Iranian Journal of Astronomy and Astrophysics
Vol. 9, No. 1, Spring 2022
©Available online at http://journals.du.ac.ir
DOI:10.22128/ijaa.2022.554.1122

Iranian Journal of Astronomy and Astrophysics

Machine Learning Method for Predicting the Merger and Morphology of Galaxies through Near-Infrared Spectroscopy

Samira Monfared[1] · Neda Abdolvand*[2] · Mohammad Taghi Mirtorabi[3] · Saeedeh Rajaee Harandi[4]

[1] Department of Management, Faculty of social Sciences and Economics, Alzahra University, Tehran, Iran;
email: samiramnr73@gmail.com

[2] Department of Management, Faculty of Social Sciences and Economics, Alzahra University, Tehran, P.O.Box 1993891176, Iran;
*email: n.abdolvand@alzahra.ac.ir

[3] Department of Physics, Alzahra University, Tehran, Iran;
email: torabi@alzahra.ac.ir

[4] Department of Management, Faculty of Social Sciences and Economics, Alzahra Univerdsity, Tehran, Iran;
email: sa.rajaeeharandi@gmail.com

Abstract. Astronomy is experiencing rapid growth in the size and complexity of data.This reinforces the development of data-driven science as a useful complement to the current model of model-based data analysis. In spite of this, traditional merger analysis of catalogs is mostly done through visual inspection by trained experts. These methods are not efficient today because, the subjectivity of manual classification has made the result of the process very dependent on the intuition of the analyst and the type and quality of the image. Hence, this study focuses on data processing methods based on Artificial Intelligence (AI) algorithms and investigates the possibility of a pattern among astronomical data to predict the merger of galaxies. The modeling is done in two phases. The first phase deals with the classification between minority and majority classes and the second phase search for any distinction between minority classes. In both phases, various algorithms such as Naive Bayes, Random Forest, and Generalized linear model (GLM) and Neural network are used to ensure the best results according to the research data. The best results for both phases were obtained from the implementation of the GLM algorithm with the accuracy of 70.28% and 76.51% for the first and second phase, respectively.

Keywords: Galaxy Morphology, Galaxy Merge, Near-Infrared Spectroscopy, Galaxy Zoo, Machine Learning

1 Introduction

The size and complexity of data in astronomy is growing rapidly. This promotes the development of data-driven science as a useful complement to the current model of model-based data analysis, where astronomers develop automated tools to manipulate data sets and extract new information from them [1,2]. Extracting knowledge from data collected by advanced tools requires the high processing power that is possible today with the help of high-tech processors and complex machine learning algorithms. The archiving of astronomical data has increased dramatically over the past decades [1,2]. Due to the advances in telescopes technologies, increased computing power, improved data collection methods, and successful theoretical simulations, large volumes of data (terabytes) are available now and will soon appear in petabytes[1,3]. Implementation of all these data manually or even by using regular available electronics is impractical for astronomers [3,4]. However, astronomy has been one of the pioneers which brings forward the notion of Big Data. These huge data sets need proper management and processing. By processing such data sets, rare and unknown phenomena can be discovered and objects are categorized in terms of their structure, shape, and any other specific observable attribute. Furthermore, some events, such as the merger of galaxies are predicted and help to accept or reject theories [3–5].

In many astronomical catalogs, containing galaxies, morphological characteristics were compiled during twentieth-century [6–8]. In most of these catalogs, traditional integration analysis is mostly done through visual inspection by trained experts. These methods are not efficient today because they requires a lot of skills and experience in the field [3]. On the other hand, due to the subjectivity of the manual classification, the outcome of the process depends highly on the analyst intuition and the type and quality of image.

For the correct classification of an object, the image quality and other systematic factors of various astronomical studies must also be considered. The use of data mining and the application of advanced image processing techniques and powerful learning algorithms make automatic classification of stars and galaxies faster and is a better alternative to the manual methods [9]. Applications of data mining in astronomy includes the classification of stars, galaxies, and planetary nebulae, with both methods of image and spectral processing, separating stars from galaxies, and identifying the structure of galaxies (morphologically) [3]. Such studies have focused mainly on the degree of merge, and relatively little effort has been made to examine their morphologies and internal properties. But, to accurately understand the process that takes place during the merging of galaxies, measuring the interaction rate alone is not enough because the processes that determine the morphological results of mergers are not fully understood yet.

