WEP Warwick Evans Publishing

Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

Application of Five Machine Learning Techniques for Stellar Classification on SDSS DR18

Haiqi Su *

College of Electronic and Information Engineering, Tongji University, Shanghai, China

* Corresponding Author Email: su_haiqi@tongji.edu.cn

Abstract. Systematic classification of celestial objects is a cornerstone of modern astronomy, as it underpins the understanding of the universe’s large-scale structure and long-term evolution. Moreover, it facilitates the discovery of rare or extreme phenomena, offering clues to the underlying physical laws that govern them. This study presents a comprehensive evaluation of five machine learning models—Random Forest, Adaptive Boosting, Extreme Gradient Boosting, Deep Neural Network, and TabNet—on the Sloan Digital Sky Survey Data Release 18 dataset for the classification of stellar objects into GALAXY, STAR, and Quasi-Stellar Object. With comparative analysis, although the models are not specially tuned, all models achieved competitive performance with accuracy above 96.8%, with Extreme Gradient Boosting reaching an accuracy 98.5%. The reasons for classification are explained by feature importance analysis with SHAP values, which revealed that redshift and the color index E(u−g) are the most influential attributes in distinguishing object classes. Additionally, the study identified several confusing features that contribute to misclassification, suggesting the need for improved data quality in these dimensions. The research highlights the dual role of machine learning in both automation and discovery-driven astronomy.

Keywords: Machine Learning; Stellar Classification; SDSS Dataset; Explainability.

1. Introduction

The accurate classification of stellar objects is of paramount importance in the pursuit of scientific understanding. By providing a foundation for the comprehension of the universe's structural evolution, this process also unveils the nature of extreme physical processes.

In recent years, breakthroughs have emerged in the field of machine learning (ML) for the classification of stellar objects.

For classification algorithms, the advent of deep learning has led to significant advancements in the ability to process non-linear data. Li et al. observed that their enhanced Xception-AS algorithm, derived from the Xception architecture, attained 90.26% accuracy in the triple classification of galaxies, stars, and quasars on SDSS-DR16 data [1]. Concurrently, the Morpheus deep learning framework has attained a hitherto unparalleled level of sophistication by attaining pixel-level classification of celestial objects [2]. This unprecedented feat has been achieved by ingeniously integrating the Hubble legacy data with CANDELS annotations. This landmark development has effectively addressed the vexing challenge of hybrid identification of complex celestial morphology.

ML classification models in the physical domain require not only superior performance, but also physical interpretability. A substantial corpus of research has been dedicated to the interpretability of ML-aided stellar classification results. Dey's team utilized SHAP values to analyse the feature extraction results of the deep capsule network, thereby elucidating the network's behavior [3]. When studying quasi-periodic oscillations in X-ray binaries, Kiker et al. also implemented SHAP values to the ML approach [4]. The aforementioned approaches are instrumental in facilitating a more profound integration of physical mechanisms with data-driven models.

However, there is still a paucity of comparative studies on the performance of traditional and emerging models on the same task, especially on the features learned by different models.

This paper employs data analysis and comparison methods to compare Random Forest (RF), Adaptive Boosting (Adaboost), Extreme Gradient Boosting (XGBoost) , Deep Neural Network (DNN), and TabNet. Furthermore, it utilises SHAP to interpret the classification results of the models. The objective of this study is to facilitate the identification of anomalous samples and to promote the advancement of data-driven astronomical research.

2. Dataset

The Sloan Digital Sky Survey (SDSS) has issued its eighteenth data release (DR18). The input catalogs and selection algorithms provided in this data release serve a wide range of scientific objectives, offering essential targeting information for two multiobject spectroscopy programs. [5]. SDSS utilizes five filters to determine the colors of star objects: ultraviolet (u), green (g), red(r), near infrared (i), and infrared (z). These filters function best at wavelengths 3543, 4770, 6231, 7625, and 9134 Å, respectively. [6]. These magnitudes can be converted into UBVRI magnitudes using a collection of algorithms [7].

The dataset is retrieved from SDSS DR18 [8]. The dataset includes 100,000 observations from SDSS DR18. Each observation is defined by 42 features and one class column, which classifies it as a GALAXY, STAR, or Quasi-Stellar Object (QSO).

2.1. Dataset Overview

The distribution of the observed targets in the all-sky star map is shown in Figure 1. A total of 42 features are used to characterize the observed targets in the data set. Figure 2 depicts a heat map representing the correlation between these 42 features.

In the obtained data set, the proportion of three types of celestial bodies, GALAXY, STAR and QSO, is shown in Figure 3. In the five bands of u, g, r, I, z, the density distribution of the three types of celestial bodies in the data set is shown in Figure 4. The extreme values of the magnitude of the u, g, r, I, z bands in the data are shown in Table 1. In particular, the minimum redshift is -0.004268 and the maximum is 6.990327.

