Feature Extraction and Selection of Optical Galaxy Data

Roy Henha Eyono
University of Cape Town
Cape Town, South Africa
hnhroy001@myuct.ac.za

ABSTRACT

Galaxy Morphology Classification remains predominately a manual task often subject to crowd sourcing for classification. Given the vast amount of galaxy image data at our disposal and the anticipated splurge of petabytes of galaxy data, this presents an even greater need for the automatic classification of these galaxies.

Classification often relies upon a good set of features to discriminate between various classes, furthermore this paper describes an approach to feature extraction and selection of galaxy images for the task of galaxy classification. We made use of popular feature extraction techniques as per literature and developed a hybrid feature selection algorithm which makes use of univariate and tree-based feature selection techniques in a recursive feature elimination setting to select the best performing features.

Upon comparing the hybrid feature selection algorithm with other preexisting techniques on a set of generic learning algorithms, the algorithm outperformed its counterparts sporting the lowest mean squared error.

CCS CONCEPTS

• **Computing Methodologies → Machine Learning**;

KEYWORDS

Galaxy Morphology, Feature Extraction and Selection, Hubble Tuning Fork

1 INTRODUCTION

Efforts are being made to enrich our understanding of the universe. These efforts often come in the form of Sky Surveys or Large Telescopes. In understanding the space around us, not only does it encourage the exploration of our universe but it also further enriches our understanding of our earth and galaxy alike.

The ability to identify the various galaxies that make up our universe is one such way to appreciate our understanding of the universe. However, for the longest time, the task of Galaxy Identification has always been performed by human experts, as the automation of this task has yet to outperform that of a human expert.

Understanding our universe is one advantage of Galaxy Identification, but the ability to identify these galaxies automatically can help in making use of the splurge of galaxy data at our disposal and in anticipation. The Square Kilometer Array[1] is one such project which promises petabytes upon petabytes of data.

Existing approaches for galaxy classification includes crowdsourcing galaxy data [6], the use of diagnostic diagrams [11] and more recently the use of machine learning techniques. This paper explores the latter, specifically feature extraction and selection, techniques often associated with the approach.

As much as crowdsourcing as well as the use of diagnostic diagrams do produce competitive results,they both struggle to deal with the volume, velocity and variety of present day galaxy data. Hence the reason why machine learning techniques are favored in tackling this task.

Machine learning techniques give computers the ability to learn and perform a task without being explicitly programmed. This can either be achieved either in a supervised fashion or unsupervised fashion. However, irregardless of the approach, the concept of features or factors play an important role in both learning and performing a task, and even more importantly in our task, Galaxy Classification.

Intuitively, the better your features are, the easier it is to discriminate between the various categories of galaxies. Hence, this paper will describe a host of appropriate feature extraction approaches for galaxy classification on optical data as well as present a feature selection model for ranking and eventual selection of the best performing features.

This project shies against choosing features subjectively as traditionally, particularly in galaxy classification, the choice of features had been heavily derived from domain knowledge or an affinity to a particular feature extraction framework. This paper aims to address this matter in answering the following research question: What features in optical galaxy data are appropriate for galaxy classification?

This project is of a three part series, whereby moving forward, the goal is to construct a cognitive vision system for galaxy classification. A cognitive vision system can be defined as any system that aims to gather expert knowledge and combine it with predictive capabilities to model and explain future events. The construction of a CVS often centers around developing and integrating or coupling various components that solve a given problem.

As far as the structure of this paper is concerned, this paper initially discusses relevant papers which pertain to the search for discriminatory features for Galaxy Classification. Thereafter, there will be a background section to equip the reader on some of the concepts discussed further in the paper. The Model section, soon after the background, will discuss and describe the feature extraction techniques explored as well as the details of the hybrid feature selection algorithm. Following the model section is the experimental design section which is intended to describe how the hybrid feature selection algorithm went under evaluation, the results of this evaluation can be found in the results section and similarly in the discussion of results. Finally the paper comes to a conclusion in the conclusions and feature works sections.

[1] http://skatelescope.org/

2 RELATED WORK

In relevant literature as far as the choice of features for galaxy classification is concerned, emphasis is placed on using a preferred set of features rather than the best set of features.

Authors champion their choice of features based on domain knowledge of the problem or a set of features that they believe generalize to the problem at hand. The selection of features is subjective rather than objective.

WND-CHARM [8] or Weighted neighbor distances using a compound hierarchy of algorithms representing morphology is one such example whereby a set of features is believed to generalize to the task at hand. Shamir et al. [9] make use of WND-CHARM for the automated classification of galaxy images achieving an accuracy of 90%.

WND-CHARM extracts a feature vector of approximately 3000 features. This feature vector includes the raw image, transforms of the raw image, transform upon transform of the raw image, as well as a host of other features aimed to generalize the task of image classification.

Shamir et al decided to extend his work in [9], by making use of WND-CHARM to quantitatively analyze the morphology of merger galaxies[10]. Unfortunately achieving an accuracy of 51% of the test simulated galaxy merger images, 1.9% than random classification. Sextractor is another example [2], like WND-CHARM, of a feature extraction package created with the purpose of generalizing to astronomical images.

Aptoula et al.[1] work on the segmentation and classification of galaxies is an example of authors who use domain knowledge to champion their choice of features. In [1], Aptoula et al. are inspired by the astronomical view point that since spiral galaxies are dominated with active star formations,spiral galaxies tend to be brighter than elliptical galaxies which in turn are reputably dominated by old and cold stars.

