ASTCVS*

Towards a cognitive vision system for optical astronomical image data

Victor Gueorguiev
CSIR - CAIR & Meraka
P.O. Box 395
Pretoria, Gauteng 0001
grgvic001@myuct.ac.za

Julius Stopforth
University of Cape Town
Private Bag X3
Cape Town, Western Cape 7701
stpjul004@myuct.ac.za

Roy Henha Eyono
University of Cape Town
Private Bag X3
Cape Town, Western Cape 7701
hnhroy001@myuct.ac.za

CCS CONCEPTS

• **Computing methodologies → Knowledge representation and reasoning**; **Computer vision**; **Machine learning approaches**; *Control methods*; • **General and reference** → *Surveys and overviews*;

KEYWORDS

project proposal, artificial intelligence, cognitive vision systems, knowledge representation and reasoning, astronomy, feature selection, hyperparameter optimization, model selection

1 PROJECT DESCRIPTION

Astronomy is a field that actively accumulates and processes vast amounts of data with structure and information that may only be extracted using computational means [9]. For instance, large sky surveys such as the Sloan Digital Sky Survey (SDSS) [24] and the LINEAR survey [21] produce datasets many terabytes in size with tens of millions of data points [11] that require intense computer pre-processing and analysis [9]. South Africa will soon be one of the leading contributors to the field of radio astronomy with the construction of the Square Kilometre Array (SKA)[5]. The telescope will produce data on the order of exabytes in a single year of operation, at a rate of 2 terabytes per second during operation [5]. Therefore, analysis of this data will require the use of automated techniques such as machine learning in order to reduce data into consumable and useful knowledge as well as to perform analyses such as astronomical phenomenon classification[1] [5, 23] in order to assist human processing of the data.

To solve these important problems, astronomy has been the focus of many statistical machine learning approaches [9] and most recently of deep-learning approaches [8]. It has also seen top-down expert knowledge representations applied for efficient data set storage and recollection [17]. It is therefore a prime candidate field for the application of a cognitive vision system (CVS) [2], that can harness expert knowledge representation, feature selection, and machine learning techniques to perform classification tasks such as galaxy classification in optical and radio astronomy settings. Such a cognitive system, with learning and inference techniques in synergy with one another, will be applicable across a variety of different astronomy classification tasks with different datasets. Hence, a CVS for astronomy will have plasticity across such classification tasks.

The construction of a CVS often centers around developing and integrating/coupling various components that solve a given problem. In the case of astronomy, this means that techniques such as feature selection, various machine learning algorithms, and knowledge representation and inference may be applied for automated phenomenon identification and classification, with the goal of data reduction and labeling to streamline astronomy observation workflows. In what will follow in this project proposal, an outline will be discussed of how a three-stage system will be developed for galaxy classification from optical data as a means for moving towards a more general cognitive vision system for astronomy.

2 PROBLEM STATEMENT

This project will tackle astronomical phenomenon classification by examining and developing specific components of a cognitive vision system. These components entail a feature selection step for the extraction of a feature representation from a dataset, a machine learning approach harnessed by model selection and hyperparameter optimization techniques, and inference capabilities embodied by expert domain knowledge of astronomy for spatial and temporal reasoning for identification of phenomenon in optical images. The development and examination of these components will form part of a three-stage pipeline. This pipeline/system will be adaptable to dataset choice, and able to process large tracts of data from astronomical sky surveys. In building such as system, the project will construct a piece of work that moves towards the final objective of a cognitive vision system for astronomy.

To accomplish this task of constructing this pipeline, the developed components will need to interface with each other. This can be udnerstood by considering the following aspects: The first aspect is the application of image processing techniques for the general problem of feature extraction and detection. This may also be supplemented with approaches from machine learning to get the job done; The second aspect is applying various machine learning techniques to the problem of image classification- with the added benefit of well-defined features from the prior step. In this aspect, the problem of CASH will also become prevalent: That is, the problem of simultaneous model selection and hyperparameter optimization. This will be further discussed below; The third aspect is that of expert systems and knowledge representation. This

*Prepared for A/Prof. Deshen Moodley (Supervisor), A/Prof. James Gain (Second Reader)

[1] Phenomenon classification involves classifying astronomical objects and events into various classes. Examples include galaxy classification, asteroid and exoplanet detection, and supernovae classification.

[2] a cognitive, or intelligent, system is any system that aims to gather expert knowledge and combine it with predictive capabilities to model and explain future events [18]. By incorporating expert knowledge, the cognitive system not only learns to predict future events once an action has been selected, but it also learns to rationalize and conceptualize about choices [20]

will hope to encapsulate expert domain knowledge in astronomy using techniques such as Bayesian networks to reason and infer both temporal and spatial information. This will then be used as a top-down layer to the bottom-up machine learning approach.

