the closer to the original distribution of the class it represents it is. Highest magnitude (as negative values are possible) and closest to one or zero are all possible ways to rank the outputs as well. However, on all our experiments, the highest value provided the best accuracy, empirically corroborating our assumption.

5.5.2 The necessary variability of data

This process of building matrix A can be viewed as an extreme data reduction process where we are left only with k datapoints per feature. This means that this method is more suited

for multiclass classification, where there are a large number of non-sparse features. If these conditions are not met, some problems may arise, like the one we observed with the Iris dataset.

We cannot guarantee that A consists of linearly independent vectors because A depends on the distributions of the original data. For example, if the data's features are all normally distributed, then after standardizing, the mode of each feature is the same as the mean. A will consist of 0's and hence will not have a pseudo-inverse, which means we would not be able to solve the linear system. Let us then assume that the data's features have sufficiently 'different' distributions. In this case, if the rank of A (i.e., number of linearly independent rows) is the same as one of A's dimensions (i.e., rank = number of rows, or rank = number of columns), then A will have a pseudo-inverse. Once we have a pseudo-inverse for A, we can compute the solution for x in $A * x = b$.

Suppose b is a vector whose coordinates are the modes of each feature, i.e., b is exactly one of the columns of A. Then x will be the unit vector with a 1 in the coordinate of the class of b and 0's otherwise. If all the feature vectors of each class are 'sufficiently distinct', in other words, if the modes of each feature is 'sufficiently' distinct between classes, then if we perturb b to be $b + \epsilon$, then since $Ax = \epsilon$ has a solution $x = x_\epsilon$, then we can find $x = x + x_\epsilon$.

5.6 Conclusion

This method is a simple approach to the classification problem in machine learning. Our tests demonstrate that it works very well on the datasets we tested on, but more extensive testing on harder datasets are necessary in order to draw more meaningful conclusions.

Improving the KDE calculation part would immensely benefit this method as the amount of computation grows linearly with the number of features and classes. Even testing on a

dataset with 10 classes and 784 features like the MNIST dataset [Lecun et al., 1998] would take a prohibitive amount of time while the same task can be achieved much faster by other methods. The same principle applies to the computations used to solve the linear system.

Our original intention with developing this technique was detecting when out-of-class samples and samples with characteristics from multiple classes are present. While the Wine and Iris datasets were suitable to evaluate general performance, we need more testing to make sure we can identify these other cases as well.

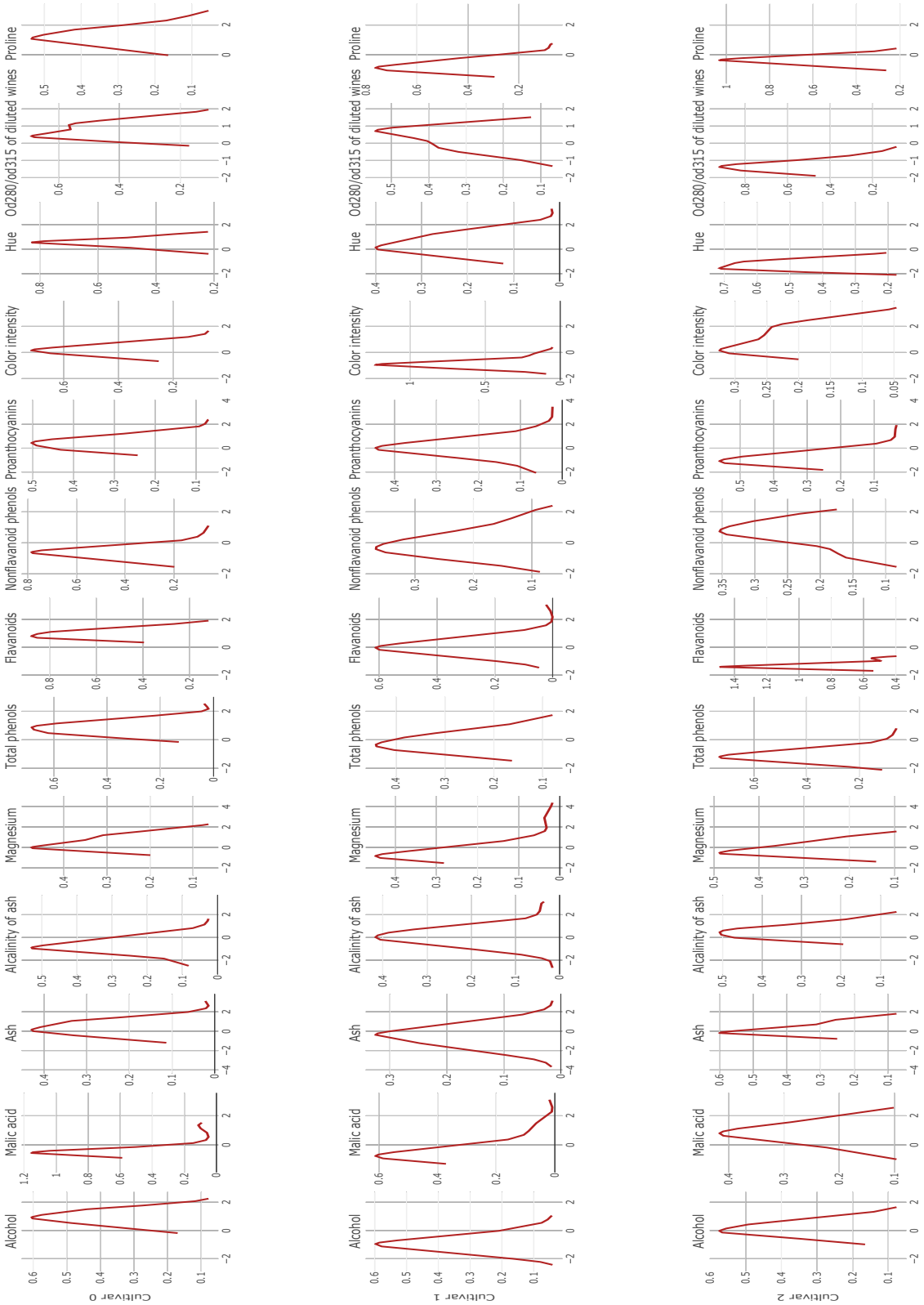


Figure 5.3: The feature distribution on the training set to build the A matrix for the Wine dataset.