Aptoula et al. leverage this knowledge by producing a feature set based on the top-hat operator, for the distinction of spiral galaxies from elliptical galaxies. The resulting model had an accuracy of of 79% whereby classification errors were mainly attributed to the noise levels of the images and normalization issues.

Like Aptoula et al., Moore et al. [7] leveraged domain knowledge by using the average size and average roughness of a galaxy to differentiate between elliptical and late-type spiral galaxies. They applied both features to a multilayer perceptron where they reached an accuracy of 91.4%.

In the work of Goderya et al[5], they too make use of shape descriptors such as elongation and convexity as features for galaxy classification. Achieving an accuracy of 57% on a test set of 37 images. Literature suggests that the appropriate use of features can lead to promising results, and that the extraction of features is still a popular in galaxy classification.

There has been work published by Dieleman et al. [4], where they presented a convolution neural network trained on galaxy images from the Sloan Digital Sky Survey[2]. Convolution Neural Networks perform their own internal feature extraction on raw images, often alluding to a "black-box" effect. Dielemen et al's results in [4] achieved a near-perfect accuracy of 99%. Irregardless of the result, due to this black-box effect in this particular model, researchers are still left puzzled on what features are discriminatory in galaxy classification.

Taking into consideration the state of existing literature, it is evident that there is a limitless amount of features that one can choose from for galaxy classification. However, it is possible to perform feature selection, given a finite set of features, in order to maximize the performance of the classifier, hence the premise of this paper.

[2]http://www.sdss.org/

3 BACKGROUND

In order to perform the necessary feature extraction and selection techniques, background knowledge of galaxies and their properties should be understood. An understanding of the different types of feature selection techniques at our disposal should also be established as it will reinforce ones understanding of the feature selection algorithm presented in this paper.

3.1 Hubble's Tuning Fork

Edwin Hubble's Classification scheme also known as Hubble's Tuning Fork[3] is the standard agreed upon by astronomers alike to categorize galaxy types. In this scheme, galaxies are categorized according to their morphology, namely into two main groups: elliptical galaxies and spiral galaxies.

Elliptical galaxies tend to be round and flat, varying according to their roundness and flatness, whereas spiral contain a central bulge in their centers. The bulge in spiral galaxies is surrounded by stars in a separated fashion in which the stars separate themselves into two or more spiraling arms.

Elliptical galaxies are further sub-categorized into four categories namely E0, E3, E5 and E7. E0 being more spherical in shape, E3 less spherical than E0 but more ellipsoidal than E0 and the trend applies with E5 and E7. Hence, chronologically E7 being the most ellipsoidal and flatter galaxy in comparison to the other elliptical galaxies. Hence, E0 being the most spherical galaxy of the elliptical galaxies.

Spiral Galaxies, like elliptical galaxies, are further subdivided into two categories, regular and barred spiral galaxies. Regular Spiral Galaxies, have more natural spirals projecting outwards of the central bulge, whilst barred galaxies have a protruding bar within their central bulge with spirals are projecting outwards from the protruding bar.

Both regular spiral galaxies and barred spiral galaxies are further divided into three categories: Sa, Sb and Sc for regular spirals and SBa, SBb and SBc for barred spirals. Each of which is a scale in chronological order of the compactness of the spirals in relation to the central bulge. Sa and SBa being the most compact whilst Sc and SBc being the least compact.

Apart from elliptical and spiral galaxies, there are other types of galaxies. These galaxies include irregular galaxies and lenticular galaxies. Irregular Galaxies are galaxies that donâĂŹt contain any general structure and Lenticular galaxies are essentially an intermediary between spiral and elliptical galaxies. Figure 1 illustrates the different types of galaxy morphologies.

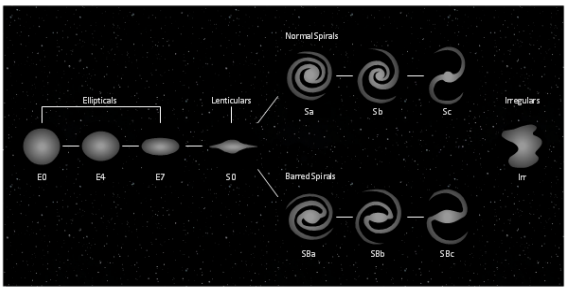

Figure 1: Hubble's Tuning Fork

3.2 Galaxy Zoo Dataset

The Galaxy Zoo Project [6] is a project which aims to accelerate the task of galaxy classification through the method of crowd-sourcing. This crowd-sourcing is achieved by allowing volunteers to classify galaxies online[3]. The images used in the Galaxy Zoo Project have been retrieved from the Sloan Digital Sky Survey. The classification of these galaxies are achieved through a series of questions and selecting a corresponding image, instead of an explicit classification of the galaxies relative to the Hubble Tuning Fork.

One of the questions asked in the classification scheme makes reference to the image of interest and subsequently asks the user if the galaxy is 'simply smooth and rounded with no sign of a disk'. Depending on the response of the user, more questions will be asked until the scheme has exhausted all the possibilities. These possibilities can be represented as a decision tree as illustrated in Figure 2.