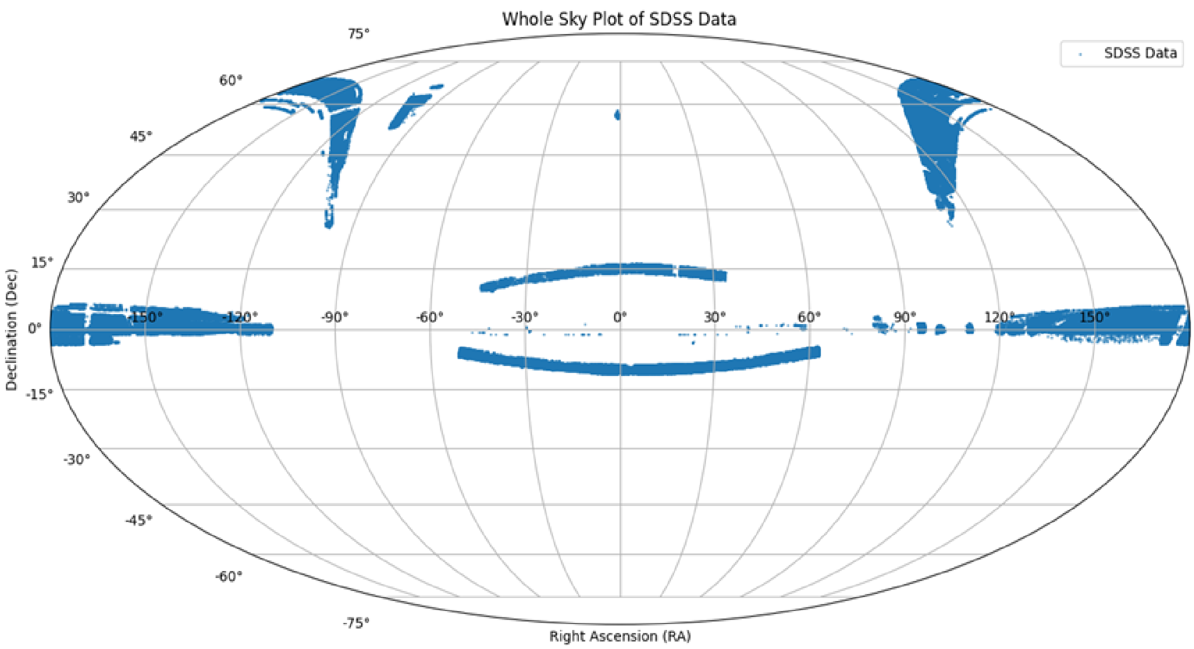


Figure 1. Whole Sky Plot of the Dataset (Picture credit: Original)

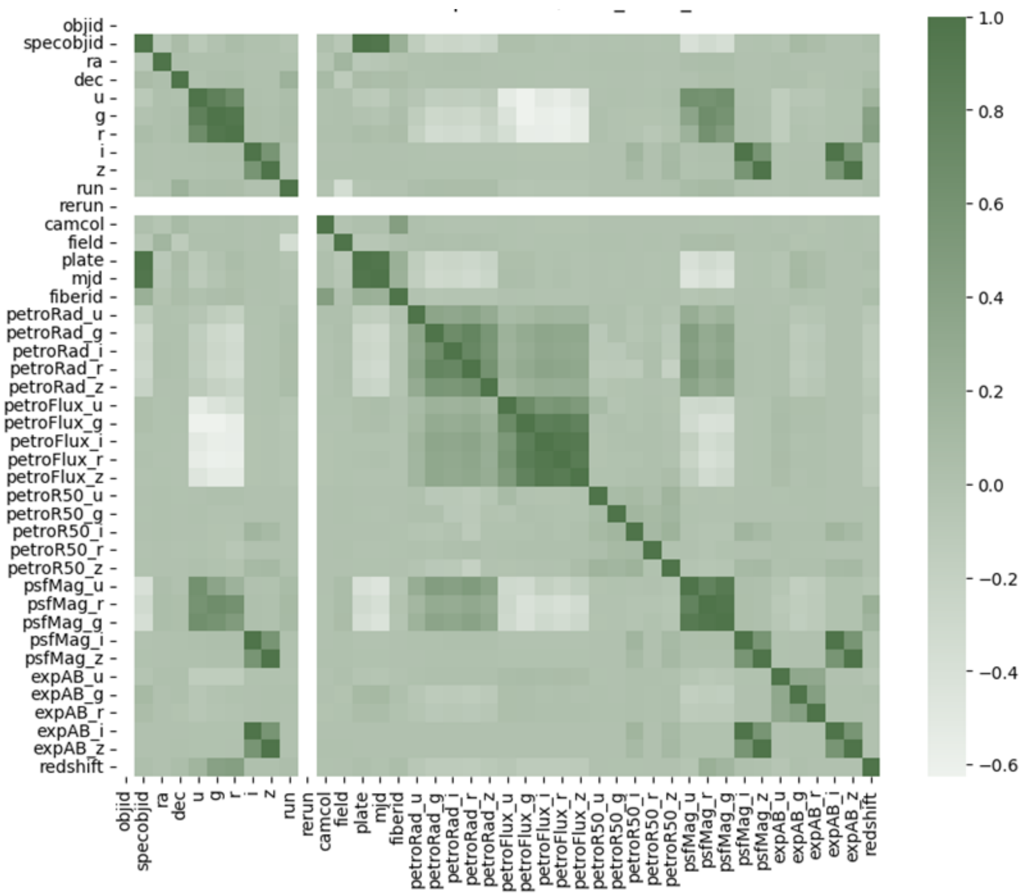


Figure 2. Correlation Heatmap for the Dataset (Picture credit: Original)

