ASYMMETRY BETWEEN GALAXIES WITH CLOCKWISE HANDEDNESS AND COUNTERCLOCKWISE HANDEDNESS

Lior Shamir

Department of Computer Science, Lawrence Technological University, Southfield, MI 48075, USA; lshamir@mtu.edu

Received 2015 December 29; accepted 2016 March 14; published 2016 May 19

ABSTRACT

While it is clear that spiral galaxies can have different handedness, galaxies with clockwise patterns are assumed to be symmetric to galaxies with counterclockwise patterns in all of their other characteristics. Here, we use data from SDSS DR7 to show that photometric data can distinguish between clockwise and counterclockwise galaxies. Pattern recognition algorithms trained and tested using the photometric data of a clean, manually crafted data set of 13,440 spiral galaxies with $z < 0.25$ can predict the handedness of a spiral galaxy in $\sim 64\%$ of the cases, which is significantly higher than the mere chance accuracy of 50% ($P < 10^{-5}$). Experiments with a different data set of 10,281 automatically classified galaxies showed similar results of $\sim 65\%$ classification accuracy, suggesting that the observed asymmetry is also consistent in data sets annotated in a fully automatic process, without human intervention. That shows that the photometric data collected by SDSS is sensitive to the handedness of the galaxy.

Analysis of the number of galaxies classified as clockwise and counterclockwise by crowdsourcing shows that manual classification between spiral and elliptical galaxies can be affected by the handedness of the galaxy, and therefore the galaxy morphology analyzed by citizen science campaigns might be biased by the galaxy handedness. The code and data used in the experiment are publicly available, and the experiment can be easily replicated.

Key words: galaxies: photometry – galaxies: spiral – Galaxy: general

1. INTRODUCTION

A highly noticeable morphological property of a spiral galaxy is its handedness. Spiral galaxies can be broadly separated into galaxies that seem to an Earth-based observer to have clockwise patterns and galaxies that seem to have counterclockwise patterns. Since clockwise galaxies are expected to be symmetric to counterclockwise galaxies, this morphological difference is not expected to be reflected by other physical characteristics. The symmetry is also expected because the handedness of a galaxy is merely a matter of the location of the observer, and a galaxy that would seem to rotate clockwise to an Earth-based observer might seem to rotate counterclockwise to an observer placed elsewhere in the universe.

Some evidence shows that the distribution of clockwise and counterclockwise spiral galaxies changes between different R.A. ranges, and may therefore violate the cosmological assumption of isotropy (Longo 2011; Shamir 2012). Other studies showed some mild photometric differences between clockwise and counterclockwise galaxies (Shamir 2013). However, unlike stars, galaxies cannot be considered as a one-parameter family (Djorgovski & Davis 1987), and therefore a galaxy can be described by a set of multiple physical measurements (Brosche 1973; Djorgovski & Davis 1987).

When testing a large number of different measurements, the probability that a certain test exhibits a difference by mere chance increases as the total number of measurements gets higher. For instance, assuming that a single hypothesis can be considered to be statistically significant if the probability of false positive is smaller than 0.05, then when testing multiple different hypotheses the probability that one of them exhibits a difference with a statistical significance of $P < 0.05$ is clearly higher than 0.05, and increases as the number of hypotheses being tested becomes higher. Therefore, when multiple different hypotheses are being tested, the threshold of 0.05 must be corrected to avoid false positives.

A mature method of avoiding false positives when testing a large number of hypotheses is the Bonferroni correction (Goeman & Solari 2014), which provides the threshold of statistical significance that each specific hypothesis needs to meet when tested as part of an experiment which involves multiple hypotheses (Goeman & Solari 2014). While the Bonferroni correction reduces the possibility of false positives, applying it makes it more difficult to identify statistically significant differences between clockwise and counterclockwise galaxies when comparing a very large number of different photometric measurements (Hoehn & Shamir 2014).

2. DATA

The first data set of galaxies used in this study contained galaxies from the Sloan Digital Sky Survey (York et al. 2000) which were annotated manually by Galaxy Zoo 2 (Willett et al. 2013) as galaxies that were not smooth and round (Question 1 in the Galaxy Zoo 2 user interface). Since Galaxy Zoo is based on the annotations of non-experts, it cannot be assumed that all of the annotations are necessarily correct. To filter misclassified galaxies, we used only those galaxies which 90% or more of the voters agreed upon. This threshold provided a data set of 19,693 galaxies (Kuminski et al. 2014), which are likely to be galaxies that are not smooth and round, and therefore are potentially spiral galaxies with identifiable handedness.