For each of these components envisioned, the following may be asked:

- **Feature Selection and Extraction:** What image processing techniques are appropriate for galaxy morphology classification?
- **Hyperparameter Learning and Model Selection:** Can state-of-the-art Bayesian Optimization (BO) for the problem of Combined Algorithm and Hyperparameter Selection (CASH) be used to optimize classification accuracies of various machine learning approaches for galaxy morphology classification? Can feasible extensions be made to increase performance of BO on large astronomy data sets, while maintaining model classification accuracy?
- **Expert Knowledge Representation:** Can expert domain knowledge be used to construct a model that can reason and infer information to guide machine learning classification of galaxy morphology classification?

Following from these three questions, one can form an overall question regarding the integration of these components into a well-defined system: How will these three components be integrated/interfaced with one another to improve overall classification accuracy on datasets, and to facilitate the end project goal of moving towards a cognitive vision system for astronomy? For example, in what ways can feature selection and extraction be used to guide machine learning classification and hyperparameter learning of models, and how can expert knowledge reasoning be interfaced with feature selection to choose adequate features given the context from the reasoner?

The project will adopt an approach comprising of two phases. The first phase seeking to answer the component-based research questions, whilst the second phase, subsequent to the first, seeking to answer the overall research question. Solving these sub-problems will lead to the development of the components of an end-to-end integrated system as discussed in previous sections, with the goal of moving towards a more general cognitive vision system in future work.

3 PROCEDURES AND METHODS

The development of this model will undergo two phases. The first phase will deal with the development of the components that make up the cognitive vision system and loosely integrating them to make a unit. Whilst, the second phase will be dedicated to leveraging each component to form a robust system, and furthermore, evaluating our solution. Leveraging will entail tuning parameters in each of the respective components to produce favorable results.

The components are as follows:

- **Feature Extraction and Selection**
- **Machine Learning with Hyperparameter Optimization**
- **Knowledge Reasoner**

These will be discussed below in further detail. The project will also need to constrain itself in scope, since the construction of a full cognitive vision system is currently outside the scope given the time frame of this project. To this effect, the problem will examine a single dataset in the first phase of development. This dataset is the optical image galaxy dataset available from the Galaxy Zoo crowdsourcing initiative [11] and the SDSS, and has already been curated and secured. Only in the second phase will the robustness of the integrated and leveraged components be tested against other datasets and measurement schemes (i.e. optical versus radio versus X-ray data)

3.1 Feature Extraction and Selection

Feature extraction in image processing is a method of transforming large redundant data into a reduced data representation. In our model, we'll perform the WND-CHARM [19] scheme for feature extraction. Weighted neighbor distances using a compound hierarchy of algorithms representing morphology or WND-CHARM is a multi-purpose classifier which can be applied to a variety of image classification tasks without modifications or fine-tuning.

The features extracted from the WND-CHARM scheme will be accompanied by features that describe the brightness, shape, average size and roughness of the galactic images. Hence, amounting to the extraction of approximately 3010 features. The choice of the described feature set is motivated by literature as this is merely a combination of the different feature sets used in

Once extracted, the features will undergo evaluation, namely feature selection, whereby the best features will be selected. Feature selection will occur in two parts. The first part will focus on selecting features that are indicative to the relevant class independent of a learning algorithm, whilst the second part will select features tailored to the learning component of our cognitive vision system.

Filter methods such as the Chi Squared Test, Correlation Coefficient Scores and Information Gain will be applied in the first part of the feature selection process. For the second part, wrapper methods such as the recursive feature elimination algorithm will be explored.

Dimensionality reduction techniques will be considered to reduce the number of features under consideration, reducing the time and space required, possibly resulting in improved performance of the model.

3.2 Combined Algorithm and Hyperparameter Selection for Machine Learning

The machine learning component addresses the problem of combined algorithm selection and hyperparameter optimization (CASH), which aims to both select among a set of algorithms one most appropriate for a given task and set its algorithm parameters to some optimal configuration. Choosing between different machine learning algorithms, and adjusting the manually tunable parameters-such as the number of neurons per layer and number of layers for a neural network- is a cumbersome task [2, 22] and is amicable to be automated through various optimization means.

The method that will be used for CASH and building the machine learning component will be Bayesian Optimization as outlined by Thornton et al [22]. This method is chosen for its efficiency on expensive to evaluate black-box functions such as deep neural networks, and for the few hyperparameters that this method has in its core implementation.

Various tools for this component of CASH are listed in Appendix A under Table 2. The favoured tool emerges as Auto-WEKA for its integration with the full WEKA classification suite in Java, for the availability of comparison metrics such as in Bischl et al. [3], and for its simple and refined documentation and interface.