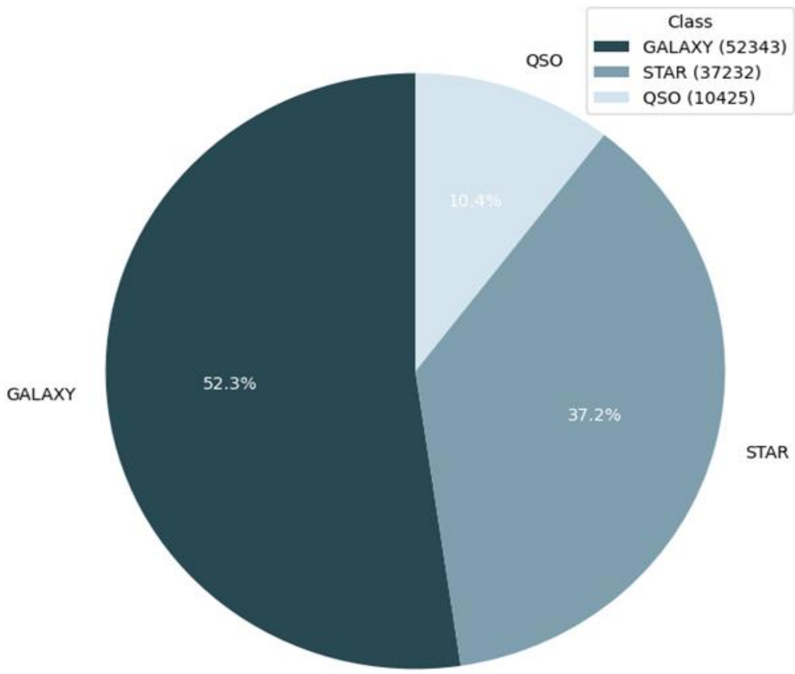


Figure 3. Distribution of the three classes (GALAXY, STAR. QSO) in the dataset (Picture credit: Original)

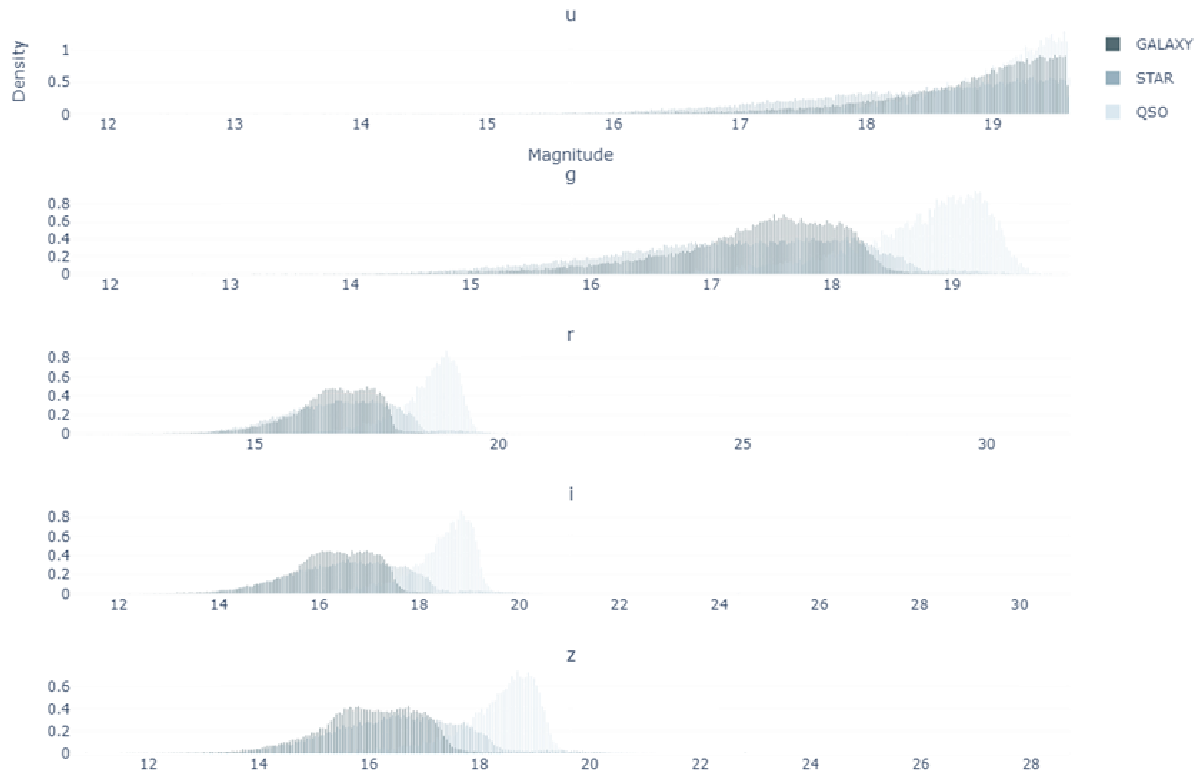


Figure 4. Distribution of u, g, r, I, z colored by class (GALAXY, STAR, QSO) (Picture credit: Original)

