Star-Galaxy-Quasar Classification Using a Novel Stacked Generalization Technique

MSc Research Project

Data Analytics

Rohan Vijay Dongare

Student ID: x18120199

School of Computing

National College of Ireland

Supervisor: Prof. Noel Cosgrave

National College of Ireland

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Rohan..Vijay…Dongare……

Student ID: …..x18120199…………..…….

Programme: ....MSc. Data Analytics……. **Year:** …2019…

Module: ……Research..Project………..

Supervisor: …….Prof. Noel Cosgrave…..

Submission Due Date: …….12th August, 2019……….

Project Title: …Star-Galaxy-Quasar Classification using a Novel Stacked Generalization Technique….

Word Count: ……………………………………… **Page Count**…………………21..…..……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ………

Date: ………

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Contents

List of Tables

List of Figures

Star-Galaxy-Quasar Classification Using a Novel Stacked Generalization Technique

Rohan Vijay Dongare

x18120199

Abstract

Since the discovery of dark energy which is responsible for the accelerated growth of the universe, the astronomical community is focused on uncovering its mystical nature. To understand the nature of this energy which is present in celestial bodies like Quasars, Stars and Galaxies and causes interactions between them, it is first crucial to classify them. Deep astronomical surveys like the Sloan digital Sky Survey (SDSS) are already in place that provide information of these bodies but require a pipeline that accurately classifies them. This study presents a Star-Galaxy-Quasar classification framework that uses a novel stacked generalization model with five heterogeneous base learners and a meta-learner using the photometric and spectroscopic data from the data release 15 of SDSS. The hyperparameters of the base learners as well as the meta-learner were tuned using Bayesian optimization. Feature selection for this study was carried out using Maximum Relevance Minimum Redundancy (MRMR) technique and the selected features were decomposed using Non-Negative Matrix Factorization. Performance of the stacked model was evaluated using the Purity, Completeness and Contamination measures. The state of art perofrmance for the Star-Galaxy-Quasar classification task was achieved by Bai *et al.* (2018) with an overall accuracy of 95% .This study demonstrates that the defined framework exceeds the state of art performance obtaining a 98% overall accuracy. This research finds that galaxies have a higher contamination rate as opposed to stars and quasars, as a high number of quasars were misclassified as galaxies. This was previously not observed in the studies conducted in this domain as majority of the quasars were unresolved and were always contaminated with stars. This indicates that since the latest SDSS data release which was used in this research a greater number of quasars now appear resolved. The future work in this domain should specifically address the galaxy-quasar classification problem which has been left unaddressed in this field.

Keywords- stars, galaxies, quasars, deep astronomical surveys, stacked generalization, Bayesian optimization, maximum relevance minimum redundancy, non-negative matrix factorization

1 Introduction

When astronomical surveys first observed a new class Nebulae(interstellar dust clouds made up of plasma, helium and hydrogen) which could not fit into the conventional classes a lot changed in astronomy ever since(Machado *et al.*, 2016). This class was later identified as an extended extra-galactic source linked to galaxies. And this distinction until recently between extended and point sources was enough but the modern astronomical surveys keep pushing the boundaries of the observable universe. Surveys like the Sloan Digital Sky Survey (SDSS)[1] cover 31,637 square degrees of the sky. Research conducted by Ramió *et al.*(2019) , Bai *et al.*(2018) Machado *et al.*(2016) and Krakowski *et al.*(2016) show that distinction between celestial bodies has become fuzzier as the ground-based telescopes cannot spatially resolve objects such as Quasars and as a result they are always confused with other point sources like stars.

Quasars are active galactic nuclei first discovered in 1960 and have characteristic high luminosities(Tadhunter, 2008). Quasars can explain the early stages of the universe which is the reason why their study has gained tremendous importance in Astronomy(Viquar *et al.*, 2019). For Astronomers to study the nature of dark energy released due to the interactions between Stars, Galaxies and Quasars it is first essential to have a clear distinction between them(Machado *et al.*, 2016).

Bai *et al.* (2018), Krakowski *et al.* (2016) and Vasconcellos *et al.*(2011) in their research find that using only the imaging catalogoues for classifying celestial bodies quickly reaches the limit as the telescopes lack resolution and the distant resolved sources such as galaxies appear as point sources which makes it difficult to distinguish from other point sources such as Stars and Quasars. Additional information such as spectroscopic data is required to

[1] https://www.sdss.org/dr15/scope/

effectively classify these bodies. Sevilla-Noarbe *et al.* (2018), Anjum *et al.*(2018), Jan-Torge Schindler *et al.* (2019) and Bai *et al.*(2018) find that using data from the infrared spectrum is very also valuable for the celestial body classification problems. Sevilla-Noarbe *et al.* (2018) also suggest the use of location data of the objects in their future work. This study in addition to both the photometric and spectroscopic data for the Star-Galaxy-Quasar classification will also use data from the infrared spectra, location data of the objects and the redshift of the celestial bodies.

Utilizing both the spectroscopic magnitudes and the photometric values the criteria for separation becomes very complicated to be represented with functions in a multi-dimensional parameter space(Bai *et al.*, 2018). This multi-dimensional parameter space has been effectively used to distinguish various celestial bodies using machine learning. (Fadely, Hogg and Willman, 2012; Solarz *et al.*, 2012; Kovács and Szapudi, 2015; Krakowski *et al.*, 2016; Machado *et al.*, 2016; Kim and Brunner, 2017; Ramió *et al.*, 2019) have used various machine learning algorithms for the Star-Galaxy classification problem. While (Elting *et al.*, 2008; Peng, Zhang and Zhao, 2013; Peters *et al.*, 2015; Bai *et al.*, 2018; Viquar *et al.*, 2019) have also used various machine learning techniques for the Quasar classification task. Machine learning algorithms without being explicitly programmed find natural patterns in the underlying data.

Most studies carry out only the Star-Galaxy classification and consider quasars as stars. This is because of two reasons, firstly until recently the number of confirmed Quasars were very low and secondly because stars and quasars are both point sources and it's very difficult to differentiate them. This research will use the latest data release of the Sloan Digital Sky Survey, Data release 15 (Aguado *et al.*, 2018) which was released in December, 2018. This data release contains data from extragalactic spectra as well as the infrared spectra and has over 500,000 spectroscopically confirmed Quasars. This data release has not been used in any of the studies yet.

In this research, a novel star-galaxy-quasar classification framework which combines and takes advantage of five diverse classification techniques to produce a more robust classification using stacked generalization is proposed. The base classifiers for the proposed stacked model are Quadratic Discriminant Analysis, Extra tree classifiers, XGBoost, Support vector machines with RBF kernel and Feed Forward Neural Networks with backpropagation. A separate Feed Forward Neural Network with backpropagation is used as a meta learner to combine the outputs of the base learners. The research in the entire celestial body classification field lacks a study that incorporates an optimization algorithm like Bayesian optimization for tuning the hyperparameters of the machine learning models. This study in addition to using Bayesian optimization for tuning hyperparameters will also implement Maximum Relevance Minimum Redundancy Feature selection technique which previously has not been used in this domain. The selected features will further be decomposed using non-Negative Matrix factorization. Kaiser-Meyer-Olkin (KMO) test and Bartlett's Test of Sphericity will be conducted on the features before carrying out feature decomposition. A novel stacking technique will be used in this study which combines stratified K-Fold cross validation with holdout sampling technique for feeding the outputs of the base learners to the meta-learner.

The research question which the study will answer is,

'Can the use of a stacking ensemble technique with multiple heterogeneous base learners and features selected by Maximum Relevance Minimum Redundancy, followed by decomposition using Non-Negative Matrix Factorization significantly improve upon the performance of the state of art for the Star-Galaxy-Quasar classification problem? '

The research objectives are,

- Compare the performance of base learners using the purity, completeness and contamination measures which are defined later in the study.
- Perform Kruskal Wallis H test on the outputs of all the base learners to check if their errors are uncorrelated. If a significant p-value is observed for the Kruskal Wallis H test, perform a post hoc paired comparison using Wilcoxon signed rank test.
- Compare the performance of the stacked model with the best performing base models using the purity, completeness and contamination evaluation metrics to check if a significant improvement was obtained.
- Perform Kruskal Wallis H test on the outputs of the stacked model and the base learners to test if their errors are uncorrelated. Perform a post hoc paired comparison between the output of every base learner and the stacked model using Wilcoxon signed rank test if a significant p-value is obtained for the Kruskal Wallis H test.

The hypothesis for this study is that the diverse base learners are specialist classifiers and have uncorrelated errors, which will enable the stacked model to perform significantly better than any one individual model.

The study is organized as follows: Section 2 gives an overview of the relevant work that has been carried out in the celestial body classification domain. Section 3 briefly explains the research methodology whereas section 4 discusses the design specification. Implementation is covered in section 5 and experimental evaluation is presented in section 6. Finally, the conclusion and future work are discussed in section 7.

2 Related Work

Source identification and classification problem is fundamental to astronomy, and various approaches and strategies have been implemented to tackle it. However, there is no consensus on which classification strategy is the most effective. While research related to Quasar identification using machine learning is relatively recent, work related to the Star-Galaxy separation problem has been carried out for two decades. But the Star-Galaxy-Quasar multiclass classification problem is new in astronomy and less research is conducted in this area as distant Stars, Quasars and Galaxies appear as point sources and hence it becomes extremely difficult to distinguish them. The literature review is divided into subsections based on the type of celestial bodies that were classified and the basis on which the classification was carried out (Morphological/Photometric/Spectroscopic).

2.1 Morphology based Star-Galaxy Classification

Star-Galaxy separation based on morphology is one of the most common approaches towards classification. Morphological classifications are based on different measurements of the shape and size of an object. Bayesian classifiers were used by Ramió *et al.* (2019) and Henrion *et al.*(2011) to classify stars and galaxies based on morphological values. Henrion *et al.* (2011) obtain 121,000 samples from the United Kingdom Infrared Telescope Deep Sky Survey (UKIRTDSS) with morphological measurements based on near infrared readings from the telescope. The star and galaxy completeness reported for objects classified with posterior probability greater than 90% was 70.7%.Research conducted by Ramió *et al.* (2019) is very similar and also uses a Bayesian classifier to classify stars and galaxies based on morphological measurements using 251,000 samples from the J-Plus astronomical survey. They achieve an overall completeness of 95% which is considerably higher than the work conducted by Henrion *et al.*(2011). The input features for both the studies were continuous and Bayesian classifiers require their input features to be categorical, so continuous features require binning. In contrast to the research by Henrion *et al.* (2011), Ramió *et al.*(2019) conducted a very rigorous study on the distribution of Star and Galaxies with respect to all input parameters before binning the data and could possibly be an explanation as to why their model performed better. They used a block fitness function to check how well a value in a bin represents the data in that bin. It is also important to note that study conducted Ramió *et al.* (2019) used bootstrapping. This could induce overfitting in the model due to data leakage and is not the best sampling technique especially when there is abundant data available. A drawback with using Bayesian classifiers is that they assume independence between the input features, although they perform decently even when the assumption is violated. But the major drawback is that they require binning of the continuous features which leads to data loss and improper binning strategy can vastly affect the model accuracy.

Random Forest and Decision trees were used by Vasconcellos *et al.* (2011) for morphological Star-Galaxy separation using the 7th SDSS data release. They obtained 884,000 samples with 12 features. The authors divide the data into three sets, first one consisting magnitudes of bright objects, second with comparatively less bright objects and third with magnitudes of faint objects. The hyperparameters for the algorithms were chosen using grid search which exhaustively searches a predefined space and may not always achieve the best results. The average completeness for Random forest over the three sets of data was reported to be 86.46% and 69.85% for decision trees.Decision trees tend to overfit the data and have high variance. As a result, they perform poorly on test datasets compared to Random forest which is an ensemble classifier that uses bootstrapping to sample the data during training. Both Decision Trees and Random Forests were reported to have a very low accuracy of 42.29% and 70.48% respectively in the faint magnitude regions.

Star-Galaxy classifications based on morphology have an assumption that galaxies appear as resolved sources and stars as point sources(Henrion *et al.*, 2011). However at fainter magnitudes galaxies become unresolved and appear as point sources and are hence confused with stars(Krakowski *et al.*, 2016). This explains why the classifiers perform poorly in the faint magnitude regions.

2.2 Photometric Star-Galaxy Classification

Research in astronomical classification has recently seen a shift towards photometric classification from morphological classification, as it does not have an assumption of galaxies appearing as resolved sources and stars as point sources. Photometric classification uses the color band data of the bodies rather than their morphological data (shape and size data of the object) for that reason the assumption does not apply to photometric classification. Deep neural networks were used by Cabayol *et al.* (2018), Kim and Brunner(2017) and Soumagnac *et al.*(2015) for classifying stars and galaxies using photometric data obtained from Physics of Accelerated Universe Survey (PAUS), Sloan Digital Sky Survey(SDSS) and Dark Energy Survey (DES) respectively. Cabayol *et al.*(2018) use Convolutional Neural Networks (CNN) for the classification task and compare their performance with

Feedforward Neural Networks and Random Forest. They use a training set of 20,000 samples with 15,000 galaxies and 5,000 stars and validation sample consisting of 5,000 galaxies and 1,000 stars. Neural Networks are generally prone to overfitting with small datasets and the choice of not including dropout or other regularization techniques in the architecture is worrying. Feed forward Neural Networks, CNN's and Random Forests come with many hyper-parameters and the authors do not mention the use of any hyper-parameter optimization techniques. They also do not discuss how the parameters were chosen for their study. The authors tackle class imbalance by using completeness and purity along with ROC-AUC curves as a performance measure. Their research finds that CNN's outperform Random Forest and Neural Networks, achieving galaxy purity of 98% compared to 91.3% and 95% of Random Forest and Neural Networks respectively. Kim and Brunner (2017) also use CNN's but achieve galaxy purity of 99.71 which is marginally higher than the study conducted by Cabayol *et al.* (2018). Kim and Brunner (2017) manually search 200 combinations of hyperparameters on the training data to find the best architecture which could possibly explain better results. Although it is a primitive hyper-parameter tuning technique which on very rare occasions will come close to finding the most optimal results, it is still better than using the default parameters. They use training data for hyper-parameter optimization which leads to information leakage and this could have been avoided by conducting it on a data partition of its own. They also use weights for their architecture from VGG16, which is a pretrained Neural Network architecture. VGG16 like other pretrained networks GoogleNet, ResNet, AlexNet etc. are trained on random images which belong to 1000 different classes. For that reason, using them for a Star-Galaxy classification task is not the most appropriate use of the weights.

Feature decomposition techniques can vastly affect the performance of the star-galaxy classification task and is proven by the research conducted by Soumagnac *et al.* (2015). They find that using principle component analysis before classification increases the purity of stars by 20% and of galaxies by 12%. They decompose the five band photometric data into three bands based on Fisher discriminant analysis and classify the objects using Feedforward Neural Networks. They achieve galaxy purity of 97.8% which is 2.8% higher than the work conducted by Cabayol *et al.*(2018) using Feedforward neural networks with a similar architecture. However like Cabayol *et al.*(2018) they too do not discuss hyperparameter optimization in their studies which makes it difficult to replicate their results. Tuning the hyperparameters can extensively change the results and is a vital step in modelling machine learning algorithms.

Sevilla-Noarbe *et al.* (2018) compare the performance of SVM's, Neural Networks and Adaboost classifier for the Star-Galaxy separation problem using photometric data obtained from the Dark Energy Survey. They use cross validated grid-search on the training data for selecting the optimal hyperparameters which is not a good practice as it leads to information leakage. The authors present their results in area under the ROC curves. They find that Adaboost perform the best with 96.7% area under the curve compared to SVM's 96.2%. Neural Networks perform the worst with 88.5% area under the curve and could be a result of poor tuning of the hyperparameters and the lack of regularization component during training to avoid overfitting of the model. Sevilla-Noarbe *et al.* (2018) in their research find that use the infrared data was valuable for their research and suggest the use of location data of the objects in their future work. Machado *et al.*(2016) in their study compare the performance of Neural Networks, Naïve Bayes, kNN, SVM and Random Forests using the photometric data obtained from the COSMOS survey. The authors do not justify the use of kNN for classification which has a low predictive power and requires high computational power especially for large datasets. kNN are lazy learners which memorize the entire data and skip the abstraction process altogether and consequently are a bad choice for this classification problem. The performance of Bayesian classifiers as observed from the studies conducted by Ramió *et al.*(2019) and Henrion *et al.*(2011) is immensely dependent on the data bins. And all the independent features in this classification problem would require binning since photometric data is continuous which would lead to data loss. Machado *et al.* (2016) use a polynomial kernel for SVM's for this problem which is not the most appropriate choice as the underlying data does not follow a polynomial distribution. The authors use cross-validated grid search for tuning the hyperparameters which has been used in majority of the studies in this area. In their study Neural Networks outperform the aforementioned classifiers with 98.4 % area under the curve. Naïve Bayes and kNN as expected performed the worst compared to other classifiers. In their future work they suggest the use of feature selection and feature decomposition techniques.