Given the increasing and rapid growth of astronomical data, this question arises: how can machine learning be used to propose a model for predicting the galaxies merger? To answer this question, this study aims to provide a model for predicting the merger and morphology of galaxies using their spectroscopic characteristic in the near-infrared region by proper implementation of machine learning algorithms. By increasing the sample size or changing the variables used in previous studies, as well as using and testing the performance of several different algorithms, an attempt was made to increase the accuracy of the prediction. The results of the research can help to discover those rare and unknown objects or phenomena and to predict some events such as the galaxies merger and to accept or reject theories in this field.

The rest of the study is organized as follows: After reviewing the research literature in this field, the research method is proposed. Then, the data analysis and its results are presented. Finally, conclusions and suggestions for future research are provided.

2 Literature Review

Galaxies are celestial bodies made up of gas, dust, and billions of stars that form over billions of years, and their morphology provides astronomers with a wealth of information about their composition and evolution [10–12]. The classification of galaxies can be important because physicists often use large catalogs of information to test existing theories or to propose hypotheses to explain the physical processes that make up galaxies, stars, and the nature of the universe [12]. Through the merger, galaxies alter their own content of stars, gas, and dark matter [13]. Mergers can also affect the formation of black holes and stimulate the activity of AGNs (active galactic nuclei) [13,14]. Morphology can be considered as a powerful indicator of the history of galaxies and mergers where it is strongly related to many physical parameters, including mass, star formation history, and mass distribution [15–17]. Galactic morphology can be used to study how galaxies evolve and to identify the predominant mechanism of their formation [18]. As telescopes, detectors, and computers become more powerful, the amount of data available to astronomers is entering the petabyte realm, providing accurate measurements for billions of celestial bodies [19,20]. These automated tools, increased computing capabilities, improved data collection methods, and provided successful theoretical simulation applications. Large volumes of data [1–3] have made the use of more advanced analytical methods, including data mining, inevitable.Data mining includes procedures for finding designs or patterns in a large dataset, and includes strategies for converging machine learning techniques and the database framework [21].

Being aware of the importance of the morphology of galaxies, many researchers have tried to predict the structure of galaxies using data mining techniques [16]. For example, [22] used SDSS DR9 labeling and machine learning algorithms such as simple Bayesian, logistic regression, SVM vector machine, random forest, and nearest neighbor to find the best machine learning methods to detect unknown morphological types of galaxies. The results of their research showed that support vector machines (SVM) and random forest methods provide the highest accuracy for morphological classification of binary galaxies. In another study, [16] used clustering to morphologically classify galaxies and demonstrated that unsupervised machine learning algorithms were powerful in performing accurate morphological analyses. The study [23] used the CNN-based regression model to predict the stage of galaxy integration using images. They showed that their proposed model provides a reasonable estimate based on actual observations, which is almost consistent with previous estimates provided by detailed dynamical modeling. In another study, [24] used machine learning to classify galaxies using images taken from an SDSS source. The results of their study indicated that morphologies based on over-trained traits such as colors, shapes and concentrations with machine monitoring, showed less bias than morphologies based on human recognition. This result is maintained even when there is a fundamental bias in the training sets used in the machine learning process with the observer. The study [25] used simple Bayesian algorithms, random forest, and SVM to classify the morphology of galaxies, which ultimately yielded the best results from random forest. The study [26] also sought to classify galaxies as spiral, elliptical, disk, or other using random forest, decision tree, KNN, and SVM algorithms. The results of their study indicated that the random forest algorithm was more accurate. In another study, [27] used machine learning to classify the morphology of galaxies from visual data of galaxies classified by citizen-scientists and indicated that machine learning performance could be dependent on data quality and can be improved by using examples that have high agreement among citizen scientists. Some of the research on the morphology of galaxies using data mining are presented in Table 1.