3.3 Knowledge Representation and Reasoning using Bayesian Networks

The knowledge representation and reasoning will be implemented as a Bayesian Network (BN) using the Edward (ref this) Python library built on top of TensorFlow. This will be done in two distinct stages: implementation and integration. However, throughout this process, the system will be evaluated in terms of the chosen metrics (decide on good metrics to use and ref)

Initially, the system will be constructed so that it is able to process the extracted features and visual concepts from an image in order to correctly identify galaxies contained within an image.

In the second phase, the system will be integrated with the image processor and hyperparameter modules in order to complete the vision system. This will require ensuring that the system is correctly receiving and handling the input received from the other two modules in the vision system's pipeline.

Implementing the inference engine requires specifying what domain knowledge should be included in the representation, what form the input concepts will take, and what form the output of the Bayesian network will take.

3.4 Evaluations

Evaluating the potential for a cognitive vision system and hence the evaluations of the 3 individual components developed will form a large part of the project as there are a couple of criteria to consider for each method, and for each phase:

3.4.1 Phase 1 Evaluations. In this phase, individual component evaluations will be conducted. In this initial phase, only one dataset will be used for evaluation: The Kaggle Galaxy Zoo dataset [11], which is a crowd-sourced classification dataset from the SDSS. For the component examining CASH and the construction of an ensemble machine learning method, the task of evaluation will be tackled by measuring the overall accuracy of final classifications on the full Galaxy Zoo dataset. Performance metrics will also be important in measuring the efficiency of the CASH step of the method, and to this effect works such as those by Bischl et al. [3] will be used as benchmark libraries to compare both performance and accuracy of this component. Evaluation of the Bayesian Network will be conducted by analyzing the performance of the network using the metrics put forth in [14] as well as how accurately it is able to identify galaxies from the Galaxy Zoo dataset. These evaluations will then be recorded as a benchmark to compare against in Phase 2. When components are integrated, i.e. leverage each others' results, the evaluations will be take form of comparing the new performance of the integrated and coupled end-to-end system against the results obtained from individual evaluations

3.4.2 Phase 2 Evaluations. The second phase of evaluation will focus on assessing how the model fairs on astronomical datasets other than the Kaggle Galaxy Zoo dataset. Cognitive Vision Systems are scarce in the field of astronomy, hence the motivation to conduct our evaluations from dataset to dataset. The ability of the model to generalize will be tested through this means.

4 PROFESSIONAL, AND LEGAL ISSUES

Conducting interviews/meetings with astronomy domain experts may require obtaining ethical clearance; however, the nature of these informal interviews is such that there is no personal data being gathered or at risk of being lost. Thus, there is no ethical dilemma. Professional issues involve organizational aspects of organizing collaborations with both domain experts for information gathering (mentioned above) for incorporating expert knowledge, and potential post-phase 2 interfacing of the project with data streams from the SKA [5]. Legal issues may arise with regards to intellectual property of the final system deliverable considering the complex net of stakeholders that are potentially involved such as The Council for Scientific and Industrial Research (CSIR), the University of Cape Town and the Department of Higher Education, and the Square-Kilometre Array. In all these cases however, the common component to deal with is the government; therefore, the legal issues in interfacing between different stakeholders are mitigated by the fact that long-term interests in the project are aligned with all parties.

Tools used for the system are open-source and published under MIT Creative Common licenses, and any implemented methods will be derived from knowledge in the public domain. The work itself will be published under an MIT Creative Common license.

5 RELATED WORK

General cognitive vision systems in astronomy are scarce and not used for galaxy classification. Current state-of-the-art approaches at the SKA [5] involve using the state-of-the-art SKYNET deep neural network and gradient descent on the hyperparameter configuration [8]. The deep learning algorithm is responsible for handling radio wavelength data and processing it for object classification. While it is a state-of-the-art approach used at the SKA, it lacks in the following areas: It does not incorporate expert domain knowledge about astronomical objects in order to perform inference tasks or data-retrieval tasks, and it only uses one method, thus lacking the same generality of a method incorporating algorithm selection that selects algorithms suited for different problems.

Another tool that was used for the related problem of identifying faint sky objects is the SKICAT tool [6] which processed 3 terabytes of raw data containing half a billion images of astronomical objects from the catalog of the Second Palomer Observatory Sky Survey. The tool used decision trees and O-B trees to perform dataset reduction by classifying faint sky objects presented to it. However, while being an example of end-to-end systems for automated data reduction and machine learning in astronomy, it does not use expert astronomy knowledge to conduct spatial inference and identification of other phenomena in images. Moreover, it does not address the more general problem of identifying candidate phenomena based on feature selection as a prior step to classification. As with SKYNET, it solely relies on a single algorithm scheme for classification, this method is further limited to the dataset it was used for and limits its generalization capabilities that would be associated with a more general cognitive system.