Fadely, Hogg and Willman (2012) in their study used SVM's with a gaussian kernel on the photometric data obtained from the COSMOS survey for the Star-Galaxy separation problem. They compare the performance of SVM's with a template fitting approach and find that SVM's perform better and achieve 96% area under the curve compared to 90% of the template fitting method. SVM's with gaussian kernels were also used by Khramtsov and Akhmetov (2018) for their celestial body classification problem using the SDSS photometric data. They however tackle a simpler problem of classifying galactic and extra-galactic objects and tune the model using grid search cross validation. They achieve extra-galactic object completeness of 99.8% and galactic object completeness of 99.2%. The gaussian kernel's ability to map the non-linear original input space to a higher dimensional feature space where the classes become linearly separable is the reason why they perform well for the celestial body classification task.

A comparison of four ensemble models, Weighted Average, Bucket of Models (BoM), Stacking and Bayesian Model combination(BMC) for classifying stars and galaxies was carried out by Kim, Brunner and Carrasco Kind (2015) with three base models. They use random forest, self-organizing maps and template fitting as their base models. They consider two cases for classification one where objects are spectroscopically confirmed hence available less in number and the other where objects are spectroscopically not confirmed and available in large numbers. They choose 66,388 objects (8545 stars and 57,843 galaxies) for the first scenario when abundant data is available and observe that Random Forest outperform other models including ensembles with 98.7% area under the curve. The authors in their study do not address the issue of class imbalance between stars and galaxies which potentially made the models biased towards the galaxy class. In the second scenario where they use only 1365 objects which were spectroscopically confirmed, BMC outperforms the rest of the models with 97.3% area under the curve. The distribution of classes in this training set where the objects are spectroscopically confirmed was not revealed. 1365 objects are very low for training machine learning algorithms and it becomes very difficult to avoid over-fitting of the data. The authors do not mention the use any optimization technique for tuning the hyperparameters of the base models. Since the latest data release the number of spectroscopically confirmed objects has dramatically increased, with effective hyperparameter optimization of the base learners along with feature selection and decomposition techniques ensemble classifiers have immense potential to achieve state of the art accuracies.

Adaboost, Gradient Boosted trees, Random Forest and Extra Trees were compared in a study conducted by Morice-atkinson, Hoyle and Bacon(2017) using the SDSS photometric data from their 12th data release. They work on the classifying extended sources(galaxies) from the point sources (stars and quasars combined). Like other works previously conducted they use grid search cross validation for tuning the hyperparameters. They use Mutual Information based Transductive Feature Selection (MINT) for selecting the best features. Their research shows that Gradient boosted machines and Extra trees perform marginally better than the other tree-based models when all the features are selected, and both the models are reported to have an accuracy of 98.1%. The accuracy of all the classifiers drops to 97% after using only the feature selected by MINT. They also test only the random forest classifier for Star-Galaxy-Quasar classification and achieve an accuracy of 89.6%, which is significantly lower than the previously obtained Star-Galaxy classification accuracies and was reported to be due to the inherent similarities between Stars and Quasars.

2.3 Star-Quasar classification and Quasar Selection

Star-Quasar classification problem is relatively new in the astronomical community and less amount of research has been conducted in this field. One of the earliest studies on the Star-Quasar classification task was conducted by Gao, Zhang and Zhao(2008) using the 5th data release of SDSS. They compare the performance of kdimensional(kd)-trees which is a distance-based tree classification method with SVM. After manually tuning the hyperparameters they find that SVM's perform better than the kd-trees. Peng, Zhang and Zhao(2013) follow up on their research using the photometric data obtained from the 7th SDSS data release. They find SVM's are reliable only when the objects are resolvable however fail to perform well for unresolvable/faint magnitude sources. For that reason, they use a combined SVM-kNN classifier for classifying stars and Quasars, where the objects are first classified by the SVM classifier (RBF kernel) and kNN is applied to the classified samples to correct any errors made by SVM. They find that this approach is more generalizable than using just one of the two classifiers. They optimize the hyperparameters of SVM using grid search while they choose k=9 for kNN without providing any justification. Their study attains an accuracy of 97.9% however their data is imbalanced, and they do not take any measures such as weighted classification, under-sampling the majority class or oversampling the minority class to address it.

Viquar *et al.*(2019) build up on the research conducted by Peng, Zhang and Zhao (2013) however argue that the use of SVM with RBF kernel by Peng, Zhang and Zhao (2013) was not justified. They perform a two-dimensional visual analysis of the star-Quasar feature space and find that the classes are mostly linearly separable. The authors use the same SVM-kNN method to classify stars and quasars but use SVM's with a linear kernel instead of RBF kernel. They also address the data imbalance problem which was unaddressed in the study conducted by Peng, Zhang and Zhao (2013), by under-sampling the Star class. An accuracy of 97.8% for the classification task was achieved in their study. They use Adaboost to compare the results of the SVM-kNN method and achieve an accuracy of 96.5%. The authors do not mention the use of any hyperparameter tuning technique and do not disclose the hyperparameters used in their study for both the models.

The drawback of most boosting algorithms like Adaboost and Gradient Boosted Machines is that they lack a regularization term which can make them perform exceptionally well for training sets but struggle to perform on the test dataset.

XGBoost is boosting algorithm that overcomes the drawbacks these algorithms as it comes with a regularization parameter. XGBoost was recently used for the first time in the celestial body classification task by Nakoneczny *et al.*(2019). They compare Random Forest, XGBoost and Neural Networks for the Star-Quasar classification task using the photometric data from the kilo-degree survey (KiDS). They use an imbalanced dataset with 12,144 stars, 7,061 quasars and 32,542 galaxies and use holdout method for sampling the dataset. They perform feature selection using backward elimination and tune the hyperparameters of the models using grid search. Both the tasks were performed on the training set which leads to information leakage and could have been avoided by using separate samples for the tasks. Random Forests, XGBoost and Neural Networks achieve an accuracy of 96.5%, 96.44% and 96.28% respectively. The authors speculate that using near-infrared readings will improve Quasar detection and are not able to use it as is not provided by the Kilo-degree survey but is provided by the SDSS survey and will be used in this study. The study conducted by Jan-Torge Schindler *et al.* (2019) proves that using the infrared readings does benefit the quasar detection rate. They use photometric data along with near-infrared data obtained from 3∏ survey of Panaromic survey Telescope, 2 Micron All Sky Survey and SDSS data release 13 for the Star-Quasar classification and focus on extremely luminous Quasars (high quasar magniudes). They use Random Forest for the classification task and tune the hyperparameters using grid search. Their results show F1 score of 0.99 for extremely luminous objects using both photometric and infrared data while using only the photometric data yields F1 score of 0.92.

2.4 Photometric and Spectroscopic Star-Galaxy-Quasar Classification

In the study conducted by Vasconcellos *et al.*(2011) they find that that the limit for separating stars from galaxies is quickly reached using only the photometric data for classification. Krakowski *et al.* (2016) build up on the research conducted by Vasconcellos *et al.*(2011) and use photometric and spectroscopic data for both classifying stars, quasars and galaxies using support vector machines. They use a gaussian kernel for classification due to the non-linearity of the feature space with respect to the classes. Like the previous studies conducted, they use grid search cross-validation to tune the hyperparameters of SVM. They create five bins of data based on color magnitudes and train five different SVM models, one for each bin. This severely impacted the size of training sets and they were left with 1000 samples of each class for the first bin, 4000 for the second and third bin, 5000 and 2600 for the fourth and fifth bin respectively. During the testing phase outputs of the five models were merged to produce one final output. The authors claim that there were no inconsistencies for any of the classifiers, which means SVM's with RBF kernel performed the same for both the faint and bright magnitudes. The size of the test dataset and the hyperparameters used for this experiment were not mentioned in the study, both of which can greatly influence the results. The weighted mean accuracy of the classifiers was reported to be 97.3% for the star-galaxy-quasar classification. In their study to test the combined accuracy of all the classifiers they randomly choose 5000 galaxies of different magnitudes independent of the training dataset. They were unable to perform this test for stars and quasars due to the scarcity of data. They achieve mean galaxy classification accuracy of 92.5% which compared to previous studies is very low. Their approach of having specialist classifiers based on different magnitude bins reduced the generalizability of the combined model which resulted in high bias. In a real-world scenario, the objects detected by a telescope won't always belong to a particular magnitude bin, objects detected will belong to varying magnitudes. So, training a classifier on the entire magnitude range would make it more generalizable.

Template fitting method was used by Anjum *et al.*(2018) for the Star-Galaxy-Quasar classification task as it does not require spectroscopic data for classification. It is a simple yet effective classification technique, where an existing object template is compared with the unknown source and whichever template the unknown source resembles the most the object is classified into the class of the template. The authors in their study use an imbalanced dataset with 37,492 quasars, 3,374 galaxies and 23,332 stars and achieve a combined accuracy of 78.6% (89% for Quasars, 63% for Galaxies and 84% for Stars). The authors attain a higher-class accuracy for Quasars due to the imbalance in data and choice of not using completeness and purity as performance metrics despite the imbalance is concerning. They find that the star-galaxy-quasar separation problem using only the photometric color space is difficult due to the overlap of classes. A drawback to using template fitting methods is that they are dependent on the quality and availability of templates (Morice-atkinson, Hoyle and Bacon, 2017).

Bai *et al.*, (2018) build up on the research conducted by Krakowski *et al.*(2016), they use both the photometric and spectroscopic data but do not train different classifiers based on magnitudes but instead use an ensemble machine learning technique, Random Forest. They achieve test accuracy of 95% for the model trained on uniform sample with equal objects from all classes and is the highest achieved accuracy for this classification task. Random

Forest have many hyper-parameters although the authors do not mention which values were used and how they were tuned which makes it difficult to replicate their study. Both Krakowski *et al.*(2016) and Bai *et al.*(2018) find that in addition to the spectroscopic data adding infrared spectrum data increases the accuracy of the classifiers. There are many inconsistencies not just in the Star-Galaxy-Quasar but in the entire celestial body classification domain.

Most studies as observed from the literature review do not perform a rigorous search of the hyperparameter space for finding optimal hyperparameters and mostly settle for grid search for the task. Researches in this domain until now have not even considered using meta-heuristic techniques for parameter tuning. Very few studies have employed feature selection and feature decomposition techniques, and the ones who have do not comment about other aspects of their study like tuning of hyperparameters. Studies rely only on performance metrics and do not perform statistical tests to ensure if the results they obtained were not a result of pure chance. There lacks a study in this field that employs optimization of the hyper-parameters, feature selection and feature decomposition together as all of them focus only on one or the other aspect of machine learning. However, this study aims to fill that void by using state of the art machine learning model setup, hyper-parameter optimization technique, feature selection and feature decomposition technique that previously never been used in any of the studies conducted in this domain.

3 Research Methodology

The Cross-Industry Process for Data Mining (CRISP-DM) will be followed in this research which divides the methodology into six sections namely, Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment.

3.1 Business Understanding

The business problem with the SDSS data is well defined in terms of predictive analytics. The pipeline which many surveys like the SDSS follow begins with capturing data from their telescopes and spectrographs, processing it and storing it in a centrally accessible database. In the processing part, the pipeline has rule based systems that classify the objects captured by the instruments into broad categories of Galaxies and Stars. The systems however fail to accurately perform more fine-grained classification like between Quasar, Stars and Galaxies as all three of them appear as point sources due to the vast distance between them and the Earth. This research aims to build a system that would accurately classify Stars, Quasars and Galaxies with accuracies better than the state of art which can then be used by any deep astronomical sky survey.

3.2 Data Understanding

3.2.1 Photometric Data

The raw images of the distinct objects in the night sky due to their vast distances appear as point objects and it makes it very difficult to classify them. For that reason, images by the SDSS camera are captured in five distinct photometric bands: infrared (z), near infrared (i), red and yellow (r), green and blue (g) and ultraviolet (u). The SDSS pipeline refers to the measure of brightness of each of these filters as magnitude. However, this study will consider the flux values which standardizes the magnitude depending on how far the object is from the Earth.

Shapiro-Wilk Test with Bonferroni Adjustment: For the Shapiro Wilk test of normality, the significance level (α) will be set to 0.05 for this research. This would allow a maximum type I error of five percent. This test will be conducted for all the photometric, spectroscopic and location data of the objects. A total of 11 tests will be conducted on the same sample of data. This will increase the family-wise error rate to $(1\text{-}(1\text{-}0.05)^{11})$ which equals to 43%. So, the chance of committing a type I error at least once for the entire family is 43%. To avoid this a Bonferroni adjustment will be applied to the p-value. The Bonferroni adjustment will be calculated by dividing the set p-value with the number of tests to be conducted. The new p-value after the Bonferroni adjustment is 0. 0045.The Shapiro-Wilk normality test reveals that all the five bands do not belong to a normally distributed population as we obtain a significant p-value of $2.2e^{-16}$ and reject the null hypothesis.

Observations from the Cullen and Frey Graphs: All the photometric bands have a similar Cullen and Frey graph and they indicate that these bands have a beta distribution with a skewness of -0.28 and kurtosis of 2.69. For a kurtosis value of zero, a normal distribution is assumed. A value of 2.63 indicates that the distributions are leptokurtic and hence have heavy tails. This indicates that there are outliers present in the data. A more detailed

study on the outliers and their treatment will be carried out in the next section. A near zero negative skewness value indicates that the distribution is slightly left skewed.
All the Cullen and Frey graphs for individual color bands have been attached to the appendix.

Linear Separability: The density plots for each class with respect to the photometric data show an overlap among the three classes which makes the Star-Galaxy-Quasar problem non-linearly separable. A color-color 3D plot was also observed to ensure the non-linearity exists even after adding an extra dimension. All the 2D density plots have been attached to the appendix.

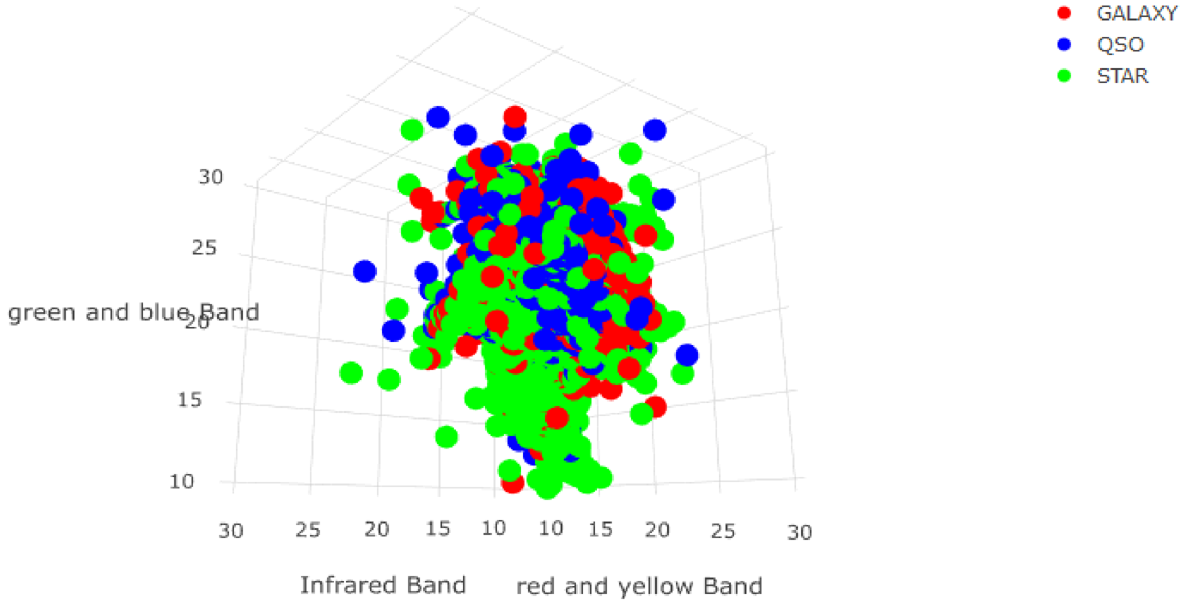


Figure 1: Three-dimensional plot for photometric observations

3.2.2 Spectroscopic Data

The SDSS spectroscopic data of an object is given by three values Plate, MJD and FibreID. The SDSS spectrograph identifies an object spectrum using these three values. Since the 9th data release of SDSS, spectroscopic data was collected from the Baryon Oscillation Spectroscopic Survey (BOSS) which in addition to the spectroscopic data also measures the redshift of the celestial objects.