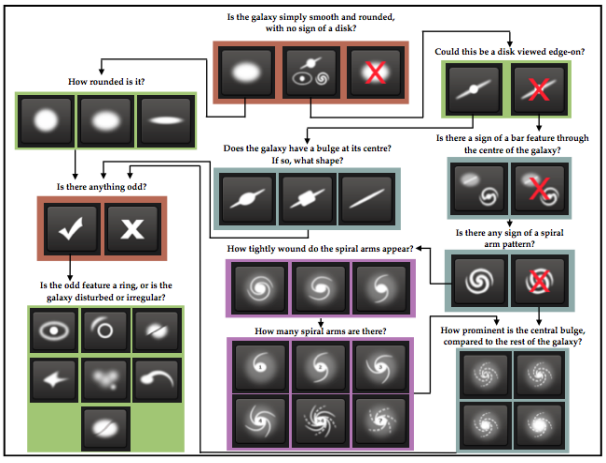

Figure 2: Galaxy Zoo Decision Tree

The Galaxy Zoo Project was a success, as it resulted in the classification of 900 000 galaxies within a timespan of several months. Given its success, Galaxy Zoo had then generously made the data freely available to the public in search for an automated galaxy morphology classifier. This dataset forms the basis of the feature extraction techniques discussed in this paper.

[3] www.galaxyzoo.org

The dataset contains raw images of various galaxies and a set of 37 labels for each galaxy, each representing the probability of a user agreeing to a particular property along the decision tree as per the historical data.

Figure 3: Sample Galaxy Zoo Raw Image

3.3 Feature Selection

Unlike feature extraction, there are finitely many methods of feature selection. Feature selection can be defined as the process of selecting a set of relevant features for use in the construction of a model. It is categorized into three types namely Filter Methods, Wrapper Methods or Embedded Methods.

3.3.1 Filter Methods. With Filter Methods, the selection of features is independent of a learning algorithm. Features are ranked relative to the dependent variable through a series of statistical tests. Through this ranking, users can select the best K features suitable for their model. Examples include the Chi-Squared Statistical Test, Linear Discriminant Analysis and Pearson's Correlation.

3.3.2 Wrapper Methods. Wrapper Methods essentially treat feature selection as a search problem. A subset of features are fed into a learning algorithm and done so continuously until the set of features with the best performance on the algorithm is reached. Examples include Forward Selection and Recursive Feature Elimination.

3.3.3 Embedded Methods. In the case of embedded methods, feature selection is performed within a learning algorithm through the process of regularization. Whereby features are penalized to reduce over fitting. Examples of embedded methods include LASSO Regression and Decision Trees.

3.4 Principle Component Analysis

The Principle Component Analysis (PCA) Statistical Procedure is a technique we use to perform dimensionality reduction on our data. PCA is a method that reduces the dimensionality of multi-dimensional data in a manner that captures the essence of the original data. It essentially achieves this reduction by repetitively

projecting n-dimensional vectors onto an (n-1)-dimensional vector in a manner that maximizes variance. Figure 4 illustrates this concept.

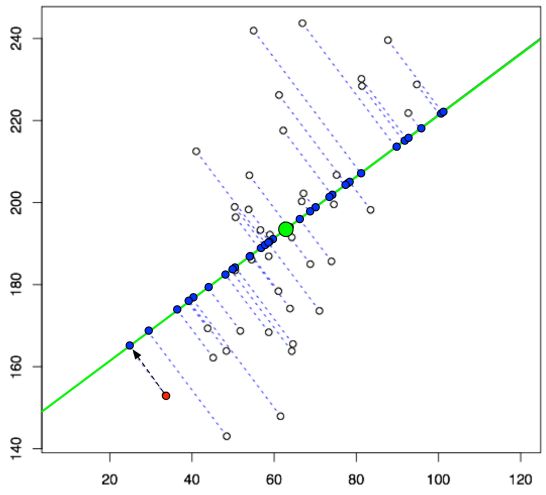

Figure 4: Principle Component Analysis

4 MODEL

The model consists of two components, the feature extraction component and the feature selection component. We'll proceed to describe the features chosen for this particular study as well as describe the underlying technicalities of the feature selection algorithm.

4.1 Extracted Features

The features chosen for the model are a compilation of the numerous features explored in literature for the problem of galaxy classification on optical images, particularly the Galaxy Zoo Dataset. These set of features amount to a total of 2928 features. This includes the set of features extracted by the WND-CHARM feature extraction package as well as a handful of shape descriptors.

The WND-CHARM package extracts a total of 2919 features. The features include the raw image, transforms of the raw image, and transforms of transforms of the raw image. Transforms include the Fourier transform, Chebyshev transform and Wavelet transform or possibly the combinations of these transforms. The extracted features also include polynomial decompositions, high contrast features, pixel statistics and texture features. The features extracted by WND-CHARM make up the bulk of the entire feature set, as it is was designed to tailor for general image classification tasks, however it has been previously applied to galaxy classification [9].

A handful of shape descriptors were extracted from the galaxy zoo images to stray away from the dependency of third party feature extraction packages. These features were inspired by the works of Goderya et al. [5]. A total of 9 features were extracted:

- Elongation
- Form Factor
- Convexity
- Bounding Rectangle to Fill Factor
- Galaxy Area
- Galaxy Bounding Rectangle Width
- Galaxy Bounding Rectangle Height
- Brightness

Elongation is defined as the measure of flatness of the object,whereas it is particularly useful in discriminating between the different classes of elliptical galaxies, as well as discriminating between normal spirals and barred spirals.