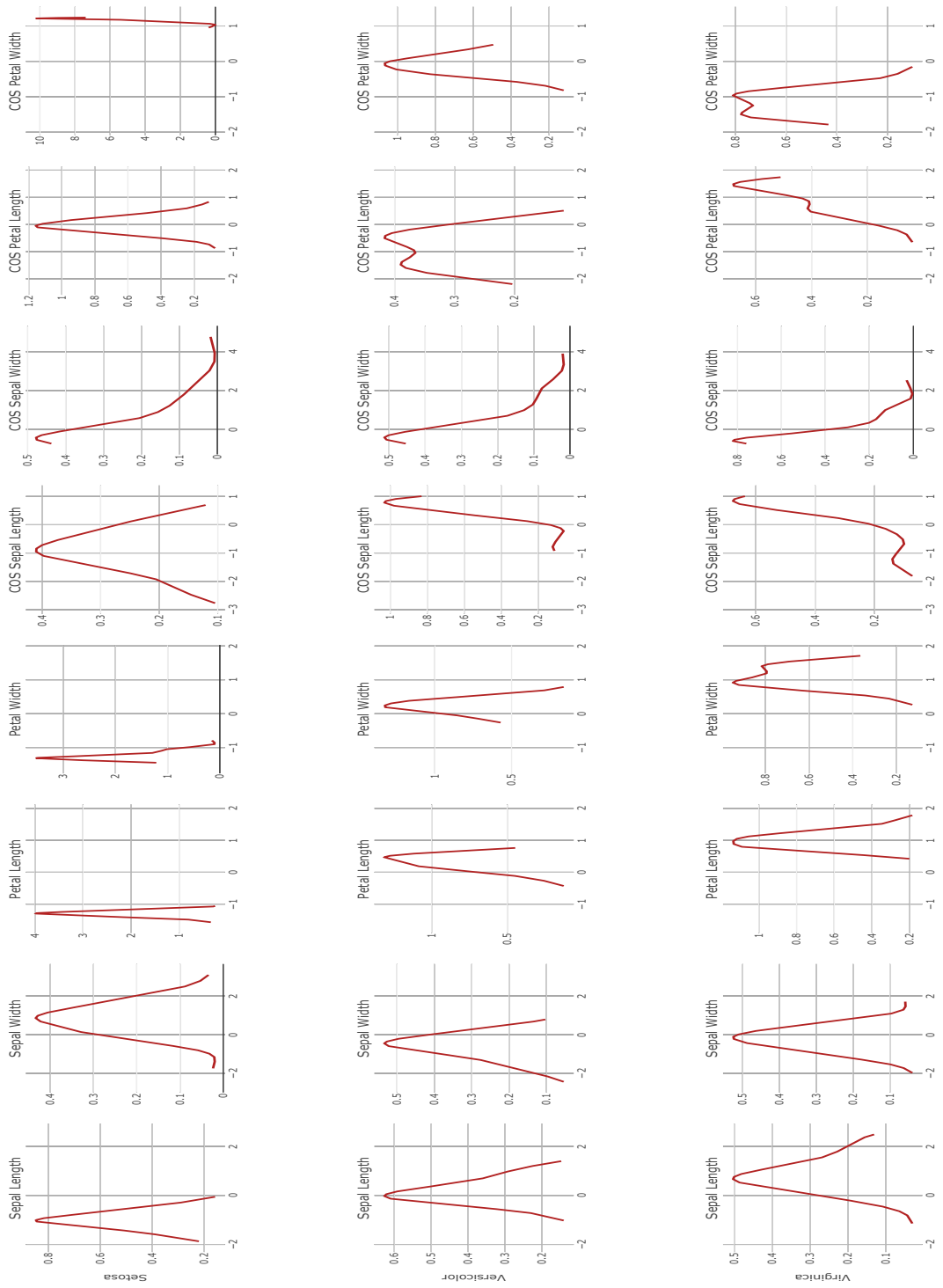


Figure 5.4: The feature distribution on the training set to build the A matrix for the Iris dataset.

Chapter 6

Future Directions

In this dissertation, we discussed how we built a random forest model that has demonstrated effective removal of bias with respect to chirality using SpArcFiRe's output and photometric data provided by SDSS, and how we used it to detect and correct a chirality bias present when selecting a sample of spiral galaxies at any spirality cut off. We also used a version of this random forest to aid in the detection of a selection bias that occurs due to reduced visibility of spiral arms with respect to their pitch angle.

Due to the size of future surveys, we believe that using machine learning to evaluate the morphology of unseen objects is the best approach to the problem. However, because of the known and unknown biases present on this dataset, we recommend caution when training such models. Ideally, we would like to be able to correct all the biases and republish this dataset with new labels, not only for the spiral class but for all the classes. As of right now, this is non-trivial because it also involves detecting and correcting biases present on other categories, but we would like this to be tackled again in the future.

In chapter 5, we introduced a new classification method based on a linear combination of feature vectors. We tested it on two traditional ML datasets, Wine and Iris, and it achieved

94% and 90% accuracies, respectively. Our goal with the introduction of this method is to identify samples that do not belong to any of the classes previously seen and samples that have characteristics of more than one of the classes. The idea is to inspect the output vector and make those decisions based on the magnitude of the coefficients and how closely related they are. This part of the work is still untested because building A and solving the linear system are still inefficient.

Some feature-class pairs may present multimodal distributions, and we believe we should take advantage of that instead of ignoring it by just using the value of the highest mode. This presents yet another challenge: how to incorporate a second or more mode into the matrix? Do we add new features (rows), or do we now break the original class into subclasses, which would also affect the distributions across the other features by adding new columns to A? These are essential questions to further the effectiveness of our method, and we leave it as future work.

We would also like to include a mechanism to help us decide when a class is not a suitable choice. Sometimes the output vector x contains more than one high value, and at the moment, we simply pick the highest value to determine as its class. However, we should also check the proximity of that value to the second and third highest to determine if the sample is not, in fact, one that presents characteristics of multiple classes or even one that does not belong to any class. One can do this by using percentile or fractional differences, or more sophisticated approaches.