Shapiro-Wilk Test: The null hypothesis of the Shapiro Wilk test of normality was rejected after a significant p-value (with Bonferroni adjustment) of nearly zero was obtained for the spectroscopic data. This indicates that the spectroscopic data does not follow a normal distribution.

Observations from the Cullen And Frey Graph: The Cullen and Frey graph shows that the spectroscopic measurements ‘fiberid’ follows a beta distribution with a skewness of -0.43 and a kurtosis of 1.84. This implies that the distribution is moderately left skewed and has heavy tails. Another feature ‘mjd’ also follows a beta distribution with a skewness of -0.43 and kurtosis of 1.84. The heavy tails for both ‘plate’ and ‘mjd’ imply presence of outliers which would require treatment. The measurement ‘plate’ however follows a uniform distribution with a skewness of 0.07 and kurtosis of 1.07 which shows that the distribution is almost symmetrical and has heavy tails.

Linear Separability: The density plots of the spectroscopic data reveal that there is an overlap between the classes in the spectroscopic data. A three-dimensional plot also shows that there is a high overlap between the Star, Galaxy and Quasar class.

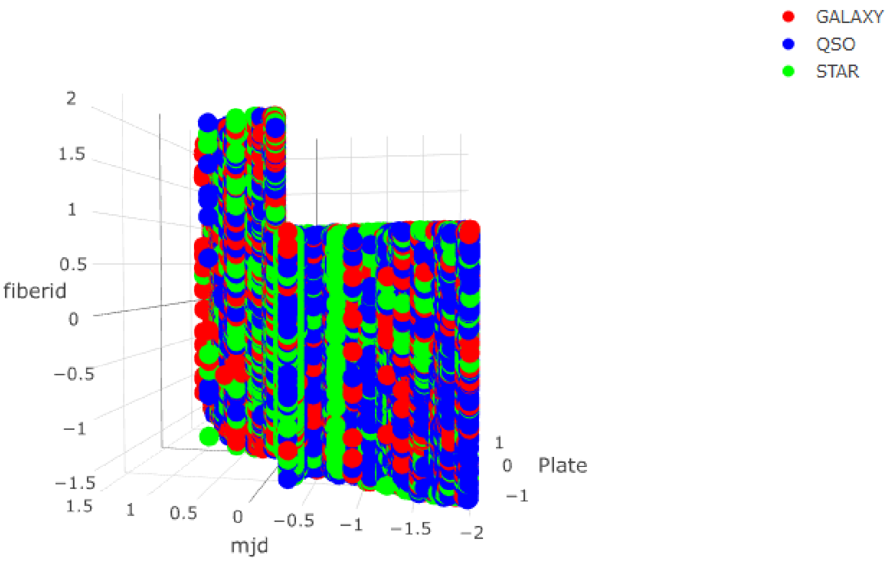


Figure 2: Three-dimensional plot for Spectroscopic Observations

3.2.3 Location Data

The location of the objects is given by two features, Right Ascension (RA) and Declination (DEC) in astronomy. These measurements are taken with respect to a hypothetical celestial sphere which has an infinite radius and the center of this sphere coincides with center of the Earth. Right Ascension corresponds to longitude and Declination corresponds to latitude and they together are an ordered pair that can be plotted on a graph[2].

Shapiro-Wilk Test: After conducting the Shapiro-Wilk test for the location data, the null hypothesis was rejected as a significant p-value (with Bonferroni adjustment) was obtained.

Observations from the Cullen and Frey Graph: The Cullen and Frey graph for RA shows that it has a logistic distribution with a kurtosis of 4 and a skewness of -0.10 which indicates presence of heavy tails and that the data is slightly left skewed. While for DEC the graphs show that it has a uniform distribution with a kurtosis of 1.5 and skewness of -0.13 indicating a light tail at the left end of the distribution.

Linear Separability: Similar to the photometric and spectroscopic values, the two-dimensional density plot of the location data indicates that there is a considerable overlap between the three classes. It was not possible to plot the location data in three dimensions as it only has two features.

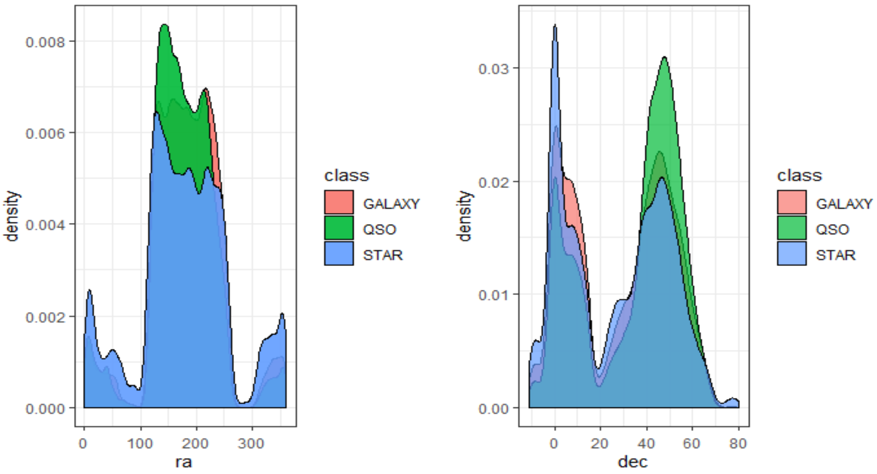


Figure 3: Density plot for the location data

A detailed data report for all the features is attached in the appendix.

[2] https://solarsystem.nasa.gov/basics/chapter2-2/

3.3 Data Preparation

During the initial data exploration phase, the Cullen and Frey graphs revealed that the photometric data was leptokurtic, indicating that the distributions had heavy tails which implied the presence of outliers. Upon further inspection of the data using boxplots it was observed that certain photometric observations had the value'-9999'. It is not possible for photometric values to be negative, hence these values were discarded. Previous researches conducted by (Vasconcellos *et al.*, 2011; Peng, Zhang and Zhao, 2013) also reported coming across these non-physical values and were discarded from their study as well.

The data obtained had no missing values present in it. Data normalization is an important task in this study as photometric data, location data, spectroscopic data and redshift data have different scales. And using the data without normalization will make the classifiers biased towards features with higher range of values. A Yeo-Johnson transformer will be used in the study to stabilize the variance in the input features and reduce the skewness in data, transforming data into a near gaussian distribution. Yeo-Johnson transformer is an extension of the Box-Cox transformer, unlike the latter Yeo-Johnson transformer works for both positive and negative data.

As number of Quasars are low compared to Stars and Galaxies there exists a problem of data imbalance in the data. In this study the majority classes which are stars and galaxies will be under sampled. The minority class was not oversampled using oversampling techniques such as synthetic minority oversampling as they induce noise in data. The machine learning models when trained in the presence of such noise can lose their generalizability. The drawback to under-sampling data is that it leads to loss of data but in applications of astronomy there is abundant data available even after randomly under-sampling it.

3.4 Modelling

3.4.1 Stacking

Stacking generalization or stacking is an ensemble technique that combines the outputs of multiple heterogeneous machine learning (base) models using a machine learning (meta) model. The base models in stacking unlike bagging or boosting are diverse and not the same model. It is advantageous to have diverse classifiers as it was observed from the literature review that different machine learning models specialize in different areas of astronomical classification. For example, Bai *et al.* (2018) find that Random Forests perform really well in terms of star and galaxy purity but struggle with quasar purity while Krakowski *et al.* (2016) find that SVM's with RBF kernel perform well for quasar purity but struggle with galaxy purity. An alternative to stacking with diverse classifiers is a voting ensemble. However, if suppose there are four base models and three of them are grossly incorrect, the voting ensemble will fail. Using a weighted voting ensemble solves the issue. Stacking is still a superior ensemble technique as it eliminates the voting procedure and instead has an entity called the meta-learner which is a machine learning model that learns which classifiers are more reliable. In addition, it also learns how the output of the base learners can be combined in the best possible way. In this study a simple perceptron will be used as meta-model (level-1-model) and the input to the meta-model will be the predictions of five distinct base models (level-0-models). The base models used in this study will be Regularized Quadratic Discriminant Analysis, Extremely Randomized trees, Extreme Gradient Boosted Machines, Support Vector Machines with Radial Basis Function (RBF) Kernel and Neural Networks.

Stratified K-Fold cross validation will be used to train and evaluate the base learners and the subsequent predictions will be used to train the meta learner. Using only stratified k-fold cross validation can lead to data leakage so this study will also incorporate an additional hold out set in the stack.

3.4.2 Base Learners

3.4.2.1 Regularized Quadratic Discriminant Analysis

Quadratic Discriminant Analysis (QDA) was chosen as one of the base learners as it accurately defines decision boundaries for data that are non-linearly separable (James et al., 2002). QDA is an extension to Linear Discriminant Analysis (LDA) and unlike LDA it does not assume all classes to have a common covariance matrix. QDA assumes data to be normally distributed which was initially not true in this study for the independent variables although in the data pre-processing stage normalization transformation was carried using the Yeo-Johnson transformer out to ensure that the independent variables are normally distributed. If suppose an observation is randomly chosen from n^th^ class, QDA represents it in the form $X \sim N\ (\mu_{n,} \sum_{n})$ where μ_n is the mean vector and $\sum_n$ is the covariance matrix of the class. QDA predicts class label for an observation based on $\partial_n(x)$ which is computed by equation (1).

$$\partial_n(x) = -\frac{1}{2} x^T \textstyle\sum_n^{-1} x + \ x^T \textstyle\sum_n^{-1} \mu_n - \mu_n^T \textstyle\sum_n^{-1} \mu_n - \frac{1}{2} \log | \textstyle\sum_n | + log\pi_n \qquad (1)$$

where π_n denotes the prior probability of class n. Whichever class has the highest ∂(x) value, the observation is classified into that class. Tests conducted by James et al. (2002) show that QDA performs better than other classifiers such as kNN, Logistic Regression and LDA and have the least test errors when data is non-linearly separable as it forms a quadratic decision boundary. And was hence chosen over these classifiers as a base learner. While LDA can suffer from high bias as it has a linear classification function, QDA has a quadratic classification function and can suffer from high variance. Regularized Quadratic Discriminant Analysis has a regularization component λ that can be tuned and can vary from zero to one, which ensures that QDA does not overfit the training data. The regularization parameter for the covariance matrices is computed using equation (2).

$$\sum_n(\lambda) = (1-\lambda)\sum_n + \lambda\sum_n \tag{2}$$

3.4.2.2 Extremely Randomized Tree Ensemble

Extremely Randomized Forests have never been used in the Star-Galaxy, Star-Quasar or Star-Galaxy-Quasar classification although Random Forests have been used before. (Kim, Brunner and Kind, 2015; Machado *et al.*, 2016; Morice-atkinson, Hoyle and Bacon, 2017; Bai *et al.*, 2018; Schindler *et al.*, 2019) have used Random Forests for classifying SDSS objects. Vasconcellos *et al.*(2011) have used decision trees for classifying Stars from Galaxies, while Zhao and Zhang(2008) used them for classifying active galactic objects from non-active galactic objects. A major drawback with simply using decision trees is that they overfit the underlying data and have extremely high variance and hence perform poorly on an unseen(test) datasets. Bagged trees are an improvement over the decision trees but usually have co-related trees which again lead to high variance. An improvement over these bagged trees are Random Forests which decorrelate the decision trees by randomly selecting 'm' input features for each tree and 'm' is approximately equal to √total features. Extremely randomized trees in addition to selecting random input features for bagged samples, randomly split nodes instead of choosing the best split (Geurts, Ernst and Wehenkel, 2006). This decreases variance and at the same time does not increase the model bias. Extremely Randomized Trees generalize better than Random Forests and are computationally faster which will be beneficial for this study due to the enormous data size. Hence, Extra trees were chosen as a base learner over Random Forests and Decision trees for this classification task.

3.4.2.3 Extreme Gradient Boosted Machines

XGBoost was recently used for the first time in the celestial body classification task by Nakoneczny *et al.*(2019) for the Star-Galaxy-Quasar classification task which achieved an accuracy of 96.4%. The authors in their study manually tune the hyperparameters however in this study Bayesian Optimization will be used to tune the hyperparameters which can vastly affect the results. Initially Gradient boosted machines (GBM), a tree-based ensemble technique that build trees sequentially instead of parallelly like Extra trees or Random Forests were chosen as a base learner. They were chosen as in these sequential models each subsequent tree learns from the errors of the previous tree. It reduces the misclassifications of the previous tree by increasing the weights associated with the misclassified samples. A drawback to using GBM's and other boosting algorithms like Adaboost, is that they lack a regularization component which can lead to overfitting of the training data and as a result perform poorly on the test dataset. However, XGBoost which is an advanced version of the gradient boosted machines has a regularization component to avoid over-fitting, which makes them more generalizable than other boosting algorithms(Chen and Guestrin, 2016) .The predictions of gradient boosting machines for ith observation of bth boost can be denoted as,

$$\hat{y} = \sum_{n=1}^{B} f_b(x_i) \tag{3}$$

The loss function for gradient boosted machines is given by equation (4),

$$L_b = \sum_{i=1}^{n} l(y_i, \hat{y}_i) \tag{4}$$

The main idea is to minimize this loss function where $l(y_i, \hat{y}_i)$ denotes the difference between the real value and the predicted value. XGBoost on top of the GBM model has a regularization term $\Omega(f_b)$ which penalizes the model and hence avoids overfitting of the model. The regularization term has two components γT and 0.5λ||w||, where T is the number of leaves in each tree and γ is the minimum required reduction in loss before the model partitions further. XGBoost penalizes the model if the loss reduction is less than γ. ||w|| is the l2 norm or the Euclidean norm and λ is a fixed coefficient value. Loss function for XGBoost is given as,

$$L_b = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \gamma T + 0.5\lambda\, ||w|| \tag{5}$$

XGBoost due to their capability of having uncorrelated errors for each subsequent learner result in a good performance while the regularization term ensures that the model does not overfit the data.

3.4.2.4 Support Vector Machines with Radial Basis Function Kernel

Peng, Zhang and Zhao (2013) and Viquar *et al.* (2019) have used SVM's for the Star-Quasar classification task while Soumagnac *et al.* (2015) and Machado *et al.* (2016) have used them for the Star-Galaxy classification task. Krakowski *et al.*, (2016) used SVM's for the Star-Galaxy-Quasar task and Khramtsov and Akhmetov (2018) have used them for classifying extra galactic and galactic objects.

The Star-Galaxy-Quasar data is not linearly separable in the original feature space, but this may not be true for a higher dimensional space. SVM's with RBF kernel construct an optimal hyperplane by mapping the input features to a n-dimensional feature space making a non-linearly separable problem linearly separable and for that reason were chosen for this task.

SVM with a linear kernel would perform well for the classes are linearly separable and polynomial kernel assumes features to have a polynomial distribution and neither of that is true for this data and so were not chosen for this classification task.

3.4.2.5 Deep Neural Networks

Deep neural networks have been used for the Star-Galaxy classification problem by Cabayol *et al.* (2018), Hoyle *et al.*(2018), Kim and Brunner (2017) and Soumagnac *et al.* (2015). Nakoneczny *et al.*(2019) used deep neural networks for the Star-Quasar classification task. One common problem that all these studies faced is the dealing with the large number of hyper-parameters and designing a network architecture. A common approach taken by these studies was manually tuning the hyper-parameters which is an inefficient and tedious process. This study will however use Bayesian optimization for tuning the hyperparameters.

While a perceptron which is a fundamental unit of a deep neural network, it can only handle linearly separable data. Multi-level perceptron/deep neural networks can handle multiclass classification when the data is non-linearly separable and thus chosen for this task. Additionally, neural networks do not have any assumption about the underlying data and generalize well for unseen data. This study will use feedforward neural networks with backpropagation for the Star-Quasar-Galaxy classification problem.

Dropout Regularization: To reduce the overfitting of data a regularization technique called drop-out will be employed where a randomly chosen neuron's output is set to zero.

Early Stopping: Having a high number of epochs can also lead the neural network model to have high variance while using a smaller number of epochs can lead to high bias. To tackle this problem a method called early stopping will be used in this study where the model stops training when its performance on the validation set stops improving.