Table 1. The extreme magnitudes at filters u, g, r, I, z

	Min	Max
u	11.72647	19.59999
g	11.69617	19.97727
r	11.27709	31.69816
i	11.05139	30.98087
z	10.61626	28.66870

2.2. Processing

The dataset under consideration characterizes the target with 42 features, which can be divided into six categories. The initial category of data encompasses the target's right ascension, declination and redshift, which are associated with the object's inherent characteristics, thus retained. The subsequent category of data comprises the target's AB magnitude in the u, g, r, I, and z bands, which are likewise retained. The third category of data is technical features, including "objid", "rerun", "run", "camcol", "specobjid", "plate", "mjd", "fiberid" and "field". They pertain to the acquisition of the data, rather than the objects themselves, and are designated for removal. The fourth type of data is the Petrosian magnitude, which is primarily utilised to assess the characteristics of galaxy photometry. However, it should be noted that the primary data has already been optimised, and as such, its removal is recommended. The fifth type of data is PSF magnitudes, which are principally utilised for the characterization of quasars. However, it should be noted that the primary data has already undergone optimization, and as such, its removal is recommended. The sixth type of data is exp, which can be inferred from the second type of data, therefore it is dropped.

As demonstrated in Figure 3, the dataset is imbalanced across categories, with the GALAXY category accounting for more than half of the total data. Considering the detrimental effect brought by an imbalanced dataset, the research employs under sampling to balance the proportion of each class in the dataset. The paper calls the sklearn tool for data processing, and performs no-return sampling on the samples, setting the random seed as 42. Following the process of undersampling, the number of items in each class (i.e. GALAXY, STAR and QSO) is found to be 10,425.

In the uniformly isotropic Friedmann-Lemaître-Robertson-Walker universe, the photometric distances of objects are a function of redshift:

$$d_L(z) = (1 + z) \cdot \frac{c}{H_0} \int_0^z \frac{dz'}{E(z')} \tag{1}$$

With

$$E(z) = \sqrt{\Omega_m(1 + z)^3 + \Omega_\Lambda + \Omega_k(1 + z)^2 + \Omega_r(1 + z)^4} \tag{2}$$

Where, c is the light speed, H0 is the current Hubble constant, E(z) is the dimensionless Hubble parameter, and the cosmological constants Ω_m,Ω_Λ,Ω_k,Ω_r are the matter density parameter, the dark energy density parameter, the space curvature parameter, and the radiation density parameter, respectively [9].

The redshift column is already present in the dataset, and so the redshift is primarily employed to denote the distance factor of the celestial body. The AB magnitudes of the u, g, r, i, z bands are substituted with the magnitude difference of adjacent bands, that is, the color index, in order to facilitate feature discovery by the ML model.

The present study employs label encoding to reduce feature dimensions and enhance model efficiency. Following the process of data handling, the final features of the data are as follows: r, a, z, u_g, g_r, r_i and i_z.

3. The Methods

3.1. Random Forest

RF is an ensemble learning technique [10]. In the context of classification problems. The model assumes that a strong learner, comprising multiple decision trees, exhibits superior performance in the classification of samples in comparison to a single decision tree. Therefore, the model employs the algorithmic approach of bagging, which involves constructing multiple decision trees to classifying the samples individually. Ultimately, the final category of the samples is determined by a voting process.

In this research, Gini impurity is to be minimized when dividing the dataset in the weak learners' internal node. This method has been demonstrated to offer a more objective approach, as it is not subject to the potential biases that can arise from manual screening.

RF offers both temporal and spatial advantages for stellar classification tasks. Temporally, the parallelized training of RF can significantly reduce computation time, allowing for efficient completion of stellar classification. Spatially, the seven-dimensional features are much smaller than the number of samples. The decision tree structure of RF effectively captures interactions between features, helping to avoid dimensionality explosion.

3.2. Adaptive Boosting

AdaBoost is an ensemble learning method that improves classification by focusing on hard-to-classify samples [11]. Unlike RF, it updates sample weights after each iteration, where misclassified samples receive higher weights, while correctly classified ones are down-weighted. Each new weak learner is trained on the reweighted data. The resulting model a weighted majority vote of all of them, where weights reflect individual accuracy.

This research addresses the multiclassification problem by employing a stagewise additive modeling approach with a multiclass exponential loss function (SAMME), an advanced algorithm based on Adaboost. The research proposes a default model with a random state of 42 to ensure replicability.

The Adaboost algorithm is of particular significance due to its capacity to integrate other algorithms, such as DNN. In this study, the decision boundary undergoes a gradual refinement through boosting, which is more likely to effectively distinguish AGN from QSO.