According to the studies, the researches have been more focused on the degree of integration, and relatively little effort has been made to study their morphologies and internal characteristics. To accurately understand the process that takes place during the merging of galaxies, measuring the interaction rate alone is not enough because the processes that determine the morphological results of mergers are not yet fully understood. Therefore, due to the increasing ability to collect large volumes of galactic data, the importance of studying this type of data, the problems and shortcomings of traditional methods, and ultimately the need for automated analysis of galactic data, new methods of processing and studying this type of data should be provided with the help of data mining. Hence, this study aims to use various algorithms to improve the accuracy of morphological prediction and merger of galaxies as well as providing a model using near-infrared spectroscopic data with the help

Table 1: Some Studies on the Morphology of Galaxies.

Resource	Objective	Method	Data Resource
[25]	To find the best machine learning methods for detecting unknown morphological types of galaxies from SDSS DR9	Labeling and machine learning algorithms such as Naive Bayes, logistic regression, SVM vector machine, random forest and nearest neighbor	A sample of galaxies from the SDSS DR9 catalog with redshifts of $0.02 < Z < 0.1$ and absolute stellar magnitudes of -24m $<$ Mr $<$ -19.4m
[19]	Morphological classification of galaxies	Clustering	Images of celestial bodies containing objects, including galaxies
[26]	To predict the integration step using the image	CNN-based regression model	Mergers' simulated data
[27]	Classification of galaxies	Machine learning	Images of galaxies from SDSS source
[28]	Morphological classification of galaxies	Naive Bayes, Random Farest and SVM	I-g band color indices Inverse concentration indices from SDSS source
[14]	To classify galaxies as spiral, elliptical, round, disk	Random forest, decision tree, KNN and SVM	Pre-categorized data and Galaxy Zoo images
[29]	Morphological classification of galaxies from the field visual data of galaxies classified by citizen-scientists.	Machine learning	Images of galaxies from SDSS source

of machine learning. Examining various algorithms makes it possible to achieve the highest possible accuracy in prediction by using the appropriate algorithm.

3 Methodology

This study aims to provide a model for predicting the merger and morphology of galaxies through near-infrared spectroscopic data using machine learning. The study was done on real data from the Galaxy Zoo website, which is specifically a collection of galaxy mergers and covers a range of spectral and morphological features of galaxies. This sample of galaxies aggregated from 2010 SDSS Galaxy Zoo data, that is a sample of galaxies which their merging was detected by spectroscopy of at least one of the two galaxies. Galaxy Zoo-data includes the morphologies of the merged galaxies as well as the relative phase of the merger. It is understandable that compiling such a collection would require years of effort and collaboration by large teams of astronomers, and it would not be possible to collect and study such dataset of several thousand galaxies in a short period of time. As a result, no new datasets have been released since 2010 to examine the merger of galaxies on the Galaxy Zoo website and other astronomical databases. Therefore, the studied dataset is the most complete dataset containing the spectroscopic characteristics of merging galaxies.

This dataset contains 3003 records or rows, and each row represents information about a galaxy. It also contains 56 properties that represent the spectroscopic feature of each of the merging galaxies. A description of each feature is summarized in Table 2. After examining and recognizing each of the features, the features that had an ID nature and had no effect on the process and accuracy of the model prediction including 12 features including: of OBJECT (1, 2), PLATE (1, 2), FIBERID (1, 2), MJD (1, 2), RA (1, 2), DEC (2, 1) were removed and only one feature was retained as an ID to identify and differentiate between galaxies.

Based on related studies the variables that indicate the size of the measurement error, including all variables with the extension (ERR) were removed and not used in modeling [28]. Two variables SPECZ2 and COMMENT, which included 2322 and 2856 missing data out of 3003 total data, respectively, were also removed due to the high number of missing values and their outdated data and irreplaceable nature. Properties such as SPECZ2, which include items such as -9999999 or NaN, appear to have correct values, but in reality these numbers represent outdated data and in some cases represent missing data [29]. These values are not incorrect in size but are clearly meaningless. After removing the above properties, 26 features remained in the data set.

Then, the Synthetic Minority Over-sampling Technique (SMOTE) was used to balance the dataset [30].The pre-processing and up-sampling time of minority data takes about 50 minutes. In the next phase, the modeling was performed to differentiate the two minority classes so that all three classes could be predicted. Then, four algorithms of Generalized Linea rModel (GLM), Neural Network, Naive Bayes and Random Forest were used to analyze the accuracy of predictive model.