Model Checkpoint: Model checkpoint basically saves the best weights during the training phase of the network. These weights act as network snapshots so in case of a training failure these saved weights can be reloaded and can be used as a starting point. The pre-trained weights in addition to replicating this study can also be used as a starting point for further researches in this field instead of randomly initializing them.

Convolutional Neural Networks which were previously used by Cabayol *et al.* (2018) and Kim and Brunner (2017) are not appropriate for this task and will not be used for this study as the input data is not pixel data obtained from images, but instead the photometric data are flux values obtained from the SDSS camera. Long Short-Term Memory (LSTM), gated recurrent networks or any other recurrent neural networks were also not chosen as the data is not sequential.

3.4.3 Meta Learner

A simple perceptron will be used as a meta-learner for this study as it is an efficient algorithm to weight every base learner and to combine their outputs. A perceptron for a given input vector X=[$x_{1,}x_2 \ldots x_n$] which are the outputs of the base learners and a vector of weights w=[$w_{1,}w_2 \ldots.. w_n$] will have a function *f(x,y)* that maps every input instance to a particular class. The learning process occurs iteratively where the weights are updated when the target value does not match the predicted value, otherwise they are kept unchanged. Although a perceptron is represented as:

$$\hat{y} = argmax_{k \in \{1..k\}} f(x, y). w \qquad (6)$$

3.4.4 Feature Selection using Maximum Relevance Minimum Redundancy

Maximum Relevance Minimum Redundancy (MRMR) is a feature selection technique based on information theory. MRMR searches for a feature set 'S' with 'k' features which have maximum mutual information value between the feature set and the target variable while having low dependency among the selected features.

A feature set with maximum relevance between the input features and the target variable can be found by (7),

$$D(S,t) = \frac{1}{|S|}\sum_{x_i \in S} I(x_i,t) \quad (7)$$

Where $I(x_i,t)$ is the mutual information between individual input feature ‘x’ and the target variable ‘t’. But these selected features can be rich in redundancy meaning they can have a high dependency between themselves. In this case the class discriminative power of the feature set would not have significant impact even if one of the features were to be eliminated. The minimum redundancy condition can be expressed by equation (8),

$$R(x_i,x_j) = \frac{1}{||S||^2}\sum_{x_i,x_j \in S} I(x_i,x_j) \quad (8)$$

Where $I(x_i,x_j)$ is the mutual information between the input features.
The MRMR technique combines criteria (5) and (6) and is denoted by equation (9).

$$\max \emptyset(D,R), \emptyset = D - R \quad (9)$$

MRMR was chosen over other feature selection techniques like forward selection, backward elimination, model-based feature selection, as in addition to finding the features that are most relevant to the target variable the algorithm also eliminates relevant features that are redundant.

3.4.5 Feature Decomposition using Non-Negative Matrix Factorization

Non-Negative matrix factorization has never been used in any of the celestial body classification problems. Principle component analysis(PCA) however has been used by Hoyle *et al.*, (2018) and Soumagnac *et al.*, (2015) in the Star-Galaxy separation problem. Non-Negative matrix factorization was chosen over PCA and other decomposition techniques like Factor Analysis, Singular Value Decomposition or Latent Dirichlet Association as these techniques are suitable for data that is linearly separable which as observed in the data exploration phase is not the case for the Star-Galaxy-Quasar classification problem. The only assumption for NMF is that it requires the features to be non-negative which holds true for both the photometric and spectroscopic data. The objective of NMF for a given non-negative feature matrix $X = [x_1, x_2, \ldots, x_3] \in \mathbb{R}^{mxn}$ is two find two matrices $U = [u_{ij}] \in \mathbb{R}^{mxk}$ and $V = [V_{ij}] \in \mathbb{R}^{nxk}$ which minimizes the objective function:

$$O = ||X - UV||_F^2 \quad (10)$$

Where F denotes Fronebius norm for the matrix.

3.4.6 Hyperparameter Tuning using Bayesian Optimization

As observed from the literature review, the celestial body classification domain lacks a disciplined approach towards tuning of the hyperparameters. Most studies in this domain use grid search for tuning the hyperparameters of the machine learning algorithms. Grid search exhaustively searches for the best hyperparameters out of a pre-defined set. It is computationally inefficient and suffers from the curse of dimensionality. Random search is a slight improvement over grid search in terms of efficiency as it randomly selects hyper-parameters out of the pre-defined search space. Meta-Heuristic optimization techniques like Particle Swarm Optimization, Ant colony optimization and Simulated annealing are significantly better techniques than Grid or Random search to tune the hyperparameters. However, these meta-heuristic techniques are just an approximation of the optimization algorithms. These algorithms can get stuck in local minima of the error surface. Optimization algorithms like Bayesian Optimization are better at finding the global minima of the error surface without getting stuck in any local minima compared to the meta-heuristic techniques. This study will use Bayesian Optimization for tuning the hyperparameters of the base learners as well as the meta learner. Bayesian Optimization aims to find a value x’ from a domain $\boldsymbol{X}$ to minimize/maximize the value of a function $\boldsymbol{f}$ (it can be any function like gaussian processes or tree parzen estimator) over the values in $\boldsymbol{X}$. The optimization framework can be expressed as:

$$Find\ input\ x' \in X, for\ a\ function\ f: X \to \mathbb{R},$$

$$such\ that\ x' \in \arg min_{x' \in X}\{f(x)\}$$

3.5 Evaluation

The performance of the stacked model and the base learners will be evaluated using the measures defined by Soumagnac *et al.*(2015): Purity(p), completeness(c) and contamination(f) for quasar, star and galaxy samples. Theses quantities are used as a standard performance metric in the astronomical classification domain and have been used by (Krakowski *et al.*, 2016; Machado *et al.*, 2016; Bai *et al.*, 2018; Hoyle *et al.*, 2018; Nakoneczny *et al.*, 2019) in their studies. The equation to derive these quantities are as follows,

$$Completeness_{Quasars}(C_Q) = \frac{True_{Quasars}}{True_{Quasars}+False_{Quasar-Stars}+False_{Quasars-Galaxies}} \tag{11}$$

$$Contamination_{Quasars}(f_Q) = \frac{False_{Stars-Quasars}+False_{Galaxies-Quasars}}{True_{Quasrs}+False_{Stars-Quasars}+Falase_{Galaxies-Quasars}} \tag{12}$$

$$Purity_{Quasars}(P_Q) = 1 - f_Q = \frac{True_{Quasars}}{True\ _{Quasars}+False_{Stars-Quasars}+False_{Galaxies-Quasars}} \tag{13}$$

where $False_{Quasar-Star}$ and $False_{Quasar-Galaxy}$ are Quasars misclassified as Stars and Galaxies respectively. $True_{Quasars}$ are the number of Quasars correctly classified by the classifier and $False_{Stars-Quasars}$,$False_{Galaxies-Quasars}$ are Stars and Galaxies misclassified as Quasars. Evaluation metrics for the galaxy and star class can be found out in an analogous way. This study will also evaluate the classifiers on sensitivity, specificity and F1 score as these measures were predominantly used before Soumagnac *et al.*(2015) defined the existing measures. This will make this study comparable with previous works conducted in this domain.

3.6 Deployment

Once the telescopic cameras capture and process the flux values for all individual bands of an object, it needs to be integrated with the spectroscopic measurements of that object. After the data integration phase, the above defined pipeline from the data pre-processing stage to the final evaluation can be merged into the pipeline of any deep astronomical sky survey.

4 Design Specification

The data will firstly be split into four randomly sampled subsets for four different tasks to prevent information leakage, each with 500,000 observations. The first one will be used for feature selection, second for optimizing the base learners, third for optimizing the meta-learner and the fourth for training and testing the entire stacked model. This is done to avoid information leakage from any of these stages. After the feature selection phase using MRMR, KMO and Bartlett's test of sphericity will be conducted on the continuous features that were selected to test if they are suitable to carry out feature decomposition using Non-Negative Matrix Factorization. After the feature decomposition phase the hyperparameters of the base learners will be optimized using Bayesian optimization. The base learners with optimal hyper parameters will then be used to optimize the hyperparameters of the meta learner. After the meta learner optimization phase the entire stacked model will be evaluated.

Stacked Model Architecture: Using only the stratified k-fold approach for training and testing the base learners can lead to information leakage. This study will use a combination of stratified k-fold and hold-out sampling technique for stacking the predictions of the base learners. The data sample reserved for evaluation will first split into two samples, train and test. On the train sample, k-fold cross validation will be carried out. At the end of every k-fold iteration, along with the kth fold predictions for the test set that was partitioned during the initial data split will also be taken for every base learner. This process goes on for k-folds and an array stores prediction of the test set for every fold. At the end of the kth iteration, mode of a single row of the array is taken that contains all the predictions taken for a single row of the test set. Predictions on the test set are taken for k iterations so that any errors that arise due to random sampling can be eliminated. The k-fold predictions and the test predictions of all the base learners were then merged to two different arrays. The array with merged k-fold predictions was then used to train the meta learner and the array with merged test set values was used to validate the meta learner.
The pseudo code for the combining stratified k-fold and holdout sample is as follows:

```
Split the dataset into train and test set
for all base learners
        for i =1 to n stratified splits in the training set
          train the base learner on n-1 folds excluding the $i^{th}$ fold
          test the base learner on the $i^{th}$ fold
          test the base learner on the hold-out test set
      end for
conduct max voting for the array containing predictions of the test set
take transpose of the k-fold test set and the holdout set
end for
concatenate k-fold predictions of all the base learner to form the training set for the meta learner
concatenate test predictions of all the base learners to form test set for the meta learner
```

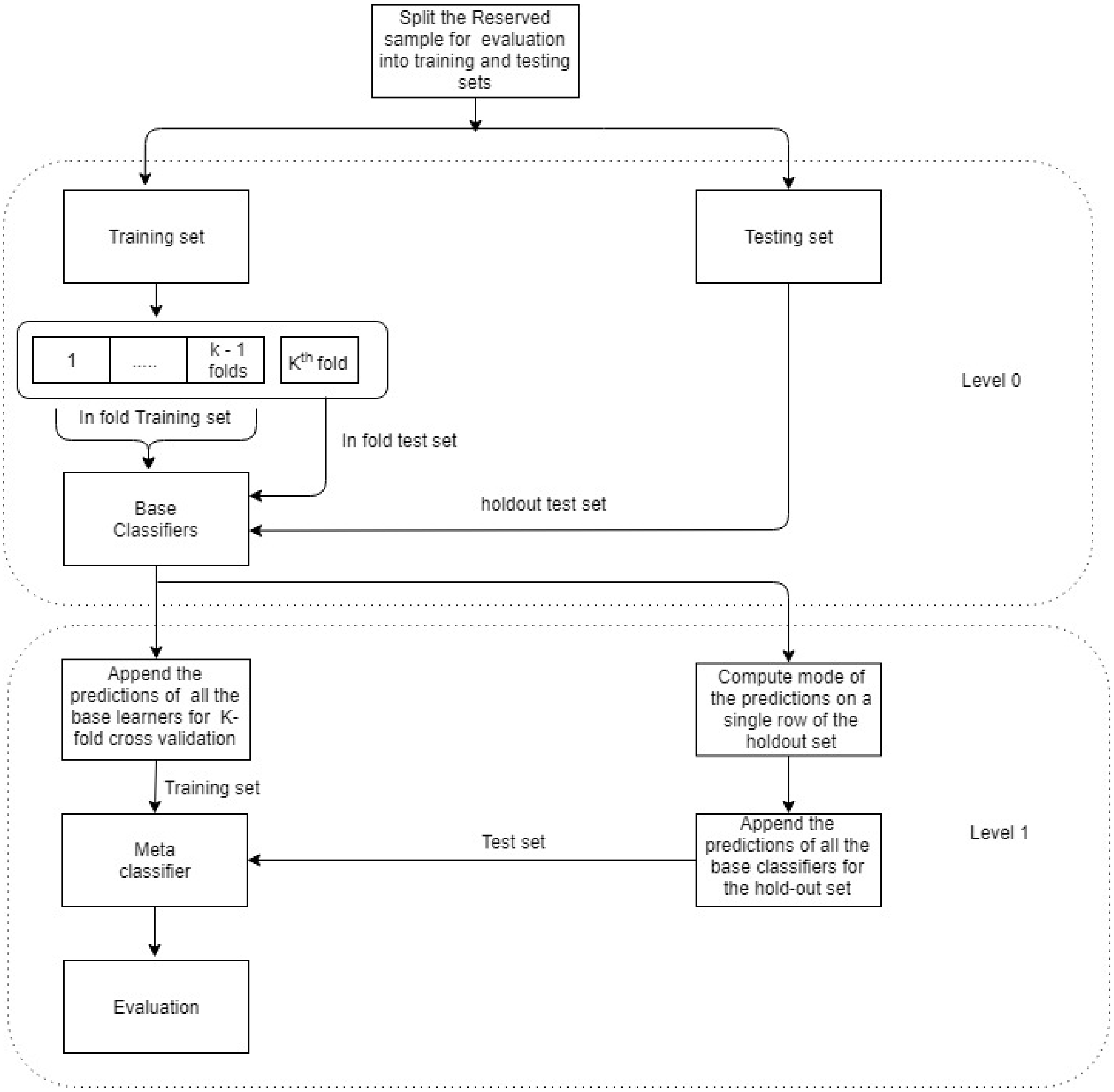

Split the Reserved sample for evaluation into training and testing sets
Training set
Testing set
1
.....
k - 1 folds
K^th fold
In fold Training set
In fold test set
Level 0
Base Classifiers
holdout test set
Append the predictions of all the base learners for K-fold cross validation
Compute mode of the predictions on a single row of the holdout set
Training set
Meta classifier
Test set
Append the predictions of all the base classifiers for the hold-out set
Level 1
Evaluation

Figure 4: Architecture Diagram for the Stacked Model Evaluation

5 Implementation

The entire implementation of the stacked model was carried out in python on a cloud-based workspace provided by Google called Google Colaboratory[3]. The data for this research was fetched from CasJobs server[4] which hosts an SQL based backend for storing data collected from the Sloan Digital Sky Survey. Photometric and spectroscopic data are stored in two different tables in the SDSS DR15 database and were retrieved using the query given below.

```
SELECT  2000000
r.objid, r.ra, r.dec, r.u, r.g, r.r, r.i, r.z, r.run, r.rerun, r.camcol, r.field,
d.specobjid, d.class, d.z as redshift, d.plate, d.mjd, d.fiberid into mydb.x18120199_research_project from
PhotoObj AS r
JOIN SpecObj AS d ON d.bestobjid = r.objid
```

[3] https://colab.research.google.com
[4] https://skyserver.sdss.org/casjobs/

5.1 Feature Selection using MRMR

The MRMR implementation in python requires the first column to be the dependent variable, there is no restriction on the data type of either the dependent or independent variable. The pyMRMR implementation has two optimization techniques for feature selection, Mutual Information Difference (MID) and Mutual Information Quotient (MIQ). The MRMR technique requires a time complexity of $O(N^{|S|})$ where N is the number of instances and |S| is the number of features. This implementation however is more efficient and is based on the study conducted by Ding and Peng (2005) who optimize the conditions for maximum relevance as

$$D(S,t) = \sum_{x_i \in S} I(x_i, t) \quad (14)$$

And minimum redundancy as

$$R(x_i, x_j) = \sum_{x_i, x_j \in D} I(x_i, x_j) \quad (15)$$

Where x_i, x_j are input features, $'t'$ is the target feature and $'S'$ is the feature set. This brings the time complexity down to $O(N.|S|)$. So instead of finding all the features with highest information gain and then separately finding features with maximum redundancy, this technique first finds all the features with maximum information gain and performs minimum redundancy test on these selected features. MID selects features based on the difference between D and R, so features are selected based on $\max \emptyset(D,R), \emptyset = D - R$ while MIQ based on $\max \emptyset(D,R), \emptyset = D/R$. The pyMRMR function returns an array with the most important features occurring first in it.

Table 1: Features selected by MRMR

MID	[redshift, ra, g, z, fiberid, i, u, mjd, r, plate, dec, run, camcol, field, rerun]
MIQ	[redshift, plate, z, u, g, fiberid, i, dec , r, ra, mjd, camcol, field, run, rerun]

Both the MID and MIQ techniques reveal that redshift, ultra-violet filter and green-blue filter are the most important features for the classification while the camera related data (camcol, rerun, run) are the least important features. A drawback of this implementation is that it does not display the information gain or redundancy values and just returns an array of features with the most important features. As an alternative the R implementation of mRMRe and mRMR was also tried but it assumes the dependent variable to be an ordered factor which is not the case for this study. Boruta package in R was then used to study the feature importance values in detail. Both Boruta and MRMR implantation in python find that in addition to the camera related features (camcol, run, rerun) the variable 'field' was also not very helpful for predicting the dependent variable. Camcol, field, run and rerun were thus eliminated from the study. The feature importance plot of Boruta is attached to the appendix.