5.1 Feature Extraction and Selection

Numerous feature extraction techniques have been applied on optical galactic images. Approaches range from applying morphological operators for feature extraction [1] [16] to the use of shape descriptors [7]. However, given the various approaches to feature extraction, finding an appropriate feature set for galaxy morphology classification comparable to that of a human expert is challenging.

5.2 Combined Algorithm and Hyperparameter Selection

The problem of CASH has been addressed using many various methods such as genetic algorithms in the EMiner tool by [15], gradient-descent methods [3], and Bayesian Optimization. Current state-of-the-art methods for CASH is dominantly Bayesian Optimization (BO) [22], and is currently implemented using Tree-Parzen Estimators or Sequential Model-Based Algorithm Configuration techniques, as in the Auto-WEKA tool by Thornton et al. [22], in the Predict-ML tool by Luo et al. [12] for biomedical applications, and the HyperOpt tool by Bergstra et al. [2]. Bayesian Optimization is the selected method for performing CASH for its few hyperparameters and its robust probabilistic treatment of model hyperparameter configuration exploration [22]. Possible extensions to this method for very large datasets include works by Li et al. [10] for performing learning curve extrapolation and Bischl et al. [3] for performing dataset sub-sampling.

5.3 Knowledge Representation and Reasoning

In terms of expert knowledge inference and reasoning within phenomenon classification and identification, a framework for using Bayesian networks in semantic image analysis using both domain knowledge and visual concepts was developed by Luo et al. [13]. However, the framework was not applied specifically to the problem of galaxy classification. Additionally, work has been done on surveying the metrics for evaluating inferences made with bayesian networks in [14] as well as an investigation into best practises when using BNs in [4]. Furthermore, evaluation metrics and strategies for BNs have also been presented in [4, 14].

6 ANTICIPATED OUTCOMES

The overall outcome of this project will be to have developed an integrated end-to-end system that is able to infer or identify phenomena in optical image data and classify them, using an ensemble machine learning approach that is outfitted with hyperparameter learning and model selection techniques, with adaptive feature selection methods, and an inference engine based on expert knowledge. Another outcome is to have benchmarked this system against current state-of-the-art systems such as SkyNet at the SKA and to initiate a process of applying it the system to radio astronomy data coming from the SKA from telescopes such as Meerkat and Meerlicht.

The feature extraction of the system will be able to successfully identify features of datasets that yield the best performance for classification algorithm metrics, and in the process reduce data dimensionality.

The hyperparameter and model selection stage of the system pipeline will be able to select among an ensemble of machine learning algorithms one that is appropriate for a given image dataset and concurrently optimize hyperparameters for that dataset.

The knowledge modeller will be able to use the visual concepts and features supplied by the image processor in order to correctly identify all galaxies within an image. It should also be able to provide supplementary information using the instilled domain knowledge in order to support the identifications it has made.

Key success factors for the project will be determined by the overall classification accuracy on the initial Galaxy Zoo dataset. This will be conducted on individual components, with the overall integration success measured by whether or not the integrated system can outperform individual components. Comparisons against current state-of-the-art methods using the Galaxy Zoo dataset will also give indication as to the the overall success of the project. The success of the CASH component will also be determined through its performance in comparison to other methods on this dataset. Should the project have time to explore other datasets for validation and evaluation, the success of the system will be determined by its ability to attain high accuracy with regard to benchmark methods on said dataset without any auxiliary modification, i.e. the system robustness to dataset will determine the success should further datasets be pursued.

7 PROJECT PLAN

7.1 Milestones and Timeline

The project plan is set out in the Gantt chart[3] allocation and Tasks and Milestones table in Appendix A Figure 1 and Table 4 respectively. The Drafting stages will contain iterations of 2 drafts of the final paper report. The iteration cycle mentioned under the Development section will be followed by each group member in development of their component of the project, with each subsequent iteration becoming shorter than its predecessor as the component becomes more well-defined. These iterations also encompass the evaluations which will be run on each of the methods, and any interplay between the methods that will be developed as a group towards a final end-to-end pipeline.

7.2 Work Allocation

For the first phase, each group member has been allocated to one of the three primary components:

- **Roy Henha Eyono** Feature Extraction and Selection
- **Victor Gueorguiev** Machine Learning with CASH
- **Julius Stopforth** Knowledge Reasoner

Once completed, group members will work collectively to integrate and leverage the components, hence accomplishing the second

[3] All durations listed in the chart are measured in days

phase of work and moving towards a single integrated pipeline for optical astronomy.