We use the mode of the distribution of the features within classes to build our matrix A, while PCA, NCA, and LDA, all use different techniques for this step. It would be interesting testing if selecting the values for A using a different technique, like eigenvectors, can improve our results while preserving the desired properties of the algorithm.

Although we have shown that the algorithm works reasonably well, we must speed up the

calculation of the KDE with a minimal loss on the accuracy, and we would like to explore some faster methods for solving the linear system before proceeding with further testing. Although it served as inspiration for the development of this technique, scaling this method to larger datasets like astronomical surveys is a direction of future research.

Bibliography

M. Abd Elfattah, N. Elbendary, H. K. Elminir, M. A. Abu El-Soud, and A. E. Hassanien. Galaxies image classification using empirical mode decomposition and machine learning techniques. In *2014 International Conference on Engineering and Technology (ICET*, pages 1–5. IEEE, 2014.

J. K. Adelman-McCarthy, M. A. Agüeros, S. S. Allam, C. A. Prieto, K. S. J. Anderson, S. F. Anderson, J. Annis, N. A. Bahcall, C. A. L. Bailer-Jones, I. K. Baldry, J. C. Barentine, B. A. Bassett, A. C. Becker, T. C. Beers, E. F. Bell, A. A. Berlind, M. Bernardi, M. R. Blanton, J. J. Bochanski, W. N. Boroski, J. Brinchmann, J. Brinkmann, R. J. Brunner, T. Budavári, S. Carliles, M. A. Carr, F. J. Castander, D. Cinabro, R. J. Cool, K. R. Covey, I. Csabai, C. E. Cunha, J. R. A. Davenport, B. Dilday, M. Doi, D. J. Eisenstein, M. L. Evans, X. Fan, D. P. Finkbeiner, S. D. Friedman, J. A. Frieman, M. Fukugita, B. T. Gänsicke, E. Gates, B. Gillespie, K. Glazebrook, J. Gray, E. K. Grebel, J. E. Gunn, V. K. Gurbani, P. B. Hall, P. Harding, M. Harvanek, S. L. Hawley, J. Hayes, T. M. Heckman, J. S. Hendry, R. B. Hindsley, C. M. Hirata, C. J. Hogan, D. W. Hogg, J. B. Hyde, S.-i. Ichikawa, Ž. Ivezić, S. Jester, J. A. Johnson, A. M. Jorgensen, M. Jurić, S. M. Kent, R. Kessler, S. J. Kleinman, G. R. Knapp, R. G. Kron, J. Krzesinski, N. Kuropatkin, D. Q. Lamb, H. Lampeitl, S. Lebedeva, Y. S. Lee, R. F. Leger, S. Lépine, M. Lima, H. Lin, D. C. Long, C. P. Loomis, J. Loveday, R. H. Lupton, O. Malanushenko, V. Malanushenko, R. Mandelbaum, B. Margon, J. P. Marriner, D. Martínez-Delgado, T. Matsubara, P. M. McGehee, T. A. McKay, A. Meiksin, H. L. Morrison, J. A. Munn, R. Nakajima, J. Eric H Neilsen, H. J. Newberg, R. C. Nichol, T. Nicinski, M. Nieto-Santisteban, A. Nitta, S. Okamura, R. Owen, H. Oyaizu, N. Padmanabhan, K. Pan, C. Park, J. John Peoples, J. R. Pier, A. C. Pope, N. Purger, M. J. Raddick, P. R. Fiorentin, G. T. Richards, M. W. Richmond, A. G. Riess, H.-W. Rix, C. M. Rockosi, M. Sako, D. J. Schlegel, D. P. Schneider, M. R. Schreiber, A. D. Schwope, U. Seljak, B. Sesar, E. Sheldon, K. Shimasaku, T. Sivarani, J. A. Smith, S. A. Snedden, M. Steinmetz, M. A. Strauss, M. SubbaRao, Y. Suto, A. S. Szalay, I. Szapudi, P. Szkody, M. Tegmark, A. R. Thakar, C. A. Tremonti, D. L. Tucker, A. Uomoto, D. E. V. Berk, J. Vandenberg, S. Vidrih, M. S. Vogeley, W. Voges, N. P. Vogt, Y. Wadadekar, D. H. Weinberg, A. A. West, S. D. M. White, B. C. Wilhite, B. Yanny, D. R. Yocum, D. G. York, I. Zehavi, and D. B. Zucker. The Sixth Data Release of the Sloan Digital Sky Survey. *The Astrophysical Journal Supplement Series*, 175(2):297–313, Apr. 2008.

E. Anderson. The irises of the gaspe peninsula. *Bulletin of the American Iris Society*, 59: 2–5, 1935. URL https://ci.nii.ac.jp/naid/10000141584/en/.

K. Applebaum and D. Zhang. Classifying Galaxy Images through Support Vector Machines. In *2015 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 357–363. IEEE, 2015.

K. Au. *Inferring galaxy morphology through texture analysis*. Phd thesis, Carnegie Mellon Univ., 2006.

K. Au, C. Genovese, and A. Connolly. Inferring galaxy morphology through texture analysis. Technical report, 2006. URL http://www.stat.cmu.edu/~inca/Pubs/tr843.pdf.

S. P. Bamford, R. C. Nichol, I. K. Baldry, K. Land, C. J. Lintott, K. Schawinski, A. Slosar, A. S. Szalay, D. Thomas, M. Torki, et al. Galaxy Zoo: the dependence of morphology and colour on environment. *Monthly Notices of the Royal Astronomical Society*, 393(4): 1324–1352, 2009.

M. Banerji, O. Lahav, C. J. Lintott, F. B. Abdalla, K. Schawinski, S. P. Bamford, D. Andreescu, P. Murray, M. J. Raddick, A. Slosar, A. Szalay, D. Thomas, and J. Vandenberg. Galaxy Zoo: reproducing galaxy morphologies via machine learning. *Monthly Notices of the Royal Astronomical Society*, 406(1):342–353, 07 2010. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2010.16713.x. URL https://doi.org/10.1111/j.1365-2966.2010.16713.x.

J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017. doi: 10.1137/141000671. URL https://doi.org/10.1137/141000671.

J. Binney and S. Tremaine. *Galactic Dynamics*. Princeton Series in Astrophysics. Princeton University Press, 2011.

C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

S. Chen, C. F. N. Cowan, and P. M. Grant. Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on Neural Networks*, 2(2):302–309, March 1991. ISSN 1941-0093. doi: 10.1109/72.80341.