Data analysis was performed using the Python programming language. The specifications of the hardware used for analysis were: Intel(R) Core(TM) i5-8250U CPU@1.60GHz 1.80 GHz, and 8.00 GB RAM.

Table 2: Features of Galaxies.

Feature	Description Feature	Features Abbreviation
OBJECT1 OBJECT2	SDSS DR7 objID for the first galaxy; SDSS DR7 objID for the second galaxy	The ID of the first galaxy in a pair of galaxies; The ID of the second galaxy in a pair of galaxies
STAGE	visually-classified stage of the merger (1 = ”separated”, 2 =”interacting”, 3 = ”approaching post-merger”)	This variable is the label variable. The phase of merging and interaction of two galaxies includes: 1 = two separate galaxies, 2 = two galaxies are affected by each other (either merging or due to the gravitational pull of each other) and 3 = close to the evolutionary stage of merger

Feature	Description Feature	Features Abbreviation
U-APP-1 G-APP-1 R-APP-1 I-APP-1 Z-APP-1 U-APP-2 G-APP-2 R-APP-2 I-APP-2 Z-APP-2	apparent U-G-R-I-Z band magnitude of the first and second galaxy in the pair	The apparent magnitude of a pair of galaxies is close to infrared. The word appearance means that these numbers do not represent the actual brightness of the galaxy, but the visible light of each galaxy from Earths surface. These numbers are not independent of the distance and the brightness of any object depends on its distance from the earths surface. For example, there are millions of brighter stars around us than the Sun, but what causes the Sun to glow is less distant from Earth than other stars [31].
U-APP-ERR-1 G-APP-ERR-1 R-APP-ERR-1 I-APP-ERR-1 Z-APP-ERR-1 U-APP-ERR-2 G-APP-ERR-2 R-APP-ERR-2 I-APP-ERR-2 Z-APP-ERR-2	Measured uncertainty in apparent U-G-R-I-Z BAND magnitude of the first AND SECOND galaxy in the pair	Uncertainty of measured apparent limited data
U-ABS-1 G-ABS-1 R-ABS-1 I-ABS-1 Z-ABS-1 U-ABS-2 G-ABS-2 R-ABS-2 I-ABS-2 Z-ABS-2	Absolute U-G-R-I-Z BAND magnitude o THE FIRST AND SECOND GALAXY IN THE PAIR in the pair, based on spectroscopic red-shift	The true magnitude of the pair of galaxies at close infrared wavelengths. Here the effect of distance (the case raised in apparent magnitude) is eliminated and the true luminosity of the mass is measured. This means that if an object is known to be bright in this data, it is really bright, and its brightness is not due to its short distance from the Earth. These data primarily represent the temperature of objects, because the colder the mass, the brighter the spectrum. Also, the spectrum accurately indicates how much intensity is received from the galaxy at each wavelength. Finally, the chemical ompositions represent the various elements of the galaxy [28].

Feature	Description Feature	Features Abbreviation
PLATE1 PLATE2 MJD1 MJD2 FIBERID1 FIBERID1	SDSS plate number for the observation of the first AND SECOND GALAXY; SDSS Modified Julian Date for the observation of the first galaxy in the pair; SDSS fiber ID for the spectroscopic observation of the second galaxy in the pair	The number of spectrum extraction tools for each galaxy (they have an ID nature) and the observation date of each galaxy.
SPECZ1 SPECZ2	Spectroscopic redshift for the first AND SECOND galaxy in the pair	Spectral red transfer rate
PHOTOZ1 PHOTOZ2	Photometric redshift for the first AND SECOND galaxy in the pair	The amount of redshift for galaxies that are so dim that it is not possible to measure the amount of redshift by spectroscopy. Consequently, where the spectrum could be measured, it was used, and otherwise, photometry was used. Spectroscopy and photometry are two different methods that measure the amount of red light transmission, differing only in technology and are the same in nature. Photometry is less accurate than spectroscopy but can be measured for all objects.
RA1 RA2 RA1 RA2	Right ascension (J2000, decimal degrees) for the first AND SECOND galaxy in the pair; Declination (J2000, decimal degrees) for the first AND SECOND galaxy in the pair	Dimension and desire to locate and coordinate each galaxy in the sky.
KMASS1 KMASS2	stellar mass (log M/M_sun) of the first AND SECOND galaxy in the pair	The logarithm of the mass of a galaxy relative to the mass of the sun (the mass of a galaxy is many times the mass of the sun). This variable is used to measure the mass and population of stars in a galaxy.
KMASS_ERR1 KMASS_ERR2	uncertainty in stellar mass (log M/M_sun) of the FIRST AND SECOND galaxy in the pair	Uncertainty of galaxy mass logarithm measurement