V. Gueorguiev will assume the role of project leader, maintaining technical and professional cohesion within the group. J. Stopforth's role entails ensuring that deadlines are met, and performing administrative duties such as recording meeting minutes. R. Henha Eyono will assume the role of a facilitator, working between V. Gueorguiev and J. Stopforth, assisting both to achieve their respective objectives.

7.3 Deliverables

The main deliverable will consist of the completed three components of the envisioned cognitive vision system, each exposed to each other via some API, such as a web service. This will be delivered via the final code submission and final project paper. An exhaustive list of the other project deliverables can be found in Table 5 in Appendix A. The major deliverables are listed in bold.

7.4 Risks, and Risk Management

See Table 6 in Appendix D for a list of possible risks to the project and how those can be mitigated or otherwise managed. Overall, the risks mentioned have either been mitigated or constitute an unlikely scenario.

7.5 Resources Required

All members of the project will need access to the following resources in order to complete the project:

- A computer, with high-performance CPU power and a GPU for the machine learning component.
- The Galaxy Zoo dataset, available online
- The frameworks and tools listed in Appendix A (also mentioned in detail in Procedures and Methods) and their dependencies
- A LaTeX distribution for typesetting the final paper
- An appropriate development environment including compilers/interpreters for the languages chosen for the components, those being Python, Java, Matlab, and C++.

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] James Bergstra, Brent Komer, Chris Eliasmith, Dan Yamins, and David D. Cox. 2015. Hyperopt: a Python library for model selection and hyperparameter optimization. *Comput. Sci. Disc.* 8, 1 (2015), 014008. https://doi.org/10.1088/1749-4699/8/1/014008

[3] Bernd Bischl, Pascal Kerschke, Lars Kotthoff, Marius Lindauer, Yuri Malitsky, Alexandre FrÃľchette, Holger Hoos, Frank Hutter, Kevin Leyton-Brown, Kevin Tierney, and Joaquin Vanschoren. 2016. ASlib: A benchmark library for algorithm selection. *Artificial Intelligence* 237 (Aug. 2016), 41–58. https://doi.org/10.1016/j.artint.2016.04.003

[4] Serena H. Chen and Carmel A. Pollino. 2012. Good practice in Bayesian network modelling. *Environmental Modelling and Software* 37 (November 2012), 134–145. https://doi.org/10.1016/j.envsoft.2012.03.012

[5] P. E. Dewdney, P. J. Hall, R. T. Schilizzi, and T. J. L. W. Lazio. 2009. The Square Kilometre Array. *Proc. IEEE* 97, 8 (Aug. 2009), 1482–1496. https://doi.org/10.1109/JPROC.2009.2021005

[6] Usama M. Fayyad, Nicholas Weir, and S. George Djorgovski. 1993. SKICAT: A Machine Learning System for Automated Cataloging of Large Scale Sky Surveys. In *ICML*.

[7] 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.

[8] P. Graff, F. Feroz, M. P. Hobson, and A. Lasenby. 2014. SKYNET: an efficient and robust neural network training tool for machine learning in astronomy. *Monthly Notices of the Royal Astronomical Society* 441, 2 (May 2014), 1741–1759. https://doi.org/10.1093/mnras/stu642

[9] Jan Kremer, Kristoffer Stensbo-Smidt, Fabian Gieseke, Kim Steenstrup Pedersen, and Christian Igel. 2017. Big Universe, Big Data: Machine Learning and Image Analysis for Astronomy. *IEEE Intelligent Systems* 32, 2 (March 2017), 16–22. https://doi.org/10.1109/MIS.2017.40

[10] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. 2016. Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *arXiv:1603.06560 [cs, stat]* (March 2016). http://arxiv.org/abs/1603.06560 arXiv: 1603.06560.

[11] Chris J. Lintott, Kevin Schawinski, AnÅ¿e Slosar, Kate Land, Steven Bamford, Daniel Thomas, M. Jordan Raddick, Robert C. Nichol, Alex Szalay, Dan Andreescu, Phil Murray, and Jan Vandenberg. 2008. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey $^{\text{âŸĚ}}$. *Monthly Notices of the Royal Astronomical Society* 389, 3 (Sept. 2008), 1179–1189. https://doi.org/10.1111/j.1365-2966.2008.13689.x

[12] Gang Luo. 2016. PredicT-ML: a tool for automating machine learning model building with big clinical data. *Health Information Science and Systems* 1, 4 (2016), 1–16. https://doi.org/10.1186/s13755-016-0018-1

[13] Jiebo Luo, Andreas E. Savakis, and Amit Singhal. 2005. A Bayesian network-based framework for semantic image understanding. *Pattern Recognition* 38, 6 (June 2005), 919–934. https://doi.org/10.1016/j.patcog.2004.11.001

[14] Bruce G. Marcot. 2012. Metrics for evaluating performance and uncertainty of Bayesian network models. *Ecological Modelling* 230 (April 10, 2012), 50–62. https://doi.org/10.1016/j.ecolmodel.2012.01.013

[15] Rayrone Zirtany Nunes Marques, Luciano Reis Coutinho, Tiago Bonini Borchartt, Samyr BÃľliche Vale, and Francisco JosÃľ da Silva e Silva. 2015. EMiner: A Tool for Selecting Classification Algorithms and Optimal Parameters. *Polibits* 52 (Dec. 2015), 17–24. https://doi.org/10.17562/PB-52-2

[16] 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.

