Section A: Overview of the Research Project Proposal

1. Academic level of research project (Masters or Doctoral): Masters
2. Broad field of research (Engineering or Astronomy/Astrophysics): Engineering
3. Title of the research project: Automatic classification of radio glaxies using traditional machine learning techniques.
4. Research project abstract/summary (max 250 words):

Often within specific sub-field of science, a specialist distinction is made in datasets with certain characteristics (such as the Fanaroff-Riley morphological classification scheme). This usually requires highly specialised domain knowledge, that is often accompanied by years of study in the given field as well as vast amounts of time spent examining the data. This in itself is not a problem. However, when the classification of data becomes the focal point, it becomes a problem, especially with the advent of projects such as the Square Kilometer Array (SKA) radio telescope which is expected to produce nearly 300 PB of data in a year, manual classification by experts is no longer a viable solution. Automatic classification methods is the most obvious solution. The main problem with the Fanaroff-Riley (FR) dichotomy is the morphological feature space used during classification. Even though one of the official features in the literature seems to be the FR ratio, very few astronomers and astrophysicist use this as the main morphological feature [Dr Imogen Whittam and Dr Mattia Vicari from UWC] rather this is but one of a set of features being used during manual classification, and that even from the literature there is no clearly defined set of features for FR classification. The focus of this project is to examine the role the FR ratio plays in classification, to clarify the additional feature space being examined by the experts, and to attempt to set up classifiers based on those features examined.

Section B: Supervisor(s) Details

1. Primary supervisor's details
    a. Title and full name: Dr. Trienko Lups Grobler
    b. Name of South African or SKA Partner Country university at which the primary supervisor is a permanent academic staff member: University of Stellenbosch
    c. Email address and/or contact telephone number: tlgrobler@sun.ac.za
    d. Supervision of postgraduate students:
        i) Doctoral: N/A

        ii) Maters:

Name	**Nationality**	**Start Date**	**Completion Date**	**Title**	**Co-Supervisor**
Chuneeta Nunhokee	Mauritian	Jan 2013	Dec 2014	Link between ghost artefacts, source suppression and incomplete calibration sky models	O.M. Smirnov
Ulrich Armel Mbou Sob	Cameroonian	Jan 2015	Dec 2016	Calibration and imaging with variable radio sources	S.K. Sirothia

Section C: Full Research Project Proposal

1. Scientific Merrit:

LITERATURE REVIEW

Radio galaxies are defined generally as active galaxies that are strong emitters of radio waves and more technically as radio sources powered by accretion onto a super massive black hole that produce extended structures called radio jets and lobes (Kozieł-Wierzbowska & Stasińska, 2011). Radio galaxies are divided into two classes based on the morphology of their radio structure (Fanaroff & Riley, 1974) namely FRI and FRII. The classification scheme proposed uses the Fanaroff-Riley (FR) ratio of the distance between the regions of highest brightness (hotspots) on opposite sides of the active galactic nucleus (AGN) and the distance between the farthest edges of the lobes. All sources where the ratio is less than 0.5, are classified as FRI and all sources greater than 0.5 is placed in FRII. Alternative classification schemes based with the ratio at 0.8 have been proposed (Lin et al., 2010) due to increased resolution with the advent of more powerful radio telescopes. All the ratios calculated in these studies have been done by hand, but no large scale computer vision and machine learning study has been performed that rigorously and statistically verifies the accuracy of the ratio as a useful morphological feature. Lin et al. (2010) also proposes " to define an objective measure (or measures) that allows us to trace the galaxy population smoothly from FR type-I-like sources to type-II-like ones (as opposed to a sharp and perhaps arbitrary type I versus type II division). With the aid of such a measure, we hope to reduce the subjectiveness inherited in the traditional ways of classification, thus increasing the repeatability of our results by other researchers". While their study has provided ground work for such a scheme and laid out several measures to perform classification, no uniform tool-set has been developed for this purpose and their own methodology has the similar issues as to what they have laid out to previous schemes: that is a lack of an objective measure and subjectiveness inherited due to an undefined feature space. Other sources in the literature provide other potential morphological features, such as the presence of jets (Owen & Laing, 1989) or hotspots towards the edges of the lobes (Gendre & Wall, 2008) as well as the number of hotspots present (Lukic et al. 2018). Well established machine learning methods such as the k-means algorithm (Wagstaff et al. 2001) or Gaussian mixture models can be useful ways to segment these images and easily extract the features automatically. The k-means algorithm can be made more efficient by choosing the initial positions intelligently (Bradley et al. 1998). Further analysis of useful features will be extremely beneficial in the creation of accurate automatic classifiers. A meta-study done by Feigelson et al (2006, p 252) with regards to classification methods used in astronomy found that the majority of studies used neural networks (~150 studies, 1990-2006), commenting that although this method is quite effective, it gives 'black box' results: difficult to reproduce and very little insight into the feature space used. The majority of classifiers set up between then and now have been neural networks, albeit these were the more recently developed Convolutional Neural Networks (CNN) (Aniyan & Thorat, 2017). Very few studies used Bayesian classifiers (~30 studies, 1985-2006) or decision trees (~20 studies, 1994-2006) and none used Classification and Regression Trees (CART) (Breiman, 1984).