The form factor is the ratio of the area of the galaxy and the square of the perimeter of the galaxy. Goderya [5] suggest that elliptical galaxies tend to have somewhat higher values in form factor due to their luminosity being more symmetrical distributed in the image. Whilst barred spiral galaxies tend to show smaller values in form factor as their perimeter per unit are considerably large.

Convexity is defined as the ratio of the perimeter of the galaxy and two multiplied by the sum of the height and width of the minimum bounding rectangle around the galaxy. For elliptical galaxies, there is a decreasing trend from E0 to E7, while for simple galaxies there is an increasing trend from Sa to Sc and finally for barred spirals, there is a slight decreasing trend from SBa to SBc.

Whilst, bounding rectangle to fill factor (Bx) is defined as the ratio of the area of the galaxy to the area of the bounding rectangle. Bx measures how much space the galaxy occupies within the bounding rectangle. Simple and barred spiral galaxies show strong decreasing trends from Sa to Sc and SBa to SBc.

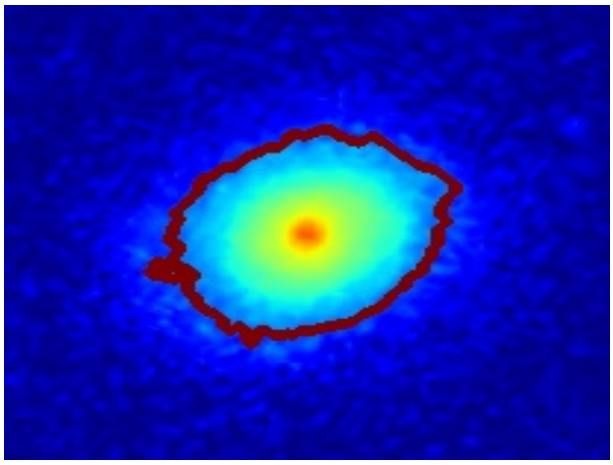

Figure 5: Contoured Galaxy Image in Grayscale Jet ColorMap

The features were extracted by cropping the galaxy image and extracting the contour of the galaxy of interest as illustrated in Figure 5. A gaussian blur was applied onto the image for a smoothening

effect, as contouring was achieved using the opencv python package [4]. Once the contouring was achieved, the remaining features were subsequently derived.

4.2 Feature Selection Algorithm

The feature selection algorithm presented in this paper makes use of the random forest feature selection technique, the pearsons correlation feature selection technique as well as recursive feature elimination. This forms essentially a hybrid feature selection technique which leverages on all three methods of feature selection.

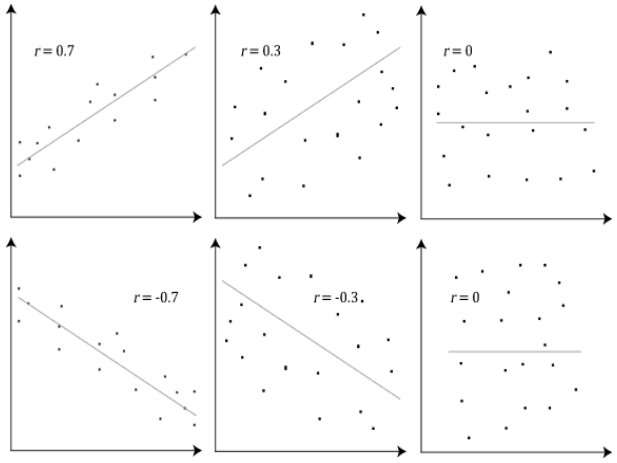

Figure 6: Different measures of Pearson's correlation

The Pearson's Correlation Feature Selection Technique falls under the category of Filter Methods. It is the measure of linear correlation between two variables. Also called the correlation coefficient, it ranges from -1 to 1. Whereby the extreme -1 represents a strong negative correlation between two variables and 1 represents a strong positive correlation. A correlation coefficient of 0 represents no correlation. As illustrated in Figure 6.

The correlation coefficient can be interpreted as the mean of the line of regression between both variables. Pearson's correlation can be applied to continuous data and it doesn't assume normality, although it does assume a finite variance and covariance.

The random forest feature selection technique is an embedded method of feature selection. A random forest is an ensemble learning algorithm for both classification and regression. It trains a multitude of decision trees and uses these decision trees for prediction, either averaging or getting the mode across all the trees. Figure 7 is an illustration of a random forest.

Apart from its predictive abilities, it also possesses a feature selection component, whereby it ranks the most important features once training is complete. This ranking is often used as a metric to assess how predictive a feature is.