C. J. Conselice, A. Wilkinson, K. Duncan, and A. Mortlock. The Evolution of Galaxy Number Density at $z8$ and its Implications. *Astrophysical Journal*, 830(2):83, Oct. 2016.

D. Darg, S. Kaviraj, C. J. Lintott, K. Schawinski, J. Silk, S. Lynn, S. Bamford, and R. Nichol. Galaxy Zoo: Multimergers and the millennium simulation. *Monthly Notices of the Royal Astronomical Society*, 416(3):1745–1755, 2011.

B. L. Davis, J. C. Berrier, D. W. Shields, J. Kennefick, D. Kennefick, M. S. Seigar, C. H. Lacy, and I. Puerari. Measurement of galactic logarithmic spiral arm pitch angle using two-dimensional fast fourier transform decomposition. *The Astrophysical Journal Supplement Series*, 199(2):33, 2012.

B. L. Davis, D. Kennefick, J. Kennefick, K. B. Westfall, D. W. Shields, R. Flatman, M. T. Hartley, J. C. Berrier, T. P. Martinsson, and R. A. Swaters. A fundamental plane of spiral structure in disk galaxies. *The Astrophysical Journal Letters*, 802(1):L13, 2015.

D. R. Davis. *Fast Approximate Quantification of Arbitrary Arm-Segment Structure in Spiral Galaxies*. Phd thesis, University of California, Irvine, 2014.

D. R. Davis and W. B. Hayes. Automated quantitative description of spiral galaxy arm-segment structure. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1138–1145, 2012. doi: 10.1109/CVPR.2012.6247794.

D. R. Davis and W. B. Hayes. SpArcFiRe: Scalable automated detection of spiral galaxy arm segments. *The Astrophysical Journal*, 790(2):87, 2014.

G. De Vaucouleurs. General physical properties of external galaxies. In *Astrophysik IV: Sternsysteme/Astrophysics IV: Stellar Systems*, pages 311–372. Springer, 1959.

S. Dieleman, K. W. Willett, and J. Dambre. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society*, 450(2):1441–1459, 2015.

D. Dua and C. Graff. UCI machine learning repository, 2019. URL `http://archive.ics.uci.edu/ml`.

D. J. Eisenstein, D. H. Weinberg, E. Agol, H. Aihara, C. A. Prieto, S. F. Anderson, J. A. Arns, É. Aubourg, S. Bailey, E. Balbinot, R. Barkhouser, T. C. Beers, A. A. Berlind, S. J. Bickerton, D. Bizyaev, M. R. Blanton, J. J. Bochanski, A. S. Bolton, C. T. Bosman, J. Bovy, W. N. Brandt, B. Breslauer, H. J. Brewington, J. Brinkmann, P. J. Brown, J. R. Brownstein, D. Burger, N. G. Busca, H. Campbell, P. A. Cargile, W. C. Carithers, J. K. Carlberg, M. A. Carr, L. Chang, Y. Chen, C. Chiappini, J. Comparat, N. Connolly, M. Cortes, R. A. C. Croft, K. Cunha, L. N. da Costa, J. R. A. Davenport, K. Dawson, N. D. Lee, G. F. P. de Mello, F. de Simoni, J. Dean, S. Dhital, A. Ealet, G. L. Ebelke, E. M. Edmondson, J. M. Eiting, S. Escoffier, M. Esposito, M. L. Evans, X. Fan, B. F. Castellá, L. D. Ferreira, G. Fitzgerald, S. W. Fleming, A. Font-Ribera, E. B. Ford, P. M. Frinchaboy, A. E. G. Pérez, B. S. Gaudi, J. Ge, L. Ghezzi, B. A. Gillespie, G. Gilmore, L. Girardi, J. R. Gott, A. Gould, E. K. Grebel, J. E. Gunn, J.-C. Hamilton, P. Harding, D. W. Harris, S. L. Hawley, F. R. Hearty, J. F. Hennawi, J. I. G. Hernández, S. Ho, D. W. Hogg, J. A. Holtzman, K. Honscheid, N. Inada, I. I. Ivans, L. Jiang, P. Jiang, J. A. Johnson, C. Jordan, W. P. Jordan, G. Kauffmann, E. Kazin, D. Kirkby, M. A. Klaene, G. R. Knapp, J.-P. Kneib, C. S. Kochanek, L. Koesterke, J. A. Kollmeier, R. G. Kron, H. Lampeitl, D. Lang, J. E. Lawler, J.-M. L. Goff, B. L. Lee, Y. S. Lee, J. M. Leisenring, Y.-T. Lin, J. Liu, D. C. Long, C. P. Loomis, S. Lucatello, B. Lundgren, R. H. Lupton, B. Ma, Z. Ma, N. MacDonald, C. Mack, S. Mahadevan, M. A. G. Maia, S. R. Majewski, M. Makler, E. Malanushenko, V. Malanushenko, R. Mandelbaum, C. Maraston, D. Margala, P. Maseman, K. L. Masters, C. K. McBride, P. McDonald, I. D. McGreer, R. G. McMahon, O. M. Requejo, B. Ménard, J. Miralda-Escudé, H. L. Morrison, F. Mullally, D. Muna, H. Murayama, A. D. Myers, T. Naugle, A. F. Neto, D. C. Nguyen, R. C. Nichol, D. L. Nidever,