4 Data Analysis And Results

4.1 Modeling And Model Evaluation

At this phase, first minority and the majority classes were identified. The STAGE feature had three classes with unbalanced distribution that represented the phase of merging and interaction of two galaxies, including: 1 = two galaxies are apart, 2 = two galaxies are under the influence of each other (either merging or due to gravitational pull of each other) and 3 = they are close to the evolutionary stage of merging. Class 1 had 167 sample data, Class 2 had 2526 sample data and Class 3 had 310 samples from the total sample of this dataset. Since most classification algorithms focus on the more frequent sample and either ignore or incorrectly classify the minority sample in the unbalanced dataset, resolving the imbalance problem is one of the most important steps in processing this type of data [33]. To solve the problem of unbalanced data, modeling was performed in two separate phases. First, by merging the two minority classes, the first three data classes were reduced to two classes to increase the ability to detect and process the algorithm between the minority classes (labels 1 and 3) and the majority class (label 2) and the algorithm error (due to Unbalanced data) to be reduced to a minimum. The SMOTE method was used to balance the data set. This method uses the original data to generate and simulate new data and at each stage is able to generate a new sample for only one of the minority classes. This method was applied in parallel to each class, and then the new data generated by this method were integrated into a table.

After implementing different classification models on the balanced data using SMOTE, acceptable accuracy was not obtained in the results. The average accuracy was equal to 45.19%. According to [33] to increase the accuracy of modeling, it is better to use both sampling methods (over sampling and sampling) simultaneously. For this purpose, in the next step, using the module (SAMPLE), the class (2) sample was reduced and so-called sampling was performed. It should be noted that during sampling, all minor data in each class were sampled from minority classes. Classification models appropriate to this dataset were also applied to the obtained data. After performing the classification models, the average accuracy was equal to 48.20%.

The next step includes the increasing the accuracy of the model, changing the classification and data class. By merging and assigning the label (yes) to two classes 1 and 3 and the label (no) to the majority class (2), two new labels were created. By merging the two minority classes, the frequency of the Yes class reached 477 in total, and the No class (label 2) with 2526 samples remains the majority class. The problem of imbalance between classes remains strong at this phase. Then, in order to bring the frequencies of the two classes closer to each other, oversampling and undersampling methods were performed on the new dataset. The sampling was repeated several times with different numbers and tested with modeling to find the optimal value after reducing the frequency of class No. Finally, the number of obtained samples in the final classification with the best accuracy in modeling was 1650 samples from the majority class and 1477 samples from the minority class.

Then, four algorithms of GLM, Neural network, Naive Bayes and Random Forest were used to investigate the accuracy of predictive models.Algorithms used in this study were not meta-heuristic except for neural network, therefore they did not have the problem of convergence and local minimum, but hyperparameters were set for neural network in the software and were estimated based on model evaluation criteria.

The maximum cost of the algorithms in terms of run time was for Neural Network with about 80 minutes, followed by Random Forest (50 minutes), GLM (40 minutes), and Naive Baise (about 30 minutes). The results of the implementation of these four algorithms are given in Table 3.

As Table 3 indicates, the accuracy of the random forest is higher than all other models, but this accuracy is due to the overfitting of the model on Class 2 data. Thus, the best result is obtained from the GLM algorithm, which predicts both classes to an acceptable level.

After distinguishing between minority and majority classes another model is required for the minority class data that is able to distinguish between class 1 and 3. For this purpose,

Table 3: Accuracy of Running Four Machine Learning Algorithms on All Classes.