3.3. Extreme Gradient Boosting

XGBoost is an optimization algorithm based on gradient boosting [12]. It models predictions by adding a series of decision trees one after the other, each of which fixes the residual mistakes of the ones before it. XGBoost improves upon traditional gradient boosting by using a second-order Taylor expansion of the loss function, incorporating both the first-order gradient and second-order Hessian, which enables more precise optimization. Additionally, it applies L1 and L2 regularization to the model's leaf weights, which helps control complexity and reduce overfitting. These innovations make XGBoost both highly accurate and computationally efficient for structured data tasks.

The goal function of XGBoost is:

$$\mathcal{L}(\phi) = \sum_{\{i=1\}}^{n} L(y_i, \hat{y}_i) + \sum_{\{k=1\}}^{K} \Omega(f_k) \quad (3)$$

Where, $L(y_i, \hat{y}_i)$ is the loss function. This paper uses logistic loss. $\Omega(f_k)$ is the regularization term penalizing complexity in the kth tree.

The prediction in iteration t is

$$\hat{y}_i^{\,t} = \hat{y}_i^{\,t-1} + \eta f_t(x_i) \quad (4)$$

Where, $\hat{y}_i^{\,t}$ and $\hat{y}_i^{\,t-1}$ are the predictions from the model after t-1and t trees respectively, η is the learning rate, $f_t(x_i)$ is the function to be added in iteration t.

3.4. Deep Neural Network

DNN has strong nonlinear modeling ability when dealing with complex data. Through multiple layers of transformations, it learns a differentiable mapping from input features to output categories. The forward propagation of input data occurs in layers, with each layer contributing to the computation of predictions. Then, the model uses backpropagation to calculate the gradients of the loss function with regard to its parameters. These gradients are used in a gradient descent technique to iteratively update the weights and reduce prediction error. The training approach relies on differentiable activation and loss functions, allowing for efficient learning via gradient-based improvements.

In this study, a DNN was constructed. The model comprises three layers: the input layer which is also the first hidden layer, the second hidden layer, and the output layer. The first layer contains 64 neurons and utilizes the rectified linear unit (ReLU) activation function. The second layer, which consists of 32 neurons, likewise uses ReLU activation function. 3 neurons are contained in the output layer, which utilizes the Softmax activation function. The efficacy of DNNs is dependent on the consistency and independence of data distribution. Consequently, data regularization is required before beginning the training procedure.

3.5. TabNet

TabNet is a deep learning network intended for tabular data that employs an attention mechanism. Its core innovation lies in its sequential attention, which progressively selects and refines feature representations through multiple decision steps. At each step, the model applies masks to focus on different subsets of features, promoting interpretability and efficient learning. Finally, the refined representations are passed to a decision layer to produce a multiclass classification output.

This study constructs TabNet in the structure shown in Figure 5. First, the input is mapped into vectors and normalized for each feature dimension using embedding generator and initial batchnorm. Then, in the encoding stage, the feature representation is first extracted using a feature transformer, and then dynamic feature selection and information reuse are realized by attentive transformer. Finally, the feature representations from each step are weighted and aggregated, and mapped to the category space through the fully-connected layer, completing the final mapping. the figure is drawn based on the structural code of the output of the trained model.

TabNet's Feature Transformer employs Gated Linear Unit (GLU) blocks to perform nonlinear transformations, which help capture complex interactions among astrophysical features. However, the sequential nature of TabNet's decision-making can lead to cascading errors, where suboptimal feature selections in early steps influence the effectiveness of later steps.