R. W. O'Connell, R. L. C. Ogando, M. D. Olmstead, D. J. Oravetz, N. Padmanabhan, M. Paegert, N. Palanque-Delabrouille, K. Pan, P. Pandey, J. K. Parejko, I. Pâris, P. Pellegrini, J. Pepper, W. J. Percival, P. Petitjean, R. Pfaffenberger, J. Pforr, S. Phleps, C. Pichon, M. M. Pieri, F. Prada, A. M. Price-Whelan, M. J. Raddick, B. H. F. Ramos, I. N. Reid, C. Reyle, J. Rich, G. T. Richards, G. H. Rieke, M. J. Rieke, H.-W. Rix, A. C. Robin, H. J. Rocha-Pinto, C. M. Rockosi, N. A. Roe, E. Rollinde, A. J. Ross, N. P. Ross, B. Rossetto, A. G. Sánchez, B. Santiago, C. Sayres, R. Schiavon, D. J. Schlegel, K. J. Schlesinger, S. J. Schmidt, D. P. Schneider, K. Sellgren, A. Shelden, E. Sheldon, M. Shetrone, Y. Shu, J. D. Silverman, J. Simmerer, A. E. Simmons, T. Sivarani, M. F. Skrutskie, A. Slosar, S. Smee, V. V. Smith, S. A. Snedden, K. G. Stassun, O. Steele, M. Steinmetz, M. H. Stockett, T. Stollberg, M. A. Strauss, A. S. Szalay, M. Tanaka, A. R. Thakar, D. Thomas, J. L. Tinker, B. M. Tofflemire, R. Tojeiro, C. A. Tremonti, M. V. Magaña, L. Verde, N. P. Vogt, D. A. Wake, X. Wan, J. Wang, B. A. Weaver, M. White, S. D. M. White, J. C. Wilson, J. P. Wisniewski, W. M. Wood-Vasey, B. Yanny, N. Yasuda, C. Yèche, D. G. York, E. Young, G. Zasowski, I. Zehavi, and B. Zhao. SDSS-III: MASSIVE SPECTROSCOPIC SURVEYS OF THE DISTANT UNIVERSE, THE MILKY WAY, AND EXTRA-SOLAR PLANETARY SYSTEMS. *The Astronomical Journal*, 142(3):72, aug 2011. doi: 10.1088/0004-6256/142/3/72. URL `https://doi.org/10.1088/0004-6256/142/3/72`.

F. Ferrari, R. R. de Carvalho, and M. Trevisan. Morfometryka - A New Way of Establishing Morphological Classification of Galaxies. *Astrophysical Journal*, 814(1):55, 2015.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936. doi: 10.1111/j.1469-1809.1936.tb02137.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1936.tb02137.x`.

M. Forina, R. Leardi, A. C, and S. Lanteri. *PARVUS: An Extendable Package of Programs for Data Exploration*. 01 1998. ISBN 0-444-43012-1.

J. Goldberger, G. E. Hinton, S. T. Roweis, and R. R. Salakhutdinov. Neighbourhood components analysis. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 513–520. MIT Press, 2005. URL `http://papers.nips.cc/paper/2566-neighbourhood-components-analysis.pdf`.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, Nov. 2009.

R. E. Hart, S. P. Bamford, K. W. Willett, K. L. Masters, C. Cardamone, C. J. Lintott, R. J. Mackay, R. C. Nichol, C. K. Rosslowe, B. D. Simmons, et al. Galaxy Zoo: comparing the demographics of spiral arm number and a new method for correcting redshift bias. *Monthly Notices of the Royal Astronomical Society*, 461(4):3663–3682, 2016.

W. B. Hayes, D. R. Davis, and P. Silva. On the nature and correction of the spurious S-wise spiral galaxy winding bias in Galaxy Zoo 1. *Monthly Notices of the Royal Astronomical Society*, 466(4):3928–3936, Dec. 2016. doi: 10.1093/mnras/stw3290.

J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982. ISSN 0027-8424. doi: 10.1073/pnas.79.8.2554. URL `https://www.pnas.org/content/79/8/2554`.

B. Hoyle, K. L. Masters, R. C. Nichol, E. M. Edmondson, A. M. Smith, C. Lintott, R. Scranton, S. Bamford, K. Schawinski, and D. Thomas. Galaxy Zoo: bar lengths in local disc galaxies. *Monthly Notices of the Royal Astronomical Society*, 415(4):3627–3640, 2011.

M. Huertas-Company, J. A. L. 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. *Astronomy & Astrophysics*, 525:A157, Dec. 2010.

Ž. Ivezić, S. M. Kahn, J. A. Tyson, B. Abel, E. Acosta, R. Allsman, D. Alonso, Y. AlSayyad, S. F. Anderson, J. Andrew, J. R. P. Angel, G. Z. Angeli, R. Ansari, P. Antilogus, C. Araujo, R. Armstrong, K. T. Arndt, P. Astier, É. Aubourg, N. Auza, T. S. Axelrod, D. J. Bard, J. D. Barr, A. Barrau, J. G. Bartlett, A. E. Bauer, B. J. Bauman, S. Baumont, E. Bechtol, K. Bechtol, A. C. Becker, J. Becla, C. Beldica, S. Bellavia, F. B. Bianco, R. Biswas, G. Blanc, J. Blazek, R. D. Bland ford, J. S. Bloom, J. Bogart, T. W. Bond, M. T. Booth, A. W. Borgland, K. Borne, J. F. Bosch, D. Boutigny, C. A. Brackett, A. Bradshaw, W. N. Brand t, M. E. Brown, J. S. Bullock, P. Burchat, D. L. Burke, G. Cagnoli, D. Calabrese, S. Callahan, A. L. Callen, J. L. Carlin, E. L. Carlson, S. Chand rasekharan, G. Charles-Emerson, S. Chesley, E. C. Cheu, H.-F. Chiang, J. Chiang, C. Chirino, D. Chow, D. R. Ciardi, C. F. Claver, J. Cohen-Tanugi, J. J. Cockrum, R. Coles, A. J. Connolly, K. H. Cook, A. Cooray, K. R. Covey, C. Cribbs, W. Cui, R. Cutri, P. N. Daly, S. F. Daniel, F. Daruich, G. Daubard, G. Daues, W. Dawson, F. Delgado, A. Dellapenna, R. de Peyster, M. de Val-Borro, S. W. Digel, P. Doherty, R. Dubois, G. P. Dubois-Felsmann, J. Durech, F. Economou, T. Eifler, M. Eracleous, B. L. Emmons, A. Fausti Neto, H. Ferguson, E. Figueroa, M. Fisher-Levine, W. Focke, M. D. Foss, J. Frank, M. D. Freemon, E. Gangler, E. Gawiser, J. C. Geary, P. Gee, M. Geha, C. J. B. Gessner, R. R. Gibson, D. K. Gilmore, T. Glanzman, W. Glick, T. Goldina, D. A. Goldstein, I. Goodenow, M. L. Graham, W. J. Gressler, P. Gris, L. P. Guy, A. Guyonnet, G. Haller, R. Harris, P. A. Hascall, J. Haupt, F. Hernand ez, S. Herrmann, E. Hileman, J. Hoblitt, J. A. Hodgson, C. Hogan, J. D. Howard, D. Huang, M. E. Huffer, P. Ingraham, W. R. Innes, S. H. Jacoby, B. Jain, F. Jammes, M. J. Jee, T. Jenness, G. Jernigan, D. Jevremović, K. Johns, A. S. Johnson, M. W. G. Johnson, R. L. Jones, C. Juramy-Gilles, M. Jurić, J. S. Kalirai, N. J. Kallivayalil, B. Kalmbach, J. P. Kantor, P. Karst, M. M. Kasliwal, H. Kelly, R. Kessler, V. Kinnison, D. Kirkby, L. Knox, I. V. Kotov, V. L. Krabbendam, K. S. Krughoff, P. Kubánek, J. Kuczewski, S. Kulkarni, J. Ku, N. R. Kurita, C. S. Lage, R. Lambert, T. Lange, J. B. Langton, L. Le Guillou, D. Levine, M. Liang, K.-T. Lim, C. J. Lintott, K. E. Long, M. Lopez, P. J. Lotz, R. H. Lupton, N. B. Lust, L. A. MacArthur, A. Mahabal, R. Mand elbaum, T. W. Markiewicz, D. S. Marsh, P. J. Marshall, S. Marshall, M. May, R. McKercher, M. McQueen, J. Meyers, M. Migliore, M. Miller, D. J. Mills, C. Miraval, J. Moeyens, F. E. Moolekamp, D. G. Monet, M. Moniez, S. Monkewitz, C. Montgomery, C. B. Morrison, F. Mueller, G. P. Muller, F. Muñoz Arancibia, D. R. Neill, S. P. Newbry, J.-Y. Nief, A. Nomerotski, M. Nordby, P. O'Connor,