[17] SS Murray, EW Brugel, G Eichhorn, A Farris, JC Good, MJ Kurtz, JA Nousek, and JL Stoner. 1992. The NASA Astrophysics Data System: A heterogeneous distributed processing system application. In *European Southern Observatory Conference and Workshop Proceedings*, Vol. 43.

[18] S. Nefti and J. O. Gray (Eds.). 2010. *Advances in Cognitive Systems.* IET, The Institution of Engineering and Technology, Michael Faraday House, Six Hills Way, Stevenage SG1 2AY, UK. http://digital-library.theiet.org/content/books/ce/pbce071e DOI: 10.1049/PBCE071E.

[19] 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.

[20] Robert J. Schalkoff. 2011. *Intelligent systems: principles, paradigms, and pragmatics.* Jones and Bartlett Publishers, Sudbury, Mass. OCLC: ocn406119763.

[21] Grant H Stokes, Frank Shelly, Herbert EM Viggh, Matthew S Blythe, and Joseph S Stuart. 1998. The Lincoln Near-Earth Asteroid Research (LINEAR) Program. *LINCOLN LABORATORY JOURNAL* 11, 1 (1998).

[22] Chris Thornton, Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. 2013. Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '13)*. ACM, New York, NY, USA, 847–855. https://doi.org/10.1145/2487575.2487629

[23] Jake Vander Plas, A. J. Connolly, and Z. Ivezic. 2014. AstroML: Python-powered Machine Learning for Astronomy, Vol. 223. 253.01. http://adsabs.harvard.edu/abs/2014AAS...22325301V

[24] D. G. York. 2000. The Sloan Digital Sky Survey: Technical Summary. *The Astronomical Journal* 120, 3 (Sept. 2000), 1579–1587. https://doi.org/10.1086/301513 arXiv: astro-ph/0006396.

A TABLES

Table 1: A list of possible tools and frameworks for working with Bayesian Networks

Name	Link	Description
Elbow	https://github.com/davmre/elbow	A framework built on Google's TensorFlow.
Jayes	http://www.eclipse.org/recommenders/jayes/	A Bayesian Network framework for Java
Edward	http://edwardlib.org/	A Python library for Bayesian Neural Networks that is capable of modeling, and inference. Uses TensorFlow libraries for neural networks.