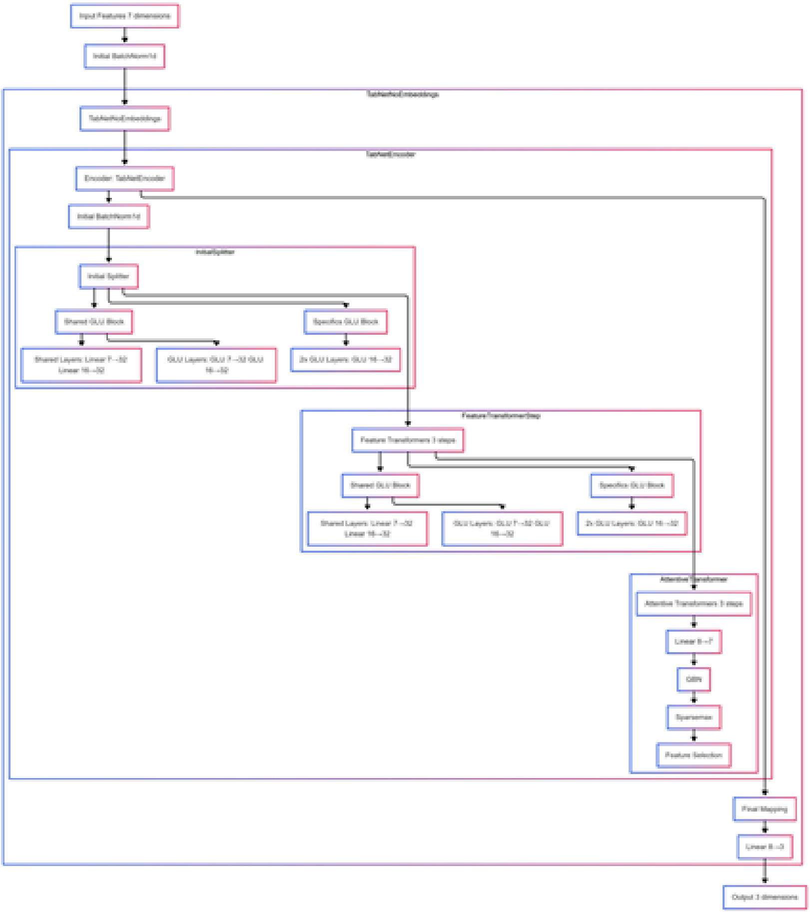


Figure 5. The structure of TabNet (Picture credit: Original)

4. Performance of the Models

The research employed RF, Adaboost, XGBoost, DNN and TabNet to classify stellar objects into galaxies, stars and QSOs, based on the location, redshift and color index of them. The score of the models is shown in Table 2.

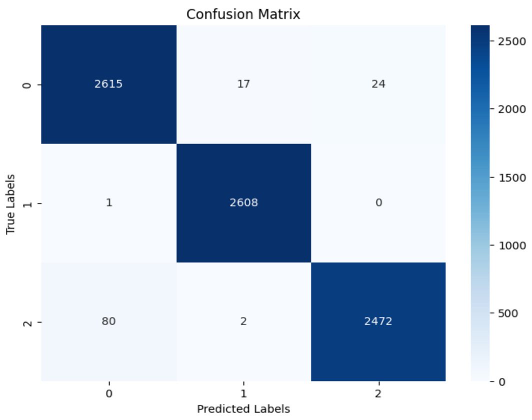
Table 2. Comparison of five models

Model	Accuracy	Precision	Recall	F1 Score
RF	0.9841	0.9842	0.9841	0.9841
Adaboost	0.9767	0.9769	0.9767	0.9767
XGBoost	0.9850	0.9851	0.9850	0.985
DNN	0.9688	0.9689	0.9688	0.9688
TabNet	0.9832	0.9834	0.9832	0.9832

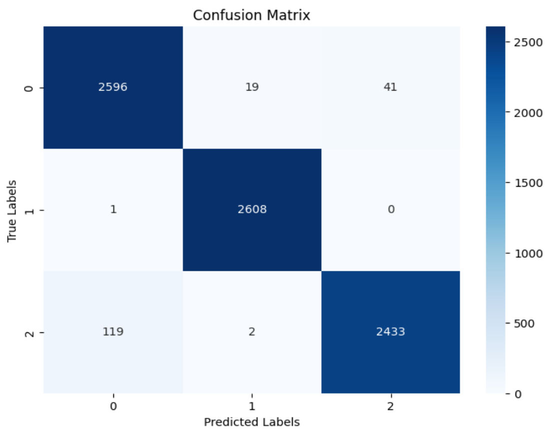
As illustrated in Table 1, XGBoost demonstrates the highest performance metrics in Accuracy, Precision, Recall, and F1 Score, with values of 0.9850, 0.9851, 0.9850 and 0.9850, respectively. Indirect evidence suggests a substantial nonlinear relationship between the dataset's characteristics and the prediction outcomes, which are predominantly influenced by a limited number of features. In the endeavor to categorize objects according to color index, position, and redshift, XGBoost evinces its remarkable aptitude for discerning salient features and its resilience, which is attributable to methodologies such as gradient boosting.

Although XGBoost achieved the highest scores in each category, the performance differences among the five ML models were not substantial. A closer analysis shows that RF, XGBoost, and TabNet form the top-performing group, with an average score of 0.9820, followed by AdaBoost with 0.9768, and DNN with 0.9688 across all evaluation metrics.

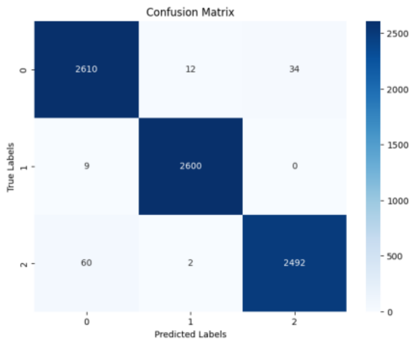
Figure 6 presents the confusion matrices for all five models. Diving into the matrices, it is revealed that RF, Adaboost, XGBoost and TabNet tend to misclassify QSO as GALAXY. In contrast, DNN is more likely to misclassify GALAXY as STAR. Notably, compared with DNN, the first four models show a lower rate of misclassification for Class1 (STAR).