J. Oliver, S. S. Olivier, K. Olsen, W. O'Mullane, S. Ortiz, S. Osier, R. E. Owen, R. Pain, P. E. Palecek, J. K. Parejko, J. B. Parsons, N. M. Pease, J. M. Peterson, J. R. Peterson, D. L. Petravick, M. E. Libby Petrick, C. E. Petry, F. Pierfederici, S. Pietrowicz, R. Pike, P. A. Pinto, R. Plante, S. Plate, J. P. Plutchak, P. A. Price, M. Prouza, V. Radeka, J. Rajagopal, A. P. Rasmussen, N. Regnault, K. A. Reil, D. J. Reiss, M. A. Reuter, S. T. Ridgway, V. J. Riot, S. Ritz, S. Robinson, W. Roby, A. Roodman, W. Rosing, C. Roucelle, M. R. Rumore, S. Russo, A. Saha, B. Sassolas, T. L. Schalk, P. Schellart, R. H. Schindler, S. Schmidt, D. P. Schneider, M. D. Schneider, W. Schoening, G. Schumacher, M. E. Schwamb, J. Sebag, B. Selvy, G. H. Sembroski, L. G. Seppala, A. Serio, E. Serrano, R. A. Shaw, I. Shipsey, J. Sick, N. Silvestri, C. T. Slater, J. A. Smith, R. C. Smith, S. Sobhani, C. Soldahl, L. Storrie-Lombardi, E. Stover, M. A. Strauss, R. A. Street, C. W. Stubbs, I. S. Sullivan, D. Sweeney, J. D. Swinbank, A. Szalay, P. Takacs, S. A. Tether, J. J. Thaler, J. G. Thayer, S. Thomas, A. J. Thornton, V. Thukral, J. Tice, D. E. Trilling, M. Turri, R. Van Berg, D. Vanden Berk, K. Vetter, F. Virieux, T. Vucina, W. Wahl, L. Walkowicz, B. Walsh, C. W. Walter, D. L. Wang, S.-Y. Wang, M. Warner, O. Wiecha, B. Willman, S. E. Winters, D. Wittman, S. C. Wolff, W. M. Wood-Vasey, X. Wu, B. Xin, P. Yoachim, and H. Zhan. LSST: From Science Drivers to Reference Design and Anticipated Data Products. volume 873, page 111, Mar 2019. doi: 10.3847/1538-4357/ab042c.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014. ISBN 1461471370, 9781461471370.

R. Jensen and Q. Shen. Are More Features Better? A Response to *Attributes Reduction Using Fuzzy Rough Sets*. *IEEE Transactions on Fuzzy Systems*, 17(6):1456–1458, 2009.

J. Kalirai. Scientific discovery with the James Webb Space Telescope. *Contemporary Physics*, 59(3):251–290, 2018. doi: 10.1080/00107514.2018.1467648. URL `https://doi.org/10.1080/00107514.2018.1467648`.

G. Kendall. Would your mobile phone be powerful enough to get you to the moon?, July 2019. URL `https://theconversation.com/would-your-mobile-phone-be-powerful-enough-to-get-you-to-the-moon-115933`.

M. Kuhn and K. Johnson. *Applied Predictive Modeling*. Springer, New York, Heidelberg, Dordrecht, London, 2013.

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):959–967, Oct. 2014.

K. Land, A. Slosar, C. J. Lintott, D. Andreescu, S. P. Bamford, P. Murray, R. Nichol, M. J. Raddick, K. Schawinski, A. Szalay, D. Thomas, and J. Vandenberg. 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, Aug. 2008. doi: 10.1111/j.1365-2966.2008.13490.x. URL `https://doi.org/10.1111/j.1365-2966.2008.13490.x`.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998. doi: 10.1109/5.726791.

C. Lintott, K. Schawinski, S. Bamford, A. Slosar, K. Land, D. Thomas, E. Edmondson, K. Masters, R. C. Nichol, M. J. Raddick, A. Szalay, D. Andreescu, P. Murray, and J. Vandenberg. Galaxy Zoo 1: data release of morphological classifications for nearly 900 000 galaxies. *Monthly Notices of the Royal Astronomical Society*, 410 (1):166–178, 12 2010. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2010.17432.x. URL `https://doi.org/10.1111/j.1365-2966.2010.17432.x`.

C. J. Lintott, K. Schawinski, A. Slosar, K. Land, S. P. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu, P. Murray, and J. Vandenberg. 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, Sept. 2008.