5.2 Non-Negative Matrix Factorization (NMF)

After studying the pearsons correlation plot between the independent variables it was observed that there is high multicollinearity between the light spectrum bands (u, g, r, i, z) and this was also observed for two of the three spectroscopic observations (plate, mjd).

5.2.1 Bartlett's Test of Sphericity and KMO Test

To further confirm the findings a Bartletts test of spericity and KMO test were carried out once on an array with photometric bands and again on an array with only plate and mjd values. The significance level(alpha value) set for Bartlett's test was 0.05.

Table 2: Results obtained after KMO and Bartlett's test of sphericity

Features	Bartlett's Test of Sphericity		KMO Test
	p-value	Test Statistic	
Photometric data	0.0	5055490.04	0.80
Spectroscopic data	0.0	1440894.41	0.43

Significant p-values were observed for both the arrays which meant we could reject the null hypothesis and conclude that there exists correlation between these independent samples. KMO test tests to what extent this correlation exists and a value 0.8 was observed for the array with photometric bands which indicates that there exists a high correlation between the features. For the array with spectroscopic bands however a value of only 0.43 was observed which indicates that there is not enough correlation between the features, and it would not be appropriate to carry out feature decomposition on them.

5.2.2 Evaluating the number of factors

The number of factors unlike PCA in NMF cannot be calculated based on screeplot of eigen values. Selecting the appropriate number of ranks for NMF involves calculating the residual sum of squares (RSS) for the components and the explained variance. The number of factors with least RSS explaining most variance will be chosen for the study. The RSS and explained variance values were calculated using the Nimfa library in python.

Table 3: Residual Sum of Squares and Explained Variance of the NMF Factors

Number of factors	Residual Sum of Squares	Explained Variance
2	2482452.692	0.800
3	2145273.149	0.990
4	2811105.009	0.998

Although 4 factors explain the most variance, they also have a high residual sum of square values compared to 2 and 3 factors. When 3 factors were chosen, as they had the least residual sum of squares while the explained variance was just marginally lower than that of 4 factors. Therefore, the photometric bands were decomposed in three factors.

5.3 Bayesian Optimization

Bayesian Optimization was carried out using the hyperopt library in python. For conducting the hyper-parameter optimization hyperopt requires a search domain and an optimization algorithm to be specified for an objective function that minimizes loss. The optimization algorithm calculates the expected improvement for the parameters defined in the search space based on the previous iteration and hyperopt in addition to Gaussian Processes approach for optimization also offers Tree Parzen Estimator approach. Tree structured parzen approach is an improvement over the gaussian approach as it allows tuning of more than one hyper-parameter of a model in parallel in a single iteration which makes it more efficient. It was therefore chosen as the optimization algorithm for this study. Hyperopt requires separate ipython notebook to be saved in the local workspace which it calls iteratively during optimization. For neural networks, hyperas which is a keras wrapper for hyperopt was used while for all the other base learners the scikit learn wrapper hpsklearn was used. The optimal hyper-parameters returned by Bayesian optimization for the base learners are given in table 2.

Table 4: Hyper-parameters selected by Bayesian optimization for the base learners

Machine Learning Model	Parameters returned after Bayesian Optimization
Support Vector Machines with RBF kernel	Gamma: 0.4, Decision function shape: ‘ovr’, C: 7.64
Quadratic Discriminant Analysis	reg_param:0.005
XGBoost	min_child_weight: 89, max_depth:6, gamma: 4.117,learning rate=0.003,subsample=0.6, colsample_bylevel=0.6, colsample_bynode=1, colsample_bytree=0.7, reg_alpha= 0.169,n_estimators=5000,reg_lambda=3.06
Extratree classifier	N_estimators=29, max_features=0.95, criterion=entropy, min_samples_split=2, max_depth=50, min_samples_leaf=1
Feed Forward Neural Network (Base Learner)	Number of layers:4, layer1: Number of nodes:1024, Activation function: Relu, dropout rate:0.3, layer 2: Number of nodes: 512, Activation function: Relu, dropout rate: 0.2, layer 3: Number of nodes:512, activation function: Relu, dropout rate:0.2, layer 4: number of nodes:512, activation function: Relu, drop out rate :0.2, Optimizer: Adam, learning rate: 10**-3, batch size:300

A simple perceptron has no tunable parameters, although in the latter half of the study it was observed that using perceptron as a meta-learner did not make a significant improvement upon the performance of the base learners. For that reason, a decision was taken to use feed forward neural network as the meta-learner. The hyper-parameters for the meta-learner were tuned using the output of base learners. Hyperopt however only accepts values from a dedicated data function. Having a stacked architecture in a single function led to many complications with the base learner predictions. So, the outputs of the base learners were saved and uploaded to github, and a separate function referring to the github page was created that passed values for optimization. The hyperparameter values for the meta-learner are given in table 3.

Table 5: Hyper-parameters selected by Bayesian optimization for the meta learner

Feed Forward Neural Network (Meta-learner)	Number of layers:2, layer 1: number of nodes:512, activation:Relu, dropout:0.22,layer 2: number of nodes:256, activation:Relu, dropout:0.53,optimizer:adam,batch size: 300

Relu(Rectified Linear unit) was selected as the activation function in the case of Neural Networks for both the base as well as the meta learner by Bayesian Optimization. The draw back to using Relu is that it frequently can result in dead neurons as it sets zero value for negative instances. A more advanced implementation of Relu known as LeakyRelu mitigates this problem it has a small non-zero value for the negative values. For that reason, LeakyRelu will be used in this study instead of Relu. Currently hyperopt does not support inclusion of advanced activation functions while tuning the hyperparameters so it was not included in the hyperparameter search space.

6 Evaluation

The base learners and the entire stacked model are evaluated on the completeness, purity and contamination measures which are defined in section 3.5.

6.1 Case Study 1: Evaluating Performance of the Base Learners

Table 6: Purity, Completeness and Contamination measures of the base learners

Base Learner	Purity in %			Completeness in %			Contamination in %		
	Star	Quasar	Galaxy	Star	Quasar	Galaxy	Star	Quasar	Galaxy
Neural Network	99.15	97.90	95.74	99.99	95.71	97.12	0.85	2.1	4.26
QDA	97.94	96.22	93.62	99.98	93.62	94.23	2.06	3.78	6.38
SVM	98.89	97.98	95.29	99.92	95.27	96.95	1.11	2.02	4.71
Extra Trees	99.43	97.67	95.67	99.99	95.66	97.10	0.57	2.33	4.33
XGBoost	99.78	97.45	95.60	99.91	95.67	97.27	0.22	2.55	4.40

Table 7: F1 score, Recall and Precision of the Base Learners

Base Learner	F1 Score			Precision			Recall		
	Star	Quasar	Galaxy	Star	Quasar	Galaxy	Star	Quasar	Galaxy
Neural Network	1	0.97	0.96	0.99	0.98	0.96	0.99	0.98	0.97
QDA	0.99	0.95	0.94	0.98	0.96	0.94	1	0.94	0.94
SVM	1	0.97	0.96	1	0.97	0.96	1	0.96	0.97
Extra Trees	1	0.97	0.96	0.99	0.98	0.96	1	0.96	0.97
XGBoost	1	0.97	0.96	1	0.97	0.96	1	0.96	0.97

In case of the base learners, XGBoost has the highest Star purity while Neural Networks have the highest Galaxy and Quasar purity. This indicates that XGBoost misclassified the least number of Galaxies and Quasars as Stars. Similarly, Neural Networks misclassified least number of Galaxies and Stars as Quasars and least number of Stars and Quasars as Galaxies. SVM have the highest Quasar purity while Neural Networks have the highest Galaxy Purity.

Extra trees and Neural Networks had the highest Star completeness indicating that these algorithms misclassified least number of Stars as Galaxies or Quasars. Neural Networks also have the highest Quasar completeness while XGBoost have the highest Galaxy completeness.

Neural Networks perform the best overall and this is due to their ability to learn complex and non-linear relationships between the dependent and independent variables. Neural Networks when tuned properly with the addition of regularization components such as dropout layers, can generalize well for unseen data.

Quadratic Discriminant Analysis did not perform as well as the other classification algorithms which could be because QDA does not assume all classes to have a common co-variance matrix. So, when there are ‘n’ predictors, estimating the covariance matrix would require estimating n(n+1)/2 parameters. For three classes this value goes to 3n(n+1)/2, which in this study is estimating 198 parameters. As the number of parameters increases so does the model variance. This was known before conducting the tests so QDA with a regularization parameter was used but despite that it failed to generalize well for the unseen dataset.

Galaxies overall have a higher contamination rate compared to stars and quasars which means that more stars and quasars were misclassified as galaxies. This is an interesting observation as stars and quasars were expected to have a higher contamination rate as they are point sources and can easily be mistaken for each other. Galaxies also had the least F1 score compared to Stars and Quasars for all the base classifiers.

6.1.1 Binomial Test

Before evaluating the predictions of the base learners, it is necessary to test whether the results achieved were not merely due to chance. The null hypothesis for the experiment is that the percentage of successful predictions of the classifiers is equal to p (accuracy achieved by the classifier). The accuracy values obtained by the base classifiers were on the test dataset. Firstly, a two-tail test was carried out which indicates that the performance of the classifier is neither less than nor greater than the accuracy acquired. Secondly, a one-tail test with the null hypothesis that the successful predictions of the classifier is greater than its attained accuracy. Alpha value of 0.05 was considered for both the tests.

Table 8: Binomial test for the base learners

Classifier	Accurate predictions (p)	p-value Binomial Test(two-tailed)	p-value Binomial Test (one-tailed)
Neural Networks	0.973	0.237	0.040
Extra Tree Classifier	0.972	0.168	0.004e-1
QDA	0.957	0.096	0.002
XGBoost	0.974	0.673	0.020
SVM	0.972	0.452	0.013

For all the base classifiers we fail to reject the null hypothesis for the two-tail binomial test indicating that the successful predictions are equal to the accuracy displayed by the classifier. However, for the one-tail binomial test we reject the null hypothesis for all the base classifiers as we get significant p-values indicating that the successful predictions do not exceed the accuracies obtained by the base learners.

6.1.2 Kruskal Wallis H-Test

For successful implementation of the stacked model it is necessary for the base classifiers to be specialist classifiers and have outputs that are statistically different from each other. A Cochran's Q test was first considered for this evaluation but since this is a multiclass classification problem it would not be appropriate to use that test. A Kruskal Wallis H test was hence used to determine if the errors of the base classifiers are uncorrelated with the significance level (alpha value) set to 0.05. After conducting the test, a significant p-value equal to 0.033 was observed. Thus, the null hypothesis was rejected indicating that the error's in the base learners are uncorrelated. The Kruskal Wallis H test does not tell which classifiers were statistically different.

6.1.3 Wilcoxon Signed-Rank Test

Since we obtained a significant value for the Kruskal Wallis H test, to observe which classifiers were different a post hoc test was required to be conducted. Wilcoxon signed-rank test was carried out on output of each of the classifiers in pairs. Since there were ten separate tests to be conducted Bonferroni adjustment was applied to the alpha value which was initially set to 0.05. The significance level after applying the Bonferroni adjustment is 0.005. A significant p-value was observed for all the pairs indicating that all base classifiers had uncorrelated errors when compared with each other individually. The p-value and test statistic value for all the Wilcoxon signed-rank tests has been added to the appendix.

6.2 Case Study 2: Evaluating Performance of the Stacked Model

Table 9: Purity, Completeness and Contamination measures of the meta learner

	Purity in %			Completeness in %			Contamination in %		
	Star	Quasar	Galaxy	Star	Quasar	Galaxy	Star	Quasar	Galaxy
Stacked Ensemble	99.80	98.06	95.50	99.92	95.51	97.90	0.20	1.94	4.5

Table 10: F1 score, Recall and Precision of the Meta Learner

	F1 Score			Precision			Recall		
	Star	Quasar	Galaxy	Star	Quasar	Galaxy	Star	Quasar	Galaxy
Stacked Ensemble	1	0.97	0.97	1	0.98	0.95	1	0.96	0.98

Stacked ensemble has a higher star and quasar purity compared to all the base classifiers which implies that not only a fewer number of quasars and galaxies were misclassified as stars, but also a smaller number of stars and galaxies were misclassified as quasars. Due to the high star and quasar purity they also have the least star and quasar contamination. The stacked model has a higher galaxy completeness compared to all the base classifiers indicating that fewer number of galaxies were misclassified as stars or quasars. The stacked model achieves an overall accuracy of 98%.

6.2.1 Binomial Test

The Binomial test was carried out for the meta learner with the significance level (alpha) set to 0.05.

Table 11: Binomial test results of the meta learner

	Accurate predictions (p)	p-value Binomial Test(two-tailed)	p-value Binomial Test (one-tailed)
Stacked Ensemble	0.980	0.64	0.01

In case of the meta learner the study fails to reject the null hypothesis for the two-tailed test indicating that the number of correct predictions of the meta meta-learner is equal to the accuracy achieved by the meta-learner. However, the null hypothesis for the one-tailed binomial test is rejected indicating that the accurate predictions of the meta-learner do not exceed the achieved accuracy.

6.2.2 Kruskal Wallis H-Test

It is necessary to test if the errors of the meta-learner and the base classifiers are uncorrelated. A Kruskal Wallis H test was carried out on the outputs of the base classifiers and the meta learner. The significance level set for the test was 0.05. A test statistic of 13.61 was obtained with a p-value of 0.01, thus the null hypothesis was rejected implying that the errors of the base classifiers and the meta learner are uncorrelated.

6.2.3 Wilcoxon Signed-Rank Test

Wilcoxon Signed-Rank test was carried out on output of each of the classifiers and the meta learner individually to test which base classifier had uncorrelated errors with the meta learner. Since there were five tests to be conducted, Bonferroni adjustment was applied to the significance level which was initially set to 0.05 and was later set to 0.01. All the tests returned significant p-values, which meant all the classifiers and the meta learner had uncorrelated errors. The p-values and test statistic values for all the Wilcoxon signed-rank meta learner tests has been added to the appendix.

6.3 Discussion

The stacked ensemble and the base classifiers as well as the base classifiers themselves despite having uncorrelated errors, the stacked model did not improve significantly upon the best performing base classifier in terms of Star/Galaxy/Quasar purity and completeness. Upon further study of the errors of each of the base classifiers as well as the meta classifier, the root cause was identified to be the Galaxy-Quasar classification which was not expected. Out of a set of 60,000 samples reserved only to test the meta-learner, none of the stars were misclassified as Quasars, while only 5 Quasars were misclassified as Stars. However, 908 Quasars were misclassified as Galaxies and 386 Galaxies were misclassified as Quasars. This study was based on the hypothesis that the diverse base classifiers would specialize in classifying different pairs of classes. This was not observed as all the base classifiers struggled in the Quasar-Galaxy classification task. As a result, the base classifiers and the meta classifier obtained a particularly high contamination rate for galaxies

Studies conducted by Viquar *et al.* (2019), Nakoneczny *et al.* (2019), Guo *et al.* (2018), Bai *et al.*(2018) and Krakowski *et al.* (2016) suggest that since Stars and Quasars are point sources, Star and Quasar classes have higher contamination rates. The one aspect common in their studies is that they all worked on the old SDSS data releases. SDSS released their latest data DR15 in December 2018, and it includes data collected from Extended Baryon Oscillation Spectroscopic Survey containing spectroscopically confirmed Quasars which were left completely unexplored by all the previous surveys(Aguado *et al.*, 2018). The DR15 also for the first time includes data of the spatially resolved Galaxies from MaNGA (Mapping Nearby Galaxies at APO) survey. Since more resolved objects were present in the latest data release the Quasars and Galaxies were more difficult to classify than Stars and Quasars. This can be explained as Quasars are essentially Galaxies which have an active nucleus, but until this data release most of them remained unresolved due to limitations of the telescopic and spectroscopic instruments.