Algorithm	Precision (N)	Precision (P)	Recall (N)	Recall (P)	Accuracy
Nave Baise	81.01%	22.86%	11.64%	90.57%	29.34%
Random Forest	77.77%	26.60%	95.82%	5.24%	75.51%
GLM	86.31%	39.45%	73.03%	60.79%	70.28%
Neural Network	81.22%	30.04%	70.79%	43.40%	64.65%

four used machine learning algorithms were run again on the data of class 1 and 3 (Table 4).

Table 4: Accuracy of Running Four Machine Learning Algorithms on class 1 and 3.

Algorithm	Precision (1)	Precision (13)	Recall (1)	Recall (3)	Accuracy
Nave Baise	54.08%	78.29%	63.47%	70.97%	68.36%
Random Forest	77.77%	26.60%	67.39%	72.73%	71.66%
GLM	67.74%	80.75%	62.87%	83.87%	76.51%
Neural Network	66.67%	80.37%	62.28%	83.23%	75.91%

According to Table 4, the GLM algorithm has the best results compared to other algorithms with a very small difference from the Neural Network algorithm.

5 Conclusions

Data plays a vital role in astronomy, and its size and complexity are rapidly increasing. Given this rapid growth, astronomers are developing automated tools to identify, describe, and classify objects using rich and complex datasets collected with a variety of features. Since the processing of data in this field is very difficult and heavy and is beyond the power of human alone, machine learning algorithms have become increasingly popular among astronomers and are widely used for a variety of tasks. Despite the widespread use of machine learning in accurately understanding the process that takes place during the integration of galaxies, few studies have addressed this issue. To fill this gap, this study seeks to find and provide a solution for processing astronomical data, with the aim of predicting the merger of galaxies with the help of machine learning algorithms. The use of machine learning can be considered as one of the innovations of this research. In this study, prediction methods and classification algorithms have been used to classify galaxies based on their degree of merging. The GLM and Naive Bayes algorithms along with other algorithms used in previous studies (i.e., Neural Network and Random Forest) were used to process and classify the data extracted from the images. Using the GLM algorithm, acceptable accuracy was obtained for predicting the merger of galaxies, which can be another innovation of this study.

Algorithms do not always make moral or careful choices. There are some reasons for this. One is that algorithms make many predictions, some of which are likely to be wrong. The probability of error depends on many factors, including the amount and quality of data used to train the algorithms, the specific type of machine learning method chosen, which may not allow it to maximize accuracy. Second, the environment in which machine learning operates may be self-evolving or different from what the algorithms were developed to deal with [34]. Besides, algorithms do not necessarily work the same on all data. Based on the data, the performance of the algorithms are different [35]. Algorithms used in this study worked better on the studied data.

Besides, the best results were obtained when in the preprocessing phase, in addition to removing features and data augmentation, the SMOTE was performed on imbalanced data and both oversampling and Undersampling methods were performed simultaneously. Classification of data into two groups, minority and majority, and then prediction in two stages in order to solve the problem of severe imbalance between the data under study, along with methods such as SMOTE sampling, which were used alone in previous studies, are other innovations of this study.

In the study [28], which was performed on a similar dataset, using the tree algorithm and information gain index, 70% accuracy was obtained for classifying galaxies. In this study, the obtained accuracy was 70.28% in the first phase, which was between minority and majority classes, and 76.51% in the second phase, which was between minority classes.

Since astronomical data are very similar in nature and distribution, the preprocessing and modeling methods used in this study can pave the way for future research and guide them in achieving better results.

Astronomers always want to know the answer to the question, how did they get to this point? What made galaxies and galaxy clusters, superclusters, cavities, and strings look like this? The existence of such large strings of galaxies and orbits is an interesting mystery. The challenge for theorists is to understand how a world almost uncharacteristically transformed into the complex and massive world we see today. As data science has become an integral part of astronomy, the results of this study may eventually lead to the use of machine learning to solve a variety of unsolved astro-physical mysteries of galaxies and the universe as a whole. Besides, galactic period maps with the help of advanced technology allow scientists to identify galaxies that were difficult to observe in the past and to obtain more data on the evolution, size and shape of galaxies.