C. J. Lintott, K. Schawinski, W. Keel, H. Van Arkel, N. Bennert, E. Edmondson, D. Thomas, D. J. B. Smith, P. D. Herbert, M. J. Jarvis, S. Virani, D. Andreescu, S. P. Bamford, K. Land, P. Murray, R. C. Nichol, M. J. Raddick, A. Slosar, A. Szalay, and J. Vandenberg. Galaxy Zoo: 'Hanny's Voorwerp', a quasar light echo? *Monthly Notices of the Royal Astronomical Society*, 399(1):129–140, 10 2009. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2009.15299.x. URL `https://doi.org/10.1111/j.1365-2966.2009.15299.x`.

M. J. Longo. Detection of a dipole in the handedness of spiral galaxies with redshifts z $\sim$ 0.04. *Phys. Lett. B*, 699(4):224–229, May 2011. ISSN 03702693. doi: 10.1016/j.physletb.2011.04.008. URL `http://linkinghub.elsevier.com/retrieve/pii/S0370269311003947`.

E. E. Martínez-García, I. Puerari, F. Rosales-Ortega, R. A. González-López lira, I. Fuentes-Carrera, and A. Luna. The behavior of the pitch angle of spiral arms depending on optical wavelength. *The Astrophysical Journal Letters*, 793(1):L19, 2014.

K. L. Masters, M. Mosleh, A. K. Romer, R. C. Nichol, S. P. Bamford, K. Schawinski, C. J. Lintott, D. Andreescu, H. C. Campbell, B. Crowcroft, et al. Galaxy Zoo: passive red spirals. *Monthly Notices of the Royal Astronomical Society*, 405(2):783–799, 2010.

K. L. Masters, R. C. Nichol, B. Hoyle, C. Lintott, S. P. Bamford, E. M. Edmondson, L. Fortson, W. C. Keel, K. Schawinski, A. M. Smith, et al. Galaxy Zoo: bars in disc galaxies. *Monthly Notices of the Royal Astronomical Society*, 411(3):2026–2034, 2011.

W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, Dec 1943. ISSN 1522-9602. doi: 10.1007/BF02478259. URL `https://doi.org/10.1007/BF02478259`.

D. Mihalas and J. Binney. *Galactic Astronomy — Structure and Kinematics*. Princeton Series in Astrophysics. Freeman, 1981.

G. E. Moore. Cramming more components onto integrated circuits. *Electronics*, 38(8), April 1965.

D. Nelson, A. Pillepich, S. Genel, M. Vogelsberger, V. Springel, P. Torrey, V. Rodriguez-Gomez, D. Sijacki, G. F. Snyder, B. Griffen, et al. The illustris simulation: Public data release. *Astronomy and Computing*, 13:12–37, 2015.

A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *STOC '97 Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, pages 427–436. University of Wyoming, Laramie, United States, IEEE, 2015.

S. C. Odewahn, S. H. Cohen, R. A. Windhorst, and N. S. Philip. Automated galaxy morphology: A fourier approach. *The Astrophysical Journal*, 568(2):539–557, apr 2002. doi: 10.1086/339036. URL `https://doi.org/10.1086/339036`.

J. H. Oort et al. Problems of galactic structure. *Astrophysical journal*, page 116, 1952.

K. Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. doi: 10.1080/14786440109462720.

C. Y. Peng, L. C. Ho, C. D. Impey, and H.-W. Rix. DETAILED DECOMPOSITION OF GALAXY IMAGES. II. BEYOND AXISYMMETRIC MODELS. *The Astronomical Journal*, 139(6):2097–2129, apr 2010. doi: 10.1088/0004-6256/139/6/2097. URL `https://doi.org/10.1088/0004-6256/139/6/2097`.

T. R. Peng, J. E. English, P. Silva, D. R. Davis, and W. B. Hayes. SpArcFiRe: morphological selection effects due to reduced visibility of tightly winding arms in distant spiral galaxies. *Monthly Notices of the Royal Astronomical Society*, 479(4):5532–5543, Mar. 2018. doi: 10.1093/mnras/sty546. URL `https://academic.oup.com/mnras/article/479/4/5532/4931761`.

S. Perlmutter, M. S. Turner, and M. White. Constraining Dark Energy with Type Ia Supernovae and Large-Scale Structure. *Physical Review Letters*, 83(4):670, 1999.

H. Pour Imani, B. L. Davis, D. W. Shields, J. Kennefick, and D. Kennefick. Spiral arm pitch angle measurements of galaxies in different wavelengths of light to investigate a prediction of density wave theory. *IAU General Assembly*, 22, 2015.

H. Pour-Imani, D. Kennefick, J. Kennefick, B. L. Davis, D. W. Shields, and M. S. Abdeen. Strong evidence for the density-wave theory of spiral structure in disk galaxies. *The Astrophysical Journal Letters*, 827(1):L2, 2016.

N. A. Rahman. A course in theoretical statistics: For sixth forms. *Technical Colleges, Colleges of Education, Universities: Charles Griffin & Company Limited*, 1968.

C. R. Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):159–203, 1948. ISSN 00359246. URL `http://www.jstor.org/stable/2983775`.

P. Refaeilzadeh, L. Tang, and H. Liu. Cross-Validation. In *Encyclopedia of Database Systems*, pages 1–7. Springer New York, New York, NY, Dec. 2016.

M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939778.

A. Saabas. Selecting good features Part III: random forests, Dec. 2014. URL `http://blog.datadive.net/selecting-good-features-part-iii-random-forests/`.

J. A. Sellwood. The lifetimes of spiral patterns in disc galaxies. *Monthly Notices of the Royal Astronomical Society*, 410(3):1637–1646, Oct. 2011. ISSN 00358711. doi: 10.1111/j.1365-2966.2010.17545.x. URL `http://doi.wiley.com/10.1111/j.1365-2966.2010.17545.x`.

L. Shamir. Automatic morphological classification of galaxy images. *Monthly Notices of the Royal Astronomical Society*, 399(3):1367–1372, 10 2009. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2009.15366.x. URL `https://doi.org/10.1111/j.1365-2966.2009.15366.x`.

L. Shamir. GANALYZER: A TOOL FOR AUTOMATIC GALAXY IMAGE ANALYSIS. *The Astrophysical Journal*, 736(2):141, jul 2011. doi: 10.1088/0004-637x/736/2/141. URL `https://doi.org/10.1088/0004-637x/736/2/141`.