7 Conclusion and Future Work

The stacked generalization framework used in this research with features selected by MRMR and decomposed by NMF significantly improved upon the performance of the state of the art for the Star-Galaxy-Quasar classification task with 98% accuracy(99.80%-star purity, 95.50%-galaxy purity and 98.06%-Quasar purity). Bai *et al.* (2018) had previously obtained state of the art accuracy of 95% and purity of 96.5%, 92.5% and 97.4% for Stars, Galaxies and Quasars respectively. Viquar *et al.* (2019) achieved an accuracy of 97.05% for the Star-Quasar classification task however they did not report their results in terms of purity or completeness. The base classifiers had uncorrelated errors, but they failed to specialize for distinct pairs of classes. The performance of the base classifiers and the entire stacked model was compared and although an improvement in the performance was observed it was only marginally higher than the best performing base classifiers. Currently, the research in this field is focused on the Star-Quasar classification but findings of this study suggest that the objects in the deep astronomical surveys are now more resolved which makes it difficult to distinguish galaxies and quasars.

This research focused solely on the Star-Galaxy-Quasar classification, and current studies are focused on the Star-Quasar classification tasks. Galaxy-Quasar classification problem has been overlooked upon as they were thought to be easily distinguishable. The future work for this study would be to implement a Galaxy-Quasar classifier and study the behaviour of the classifier for luminous and faint color magnitudes. With the current data release faint magnitude regions (u, g, r, i, z > 27) have only 5,283 objects which are very low for conducting the study on faint regions. This will change for the SDSS data release scheduled for mid-2021 when more objects of the fainter magnitude regions are expected to be released. The Galaxy-Quasar classification can be carried out Generalized Additive Models (GAMs) with logistic link functions. They can provide promising results for this binary classification problem where both the classes as observed from this study are non-linearly separable.

Acknowledgement

I'm highly indebted to Professor Noel Cosgrave for his invaluable feedback, supervision and constant guidance throughout the study. I would like to express my gratitude towards my grandparents and parents for their moral support and encouragement.

References

Aguado, D. *et al.* (2019). The Fifteenth Data Release of the Sloan Digital Sky Surveys: First Release of MaNGA-derived Quantities, Data Visualization Tools, and Stellar Library. *The Astrophysical Journal Supplement Series*, 240(2), p.23.

Anjum, A., Das, M., Murthy, J., Gudennavar, S., Gopal, R. and Bubbly, S. (2018). Template-based classification of SDSS-GALEX point sources. *Journal of Astrophysics and Astronomy*, 39(5).

Bai, Y., Liu, J., Wang, S. and Yang, F. (2018). Machine Learning Applied to Star–Galaxy–QSO Classification and Stellar Effective Temperature Regression. *The Astronomical Journal*, 157(1), p.9.

Cabayol, L. *et al.* (2018). The PAU survey: star–galaxy classification with multi narrow-band data. *Monthly Notices of the Royal Astronomical Society*, 483(1), pp.529-539.

Chen, T. and Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16.*

DING, C. and PENG, H. (2005). Minimum Redundancy Feature Selection. *Journal of Bioinformatics and Computational Biology*, 03(02), pp.185-205.

Elting, C., Bailer-Jones, C., Smith, K. and Bailer-Jones, C. (2008). Photometric Classification of Stars, Galaxies and Quasars in the Sloan Digital Sky Survey DR6 Using Support Vector Machines. *AIP Conference Proceedings*.

Fadely, R., Hogg, D. and Willman, B. (2012). STAR-GALAXY CLASSIFICATION IN MULTI-BAND OPTICAL IMAGING. *The Astrophysical Journal*, 760(1), p.15.

Gao, D., Zhang, Y. and Zhao, Y. (2008). Support vector machines and kd-tree for separating quasars from large survey data bases. *Monthly Notices of the Royal Astronomical Society*, 386(3), pp.1417-1425

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014. An Introduction to Statistical Learning: With Applications in R. Springer Publishing Company, Incorporated.

Geurts, P., Ernst, D. and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1), pp.3-42.

Guo, S., Qi, Z., Liao, S., Cao, Z., Lattanzi, M., Bucciarelli, B., Tang, Z. and Yan, Q. (2018). Identifying quasars with astrometric and mid-infrared methods from APOP and ALLWISE. *Astronomy & Astrophysics*, 618, p.A144.

Henrion, M., Mortlock, D., Hand, D. and Gandy, A. (2011). A Bayesian approach to star-galaxy classification. *Monthly Notices of the Royal Astronomical Society*, 412(4), pp.2286-2302.

Sevilla-Noarbe, I. *et al.* (2018). Star-galaxy classification in the Dark Energy Survey Y1 dataset. *Monthly Notices of the Royal Astronomical Society*.

Schindler, J., Fan, X., Huang, Y., Yue, M., Yang, J., Hall, P., Wenzl, L., Hughes, A., Litke, K. and Rees, J. (2019). The Extremely Luminous Quasar Survey in the Pan-STARRS 1 Footprint (PS-ELQS). *The Astrophysical Journal Supplement Series*, 243(1), p.5.

Khramtsov, V., Akhmetov, V. and Principles, A. (2018). Machine-learning identification of extragalactic objects in the optical-infrared all-sky surveys. *2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT)*. IEEE, 1, pp. 72–75.

Kim, E. J. and Brunner, R. J. (2017). Star–galaxy classification using deep convolutional neural networks. *Monthly Notices of the Royal Astronomical Society*, 464(4), pp. 4463–4475

Kim, E. J., Brunner, R. J. and Carrasco Kind, M. (2015) 'A hybrid ensemble learning approach to star–galaxy classification', *Monthly Notices of the Royal Astronomical Society*, 453(1), pp. 507–521.

Kovács, A. and Szapudi, I. (2015). Star–galaxy separation strategies for WISE-2MASS all-sky infrared galaxy catalogues. *Monthly Notices of the Royal Astronomical Society*, 448(2), pp. 1305–1313.

Krakowski, T., Małek, K., Bilicki, M., Pollo, A., Kurcz, A. and Krupa, M. (2016). Machine-learning identification of galaxies in the WISE × SuperCOSMOS all-sky catalogue. *Astronomy & Astrophysics*, 596, p.A39.

Machado, E., Serqueira, M., Ogasawara, E., Ogando, R., Maia, M., da Costa, L., Campisano, R., Paiva Guedes, G. and Bezerra, E. (2016). Exploring machine learning methods for the Star/Galaxy Separation Problem. *2016 International Joint Conference on Neural Networks (IJCNN)*.

Morice-Atkinson, X., Hoyle, B. and Bacon, D. (2018). Learning from the machine: interpreting machine learning algorithms for point- and extended-source classification. *Monthly Notices of the Royal Astronomical Society*, 481(3), pp.4194-4205.

Nakoneczny, S., Bilicki, M., Solarz, A., Pollo, A., Maddox, N., Spiniello, C., Brescia, M. and Napolitano, N. (2019). Catalog of quasars from the Kilo-Degree Survey Data Release 3. *Astronomy & Astrophysics*, 624, p.A13.

Peng, N., Zhang, Y. and Zhao, Y. (2013). A SVM-kNN method for quasar-star classification. *Science China Physics, Mechanics and Astronomy*, 56(6), pp.1227-1234.

Peters, C., Richards, G., Myers, A., Strauss, M., Schmidt, K., Ivezic´, Ž., Ross, N., MacLeod, C. and Riegel, R. (2015). QUASAR CLASSIFICATION USING COLOR AND VARIABILITY. *The Astrophysical Journal*, 811(2), p.95.

Ramió, H. *et al.* (2019). J-PLUS: Morphological star/galaxy classification by PDF analysis. *Astronomy & Astrophysics*, 622, p.A177.

Solarz, A. *et al.* (2012). Star-galaxy separation in the AKARI NEP deep field. *Astronomy & Astrophysics*, 541, p.A50.

Soumagnac, M. *et al*. (2015). Star/galaxy separation at faint magnitudes: application to a simulated Dark Energy Survey. *Monthly Notices of the Royal Astronomical Society*, 450(1), pp.666-680.

Tadhunter, C. (2008). An introduction to active galactic nuclei: Classification and unification. *New Astronomy Reviews*, 52(6), pp.227-239.

Vasconcellos, E., de Carvalho, R., Gal, R., LaBarbera, F., Capelato, H., Frago Campos Velho, H., Trevisan, M. and Ruiz, R. (2011). DECISION TREE CLASSIFIERS FOR STAR/GALAXY SEPARATION. *The Astronomical Journal*, 141(6), p.189.

Viquar, M., Basak, S., Dasgupta, A., Agrawal, S. and Saha, S. (2018). Machine Learning in Astronomy: A Case Study in Quasar-Star Classification. *Advances in Intelligent Systems and Computing*, pp.827-836.

Zhao, Y. and Zhang, Y. (2008). Comparison of decision tree methods for finding active objects. *Advances in Space Research*, 41(12), pp.1955-19

Appendix

The table below gives a detailed description of the features extracted from the Sloan Digital Sky Survey

Feature	Description	Minimum	Maximum	Mean	Median	Standard Deviation
objid	A unique number to identify objects in the image catalog used by CasJobs	NA	NA	NA	NA	NA
ra	Right Ascension	0.0007	360.00	180.13	180.43	67.86
dec	Declination	-11.25	79.76	28.91	34.89	21.79
u	Ultraviolet light filter	6.75	32.78	21.78	21.81	2.31
g	Green and Blue visible light filter	7.46	33.72	20.22	20.61	2.05
r	Red and Yellow Visible Light Filter	8.44	33.21	19.21	19.55	1.82
i	Near Infrared Light Filter	7.65	32.34	18.56	18.96	1.71
z	Infrared Light Filter	6.48	30.84	18.35	18.60	1.71
run	Value associated with the SDSS camera strip. It is the length of the camera strip during a single scan and is bound by ra and dec	NA	NA	NA	NA	NA
rerun	Reprocessing of an imaging run	NA	NA	NA	NA	NA
camcol	A Camcol is the output of one camera column of CCDs (each with a different filter) as part of a Run	NA	NA	NA	NA	NA
field	Field is the position of the object in the sky defined by a radius and central co-ordinates	11	812	190	157	137.78
specobjid	A unique number to identify objects in the spectroscopic catalog used by CasJobs	NA	NA	NA	NA	NA
redshift	A measure indicating the velocity of an object in the sky and serves as a measure of distance.	-0.01	7.04	0.50	0.30	0.72
plate	Indicates which spectrograph plate was used to collect information. Each plate uniquely identifies a spectrum.	266	8955	3981	4052	2538.83
mjd	An integer denoting night of the observation	51578	57520	54857	55539	1746.64

Feature	Description	Minimum	Maximum	Mean	Median	Standard Deviation
fibreid	An integer denoting the fiber number used for the observation	1	1000	421	398	264.34

Individual results of the Wilcoxon Signed Rank test for the base classifiers

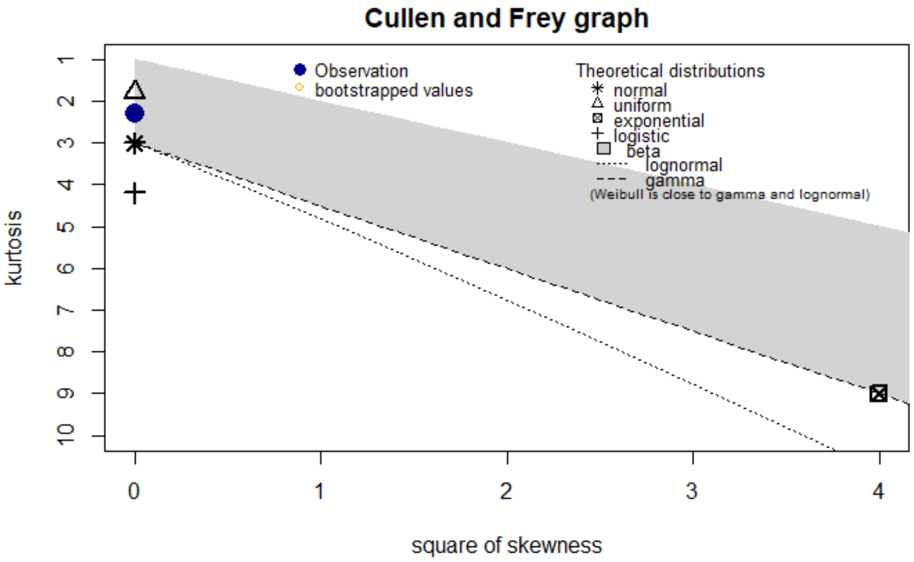
Classifiers	Test Statistic	p-value
Neural Network, Extra Trees	4297	0.0026
Neural Network, SVM	50041	0.0034
Neural Network, QDA	606225	9.42e-30
Neural Network, XGBoost	47352	3.50e-13
Extra Trees, SVM	72598	0.003
Extra Trees, QDA	586342	1.04e-35
Extra Trees, XGBoost	30754	3.12e-08
SVM, QDA	535593	6.89e-32
SVM, XGBoost	78853	2.40e-10
QDA, XGBoost	542898	7.98e-48

Individual results of the Wilcoxon Signed Rank test for the meta learner and the base learners

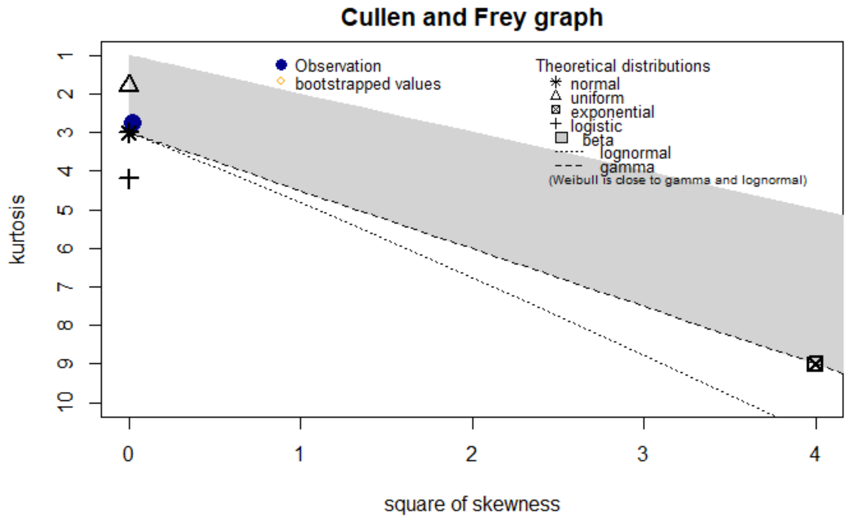
Classifier	Test Statistic	p-value
Meta learner, Neural Networks	5075	1.13e-20
Meta learner, SVM	47205	1.20e-26
Meta learner, XGBoost	6890	2.41e-24
Meta learner, QDA	475620	4.79e-40
Meta learner, Extra Trees	6272	8.75e-30