The “superclean” criterion (Lintott et al. 2011), according to which a classification is “superclean” only when it reaches 95% agreement, could not be used since using this criterion would leave only 6635 galaxies (Kuminski et al. 2014), which might not be a sufficient number of galaxies for the analysis. However, the 90% threshold is still higher than the “clean”
Galaxy Zoo criterion (Lintott et al. 2011), and therefore it can be reasonably assumed that the vast majority of the galaxies satisfying the criterion are indeed not smooth and round. Also, previous studies show that when using citizen science to classify between spiral and elliptical galaxies, the sensitivity of spiral galaxies is high, while the specificity of galaxies not annotated as spiral is lower (Dojcsak & Shamir 2014), and therefore it is expected that the galaxies annotated as “not smooth and round” are indeed not elliptical galaxies.

The fact that all of the galaxies are bright and large allows for correct classification of the galaxies regardless of their redshift, which in a randomly selected set of galaxies can lead to an inverse correlation between the classification accuracy and the $z$, regardless of whether the classification is carried out by a machine (Shamir 2011a) or by humans (Lintott et al. 2011).

The galaxies were initially classified as clockwise and counterclockwise galaxies using the Ganalyzer galaxy image analysis tool (Shamir 2011a, 2011b). Ganalyzer works by first computing the Otsu binary threshold (Otsu 1979) in order to identify the foreground pixels, and the radius is determined by the most distant foreground pixel from the center of the galaxy. Then, the galaxy image is converted to its radial intensity plot, which is an image of $360 \times 35$ pixels, such that the pixel $(x, y)$ in the radial intensity plot is the median value of the $5 \times 5$ pixels around $(O_x + \sin(\theta) \cdot r, O_y - \cos(\theta) \cdot r)$ in the galaxy image, where $(O_x, O_y)$ are the image coordinates of the galaxy center, $\theta$ is the polar angle (in degrees), and $r$ is a radial distance, which ranges over 35% of the total galaxy radius. Figure 1 shows a galaxy image and its radial intensity plot.

The horizontal lines in the radial intensity plot are searched for peaks, and the slope of the peaks determines the handedness of the galaxy. This is done by comparing each peak at coordinates $(x_0, y)$ to its closest peak in the next horizontal line $(x_1, y + 1)$. If $x_1 < x_0$, then the counter $L$ is incremented, and if $x_1 > x_0$, then the counter $R$ is incremented. For the classification of the Galaxy Zoo data used in this study, if $L > R$, then the galaxy is considered to be clockwise, and if $R > L$, then it is considered counterclockwise. That is repeated for shifted radius ranges, from 20%–55% to 50%–85% of the total radius in increments of 10%, until the sharpest slope is found. A detailed description of Ganalyzer can be found in Shamir (2011a, 2012; Hoehn & Shamir 2014).

The automatic analysis described above is crude, and was therefore followed by manual inspection and correction of all of the galaxy classifications. The galaxies were then mirrored, and were inspected again to ensure that no galaxy is misclassified. The manual classification of the galaxies provided 6941 galaxies with a clockwise pattern, and 6499 galaxies with a counterclockwise pattern. The remaining galaxies did not have a clearly identifiable handedness (e.g., edge-on), or were not spiral galaxies.

At the end of the process, 100 galaxies from each class were randomly selected and carefully inspected to ensure that they were all correctly classified. The entire process of manual classification of the galaxies required approximately 150 hr of labor, but produced a very clean data set of spiral galaxies separated by their spin direction, and no error in that data set is believed to exist.

The photometric information for each galaxy was retrieved through the Catalog Archive Server (CAS). All of the fields of the table PhotoObjAll in DR7 (Abazajian et al. 2009) were used, producing a data set of 452 variables for each galaxy. All of the galaxies were in the $90^\circ < R.A. < 270^\circ$ hemisphere, and $z < 0.25$. Figure 2 shows the histograms of the distribution of the $r$ magnitude, the Petrosian radius, and the redshift of the galaxies with clockwise and counterclockwise patterns.

3. CLASSIFICATION METHOD

One of the goals of the experiment was to test whether the handedness of a spiral galaxy can be predicted using its photometric information. This was done by using several pattern recognition algorithms, such that the label of each galaxy sample is its handedness (cw or ccw), and the variables of each sample are the photometric variables from PhotoObjAll. The purpose of the supervised machine learning was to accurately predict the handedness of the galaxy using photometric information.

For the classification, we used the Waikato Environment for Knowledge Analysis (WEKA) open source tool (Hall et al. 2009). WEKA is a comprehensive software that includes the implementation of numerous machine learning algorithms. The algorithms that were used for the classification were Random Forests (Breiman 2001), OneR (Holte 1993), Decision Table (Kohavi & John 1995), Ensembles of balanced Nested Dichotomies (Dong et al. 2005), Bayesian Network (Friedman et al. 1997), and Bagging (Breiman 1996). In addition to the algorithms provided by WEKA, the open source Weighted Nearest Distance (WND) algorithm (Shamir et al. 2008, 2013) was also used.