In this study, some features were extracted and used for modeling. Thus, it is suggested that future studies extract more and different features from the present study or combine different tables of features using astronomical knowledge. Because, in astronomical problems, there are many features that can be used to improve the prediction and classification accuracy by consulting experts and extracting optimal features with specific problem transformations. It is also suggested to use a deep learning approach to solve these problems.

6 Acknowledgment

We would like to thank Fatimah Shojai for valuable comments and useful discussions.

References

[1] Karypidou, S., Georgousis, I., & Papakostas, G. A. 2021, ICPIC, 94.

[2] Dere, S., Fatima, M., Jagtap, R., Inamdar, U., & Shardoor, N. B. 2021, ICACCS, 1, 702.

[3] Dieleman, S., Willett, K. W., & Dambre, J. 2015, MNRAS, 450, 1441.

[4] Lvdal, S. 2021, PhD thesis, Univ. Groningen.

[5] Ivezić, ., Kahn, S. M., Tyson, J. A., & *et al.* 2019, ApJ, 873, 111.

[6] Itschner, S., & Li, X. 2019, RadarConf., 1.

[7] Becker, A. B. 2022, Master Thesis, Univ. Stellenbosch.

[8] Lintott, C., Schawinski, K., Bamford, S., & *et al.* 2011, MNRAS, 410, 166.

[9] Vavilova, I., Pakuliak, L., Babyk, I., & *et al.* 2020, KDBDAO, 57.

[10] Graff, P., Feroz, F., Hobson, M. P., & Lasenby, A. 2014, MNRAS, 441, 1741.

[11] Grinin, L. E. 2020, EvolutionYearbook Editors Council, 37.

[12] Conselice, C. J. 2014, Annual Review of A&A, 52, 291.

[13] Kasivajhula, S., Raghavan, N., & Shah, H. 2007, MNRAS, 8, 1.

[14] Hani, M. H., Gosain, H., Ellison, S. L., Patton, D. R., & Torrey, P. 2020, MNRAS, 493, 3716.

[15] Bertone, S., & Conselice, C. J. 2009, MNRAS, 396, 2345.

[16] Yesuf, H. M., Ho, L. C., & Faber, S. M. 2021, AJ, 923, 205.

[17] Martin, G., Kaviraj, S., Hocking, A., Read, S. C., & Geach, J. E. 2020, MNRAS, 491, 1408.

[18] Banerji, M., Lahav, O., Lintott, C. J., Abdalla, F. B., & *et al.* 2010, MNRAS, 406, 342.

[19] Psychogyios, A., Charmandaris, V., Diaz-Santos, T., Armus, L., & *et al.* 2016, A&A, 596, 1.

[20] Baron, D. 2019, arXiv:1904.07248.

[21] Prabakaran, S., & Mitra, S. 2018, April. JPCS. IOP Publishing, 1000, 012046.

[22] Vavilova, I. B., Dobrycheva, D. V., Vasylenko, M. Y., Elyiv, A. A., & *et al.* 2021, A&A, 648, A122.

[23] Koppula, S., Bapst, V., Huertas-Company, M., Blackwell, S., & *et al.* 2021, arXiv:2102.05182.

[24] Cabrera-Vives, G., Miller, C. J., & Schneider, J. 2018, AJ, 156, 284.

[25] Dobrycheva, D. V., Vavilova, I. B., Melnyk, O. V., & Elyiv, A. A. 2017, E-print arXiv. Org.

[26] Gauthier, A., Jain, A., & Noordeh, E. 2016, Lecture Notes, Univ. Stanford, 16, 1.

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

[28] Baehr, S., Vedachalam, A., Borne, K. D., & Sponseller, D. 2010, CIDU, 133.

[29] Ball, N. M., & Brunner, R. J. 2010, IMP, 19, 1049.

[30] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. 2002, JAIR, 16, 321.

[31] Kelly, B. C., & McKay, T. A. 2004, AJ, 127, 625.

[32] Burez, J., & Van den Poel, D. 2009, Expert Syst Appl, 36, 4626.

[33] Liu, S., Ong, M. L., Mun, K. K., Yao, J., & Motani, M. 2019, CinC, 1.

[34] Babic, B., Cohen, I. G., Evgeniou, T., & Gerke, S. 2021, HARV. BUS. REV.

[35] Ackermann, A. 2022, HARV. TECH. RIGHTS.