Cullen and Frey Graphs

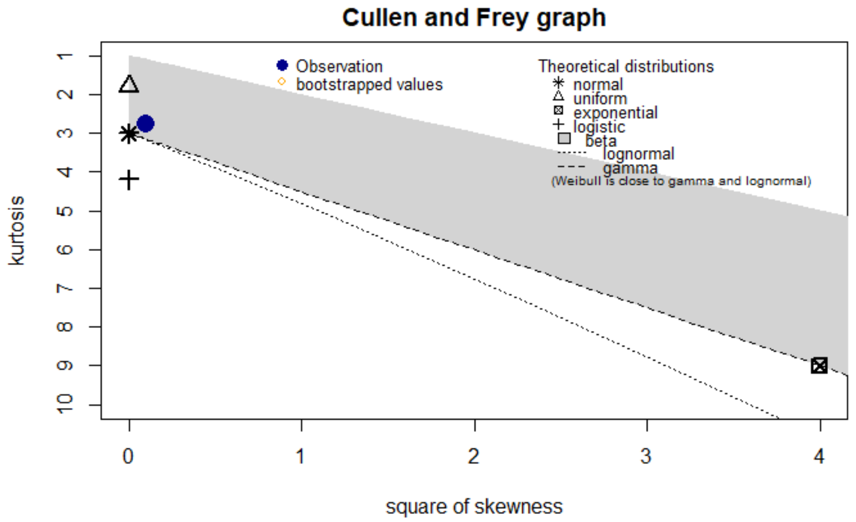
Photometric Bands



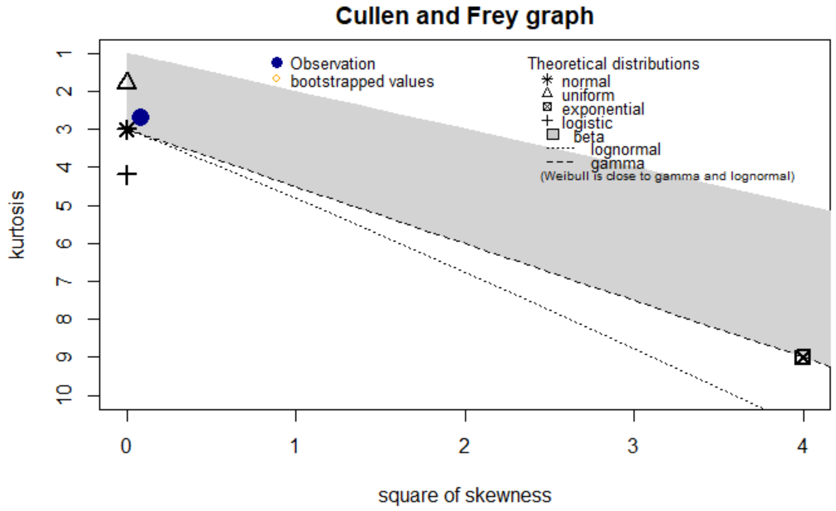
1. u-band



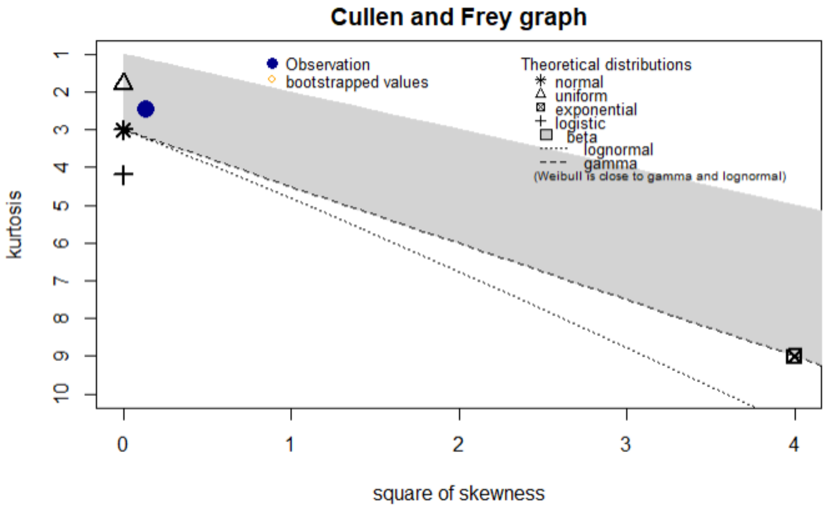
2. z-band



3. r-band

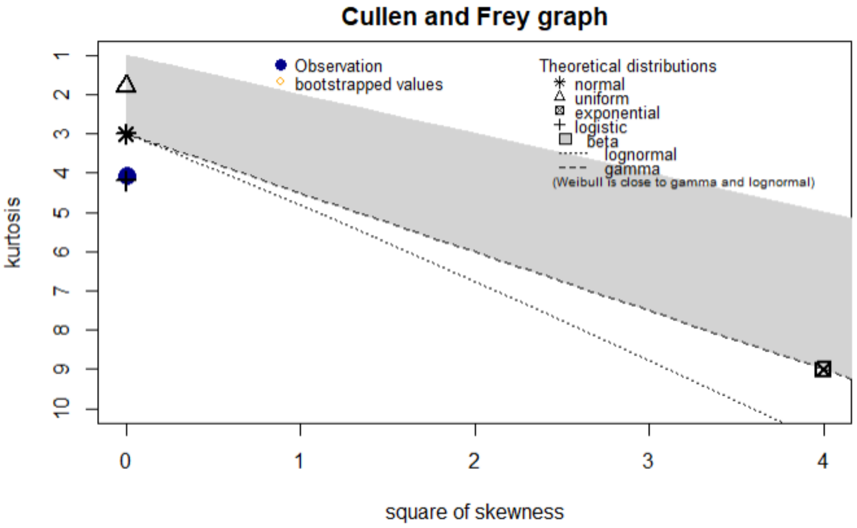


4. i-band

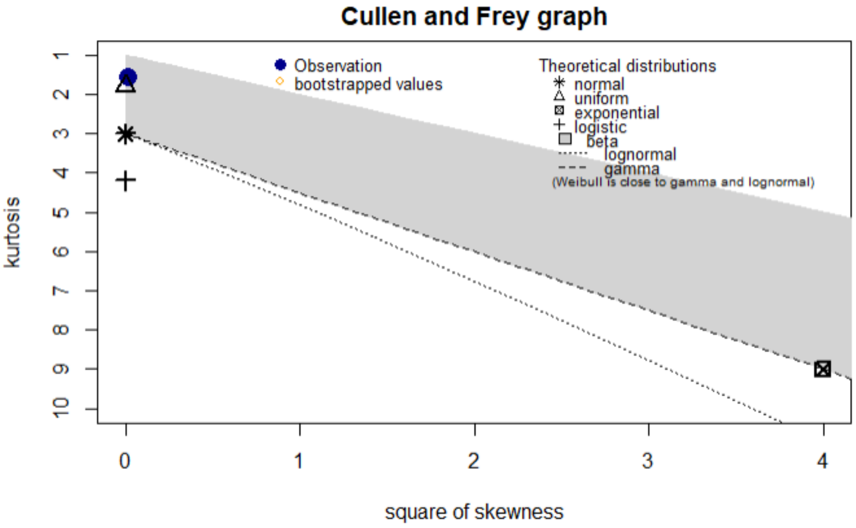


5. g-band

Location Data

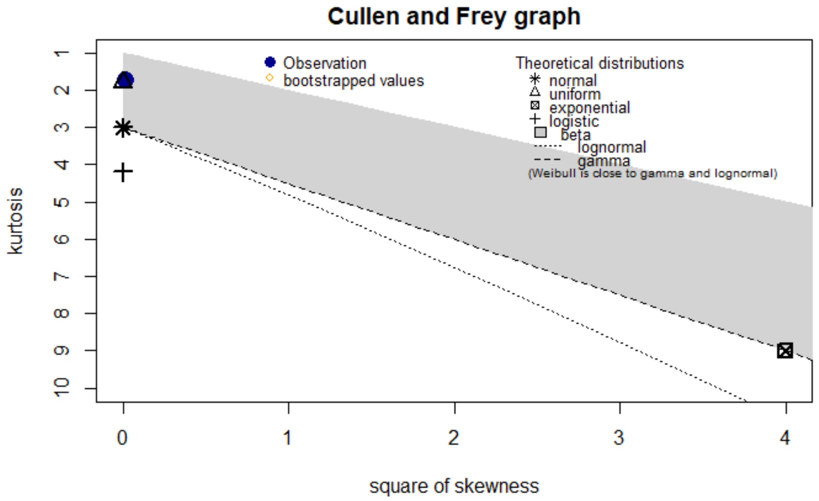


6. ra

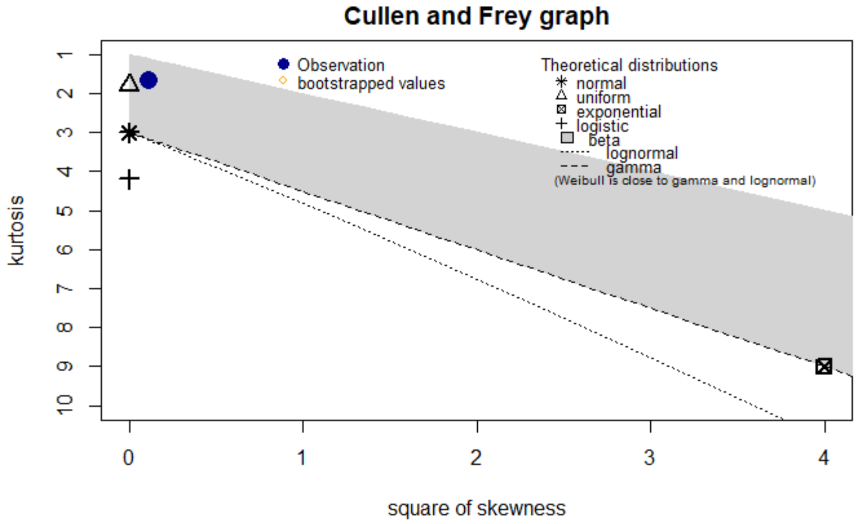


7. Dec

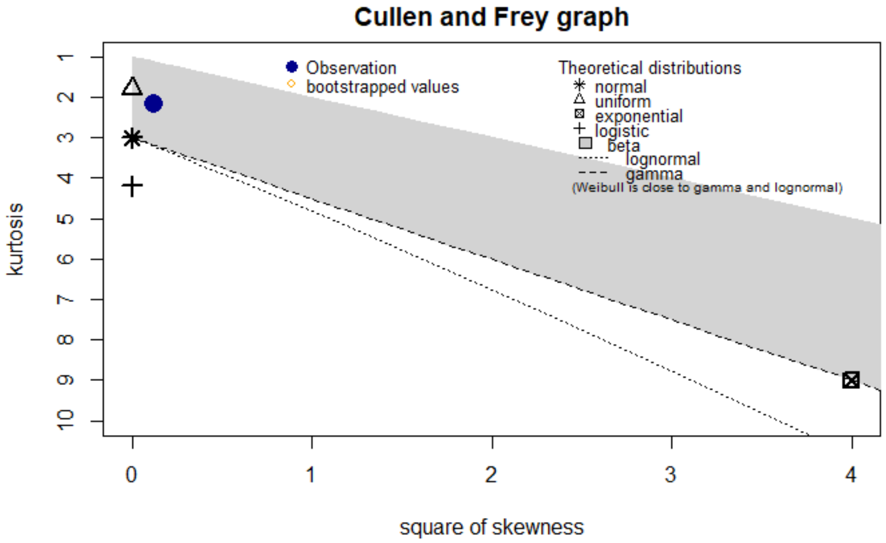
Spectroscopic Data



8. Plate

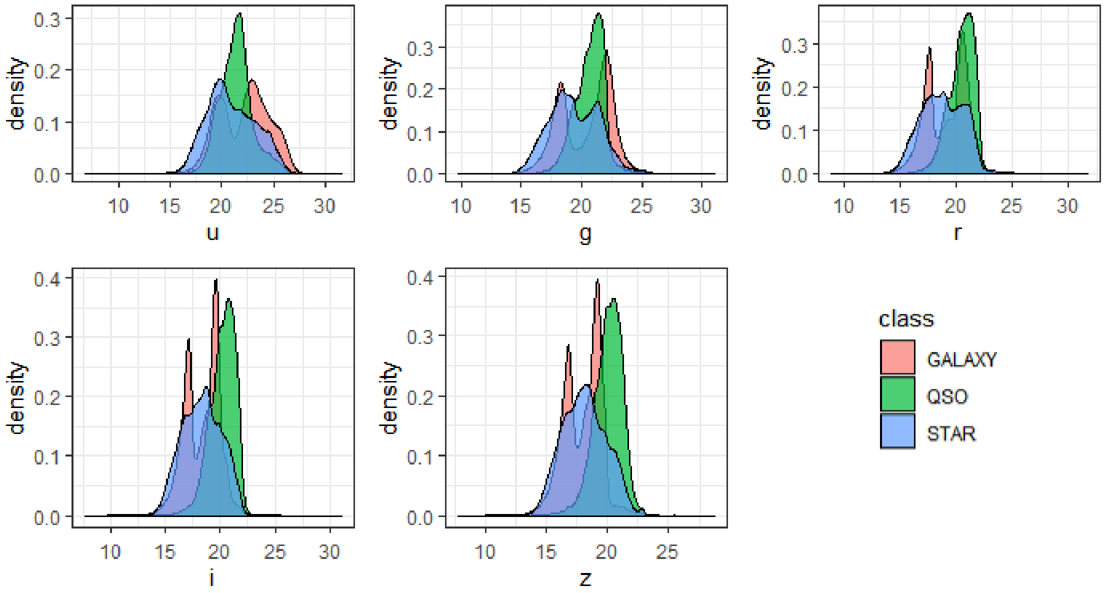


9. Mjd

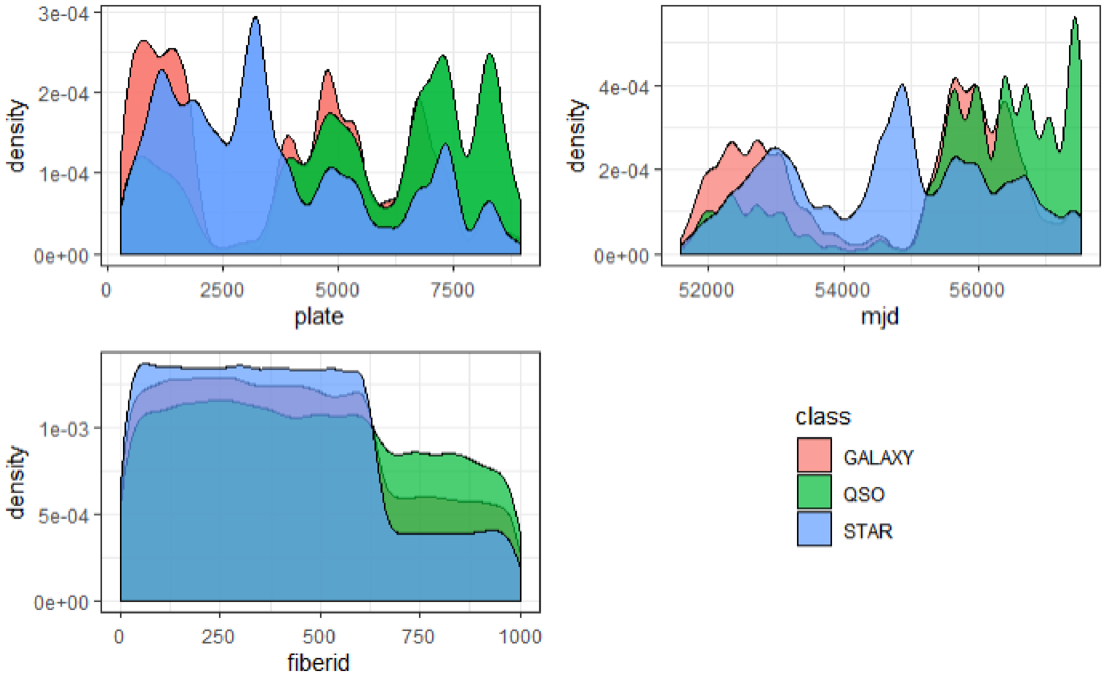


10. Fibreid

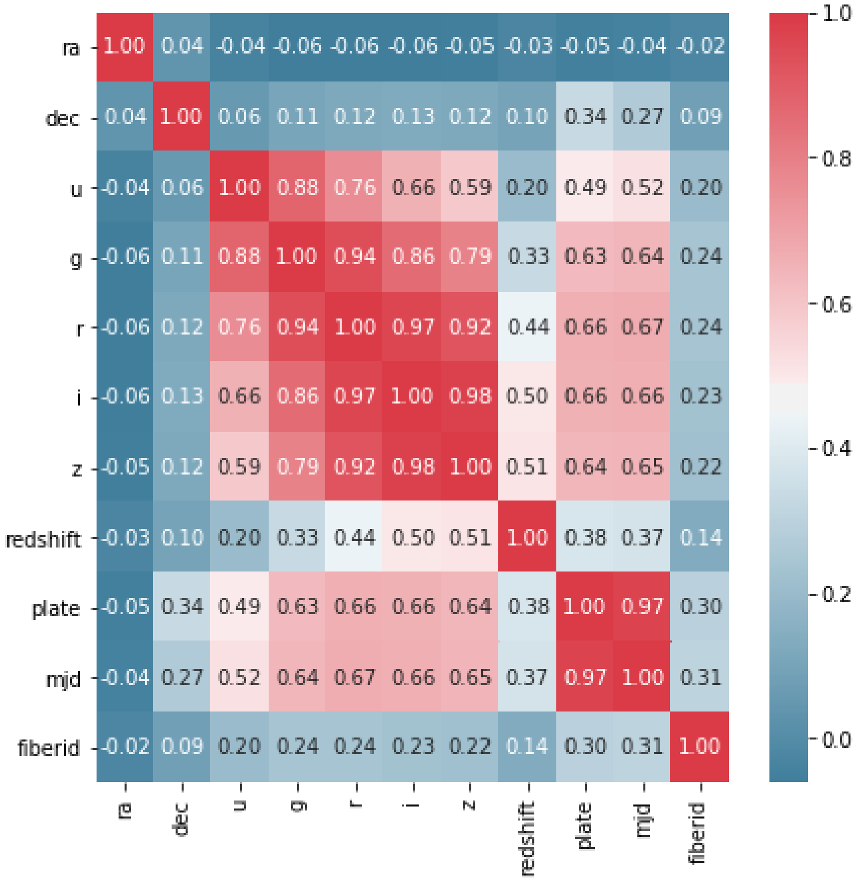
Density Plot of Photometric Data



Density plot of Spectroscopic data

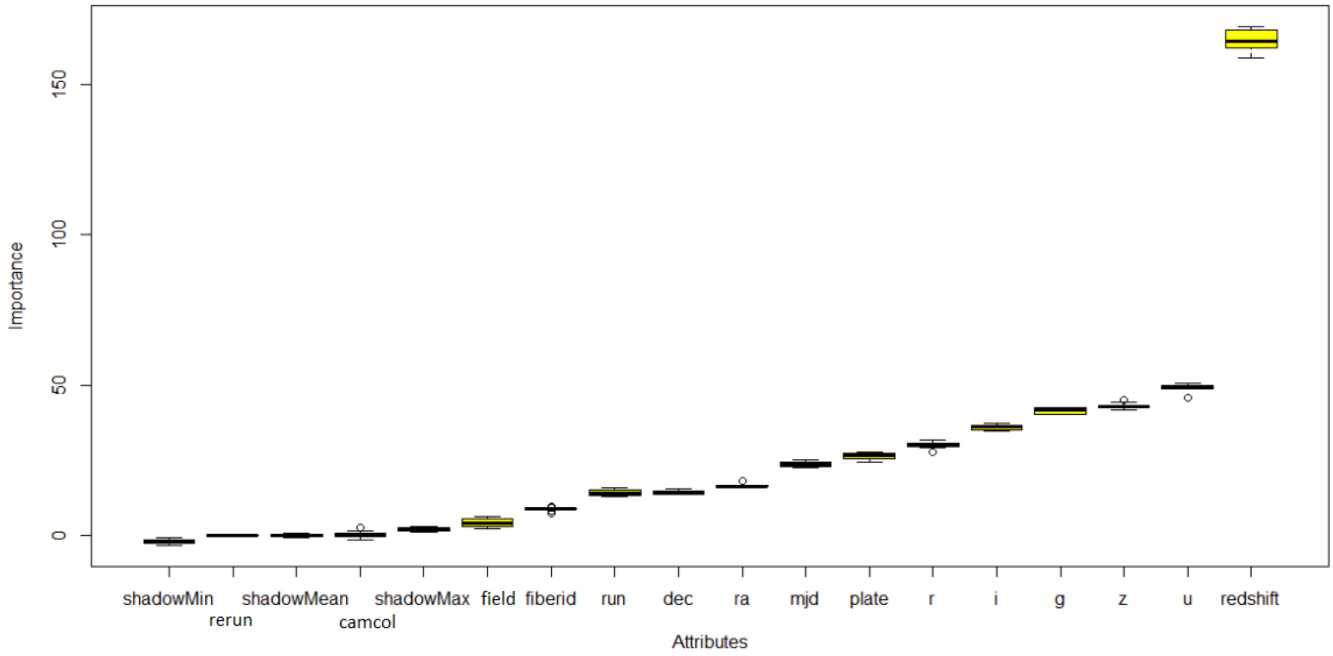


Correlation plot

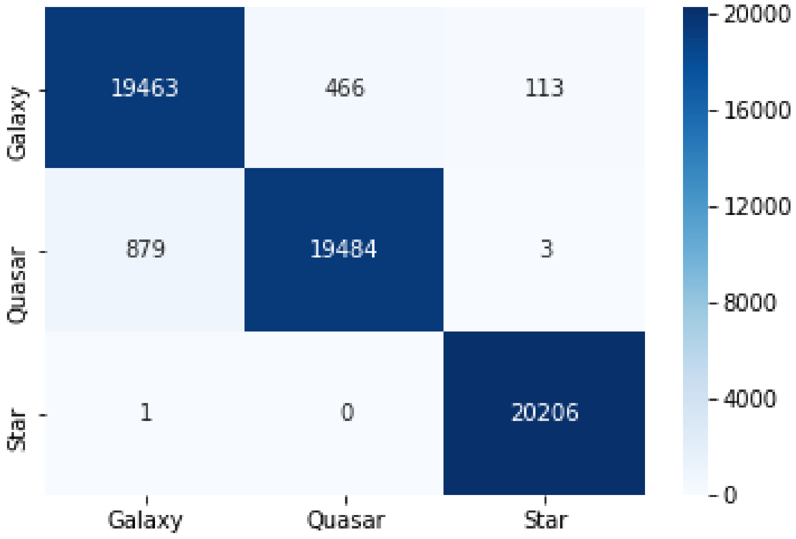


	ra	dec	u	g	r	i	z	redshift	plate	mjd	fiberid
ra	1.00	0.04	-0.04	-0.06	-0.06	-0.06	-0.05	-0.03	-0.05	-0.04	-0.02
dec	0.04	1.00	0.06	0.11	0.12	0.13	0.12	0.10	0.34	0.27	0.09
u	-0.04	0.06	1.00	0.88	0.76	0.66	0.59	0.20	0.49	0.52	0.20
g	-0.06	0.11	0.88	1.00	0.94	0.86	0.79	0.33	0.63	0.64	0.24
r	-0.06	0.12	0.76	0.94	1.00	0.97	0.92	0.44	0.66	0.67	0.24
i	-0.06	0.13	0.66	0.86	0.97	1.00	0.98	0.50	0.66	0.66	0.23
z	-0.05	0.12	0.59	0.79	0.92	0.98	1.00	0.51	0.64	0.65	0.22
redshift	-0.03	0.10	0.20	0.33	0.44	0.50	0.51	1.00	0.38	0.37	0.14
plate	-0.05	0.34	0.49	0.63	0.66	0.66	0.64	0.38	1.00	0.97	0.30
mjd	-0.04	0.27	0.52	0.64	0.67	0.66	0.65	0.37	0.97	1.00	0.31
fiberid	-0.02	0.09	0.20	0.24	0.24	0.23	0.22	0.14	0.30	0.31	1.00

Feature Importance chart obtained after performing feature selection with Boruta

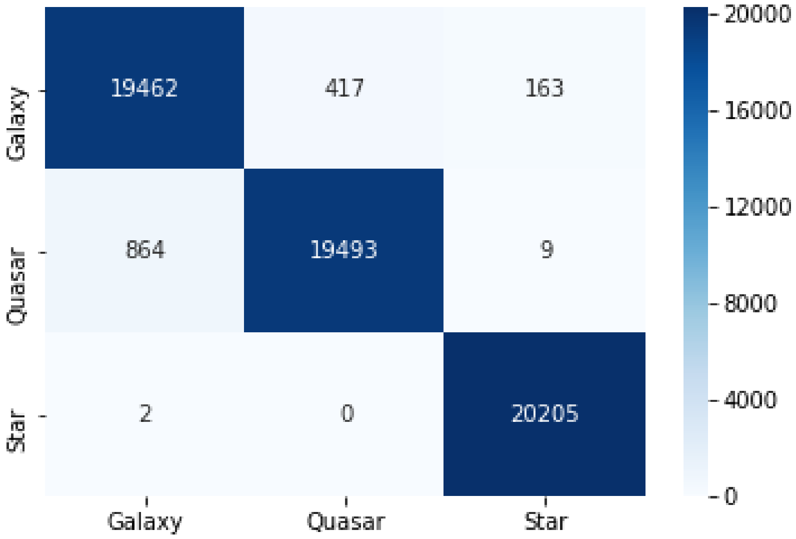


Confusion matrices of the base learners and the entire stacked model



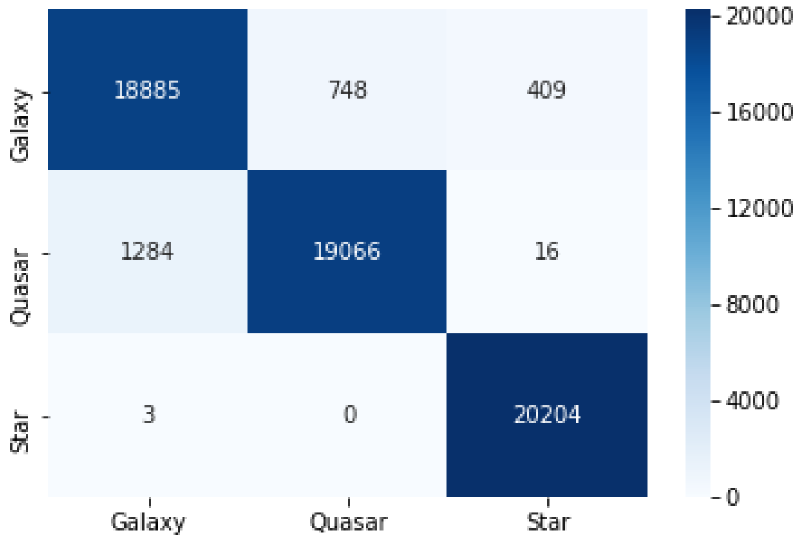
	Galaxy	Quasar	Star
Galaxy	19463	466	113
Quasar	879	19484	3
Star	1	0	20206

Confusion matrix of Extra-Tree classifier for the reserved test set

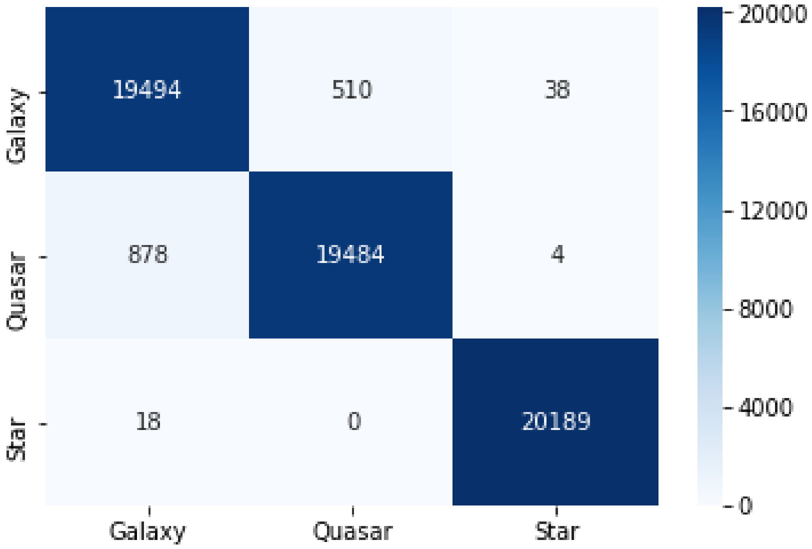


	Galaxy	Quasar	Star
Galaxy	19462	417	163
Quasar	864	19493	9
Star	2	0	20205

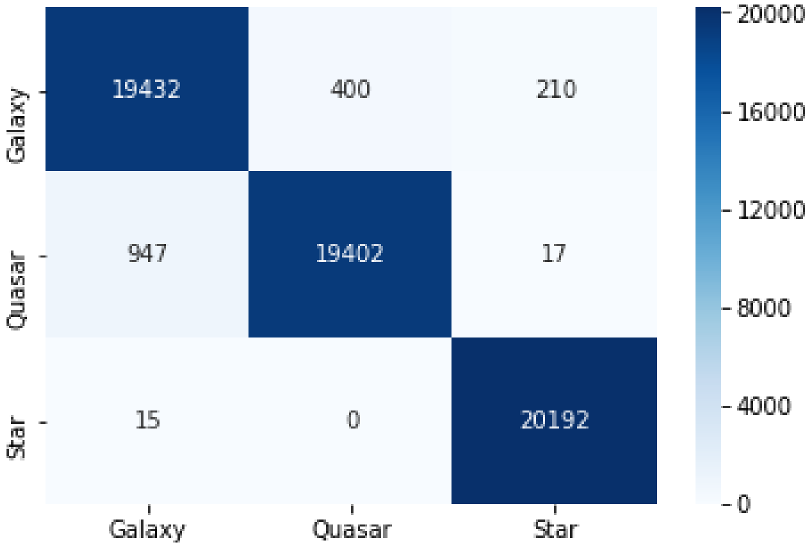
Confusion matrix of neural networks for the reserved test set



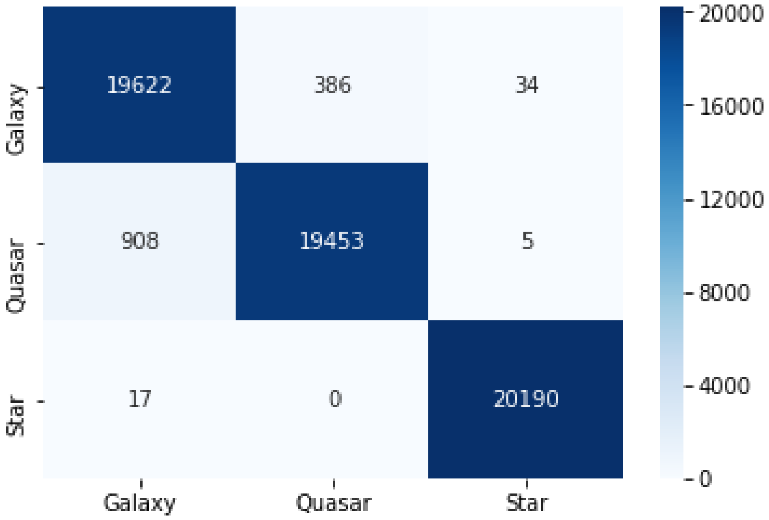
Confusion matrix of QDA for the reserved test set



Confusion matrix of XGBoost for the reserved test set



Confusion matrix of SVM for the reserved test set



Confusion matrix of the stacked ensemble for the reserved test set

A tabular comparative study of the most relevant and recent works is given on the next page.