Looking at existing CNN's trained for classification (Lukic et al, 2018) and identifying the features that it is looking at will be of extreme importance. This can be done by methods such as image retrieval by scene graph grounding (Johnson et al. 2015) and visualizing of the CNN (Zeiler & Fergus, 2014) to name a few methods. The importance of feature selection is much clearer in a case such as this. Classification without explanation or feature models to back it up leaves astronomy (or in the general case the field we're applying these methods to) in a precarious position: classification might no longer coincide with actual physical characteristics of radio galaxies, potentially undoing research from data labelled by automatic classifiers that have not undergone rigorous feature selection and engineering). One alternative method for verification and labelling for such systems is by crowd sourcing. A website is set up that any person can log into and classify the given data (after a tutorial is given). The Zooniverse project (Bonney et al, 2014) has successfully managed to do this with several projects, including radio astronomy through the Radio Galaxy Zoo (RGZ) project (Banfield et al, 2015). RGZ itself has managed to publish 5 papers from its results, 3 in the last year. Human based labelling is very useful and results much more regualrly in the identification of aberrations that are novel discoveries rather than simply interference. Objects discovered by citizen scientists include Hanny's Voorwerp (Lintott et al, 2009) and the Grean Pea galaxies (Cardamone, 2009). This provides the option to either get the South African public involved in classification to improve public engagement with science or could lead to job creation in the drive toward the knowledge-based economy by training people specifically for classification, especially where proprietary data is concerned.

Kozieł-Wierzbowska, D. and Stasińska, G. (2011). FR II radio galaxies in the Sloan Digital Sky Survey: observational facts. Monthly Notices of the Royal Astronomical Society, 415(2), pp.1013-1026.

Lin, Y., Shen, Y., Strauss, M., Richards, G. and Lunnan, R. (2010). ON THE POPULATIONS OF RADIO GALAXIES WITH EXTENDED MORPHOLOGY AT z< 0.3. The Astrophysical Journal, 723(2), pp.1119-1138.

Fanaroff, B. and Riley, J. (1974). The Morphology of Extragalactic Radio Sources of High and Low Luminosity. Monthly Notices of the Royal Astronomical Society, 167(1), pp.31P-36P.

Owen, F.N. and Laing, R.A., 1989. CCD surface photometry of radio galaxies–I. FR class I and II sources. Monthly Notices of the Royal Astronomical Society, 238(2), pp.357-378.

Gendre, M.A. and Wall, J.V., 2008. The Combined NVSS–FIRST Galaxies (CoNFIG) sample–I. Sample definition, classification and evolution. Monthly Notices of the Royal Astronomical Society, 390(2), pp.819-828.

Lukic, V., Brüggen, M., Banfield, J.K., Wong, O.I., Rudnick, L., Norris, R.P. and Simmons, B., 2018. Radio Galaxy Zoo: Compact and extended radio source classification with deep learning. arXiv preprint arXiv:1801.04861.

Wagstaff, K., Cardie, C., Rogers, S. and Schrödl, S., 2001, June. Constrained k-means clustering with background knowledge. In ICML (Vol. 1, pp. 577-584).

Bradley, P.S. and Fayyad, U.M., 1998, July. Refining Initial Points for K-Means Clustering. In ICML (Vol. 98, pp. 91-99).

Feigelson, E.D. and Babu, J. eds., 2006. Statistical challenges in astronomy. Springer Science & Business Media.

Aniyan, A.K. and Thorat, K., 2017. Classifying Radio Galaxies with the Convolutional Neural Network. The Astrophysical Journal Supplement Series, 230(2), p.20.

Breiman, L., 1984. Introduction tree classification. Classification and regression trees, pp.18-55.
Zeiler, M.D. and Fergus, R., 2014, September. Visualizing and understanding convolutional networks. In European conference on computer vision (pp. 818-833). Springer,
Cham.

Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D., Bernstein, M. and Fei-Fei, L., 2015. Image retrieval using scene graphs. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3668-3678).

Bonney, R., Shirk, J.L., Phillips, T.B., Wiggins, A., Ballard, H.L., Miller-Rushing, A.J. and Parrish, J.K., 2014. Next steps for citizen science. Science, 343(6178), pp.1436-1437.

Banfield, J.K., Wong, O.I., Willett, K.W., Norris, R.P., Rudnick, L., Shabala, S.S., Simmons, B.D., Snyder, C., Garon, A., Seymour, N. and Middelberg, E., 2015. Radio Galaxy Zoo: host galaxies and radio morphologies derived from visual inspection. Monthly Notices of the Royal Astronomical Society, 453(3), pp.2326-2340.