The experiments were performed such that 80% of the samples were used for training and the remaining 20% were used for testing. That is, the machine learning algorithms used 80% of the galaxies to automatically identify patterns that may differentiate between clockwise and counterclockwise galaxies, and the remaining 20% of the galaxies were used to predict the handedness of each of these galaxies, and count the number of correct predictions such that the classification accuracy was determined by the number of correct predictions divided by the total number of prediction attempts.

In addition to the 80/20 strategy, the classification accuracy was also tested by separating the training and test data using a 10-fold cross-validation strategy, and also by using separation by fixed numbers such that 4800 samples from each class are used for training and 1200 samples from each class are used for testing.

The same experiments were also repeated such that the label (galaxy handedness) was replaced by a random label $(0, 1)$.

4. RESULTS

Figure 3 shows the classification accuracy of the different supervised machine learning algorithms using labels that are

\[ \text{Figure 1. Galaxy image and its corresponding radial intensity plot (Shamir 2011a).} \]
the actual handedness of each galaxy, as well as the randomly assigned labels.

As expected, not all of the algorithms achieved the same classification accuracy, as not all of the supervised machine learning algorithms are equally powerful and different algorithms might perform differently on different types of data. However, it is clear that all of the algorithms classified the galaxies with accuracy higher than mere chance, whereas when the handedness was assigned randomly the classification accuracy of all algorithms was close to 50%. When using a fixed separation of 5000 samples per class for testing and 1000 per class for training, the classification accuracy of the Bayesian Network marginally dropped to $\sim 63\%$ and the classification accuracy of the WND algorithm dropped to $\sim 59\%$.

Assuming no link between the variables and the galaxy handedness, a galaxy would be randomly classified by these variables as either clockwise or counterclockwise. In that case, achieving a classification accuracy of 64.4% by chance would require 1721 or more correct classifications of the 2688 total classification attempts. Using cumulative binomial probability (Keller 2015) such that the number of trails is 2688, the minimum number of successes is 1721 and the probability of success is 0.5, while the probability to have such results by mere chance is $<10^{-5}$.

Tables 1 through 7 show the confusion matrices of each of the classifiers. In all of the cases, both classes were classified with an accuracy higher than mere chance.

The classification accuracies of the different algorithms when using a 10-fold cross-validation strategy for testing are displayed in Figure 4. As expected, the classification accuracies

### Table 1
Confusion Matrix of the Classification with a Bayesian Network Classifier

|           | Clockwise | Counterclockwise |
|-----------|-----------|------------------|
| Clockwise | 888       | 480              |
| Counterclockwise | 487     | 833              |

### Table 2
Confusion Matrix of the Classification Using Random Forests

|           | Clockwise | Counterclockwise |
|-----------|-----------|------------------|
| Clockwise | 948       | 420              |
| Counterclockwise | 652     | 668              |

### Table 3
Confusion Matrix of the Classification Using OneR

|           | Clockwise | Counterclockwise |
|-----------|-----------|------------------|
| Clockwise | 851       | 517              |
| Counterclockwise | 590     | 730              |

### Table 4
Confusion Matrix of the Classification Using a Decision Table Classifier

|           | Clockwise | Counterclockwise |
|-----------|-----------|------------------|
| Clockwise | 913       | 455              |
| Counterclockwise | 546     | 774              |

### Table 5
Confusion Matrix of the Classification Using Ensembles of Balanced Nested Dichotomies (END) Classifier

|           | Clockwise | Counterclockwise |
|-----------|-----------|------------------|
| Clockwise | 822       | 546              |
| Counterclockwise | 620     | 700              |

### Table 6
Confusion Matrix of the Classification Using Bagging

|           | Clockwise | Counterclockwise |
|-----------|-----------|------------------|
| Clockwise | 899       | 469              |
| Counterclockwise | 558     | 762              |

### Table 7
Confusion Matrix of the Classification Using a Weighted Nearest Distance (WND) Classifier

|           | Clockwise | Counterclockwise |
|-----------|-----------|------------------|
| Clockwise | 744       | 644              |
| Counterclockwise | 430     | 869              |
are similar to the classification accuracies when using 80% of the samples for training. As with the 80/20 separation, when assigning the galaxies with random handedness, the classification was close to 50% mere chance accuracy.

Because a large number of variables are being tested for the difference between clockwise and counterclockwise galaxies, the probability that one of these variables exhibits a difference by mere chance increases as the number of variables being tested becomes larger. For that reason, the Bonferroni correction is applied. Because the Bonferroni correction becomes stronger when the number of variables increases, using a large set of variables makes it less likely to identify specific variables which exhibit a Bonferroni-corrected statistically significant difference between the two classes of galaxies.