Random forests use the technique of mean decrease impurity to perform feature selection. It can be computed how much each feature reduces the weighted impurity in a tree, whereby the weighted impurity can be thought of as a measure of how much a tree is

[4]https://pypi.python.org/pypi/opencv-python

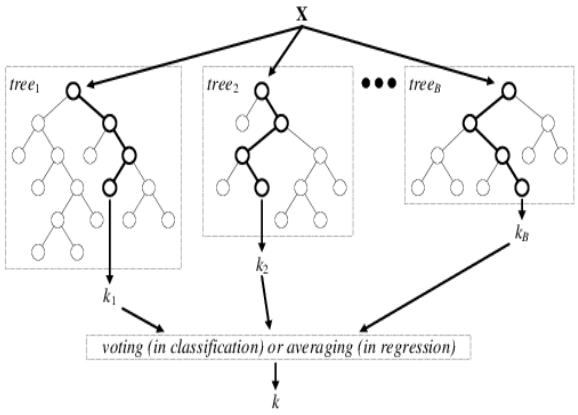

Figure 7: Random Forest

over fitting the training data. Ranking this reduction in weighted impurity across all features in a tree and averaging it across all trees in a random forest is the basis of the random forest feature selection technique.

Unlike Random Forests and Pearson's Correlation, Recursive Feature Elimination (RFE) is a wrapper method. Given a learning algorithm with the ability to measure feature performance, the RFE technique selects features by recursively considering smaller and smaller sets of features. It continuously prunes the feature set based on the performance of each feature on the learning algorithm at each iteration. The features are repeatedly pruned until the desired number of features is reached. Algorithm 1 illustrates this technique.

Algorithm 1 Recursive Feature Elimination

```
1: procedure RFE(model, minfeaturesize, step)
2:     while size(model.features) ≠ minfeaturesize do
3:         model.fit()
4:         data ← model.data
5:         ranking ← model.featureImportance
6:         indices ← getSmallestNIndices(ranking, n = step)
7:         model.data ← data.removeColumns(indices)
8:     return model
```

In constructing the hybrid feature selection algorithm, all three techniques were taken into consideration. The hybrid feature selection algorithm is essentially a wrapper of the recursive feature elimination technique. In our algorithm, recursive feature elimination is applied onto a random forest, whereby the features are recursively pruned at each iteration based on the average weighted impurity of the forest and the correlation coefficient of each feature relative to the dependent variable. Algorithm 2 is an illustration of the hybrid feature selection algorithm.

The algorithm is dependent on how the feature importance measures from both the correlation coefficient and the random forest intertwine to create a new feature importance measure. A new feature importance measure is firstly achieved by applying the softmax

Algorithm 2 Hybrid Feature Selection Algorithm

```
procedure HybridFS(rfModel, minfeaturesize, step)
    while size(rfmodel.features) ≠ minfeaturesize do
        rfmodel.fit()
        rfRanking ← model.featureImportance
        pearsScore ← pearsonsr(rfmodel.data, rfmodel.label)
        ranking ← computeRanking(rfRanking, pearsScore)
        indices ← getSmallestNIndices(ranking, n = step)
        rfmodel.data ← data.removeColumns(indices)
    return rfmodel
```

function onto the vector of correlation coefficients for the feature set. The softmax function or normalized exponential function is a normalization function, which takes in a K-dimensional vector of arbitrary real values and "squashes" it to a K-dimensional vector of real values in the range [0,1]. The function is defined by:

$$\sigma(z)_j = \frac{exp(z_j)}{\sum_k exp(z_k)}, \text{ for k=1...N}$$

Once the softmax function has been applied to the vector of correlation coefficients, we'll have a ranking of features in the ranges of [0,1]. However, before applying the function onto the vector, the absolute value of the vector should be computed as a preprocessing step. The reason being that the correlation coefficient ranges between [-1,1], with the extremes representing a strong negative or positive correlation, furthermore we are only concerned when there exists a strong correlation. Hence, the negation of the correlation coefficient has no relevance.

Once the normalized correlation coefficient feature importance vector (r) and the random forest importance vector (R) has been computed, we apply the following formula to create a new hybrid feature importance vector Ω. This new importance measure ω is defined by:

$$\omega_j = \alpha r_j + \beta R_j \text{ where } \alpha + \beta \leq 1$$

α and β are constants as they both represent the weighting of each feature selection technique. Applying softmax function on the resultant importance vector will result in the hybrid feature importance vector Ω. The derivation of Ω is the underlying computation behind the function *computeRanking()* in Algorithm 2.

5 EXPERIMENTAL DESIGN

In order to evaluate the performance of the hybrid feature selection algorithm, a number of experiments had to be performed. The basis of these experiments were to compare the hybrid feature selection algorithm against other preexisting feature selection techniques on a handful of popular learning algorithms. The random forest feature selection technique, univariate feature selection and the recursive feature elimination techniques were chosen as benchmarks for the experiment as they each represent the various types of feature selection techniques available. Hence, three learning algorithms were chosen, namely the feed forward neural network, the support vector machine and the random forest.

The comparison of the various feature selection techniques is achieved by taking the entire galaxy zoo dataset feature set and selecting the best ten for each respective feature selection technique and using these features to train a learning algorithm for the task of galaxy classification. The performance of the learning algorithm on a derived set of features as per the feature selection technique is considered as the evaluation metric of the chosen feature selection technique.

5.1 Implementation of The Feature Selection Techniques

Each feature selection technique have their different parameters and properties, this was all taken into account in implementation. Four feature selection techniques had been implemented:

- Random Forest Feature Selection
- Pearson's Correlation Feature Selection
- Recursive Feature Elimination
- Hybrid Feature Selection

For the random forest feature selection technique, a total of ten trees were trained. Each tree being trained until all the leaves of the true reached purity or until all leaves contained at most 1 tree in their sample, allowing each tree to grow to its maximum capacity. Once the random forest had been trained, the feature importance vector for each feature in the forest was retrieved and used as a ranking to get the ten best features. This was implemented in python with the scikit-learn package [5].