L. Shamir. Handedness asymmetry of spiral galaxies with $z < 0.3$ shows cosmic parity violation and a dipole axis. *Physics Letters B*, 715(1):25–29, 2012.

L. Shamir. ASYMMETRY BETWEEN GALAXIES WITH CLOCKWISE HANDEDNESS AND COUNTERCLOCKWISE HANDEDNESS. *The Astrophysical Journal*, 823(1):32, May 2016.

C. Sibley. More is always better: The power of simple ensembles, Oct. 2012. URL `http://www.overkillanalytics.net/more-is-always-better-the-power-of-simple-ensembles/`.

P. Silva, S. Akhavan-Masouleh, and L. Li. Improving Malware Detection Accuracy by Extracting Icon Information. In *IEEE Conference on Multimedia Information Processing and Retrieval*, pages 408–411, Miami, FL, USA, Apr. 2018a. IEEE. ISBN 978-1-5386-1857-8. doi: 10.1109/MIPR.2018.00088. URL `https://ieeexplore.ieee.org/document/8397044/`.

P. Silva, L. Cao, and W. B. Hayes. SpArcFiRe: Enhancing Spiral Galaxy Recognition Using Arm Analysis and Random Forests. *Galaxies*, 6(3):95, Sept. 2018b. doi: 10.3390/galaxies6030095. URL `http://www.mdpi.com/2075-4434/6/3/95`.

R. W. Smith. *The expanding universe: Astronomy's 'great debate', 1900-1931*. Cambridge University Press, 1982.

T. SunPy Community, S. J. Mumford, S. Christe, D. Pérez-Suárez, J. Ireland, A. Y. Shih, A. R. Inglis, S. Liedtke, R. J. Hewett, F. Mayer, K. Hughitt, N. Freij, T. Meszaros, S. M. Bennett, M. Malocha, J. Evans, A. Agrawal, A. J. Leonard, T. P. Robitaille, B. Mampaey, J. Iván Campos-Rozo, and M. S. Kirk. SunPy–Python for solar physics. *Computational Science and Discovery*, 8(1):014009, Jan. 2015. doi: 10.1088/1749-4699/8/1/014009.

P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar. *Introduction to Data Mining*. Pearson, 2nd edition, 2018. ISBN 0133128903, 9780133128901.

K. W. Willett, C. J. Lintott, S. P. Bamford, K. L. Masters, B. D. Simmons, K. R. V. Casteels, E. M. Edmondson, L. F. Fortson, S. Kaviraj, W. C. Keel, T. Melvin, R. C. Nichol, M. J. Raddick, K. Schawinski, R. J. Simpson, R. A. Skibba, A. M. Smith, and D. Thomas. Galaxy Zoo 2: detailed morphological classifications for 304,122 galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 435:2835–2860, Nov. 2013. doi: 10.1093/mnras/stt1458.

M. Wolff, P. Silva, X. Zhao, J. Brock, and J. Luan. Icon Based Malware Detection, U.S. Patent 10 354 173 B2, Jul. 2019.

D. G. York, J. Adelman, J. E. Anderson, S. F. Anderson, J. Annis, N. A. Bahcall, J. A. Bakken, R. Barkhouser, S. Bastian, E. Berman, W. N. Boroski, S. Bracker, C. Briegel, J. W. Briggs, J. Brinkmann, R. Brunner, S. Burles, L. Carey, M. A. Carr, F. J. Castander, B. Chen, P. L. Colestock, A. J. Connolly, J. H. Crocker, I. Csabai, P. C. Czarapata, J. E. Davis, M. Doi, T. Dombeck, D. Eisenstein, N. Ellman, B. R. Elms, M. L. Evans, X. H. Fan, G. R. Federwitz, L. Fiscelli, S. Friedman, J. A. Frieman, M. Fukugita, B. Gillespie, J. E. Gunn, V. K. Gurbani, E. de Haas, M. Haldeman, F. H. Harris, J. Hayes, T. M. Heckman, G. S. Hennessy, R. B. Hindsley, S. Holm, D. J. Holmgren, C. H. Huang, C. Hull, D. Husby, S. Ichikawa, T. Ichikawa, Z. Ivezic, S. Kent, R. Kim, E. Kinney, M. Klaene, A. N. Kleinman, S. Kleinman, G. R. Knapp, J. Korienek, R. G. Kron, P. Z. Kunszt, D. Q. Lamb, B. Lee, R. F. Leger, S. Limmongkol, C. Lindenmeyer, D. C. Long, C. Loomis, J. Loveday, R. Lucinio, R. H. Lupton, B. MacKinnon, E. J. Mannery, P. M. Mantsch, B. Margon, P. McGehee, T. A. McKay, A. Meiksin, A. Merelli, D. G. Monet, J. A. Munn, V. K. Narayanan, T. Nash, E. Neilsen, R. Neswold, H. J. Newberg, R. C. Nichol, T. Nicinski, M. Nonino, N. Okada, S. Okamura, J. P. Ostriker, R. Owen, A. G. Pauls, J. Peoples, R. L. Peterson, D. Petravick, J. R. Pier, A. Pope, R. Pordes, A. Prosapio, R. Rechenmacher, T. R. Quinn, G. T. Richards, M. W. Richmond, C. H. Rivetta, C. M. Rockosi, K. Ruthmansdorfer, D. Sandford, D. J. Schlegel, D. P. Schneider, M. Sekiguchi, G. Sergey, K. Shimasaku, W. A. Siegmund, S. Smee, J. A. Smith, S. Snedden, R. Stone, C. Stoughton, M. A. Strauss, C. Stubbs, M. SubbaRao, A. S. Szalay, I. Szapudi, G. P. Szokoly, A. R. Thakar, C. Tremonti, D. L. Tucker, A. Uomoto, D. V. Berk, M. S. Vogeley, P. Waddell, S. I. Wang, M. Watanabe, D. H. Weinberg, B. Yanny, N. Yasuda, and S. Collaboration. The Sloan Digital Sky Survey: Technical summary. *The Astronomical Journal*, 120(3): 1579–1587, Sept. 2000. URL `http://iopscience.iop.org/1538-3881/120/3/1579`.