The WND algorithm computes the Fisher discriminant (Fisher 1938) of each variable as a heuristic for determining the weight of each variable, such that higher weight indicates that the variable is assumed to be more informative for predicting the handedness of a spiral galaxy. The variables that were assigned with the highest Fisher discriminant scores are specified in Table 8. The table also displays the mean and standard error of the mean of the variables in clockwise galaxies and counterclockwise galaxies, as well as the corrected and non-corrected two-tailed $P$ value of the $t$-test of the difference between the means. Values such as $-9999$ and $-1000$ often appear in the SDSS DR7 PhotoObjAll table, but these are in fact flags and not actual measured values, and therefore these values were ignored.

When using the variables listed in Table 8 only, the Bayesian Network classifier was able to differentiate between the two classes with an accuracy of 63%, and Random Forests and Bagging were able to achieve classification accuracies of 62% and 61%, respectively. These accuracies show that this relatively small set of variables is sufficient to identify the handedness of the galaxies with accuracies higher than mere chance.

The variables that have the highest Fisher discriminant scores are the isoPhiGrad$_r$, isoPhiGrad$_g$, and isoPhiGrad$_i$, which are the isophote position angle gradients measured in the r, g, and i bands, respectively. Although their means did not exhibit a statistically significant difference between clockwise and counterclockwise spiral galaxies, the isophote position angle gradients were estimated to be the most informative according to the Fisher discriminant heuristics. These variables are often used to measure ellipticity, and are not expected to be highly accurate (and for that reason were not included in the PhotoObjAll table of SDSS data releases after DR7). The isophotal position angle itself of clockwise and counterclockwise galaxies did not show any statistically significant difference.

Of the full set of 452 variables, 10 showed a Bonferroni-corrected statistically significant difference between galaxies with clockwise patterns and galaxies with counterclockwise patterns. Out of these 10 variables, 4 are the SDSS “Stokes $U$” parameter $u\_g$, $u\_r$, $u\_i$, and $u\_z$, measured on the g, r, i, and z bands. The SDSS DR7 “Stokes $U$” parameter is measured by $U = \frac{a - b}{a + b} \sin(2\phi)$, where $a$ is the major axis, $b$ is the minor axis of the galaxy, and $\phi$ is the position angle (Abazajian et al. 2009). In all of the bands, the mean of the “Stokes $U$” parameter was negative for clockwise galaxies and positive for counterclockwise galaxies.

The “Stokes $U$” parameter measured in the u band had a much lower non-corrected $t$-test statistical significance of $\sim0.012$, and therefore cannot be considered statistically significant. The SDSS “Stokes $Q$” parameter had no statistically significant difference in any of the bands, and its non-corrected $t$-test probabilities range between 0.31 for the r band and 0.96 for the u band.

Other variables that show a statistically significant difference between clockwise and counterclockwise galaxies are $\ln L_{\text{DeV}}$ and $\ln L_{\text{Star}}$, which provide information about the reliability of the separation of the object to stars and galaxies in the SDSS pipeline based on the magnitude model fitness. $\ln L_{\text{DeV}}$ is the $\chi^2$ fitness of the de Vaucouleurs surface brightness model. The variables in the table related to that measurement are $\ln L_{\text{DeV}} \_u$, $\ln L_{\text{DeV}} \_g$, $\ln L_{\text{DeV}} \_r$, $\ln L_{\text{DeV}} \_i$, and $\ln L_{\text{DeV}} \_z$, which measure the $\chi^2$ fitness of the de Vaucouleurs surface brightness model in bands u, g, r, i, and z, respectively. $\ln L_{\text{Star}}$ measures the $\chi^2$ fitting of the point-spread function (PSF) surface brightness model. The variables in Table 8 related to $\ln L_{\text{Star}}$ are $\ln L_{\text{Star}} \_u$, $\ln L_{\text{Star}} \_g$, $\ln L_{\text{Star}} \_i$, and $\ln L_{\text{Star}} \_z$, which measure the PSF model fitness in bands u, g, i, and z, respectively.

The remaining two variables in Table 8 are petroR50Err$_\text{u}$ and petroR90Err$_\text{z}$, which are the 50% and 90% Petrosian radius measurement error in the u and z bands, respectively. Consistent differences in these variables can be the result of weak measurements or differences in resolution, but can also be affected by different morphologies. While these variables were assigned relatively high Fisher discriminant scores and affected the classification, none of them show a statistically significant difference between clockwise and counterclockwise galaxies.

Attempting to classify the galaxies with just the “Stokes $U$” parameter of the five bands provided a very low classification accuracy of $\sim50.8929$ using a Bayesian Network classifier, indicating that while the differences based on galaxy handedness exist, they are not sufficient for predicting the handedness of a galaxy just by using that parameter. On the other hand, repeating the automatic classification experiments by removing all of the “Stokes parameters” did not make any significant impact on the classification accuracy, which remained as shown in Figure 3.