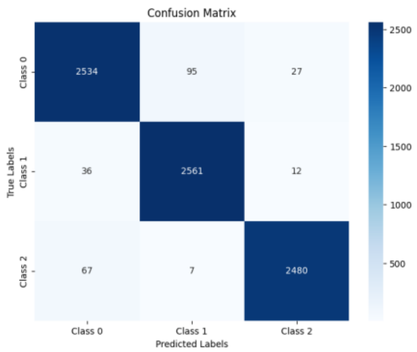
(a)



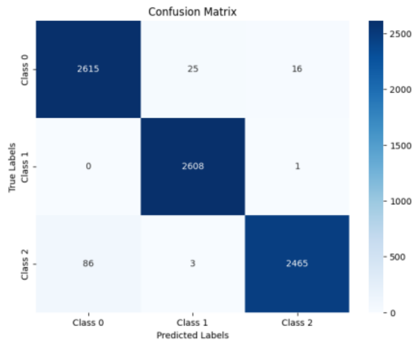
(b)



(c)



(d)



(e)

Figure 6. Confusion Matrix. (a) belongs to RF, (b) Adaboost, (c) XGBoost, (d) DNN, (e) TabNet. (Picture credit: Original)

5. Discussions

The test set includes seven features relevant to the classification task. To understand model predictions, this work uses SHAP values [13], which provide a quantitative estimate of each feature's contribution to the classification conclusion. As illustrated in Figure 7, the SHAP values of RF, XGBoost, DNN, and TabNet are summarized for all predictions in the test set. Across all four models, redshift consistently emerges as the most influential feature. This finding aligns with established astrophysical knowledge, where typical redshift values follow the order: stars < galaxies < quasars.

The results from this feature attribution analysis corroborate known physical properties of different celestial object types, reinforcing the interpretability of the ML models.

Regarding the selection of the second most important feature, notable differences emerge among the four models. A comparative analysis shows that RF, XGBoost, and TabNet consistently identify the u−g color index as the second most influential feature, whereas the DNN assigns it to the r−i index. This divergence may be due to significant flux variations among stars, galaxies, and quasars in the u and g bands, making u−g a more discriminative feature for traditional ensemble methods.

In terms of feature attribution variability, the DNN distributes its attention more uniformly across multiple features, while the other three models display a stronger focus on a few dominant features. This broader attention distribution in DNN may result from an insufficient training dataset size, which limits the network's ability to extract the most salient distinguishing characteristics. Consequently, this may contribute to DNN's higher tendency to misclassify galaxies as stars, as it fails to effectively isolate the key differentiating features between the two classes.

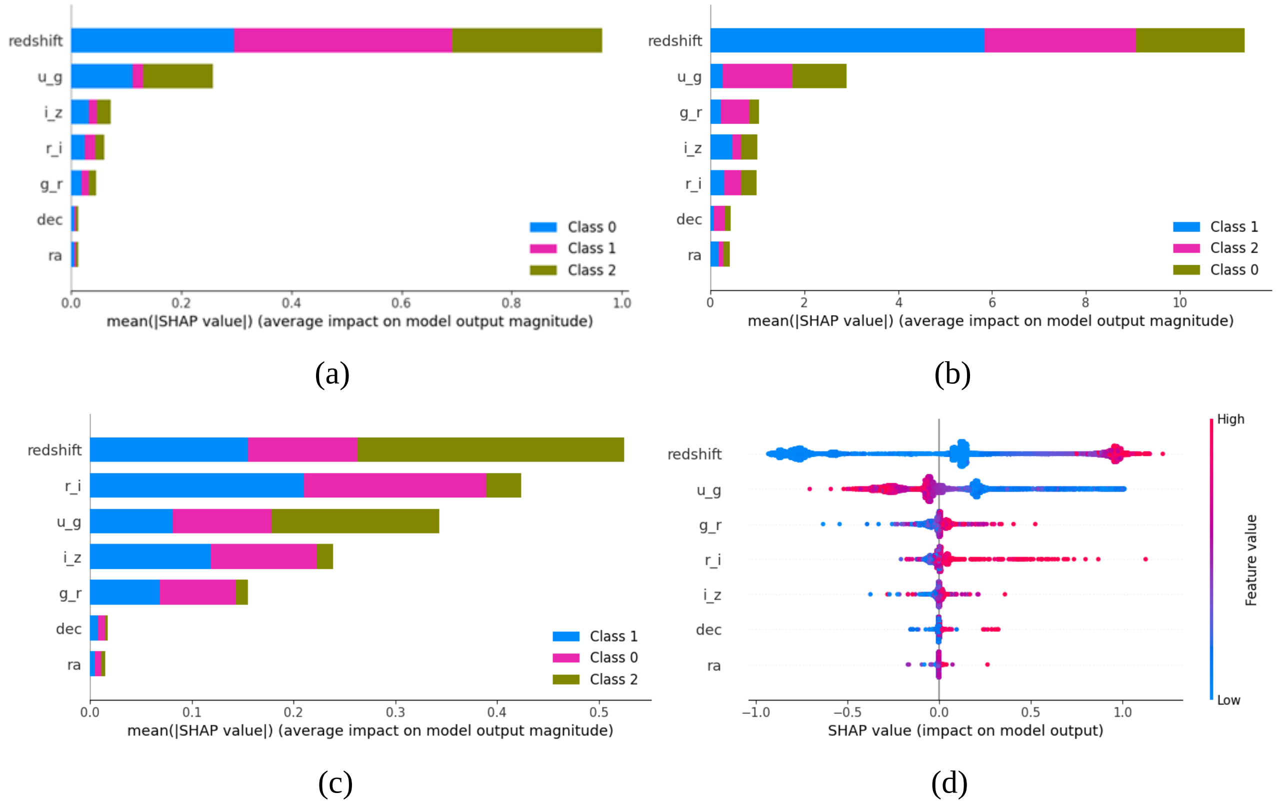


Figure 7. SHAP values summary plot, where (a) is from RF, (b) is from XGBoost, (c) is from DNN, (d) id from TabNet. (Picture credit: Original)