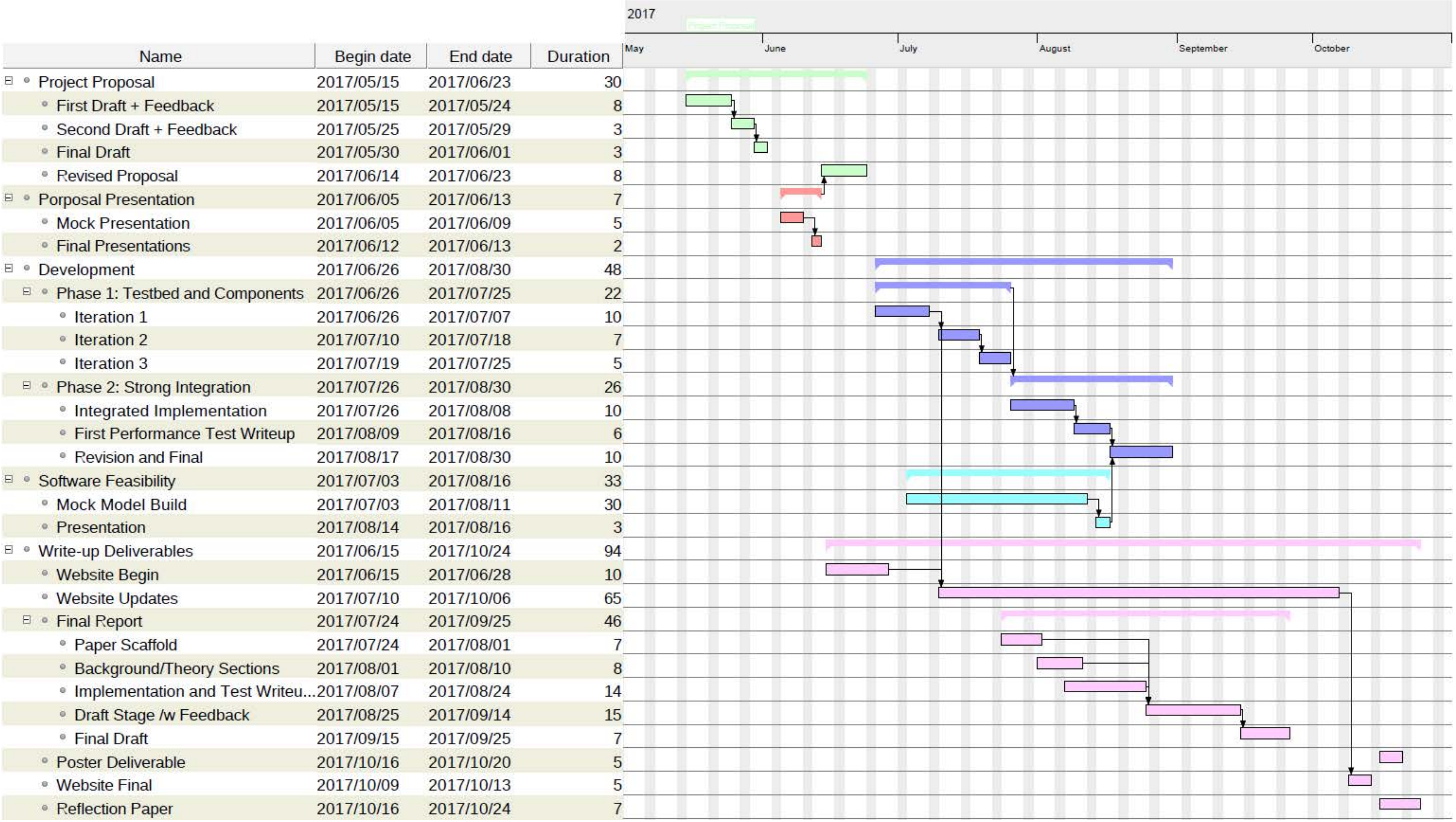
Table 2: A list tools for implementing model selection and hyperparameter optimization

Name	Link	Description
Auto-WEKA [22]	http://www.cs.ubc.ca/labs/beta/Projects/autoweka/	A hyperparameter optimizer and model selector for the WEKA ML library for Java
PredicT-ML [12]	Unavailable	A biomedical application of a SMBO CASH tool for machine learning automation
EMiner [15]	Unavailable	A hyperparameter optimizer and model selector using evolutionary algorithms developed for Java
HyperOpt [2]	https://github.com/hyperopt/hyperopt	A Python framework for hyperparameter optimization for Scikit-Learn

Table 3: A list of possible tools and frameworks for Feature Extraction and Selection

Name	Link	Description
WND-CHARM	https://github.com/wnd-charm/wnd-charm	A framework for the WND-CHARM algorithm.
Auto-WEKA [22]	http://www.cs.ubc.ca/labs/beta/Projects/autoweka/	Model selector for the WEKA ML library for Java
Scikit-learn	http://scikit-learn.org/	A Python library for ML.
OpenCV	http://opencv.org/	A multipurpose computer vision library.

B GANTT CHART AND MILESTONES



Name	Begin date	End date	Duration
⊟ ● Project Proposal	2017/05/15	2017/06/23	30
● First Draft + Feedback	2017/05/15	2017/05/24	8
● Second Draft + Feedback	2017/05/25	2017/05/29	3
● Final Draft	2017/05/30	2017/06/01	3
● Revised Proposal	2017/06/14	2017/06/23	8
⊟ ● Porposal Presentation	2017/06/05	2017/06/13	7
● Mock Presentation	2017/06/05	2017/06/09	5
● Final Presentations	2017/06/12	2017/06/13	2
⊟ ● Development	2017/06/26	2017/08/30	48
⊟ ● Phase 1: Testbed and Components	2017/06/26	2017/07/25	22
● Iteration 1	2017/06/26	2017/07/07	10
● Iteration 2	2017/07/10	2017/07/18	7
● Iteration 3	2017/07/19	2017/07/25	5
⊟ ● Phase 2: Strong Integration	2017/07/26	2017/08/30	26
● Integrated Implementation	2017/07/26	2017/08/08	10
● First Performance Test Writeup	2017/08/09	2017/08/16	6
● Revision and Final	2017/08/17	2017/08/30	10
⊟ ● Software Feasibility	2017/07/03	2017/08/16	33
● Mock Model Build	2017/07/03	2017/08/11	30
● Presentation	2017/08/14	2017/08/16	3
⊟ ● Write-up Deliverables	2017/06/15	2017/10/24	94
● Website Begin	2017/06/15	2017/06/28	10
● Website Updates	2017/07/10	2017/10/06	65
⊟ ● Final Report	2017/07/24	2017/09/25	46
● Paper Scaffold	2017/07/24	2017/08/01	7
● Background/Theory Sections	2017/08/01	2017/08/10	8
● Implementation and Test Writeu...	2017/08/07	2017/08/24	14
● Draft Stage /w Feedback	2017/08/25	2017/09/14	15
● Final Draft	2017/09/15	2017/09/25	7
● Poster Deliverable	2017/10/16	2017/10/20	5
● Website Final	2017/10/09	2017/10/13	5
● Reflection Paper	2017/10/16	2017/10/24	7