Since the galaxies that were used in the experiment are the galaxies that were initially classified by crowdsourcing as spiral, the annotation of the galaxies can be subject to human bias which, in certain conditions, might affect the results. For
instance, it has been shown that manual analysis of galaxy handedness by citizen scientists is substantially biased by human preferences (Land et al. 2008). Although the handedness identification used in this study does not rely on citizen science annotations, it is possible that the human annotators have preferences to galaxies with certain combinations of handedness and other characteristics, and that bias might be carried forward to the data set of galaxies separated by their handedness.

Such bias is expected to become weaker in galaxies on which more citizen scientists vote in the same manner, and therefore would exhibit itself in the form of smaller difference between the two classes when using galaxies on which the agreement among the citizen scientists was stronger. Table 9 shows the mean and standard errors of the variables listed in Table 8 measured using just 5132 of the galaxies that had clear handedness, and were also classified as spiral by 95% or more of the citizen scientists.

The differences between the means of the variables measured using galaxies classified as spiral by 95% or more of the voters do not show a consistent increase compared to the differences when the entire data set is used, indicating that the asymmetry does not change substantially with the voting trends of the citizen scientists.

In addition to the Fisher discriminant feature selection, several other feature selection algorithms, such as Consistency Subset Eval (Liu & Setiono 1996), Combined Feature Selection (CFS) Subset Eval (Hall 1998), and Filtered Attribute Eval, have also been used to automatically select the most informative variables, and the variables that were selected by these methods as well as the classification accuracy achieved using these variables are shown in Table 10. As also mentioned above, these variables did not exhibit statistically significant difference between clockwise and counterclockwise galaxies.

### 4.1. Analysis Using Automatically Classified Galaxies

The analysis using galaxies annotated by Galaxy Zoo involved a first step of manual classification of the galaxies as elliptical and spiral, which was performed by crowdsourcing. The involvement of citizen scientists might therefore be affected by the human bias of the manual annotations. Although classification by handedness was not performed by the citizen scientists, a certain preference of the human annotators for certain galaxies that depends on the galaxy handedness might be carried forward to produce a biased data set.

To avoid such possible bias, another experiment was performed such that all of the galaxies were classified in a fully automatic manner and without any human intervention in the analysis. That was done using galaxies from a computer-generated catalog of broad galaxy morphology (Kuminski & Shamir 2016). The catalog was generated using an automatic image classification method (Shamir 2009) applied to a large...
set of SDSS galaxies, producing a catalog of galaxies separated into elliptical and spiral galaxies as described in Kuminski & Shamir (2016). That catalog is somewhat similar in information to the Galaxy Zoo 1 catalog, but was generated in a fully automatic manner, without human intervention.

All of the galaxies with spectra classified as spirals were classified by the Ganalyzer algorithm (Shamir 2011a) described in Section 2 to determine their handedness. As discussed in Section 2, the algorithm is imperfect and requires a step of intensive manual correction to produce a clean data set. To avoid manual intervention, the algorithm was used such that the criteria for correct classification of a clockwise galaxy was an $L$ counter of 30 or higher, and $L > 3R$. Similarly, the criteria for classifying a galaxy as counterclockwise is $R > 30$ and $R > 3L$. All of the other galaxies were considered to be undecided and were excluded from the analysis. That strategy improved the correctness of the classification to clockwise and counterclockwise galaxies, but also resulted in the sacrifice of 105,078 samples out of 115,359 galaxies. That provided a fully automatically generated data set of 5139 galaxies classified as clockwise, and 5142 galaxies classified as counterclockwise. Out of the data set of 10,281 galaxies, 2280 overlap with the data set described in Section 2 (1139 clockwise and 1141 counterclockwise). The distribution of the redshift, radius, and $r$ magnitude of the galaxies are broadly consistent with those displayed in Figure 2.

To assess the consistency of the data set, 200 galaxies from each class were inspected manually. The manual inspection revealed that out of the 400 galaxies, 11 galaxies classified as clockwise and 8 galaxies classified by the algorithm as counterclockwise did not have clear handedness. One galaxy classified as counterclockwise was in fact a clockwise galaxy.

The photometric information from SDSS DR7 for each galaxy was retrieved through CAS, and the data set was classified similarly to the analysis of the Galaxy Zoo galaxies, with several different classifiers and a standard 10-fold test strategy. The classification results are displayed in Figure 5.

As the figure shows, all of the classifiers were able to identify clockwise and counterclockwise galaxies with accuracy higher than mere chance, and the results are similar to the classification accuracy when the data set was built based on the Galaxy Zoo galaxies. Tables 11 and 12 show the confusion matrices of the classifications using the Random Forest and Bagging classifiers, respectively.