The scikit-learn package was also used in implementing the pearson's correlation feature selection technique as each features correlation coefficient was computed against the dependent variables. Taking into consideration that the correlation coefficient ranges from [-1,1], with the extremes representing correlation, the absolute value was computed. This vector of correlation coefficients were ranked and subsequently allowed the selection technique to extract the ten best features.

Due to the nature of the galaxy zoo dataset and the feature selection techniques, the 37 classes in the galaxy zoo dataset was reduced to one representation. Essentially, the aim was to find features that are indicative to all 37 classes instead of each individual class. The 37 classes each represent a node in a decision tree and finding features that are able to correctly identify the probability of a human traversing a node in the galaxy zoo decision tree across 10 features is much more feasible and valuable as in comparison to getting the 10 best features for each node whereby, if distinct, can reach a total of 370 features for classification. The reduction of the 37 classes into one representation was achieved through the Principle Component Analysis (PCA) statistical procedure.

The implementation of the Recursive Feature Elimination (RFE) algorithm differs from its counterparts in the sense that it can only be applied on learning algorithms that measure feature importance, unfortunately not all algorithms do the following, neural networks being amongst that class of learning algorithms. Hence, the RFE algorithm was only applied on the random forest.

[5] http://scikit-learn.org/

The RFE algorithm implemented in this experiment has a step size of 250, eliminating the worst performing 250 features at each iteration. The RFE + Random Forest implementation makes use of the same random forest used in the Random Forest Feature Selection Algorithm. These RFE techniques were also implemented in python with the scikit-learn package.

The Hybrid Feature Selection algorithm was developed as a wrapper of the scikit-learn RFE package. The selection algorithm made use of the same random forest used in the random forest feature selection technique, as well as the pearsons correlation selection technique, both embedded into the RFE algorithm, whereby at each iteration with a step size of 250, a hybrid feature importance vector was returned. Both techniques were equally weighted.

All the feature selection algorithms were computed on the UCT High Performance Cluster [6]. Due to limited time and processing power on the cluster, the selection algorithms were ran on 1000 instances across the 2928 features.

5.2 Implementation of The Learning Algorithms

Three learning algorithms were implemented:

- Feedforward Neural Network
- Random Forest
- Support Vector Machine

All learning algorithms were trained on a total of 57971 images, in the form of the ten best features from each feature selection algorithm.

The chosen topology for the neural network was one with one hidden layer consisting of 50 nodes, an input layer of 10 nodes for the 10 best features and an output layer of 37 nodes representing the 37 nodes in the galaxy zoo decision tree. Since the problem required 37 continuous numbers as solutions, in other words, a multilabel output regression problem, the hidden layer used the relu function as its activation function with the output layer containing no activation function. The performance of the neural network was achieved by cross-validating the training data by 10-folds. Cross-validation is a technique that is used for the assessment of how the results of a model generalize to an independent data set, the folds represent the number of splits the data must undergo for evaluation. The neural network was implemented on the deep learning framework, keras[7].

The Random Forest chosen for evaluation had a total of 10 trees, however each tree couldn't grow to it's full capacity and was restricted to a max depth of 2. Naturally, the random forest cannot train on multiple outputs or targets, hence the MultiOutputRegressor class in scikit-learn helped in adapting the random forest to solve the problem at hand. As in the case of the neural network for evaluation, the Random Forest underwent 10-fold cross validation.

The Support Vector Machine (SVM) used in this experiment was of the Regressor type rather than for classification. Considering the lengthy runtimes that Support Vector Regressor Machines go through, the learning algorithm was bootstrap aggregated or 'bagged'. Bootstrap aggregating is a machine learning ensemble meta-algorithm whereby an estimator, in our case the SVM, is trained on k subsets of the original data, producing k estimators. In prediction, the input is fed separately to all k estimators and a vote is taken between the estimators on which is the appropriate output. The k subsets are in often cases mutually inclusive.

The bagged SVM consisted of 5 estimators, each with a maximum sample size of 1/10 of the training data. Once the SVM had been bagged, like the Random Forest, it was transformed into a Multi Output Regressor. Evaluation was achieved via a 2-fold cross validation of the training data.

6 RESULTS

After having applied the various feature selection techniques onto the galaxy zoo dataset and training them on the various learning algorithms, the mean squared error, trained on the chosen feature selection techniques, were measured. Table 1 reports the various scores achieved during evaluation.

FS Technique	NN(10-fold CV)	Random Forest(10-fold CV)	SVM(2-fold CV)
Random Forest	0.222	0.0228	0.0459
Pearson's Correlation	0.024	0.0238	0.040
RFE + Random Forest	0.050	0.0223	0.0458
Hybrid FS	0.038	0.0221	0.0459

Table 1: Feature Selection Technique Comparisons

Across all three learning algorithms, each feature selection technique produced relatively similar mean squared errors. The best performing techniques for each respective learning algorithm changed hands between Pearson's Correlation FS and the Hybrid FS Algorithm. Pearson's correlation produced the best results for both the 10-fold cross validated Neural Network and the 2-fold cross validated Support Vector Regressor Machine, whilst HybridFS performed best in the 10-fold Cross Validated Random Forest.