Lintott, C.J., Schawinski, K., Keel, W., Van Arkel, H., Bennert, N., Edmondson, E., Thomas, D., Smith, D.J., Herbert, P.D., Jarvis, M.J. and Virani, S., 2009. Galaxy Zoo:'Hanny's Voorwerp', a quasar light echo?. Monthly Notices of the Royal Astronomical Society, 399(1), pp.129-140.

Cardamone, C., Schawinski, K., Sarzi, M., Bamford, S.P., Bennert, N., Urry, C.M., Lintott, C., Keel, W.C., Parejko, J., Nichol, R.C. and Thomas, D., 2009. Galaxy Zoo Green Peas: discovery of a class of compact extremely star-forming galaxies. Monthly Notices of the Royal Astronomical Society, 399(3), pp.1191-1205.

PROJECT PLAN

The aim of this project is two fold:

The development of...
1.) a radio galaxy inspection tool, implementing a variety of computer vision and machine learning algorithms, that can aid astronomers with feature extraction,
classification and data mining.
2.) a general crowd sourcing classification framework. This can lead to job creation and the involvement of the general South African community in Astronomy.

PROPOSED RESEARCH PLAN

Taking the objectives set out in the Aims and Objectives section, I break each one up into a set of tasks for which I estimate a period of time, fitting into my proposed 2
year period of Masters research.

OBJECTIVES & TASKS

1. MANUAL EXTRACTION OF FANAROFF-RILEY RATIO: Manual extraction of the Fanaroff-Riley (FR) ratio from a selection of the data in order to set up testing
samples for automatic or augmented classification.
1.1 Setting up a platform in order to do manual extraction of the ratio (1 week)
1.2 Manual (by hand) extraction of the ratio on a large data set (1 week)
2. FR RATIO TEST: Set up automatic extraction of FR Ratio and compare that with the given label and manual label from the above objective.
2.1 Develop automatic extraction technique of the ratio using computer vision (2 weeks)
2.2 Test the accuracy of the technique vs. the labelled data and adjust the hyperparameters accordingly (1week)

3. CONVOLUTIONAL NEURAL NETWORK and DEEP LEARNING: Setting up a Neural Network that performs as well as the classical methods.
3.1 Training and setting up the CNN (6 weeks)
4. IDENTIFICATION OF FEATURES: Identifying useful features from the relevant literature and by analysis of what features the CNN is looking at.
4.1 Feature identification from the literature (2 weeks)
4.2 Analysis of features being used by the CNN (6 weeks)
5. FEATURE EXTRACTION: Set up a tool that can extract features from the images and make a database of that.
5.1 Setting up computer vision tools that can extract the features identified in the previous objective (8 weeks)
5.2 Developing a database of the extracted data (1 week)
6. FEATURE & DATA ANALYSIS: Perform feature and data analysis to ensure statistically independence of variables.
6.1 Data analysis on the dataset developed (2 week)
7. SUPERVISED LEARNING: Assuming the input labels as correct, we set up a conventional machine learning classifier for the given features.
7.1 Setting up several classifiers (Naive Bayes, Classification and Regression Trees, Random Forest) (12 weeks)
7.2 Compare accuracy with CNN (1 week)
8. UNSUPERVISED LEARNING: Assuming that the input labels are not necessarily correct, we treat the data as unlabelled and perform clustering on the data set. This
might lead to different class definitions, that can then be correlated to corresponding physical characteristics of the galaxies.
8.1 Setting up different clustering methods (12 weeks)
8.2 Analysis of newly defined clusters with characteristics of galaxies (12 weeks)
9. CITIZEN SCIENCE: Setting up a framework that does classification via crowd sourcing.
9.1 Development of website for classification, that is scalable to large data sets and large volume of users (4 weeks)
9.2 Proof of concept in local communities (3 weeks)
10 WRITING OF ACADEMIC PAPERS: assuming a month per write up for 3 distinct papers (12 weeks)
Total 84 weeks, leaving some additional time for unknown issues occurring.

2. Feasibility:

- We have already secured the needed data from A. Aniyun (Aniyan & Thorat, 2017).
- There is a strong radio galaxy group at UWC (University of the Western Cape) [Dr. Mattia Vaccari]. Once funding is secured every effort will be made to develop a strong research relationship with this research group (and in doing so extend our current data-set). The tools we develop will be of interest to this research group. Preliminary talks have already taken place.
- The computer science division has their own computing cluster, Big-Metal, which is availableto students to do their research.

3. Research Priority Area:
Big Data topics, Interferometric Data Processing and Analysis. This project aims to help with the automatic processing of large amounts of data and can be applied to MeerKAT data. It will become part of the reduction pipeline and slots in after imaging.

4. Ability or skills:
Student should have some programming experience and an interest in machine learning techniques.

Dr. T.L. Grobler 24 July 2018
Lecturer at Stellenbosch University