As with the previous experiment using Galaxy Zoo galaxies, the informativeness of each variable was measured using Fisher discriminant heuristics, and the variables with the highest Fisher discriminant scores are listed in Table 13.

Similar to the first experiment, the variables with the highest Fisher discriminant scores were the isophote position angle gradients measured in the $r$, $g$, and $i$ bands. Other variables are $\text{devPhi}_z$, $\text{devPhi}_i$, $\text{devPhi}_r$, and $\text{devPhi}_g$, which are the DeVaucouleurs fit position angles measured in bands $z$, $i$, $r$, and $g$, respectively, and the $\text{expPhi}_g$, $\text{expPhi}_i$, $\text{expPhi}_z$, $\text{expPhi}_r$, which are the exponential fit position angles measured in bands $g$, $i$, $z$, and $r$. $m\text{E1E2Err}_g$, $m\text{E1E2Err}_i$, and $m\text{E1E2Err}_r$ are the square roots of the covariance matrix of the intensity second moments (Bernstein & Jarvis 2002), measured on the $g$, $r$, and $i$ bands.

The observation that the variables are associated with the position angle indicates possible mild asymmetry between the morphology of these galaxies. Another explanation can be that inaccuracies in the measurement of the position angle might be sensitive to the handedness of the galaxy. That can also be related to the differences in the “Stokes U” parameter, which also depends on the position angle. However, since the position angles are randomly distributed, a consistent measurement
error might lead to difference between the means measured in the two types of galaxies, but is not expected to result in classification between the types of galaxies based on the position angle alone. Also, when removing all of the variables related to the position angle (measured with the isophote, exponential fit, DeVaucouleurs fit, and “Stokes” parameters) and their errors, the classification accuracy is still higher than mere chance. The classification accuracy when removing these variables is \( \sim 58\% \) and \( \sim 57\% \) when using the Random Forest and Bagging classifiers, respectively.

Unlike the variables in Table 8, none of the variables in the PhotoObjAll table of DR7 showed statistically significant difference between clockwise and counterclockwise galaxies. This can be explained by the fact that the automatically generated data set is not as clean as the data set that was carefully inspected and corrected by manual intervention. Another possible explanation is that the human bias of the Galaxy Zoo citizen scientists was carried forward in some way, and possible preference of the human annotators of a certain handedness when attempting to differentiate between spiral and elliptical galaxies could be reflected by these variables.

Table 14 shows the variables that were automatically selected by applying three different methods of automatic feature selection as was done in Table 10, as well as the classification accuracy achieved when using these features alone. Most of these variables are the same variables identified by the Fisher discriminant heuristics listed in Table 13. The only exception is \( cx \), which is the \( x \) of the unit vector of the R.A. and decl.

This table also shows that when using only the selected features, the classification accuracy does not change substantially. When using all of the features except for those selected by the CFS algorithm, the classification accuracy is 63.94% with Random Forest and 65.48% with Bagging. When removing the features selected by the Subset Attribute Eval and Filtered Attribute Eval, the classification accuracy is 62.15% and 68.04% using the Random Forest and Bagging classifiers, respectively.

The experiment using the Galaxy Zoo galaxies showed that some of the magnitude model fitting variables showed statistically significant difference between clockwise and counterclockwise galaxies. These variables did not exhibit a statistically significant difference when using the computer-generated data set, but the classification accuracy when using the PSF magnitude model fitting likelihoods (lnLStar), the exponential magnitude model (lnLExp), and the de Vaucouleurs magnitude model (lnLDeV) using all five bands, the classification accuracy is higher than mere chance. Using Random Forest, the classification accuracy is 56.1%, and when using Bagging the classification accuracy is 54.5%. When assigning the galaxies with random handedness, however, the classification accuracy is random. Table 15 shows the confusion table of the classification when using the magnitude model variables and Random Forest classifier.
5. DISCUSSION

Machine learning is typically used in astronomy for handling the vast pipelines of astronomical data (Fayyad et al. 1993; Djorgovski et al. 2006) and, in particular, analyses of galaxies (Ball et al. 2004; Oyaizu et al. 2008). In this study, supervised machine learning was used to show differences between patterns of photometric variables of different types of galaxies: galaxies with clockwise patterns and galaxies with counterclockwise patterns. In particular, statistically significant differences were observed in the SDSS “Stokes U” parameter, as well as the magnitude model fitting likelihoods.