Author and Year	**Classification**	**Machine Learning Technique used**	**Hyper Parameter Opt.**	**Sample Size (used)**	**Sampling Technique**	**Feature Selection**	**Feature Decomposition**	**Balance**	**Model Performance**
Viquar *et al.(* 2019)	Photometric and Spectroscopic Data, Star-Quasar	Adaboost compared with SVM-KNN	Absent	75000	Absent	Absent	Absent	Yes	Adaboost-96.54% SVM-KNN-97.8%
Ramió *et al.(* 2019)	Morphological Data, Star-Galaxy	Bayesian Classifier	Absent	251,000	Bootstrapping	Absent	Absent	No (use weighted approach for classification)	95% galaxy completeness
Schlinder *et al.*(2019)	Photometric Star-Quasar classification	Random Forest	Grid Search	500,000	Holdout(80/20)	Absent	Absent	Yes	77% quasar accuracy
Nakoneczny *et al.(* 2019)	Photometric Star-Quasar classification	XGBoost,Random Forest, ANN	Trial and Error	51,752	Training-Validation-Test	Tree based Backward elimination	Absent	Yes	XGB-96.44% ANN-96.28 RF-96.56%
Bai *et al.*(2018)	Photometric and Spectroscopic Data, Star-Galaxy-QSO	Random Forest	Absent	440000	Holdout(80/20)	Absent	Absent	Yes	99.6% star completeness 97.6% Galaxy completeness 88.9% quasar completeness
Anjum *et al.*(2018)	Photometric Data, Star-Galaxy-QSO	Template Based Classification	Absent	37,492	Holdout(80/20)	Absent	Absent	No	84% -Star Accuracy 63%- Galaxy Accuracy 89%- QSO Accuracy

Author and Year	**Classification**	**Machine Learning Technique used**	**Hyper Parameter Opt.**	**Sample Size (used)**	**Sampling Technique**	**Feature Selection**	**Feature Decomposition**	**Balance**	**Model Performance**
Cabayol *et al.*(2018)	Photometric Data, Star-Galaxy classification	CNN Random Forest Neural Network	Absent	30,000	Holdout(80/20)	Model Based Feature Selection	Absent	No	98%-Galaxy purity- CNN, 91.3%-Galaxy Purity Random Forest, 95%-Galaxy Purity- Feedforward Neural Network
Khramtsov et al.(2018)	Photometric Star-Galaxy classification extended study of Intergalactic-Galactic classification	SVM	Grid Search	600,000	K-Fold Cross Validation	Auto-encoders	Absent	Yes	99.2 Extra galactic objects 99.8 Galactic objects
Hoyle *et al.* (2018)	Photometric Data, Star-Galaxy classification	SVM Neural Network Adaboost	Grid Search	116,000	Holdout(80/20)	Absent	Absent	Yes	88.5%- NN-Accuracy 96.7%-Adaboost Accuracy 96.2%- SVM Accuracy
Guo *et al.*(2018)	Infrared Data, Quasar Selection	SVM	Absent	662,000	Choose a different catalog for testing	Absent	Absent	Yes	Quasar completeness-75%