6. Conclusion

In this study, we systematically evaluated the performance of five ML models—RF, AdaBoost, XGBoost, DNN, and TabNet—on the SDSS dataset for the task of stellar object classification. All models demonstrated strong performance in distinguishing between GALAXY, STAR, and QSO, underscoring the effectiveness of ML in astrophysical classification tasks.

Beyond performance metrics, this paper analyzed the underlying features contributing to model decisions. Notably, redshift and the color index E(u−g) emerged as the most influential features, serving as key discriminators among stellar object types. This work provides a valuable characterization of model behavior in this domain, serving both as a benchmark and a practical guide for future astronomical classification pipelines.

Crucially, this paper also identified features prone to confusion, which often lead to misclassifications. These ambiguous cases frequently correspond to outlier objects—non-typical stellar sources whose features are either buried in noise or represent astrophysically complex systems beyond the reach of shallow ML models. The presence of such outliers highlights the need for higher-fidelity data collection on specific features, and it serves as a cautionary insight for future surveys aiming to minimize classification error.

Ultimately, this study bridges automated classification with scientific inquiry, laying the groundwork for using ML not only as a filter, but as a lens for discovery in modern astronomy.

References

[1] Cheng Qiyun, Sun Caixin, Zhang Xiaoxing, et al. Short-Term load forecasting model and method for power system based on complementation of neural network and fuzzy logic. Transactions of China Electrotechnical Society, 2004, 19 (10): 53 - 58.

[2] Li Xin, Tu Liangping, Li Juan, Gao Xiang, Feng Xueqi, Zhong Zhengdi, “Xception-AS: An Automatic Object Classification Algorithm Based on the Structure of Xception Model,” Astron. Res. Technol., vol. 2023, no. 20 (3), pp. 267 - 274.

[3] Hausen, R., & Robertson, B. E., “Morpheus: A Deep Learning Framework for the Pixel-level Analysis of Astronomical Image Data,” Astrophys. J. Suppl. Ser., vol. 248, no. 1, p. 20, May 2020.

[4] Dey, B., Andrews, B. H., Newman, J. A., Mao, Y.-Y., Rau, M. M., & Zhou, R., “Photometric redshifts from SDSS images with an interpretable deep capsule network,” Mon. Not. R. Astron. Soc., vol. 515, no. 4, pp. 5285– 5305, Aug. 2022.

[5] Kiker, T. J., Steiner, J. F., Garraffo, C., Méndez, M., & Zhang, L., “QPOML: a machine learning approach to detect and characterize quasi-periodic oscillations in X-ray binaries,” Mon. Not. R. Astron. Soc., vol. 524, no. 4, pp. 4801 – 4818, Jul. 2023.

[6] Almeida, A. et al., “The Eighteenth Data Release of the Sloan Digital Sky Surveys: Targeting and First Spectra from SDSS-V,” Astrophys. J. Suppl. Ser., vol. 267, no. 2, p. 44, Aug. 2023.

[7] “SDSS Filters,” SDSS Filters. Accessed: May 02, 2025. [Online]. Available: https://skyserver.sdss.org/dr1/en/proj/advanced/color/sdssfilters.asp.

[8] “Transformations between SDSS magnitudes and other systems.” Accessed: May 02, 2025. [Online]. Available: https://www.sdss4.org/dr17/algorithms/sdssUBVRITransform/.

[9] Sloan Digital Sky Survey, “Data Release 18 [Data Set],” 2023. Accessed: May 02, 2025. [Online]. Available: https://www.sdss.org/dr18/.

[10] Planck Collaboration et al., “Planck 2018 results: VI. Cosmological parameters (Corrigendum),” Astron. Astrophys., vol. 652, p. C4, Aug. 2021.

[11] Breiman, L., “Random Forests,” Mach. Learn., vol. 45, no. 1, pp. 5 – 32, Oct. 2001.

[12] Freund, Y., & Schapire, R. E., “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting,” J. Comput. Syst. Sci., vol. 55, no. 1, pp. 119 – 139, Aug. 1997.

[13] Chen, T., & Guestrin, C., “XGBoost: A Scalable Tree Boosting System,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco California USA: ACM, Aug. 2016, pp. 785 – 794.

[14] Lundberg, S. M., & Lee, S.-I., “A Unified Approach to Interpreting Model Predictions,” in Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., 2017. Accessed: May 04, 2025.