The PSF fitting likelihood (lnLStar) variable is used in the SDSS pipeline to separate between stars and galaxy sources. When measured for galaxies, lnLStar can distinguish between flatter galaxies and galaxies which are less extended and are more point-like. The galaxies annotated by Galaxy Zoo 2 have a relatively large surface size, but the higher PSF fitting likelihood shows that counterclockwise galaxies may be somewhat more dense, or have a more dominant nucleus, allowing a better PSF fitting.

The de Vaucouleurs fit is often used to measure the variations in the surface brightnesses of galaxies, and despite normally being used to profile elliptical galaxies, it can be considered more useful for profiling spiral galaxies compared to PSF fitting. The results show that counterclockwise galaxies, on average, exhibit a higher $\chi^2$ likelihood of fitting to the de Vaucouleurs surface brightness distribution model. As mentioned above, the de Vaucouleurs surface brightness model was initially proposed for elliptical galaxies (de Vaucouleurs 1948), but in SDSS DR7 measurements were collected for all of the photometric objects. The higher likelihood of fitting the de Vaucouleurs surface brightness model might also suggest the existence of a bright nucleus and a sharper and more consistent drop in brightness in spiral galaxies that rotate counterclockwise.

The experiment was performed with two different data sets of galaxies separated by their handedness. The first was based on galaxies classified as spiral by crowdsourcing, and the second was generated in a fully automatic process, without the intervention of humans. The consistency between the experiments using the two different data sets indicates that the ability of a classifier to predict the handedness of a galaxy using its photometry data is not necessarily driven by bias of the manual galaxy annotations.

The first data set, in which spiral galaxies were selected by crowdsourcing, the number of galaxies with clockwise handedness was higher than the number of galaxies with counterclockwise handedness. That distribution of handedness disagrees with the handedness distribution in the data set in which the galaxies were classified automatically, where the number of clockwise galaxies is slightly higher, but does not exhibit a statistically significant difference. That shows that the human classification of the spiral galaxies carried out by Galaxy Zoo was biased by the handedness, making a galaxy with clockwise handedness more likely to be voted as spiral compared to a galaxy with a counterclockwise pattern. These results show that studies carried out by applying the force of citizen scientists to analyze galaxy morphology can lead to biased results.

The asymmetry discussed here between clockwise and counterclockwise galaxies is observed through those galaxies classified by Galaxy Zoo, as well as a data set of galaxies classified in a fully automatic process. As the measurements for a large population of galaxies is expected to be symmetric, it is difficult to identify specific reasons for the ability of a classifier to predict the handedness of a galaxy based on its photometry.

One possible explanation can be a consistent bias in the SDSS measurements taken from clockwise and counterclockwise galaxies. For instance, a consistent error in the measurement of the position angle that discriminates between clockwise and counterclockwise galaxies could affect several different measurements such as the SDSS “Stokes parameters.” Since in both data sets classification accuracy can be achieved by just using a subset of the variables, such possible error possibly affects several different variables. Variables that allows the classification include variables affected by the position angle, but also variables that are not affected by the measurement of the position angle, such as the magnitude model fitting likelihood variables. It will therefore require further investigation to profile the nature of the measurement bias and identify its source.

I would like to thank Dr. Noah Brosch and Dr. George Djorgovski for the helpful discussions and insightful comments that helped to improve the manuscript. The study was supported in part by NSF grant IIS-1546079.

APPENDIX

MATERIAL

The data files used in the experiments as well as computer-generated output files can be accessed at http://vfacstaff.ltu.edu/lshamir/data/assym, and permanent data files are also available at Shamir (2016).

To perform the experiment using WEKA, the CSV file with the photometry information of the galaxies should be downloaded at http://vfacstaff.ltu.edu/lshamir/data/assym/p_all_full.csv, and opened using WEKA Explorer. In the “classify” tab, the field “rotation” should be selected as the class label, and then “BayesNet” (or any of the other classifiers) should be selected as the classifier. The results in this paper were produced by selecting a percentage split of 80% and 10-fold cross-validation, but other test strategies can also be used.

A file with randomized galaxy handedness can be downloaded at http://vfacstaff.ltu.edu/lshamir/data/assym/p_all_full_randomized.csv.

The results using WND can be produced using the input file http://vfacstaff.ltu.edu/lshamir/data/assym/p_all_full_wndchrm.csv. The CSV file can be analyzed directly using the UDAT software, which implements the WND classifier and can be downloaded at http://vfacstaff.ltu.edu/lshamir/downloads/udat/. After downloading the executable and library files, the following command line should be used:

```
udat-w-r0.2p_all_full_wndchrm.csvwndchrm_output.html
```

The “-w” switch activates the WND algorithm for classification, and “-r0.2” makes it randomly allocate 20% of the samples for testing, and use the remaining 80% for training.