C TASKS AND MILESTONES

Table 4: Detailed tasks and milestones with dates for the ASTCVS project pipeline

Task Name	Start	End
Project Proposal	2017/05/15	2017/06/23
First Draft + Feedback	2017/05/15	2017/05/24
Second Draft + Feedback	2017/05/25	2017/05/29
Final Draft	2017/05/30	2017/06/01
Revised Proposal	2017/06/14	2017/06/23
Project Presentation	2017/06/05	2017/06/13
Mock Presentation	2017/06/05	2017/06/09
Final Presentations	2017/06/12	2017/06/13
Development	2017/06/26	2017/08/30
Phase 1: Testbed and Components	2017/06/26	2017/07/25
Iteration 1	2017/06/26	2017/07/07
Iteration 2	2017/07/10	2017/07/18
Iteration 3	2017/07/19	2017/07/25
Phase 2: Strong Integration	2017/07/26	2017/08/30
Integrated Implementation	2017/07/26	2017/08/08
First Performance Test Writeup	2017/08/09	2017/08/16
Revision and Final	2017/08/17	2017/08/30
Software Feasibility	2017/07/03	2017/08/16
Mock Model Build	2017/07/03	2017/08/11
Presentation	2017/08/14	2017/08/16
Write-up Deliverables	2017/06/15	2017/10/24
Website Begin	2017/06/15	2017/06/28
Website Updates	2017/07/10	2017/10/06
Paper Scaffold	2017/07/24	2017/08/01
Background/Theory Sections	2017/08/01	2017/08/10
Implementation and Test Write-ups	2017/08/07	2017/08/24
Feedback and Revised Draft	2017/08/25	2017/09/14
Final Paper	2017/09/15	2017/09/22
Final Code	2017/09/15	2017/10/02
Poster Deliverable	2017/10/16	2017/10/20
Website Final	2017/10/09	2017/10/13
Reflection Paper	2017/10/16	2017/10/24

Table 5: Deliverables for the ASTCVS project

Description	Due
Project Proposal	2017/06/02
Revised Proposal	2017/06/30
Project Presentation	2017/06/14
Software Feasibility	2017/08/14-18
Write-up Deliverables	
Web Presence	2017/06/30
Paper Scaffold	2017/08/01
Background/Theory Sections	2017/08/10
Implementation and Test Write-ups	2017/08/24
Final Report	2017/09/22
Final Code Submission	2017/10/02
Final Project Demo	2017/10/02-09
Poster Deliverable	2017/10/20
Website Final	2017/10/12
Reflection Paper	2017/10/24

D RISKS, AND RISK MANAGEMENT

Table 6: A list of possible risks, their likelihood of occurring, their impact, and management/mitigation strategies

Risk	Impact	Likelihood	Mitigation/Management
Unable to find a usable dataset of labeled astronomical images	Disastrous	Rare	Already mitigated. Using the open source Galaxy Zoo dataset.
Components are not able to integrate at all	Disastrous	Unlikely	Discussion and meeting amongst team members regarding the inputs and outputs of each component in Phase 1.
Unable to expose components via a web interface	Critical	Unlikely	Wrap component output/inputs into a format that does permit exposure via web interface.
Unable to find a second dataset to compare against	Moderate	Rare	Segmenting the Galaxy Zoo dataset and using one half for evaluation of Phase 1 and the other half for evaluation of Phase 2.
One or more components are not ready to integrate	Moderate	Unlikely	Ensure that each project member has a clear explicit idea of how their component will integrate with the other two components and meet regularly to discuss progress of component implementation.
The results of the evaluations are inconclusive	Moderate	Possible	If time permits, choose different evaluation metrics and re-evaluate. Components will be evaluated as different iterations are completed to ensure that the evaluations can pivot earlier if need be.
Implementation iteration takes longer to complete that expected	Critical	Likely	Prevent entire project from falling behind by regularly assessing where the implementation is every week in comparison with deadlines and adjust deadlines if possible/necessary.
There is not enough time to move the project into Phase 2	Trivial	Possible	Ensure that deadlines are followed as closely as possible so that time is available to move to Phase 2. The split of project into 2 phases means there will still be a demonstrable system.