Interestingly enough, both the Hybrid Algorithm and the Random Forest Learning Algorithm produced the lowest mean squared error across the experiment with a mean squared error of 0.0221, achieving the best performance between a feature selection technique and a learning algorithm.

Upon observing how each feature selection technique converges to their minimum mean absolute error on a neural network, one notices minimal difference in learning performance across all 30 epochs from pearsons feature selection algorithm. It appears that the feature selection algorithm after the first few epochs already circumvents around it's local minimum. Upon numerous trials, the feature selection technique still maintained this performance. This performance is illustrated in Figure 8.

However, the remaining feature selection techniques were conventional as they underwent a few epochs out of 30 to reach their respective plateau's, with the Hybrid Feature Selection taking the longest and the Random Forest Feature Selection the shortest.

As far as the average runtime of each respective feature selection algorithm is concerned, there is great variance between each technique. The algorithms with the largest runtime are the Hybrid and the RFE+Random Forest algorithms both coming in at 141.86s and 142.05s respectively. Pearson's FS algorithm came in at the low time of 2.81s. These algorithms were all computed on the same node on the UCT High Performance cluster, each node possessing

[6] http://hex.uct.ac.za/
[7] https://keras.io/

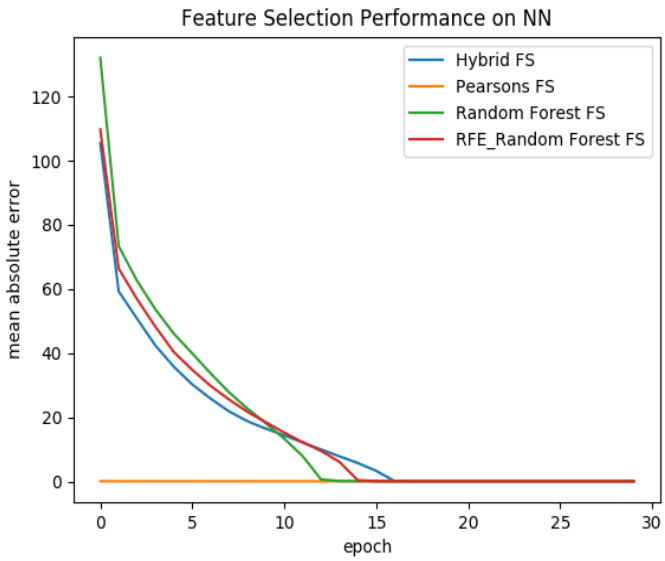

Figure 8: Performance Evaluation on Neural Network

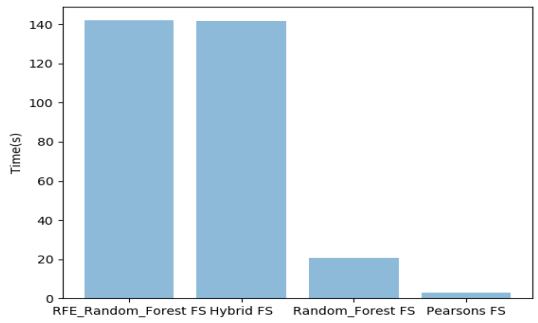

Figure 9: Feature Selection Runtime Comparisons

approximately 128GB of RAM, 4 AMD Opteron 6274 CPU's each equipped with 16 cores. Figure 9 is a bar chart which illustrates the time differences between each selection technique.

7 DISCUSSION OF RESULTS

The results produced in these experiments are testament to the importance of the feature selection whether user defined or automatically selected. This is shown by how the scores of the feature selection techniques vary both within and between learning algorithm.

However, these experiments also revealed the computational complexity of these algorithms, indicating that the task of feature selection can be computationally intensive. The need for computational power for both feature selection as well as the task of learning, which is reputably a computationally expensive task, can be discouraging, but promises results.

The pearsons correlation feature selection technique impressed with its results across all learning algorithms. The technique outperformed other selection techniques on 2 out of 3 learning algorithms. It achieved this performance in the shortest time as well, as it had the shortest runtime out of all the techniques. However, its performances on the feedforward neural network raised questions as the technique converged to a local minimum in its first few epochs and maintained this performance throughout the duration of the experiment.

The pearsons feature selection neural network underwent 10-fold cross validation suggesting that this behavior did not happen by chance. It is possible that the neural network architecture gave rise to this behavior. Had the network been deeper or wider, the result could have been different.

The standout learning algorithm of the experiment was the Random Forest as it produced the lowest mean square error, as well as maintained promising results across all feature selection techniques. The mean squared error produced by the Random Forest on all the feature selection techniques were quite closely tied, sporting one of the lowest variances, the lowest being the SVM. This variance raised the question of whether or not the Random Forest is sensitive to feature selection, and furthermore whether there exists a set of learning algorithms that are resilient to feature selection techniques.

Considering that 3 out of the 4 feature selection techniques technically resembled that of the random variance, this could have resulted in the low variance between values, but this does not apply to the SVM, leaving this as an open issue. The Neural Network on the other hand was sensitive to the various feature selection technique sporting the highest variance.

The Hybrid Feature Selection presented in this paper also produced impressive results, achieving the lowest mean squared error in collaboration with the random forest learning algorithm. It also competed with other techniques on other learning algorithms.