The resulting HTML file generated by the program can be viewed at http://vfacstaff.ltu.edu/lshamir/data/assym/wndchrm_output.html. The HTML file contains the classification accuracies, as well as other relevant information about the experiment such as the Fisher discriminant scores of the different variables that were used in the classification.
Input file with randomized handedness can be downloaded at [http://vfacstaff.ltu.edu/lshmim/data/assym/p_all_fullwndchrm_randomized.csv](http://vfacstaff.ltu.edu/lshmim/data/assym/p_all_fullwndchrm_randomized.csv), and the results when using it can be viewed at [http://vfacstaff.ltu.edu/lshmim/data/assym/wndchrm_output_randomized.html](http://vfacstaff.ltu.edu/lshmim/data/assym/wndchrm_output_randomized.html).

For cross-validation, the following command line is used: `udat-w-r0.2-n10p_all_full_wndchrm. csvwndchrm_output.html`.

The resulting report file produced when using cross-validation can be viewed at [http://vfacstaff.ltu.edu/lshmim/data/assym/wndchrm_output_10fold.html](http://vfacstaff.ltu.edu/lshmim/data/assym/wndchrm_output_10fold.html).

REFERENCES

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2009, ApJS, 182, 543

Ball, N. M., Loveday, J., Fukugita, M., et al. 2004, MNRAS, 348, 1038

Bernstein, G., & Jarvis, M. 2002, AJ, 123, 583

Breiman, L. 1996, Mach. Learn., 24, 123

Breiman, L. 2001, Mach. Learn., 45, 5

Brosche, P. 1973, A&A, 23, 259

devaucouleurs, G. 1948, AnAp, 11, 247

Djorgovski, S., & Davis, M. 1987, AJ, 313, 59

Djorgovski, S. G., Donalek, C., Mahabal, A., et al. 2006, in 18th Int. Conf. Pattern Recognition, ed. Y. Y. Tang et al. (Piscataway, NJ: IEEE), 856

Dojcak, L., & Shamir, L. 2014, NewA, 28, 1

Dong, L., Frank, E., & Kramer, S. 2005, Knowledge Discovery in Databases: PKDD 2005 (Berlin: Springer)

Fayyad, U. M., Weir, N., & Djorgovski, S. 1993, in Proc. Tenth Int. Conf. on Machine Learning, 112

Fisher, R. A. 1938, Ann. Eugenics, 8, 376

Friedman, N., Geiger, D., & Goldszmidt, M. 1997, Mach. Learn., 29, 131

Goeman, J. J., & Solari, A. 2014, Stat. Med., 33, 1946

Hall, M., Frank, E., Holmes, G., et al. 2009, ACM SIGKDD Explor. Newsl., 11, 10

Hall, M. A. 1998, PhD thesis, Univ. Waikato, Hamilton

Hoehn, C., & Shamir, L. 2014, AN, 335, 189

Holte, R. C. 1993, Mach. Learn., 11, 63

Keller, G. 2015, Statistics for Management and Economics, Abbreviated (Toronto: Nelson Education)

Kohavi, R., & John, G. H. 1995, in Int. Conf. on Machine Learning ICML, 304

Kuminski, E., George, J., Wallin, J., & Shamir, L. 2014, PASP, 126, 959

Kuminski, E., & Shamir, L. 2016, ApJS, 223, 20

Land, K., Slosar, A., Lintott, C., et al. 2008, MNRAS, 388, 1686

Lintott, C., Schawinski, K., Bamford, S., et al. 2011, MNRAS, 410, 166

Liu, H., & Setiono, R. 1997, IEEE Trans. Knowl. Data Eng., 4, 642

Longo, M. J. 2011, PhLB, 699, 224

Mak, N. 1979, ITSML, 9, 62

Oyaiu, H., Lima, M., Cunha, C. E., et al. 2008, AJ, 674, 768

Shamir, L. 2009, MNRAS, 399, 1367

Shamir, L. 2011a, ApJ, 736, 141

Shamir, L. 2011b, Ganaizer: A tool for automatic galaxy image analysis, Astrophysics Source Code Library, ascl:1105.011

Shamir, L. 2012, PhLB, 715, 25

Shamir, L. 2013, Galax, 1, 210

Shamir, L. 2016, Input file for Asymmetry between galaxies with clockwise handedness and counterclockwise handedness, v2.0, Figshare, doi:10.6084/m9.figshare.3116815

Shamir, L., Orlov, N., Eckley, D. M., et al. 2008, Source Code Biol. Med., 3, 13

Shamir, L., Orlov, N., Eckley, D. M., et al. 2013, WND-CHARM: Multi-purpose image classifier, Astrophysics Source Code Library, ascl:1312.002

Willett, K. W., Lintott, C. J., Bamford, S. P., et al. 2013, MNRAS, 435, 2835

York, D. G., Adelman, J., Anderson, J. E., Jr et al. 2000, AJ, 120, 1579