Interestingly enough, both the Hybrid Feature Selection technique and the RFE + Random Forest have similar technical makeups, the difference lies in the fact that the Hybrid selection algorithm has not only a Random Forest embedded in the RFE but also Pearson's correlation selection algorithm. The Hybrid Selection, in this particular experiment, weights both techniques equally. Regardless of their similarity, they both produced significantly different results for the Neural Network. In that particular instance, it is safe to assume that the Pearsons Selection Algorithm could have been the difference between the two seeing that it performed best for that particular experiment. Hence, perhaps weighting the algorithms differently could have changed the face of experiment. Adding another feature selection technique in this hybrid feature selection technique and weighting this appropriately could also heed interesting results.

Given the nature of the galaxy zoo crowd-sourcing experiment, the galaxy zoo dataset had multiple outputs as labels. In the experiments, these labels were reduced to one dimensional with the principle component analysis statistical procedure. Even though the reduced labels were merely representations of the true data, the selection algorithms appeared to able to see through this representation and extract features that generalized to all labels. A comparison can be explored to see which technique works well for

feature selection on multioutput labelled data but this particular approach appears feasible judging from the results.

In comparison to some of the results achieved with the galaxy zoo dataset with similar extracted features [10][9][5], the results achieved in this experiment outperform these results reported in literature. The reason possibly lies in the fact that in the experiments, emphasis was placed on strict automated feature selection rather than choosing features based on domain knowledge or an affinity to a particular feature extraction framework. The advantage of the approach described in this paper is that it is able to prune a large volume and variety of features to a smaller set of features, in this case 10, and produce promising results.

The results are still not at the standard set by [4] with their convolution neural network which achieved a mean squared error of 0.005. Regardless, the results achieved in these experiments do reveal interesting insight on the capabilities of feature selection.

8 CONCLUSION

In this project, we extracted thousands of features from galaxy optical data, the bulk of the features originating from the WNDCHARM feature extraction framework and the remainder being a host of shape descriptive features. With these features, we developed a hybrid feature selection algorithm, leveraging on the various types of feature selection techniques, to automatically select ten of the best performing features.

Upon comparing the hybrid feature selection technique presented in this paper with other preexisting feature selection techniques, the former produced the lowest mean squared error after having been trained on a Random Forest. The technique reveals some interesting insight on embedding feature selection techniques and simultaneously weighting their importance within a recursive feature elimination setting.

The results produced in this paper are still a far-cry from the benchmark for galaxy classification but they do suggest that automated feature selection is a feasible way to assess the appropriateness of a feature for the task of galaxy classification. Considering that there are possibly infinitely many feature extraction techniques at ones disposal, whether it be domain driven or through a framework, one could make use of all these techniques and undergo evaluation to ascertain which features work best considering the problem. Much can be said about the important role played by feature selection algorithms and their contribution in the machine learning pipeline.

9 FUTURE WORK

Future work will focus on further developing the hybrid feature selection technique, particularly experimenting with the weighting of each embedded algorithm and similarly adding more embedded components. Plans to use the selection technique on other problem spaces apart from image classification will also be explored. Similarly, work will also be done on using this feature selection component to develop the cognitive vision system for galaxy classification.

REFERENCES

[1] Erchan Aptoula, Sébastien Lefevre, and Christophe Collet. 2006. Mathematical morphology applied to the segmentation and classification of galaxies in multispectral images. In *Signal Processing Conference, 2006 14th European*. IEEE, 1–5.

[2] Emmanuel Bertin and Stephane Arnouts. 1996. SExtractor: Software for source extraction. *Astronomy and Astrophysics Supplement Series* 117, 2 (1996), 393–404.

[3] Christopher J Conselice. 2006. The fundamental properties of galaxies and a new galaxy classification system. *Monthly Notices of the Royal Astronomical Society* 373, 4 (2006), 1389–1408.

[4] Sander Dieleman, Kyle W Willett, and Joni Dambre. 2015. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly notices of the royal astronomical society* 450, 2 (2015), 1441–1459.

[5] Shaukat N Goderya and Shawn M Lolling. 2002. Morphological classification of galaxies using computer vision and artificial neural networks: A computational scheme. *Astrophysics and space science* 279, 4 (2002), 377–387.

[6] Chris J Lintott, Kevin Schawinski, Anže Slosar, Kate Land, Steven Bamford, Daniel Thomas, M Jordan Raddick, Robert C Nichol, Alex Szalay, Dan Andreescu, et al. 2008. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society* 389, 3 (2008), 1179–1189.

[7] Jason A Moore, Kevin A Pimbblet, and Michael J Drinkwater. 2006. Mathematical morphology: Star/galaxy differentiation & galaxy morphology classification. *Publications of the Astronomical Society of Australia* 23, 4 (2006), 135–146.

[8] Nikita Orlov, Lior Shamir, Tomasz Macura, Josiah Johnston, D Mark Eckley, and Ilya G Goldberg. 2008. WND-CHARM: Multi-purpose image classification using compound image transforms. *Pattern recognition letters* 29, 11 (2008), 1684–1693.

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

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

[11] Henrik WW Spoon, JA Marshall, JR Houck, Moshe Elitzur, L Hao, L Armus, BR Brandl, and V Charmandaris. 2006. Mid-infrared galaxy classification based on silicate obscuration and PAH equivalent width. *The Astrophysical Journal Letters* 654, 1 (2006), L49.