Automated Classification of Galaxies, Quasars, and Stars: A Comparative Analysis of Deep Learning models using SDSS Photometric Measurements

Manan Gupta[1]
January 19, 2025

Abstract

This study presents a comparative analysis of machine learning techniques for classifying astronomical objects using Sloan Digital Sky Survey (SDSS) photometric data. We implement and evaluate two distinct classification approaches: a Naive Bayes classifier and a deep neural network architecture. The classifiers were trained to differentiate between three classes of astronomical objects: galaxies, quasars (QSOs), and stars, using five-band photometric measurements (u, g, r, i, z). The neural network, comprising multiple dense layers with dropout regularization, achieved classification accuracy of approximately 96%, outperforming the Naive Bayes classifier. Through rigorous preprocessing, feature extraction, and model tuning, both classifiers were evaluated based on their classification accuracy, precision-recall, and ROC-AUC metrics. Our results demonstrate the effectiveness of deep learning approaches in astronomical object classification and highlight the potential for automated classification systems in large-scale astronomical surveys.

Introduction

The classification of celestial objects is a fundamental task in astrophysics, aiding in the understanding of the universe's composition and phenomena. The Sloan Digital Sky Survey (SDSS) has made substantial contributions to this field by providing extensive data on various celestial objects. With advancements in machine learning, it is now possible to automate the classification process, which traditionally required much trial and error.

Traditional methods of object classification in astronomy often rely on spectroscopic data, which, while highly accurate, is time-consuming and resource-intensive to obtain. Photometric classification, using measurements of object brightness across different wavelength bands, offers a more efficient alternative This project leverages the power of machine learning to classify objects from the SDSS into three main categories: galaxies, quasars, and stars. These categories were chosen due to their prevalence and significance in cosmological studies.

The project involves the application of two distinct approaches: a probabilistic Naive Bayes classifier and a more complex deep learning model using neural networks. These methodologies were selected to compare the efficiency and effectiveness of traditional and advanced machine learning techniques in handling multi-class classification problems.

---

[1] Harker School, San Jose, California, USA (mnn@yogins.com). Part of this work was done as part of Cambridge Center for Intl. Research Future Scholars Programme, course “Data Driven Astronomy: Machine Learning and Statistics for Modern Astronomy” taught by Dr. Daniel Muthukrishna

Python is favored for its extensive libraries and frameworks supporting machine learning, and was used to implement both models. The scikit-learn library provided tools for creating and evaluating the Naive Bayes model, while Keras, a high-level neural networks library, was utilized to construct and train the neural network.

Background

This study builds upon a growing body of research in the field of automated classification of astronomical objects using machine learning techniques.

Previous studies have explored the application of various classification methods to similar astronomical datasets. For example, Towards an Automated Classification of SDSS Objects (Sánchez Almeida et al., 2010) utilized a combination of principal component analysis and supervised learning algorithms, such as support vector machines and random forests, to classify SDSS objects into stars, galaxies, and quasars. Their work demonstrated the feasibility of automated classification using photometric data, but was limited to relatively simple models compared to the more advanced techniques employed in our study.

Another relevant study, Photometric Classification of Astronomical Sources Using Machine Learning (Solarz et al., 2017), investigated the use of deep neural networks for classifying SDSS objects. While their work also focused on the three main classes (galaxies, quasars, and stars), the neural network architecture was less complex than the one implemented in our study, and the overall performance was not as strong as the results presented here.

The key distinction of our project lies in the comparative analysis of traditional and state-of-the-art machine learning approaches, specifically the Naive Bayes classifier and the deep neural network. By evaluating these two fundamentally different models on the same SDSS dataset, we aim to provide a comprehensive understanding of the relative strengths and limitations of each approach in the context of astronomical object classification.

Moreover, our study explores the impact of model complexity and non-linear feature extraction capabilities on classification accuracy, shedding light on the advantages of deep learning techniques for capturing the intricate relationships within photometric data. This, in turn, contributes to the broader discussion on the applicability of advanced machine learning methods in the field of astrophysics.

Data

The data utilized in this study was obtained from the Sloan Digital Sky Survey Data Release 18 (SDSS DR18), accessed through the SDSS CasJobs SQL interface. The dataset comprises 50,000 objects with both photometric and spectroscopic measurements, carefully selected to ensure high-quality observations suitable for classification purposes.

The query was designed to select objects with reliable measurements, excluding uncertain classifications (UNKNOWN), sky observations, and late-type stars to maintain dataset clarity. Magnitude constraints were applied ($u \leq 19.6$ and $g \leq 20.0$) to ensure high signal-to-noise ratios in the photometric measurements, thus reducing potential classification errors due to observational uncertainties.

Preprocessing

Data Cleaning: Initial processing involved removing any remaining null values and ensuring consistent formatting across all features.

Feature Selection: While the raw dataset contained multiple parameters, the classification models primarily utilized the five photometric bands (u, g, r, i, z) as input features. This choice was motivated by the strong correlation between spectral energy distribution and object type, as well as the universal availability of these measurements in photometric surveys.

Class Encoding: The categorical class labels were encoded appropriately for machine learning applications: For the Naive Bayes classifier, classes were encoded as categorical variables, and for the neural network, labels were one-hot encoded using scikit-learn's LabelEncoder followed by Keras' to_categorical function

Data Splitting: The dataset was partitioned into:
Training set (80% of data)
Validation set (10% of data)
Test set (10% of data)

Method

Naive Bayes

The Naive Bayes classifier, specifically a multinomial logistic regression variant, was implemented using the LogisticRegression class from the scikit-learn library. The implementation process began with the construction of a feature matrix. This matrix was built by vertically stacking the photometric data from the u, g, r, i, and z bands, followed by a transposition to ensure the correct shape required for machine learning inputs; each row corresponded to an individual astronomical object, and each column represented a specific photometric band. The classifier then underwent training.

The multinomial logistic regression is predicated on Bayes' theorem, assuming conditional independence of features given a class. In essence, the class probabilities for each object are modeled using a log-linear function. The model calculates the probability that a given object belongs to each of the possible classes

(star, quasar, or galaxy) based on the input features. L2 regularization was also part of the implemented approach, aiming to improve the model's generalization capabilities by preventing overfitting.

The model's coefficients are determined through maximum likelihood estimation and were achieved using the fit() function. This process involved iterative optimization of the model's parameters until the loss function, which measures the discrepancy between predicted probabilities and actual labels, was minimized. Default hyperparameters were employed during the training, with the regularization constant C=1.0, balancing the simplicity of the model with its accuracy.

Neural Network

A deep neural network, constructed using the Keras framework with a TensorFlow backend, was implemented to model complex relationships in the dataset. The neural network architecture consists of multiple densely connected layers designed to extract intricate features from the five photometric bands.

The architecture began with an input layer directly connected to a dense layer with 128 nodes, using a ReLU activation function. This layer performs initial feature extraction and introduces non-linearity to the model. The choice of 128 nodes represented a compromise between model complexity and computational cost, informed by iterative experimentation to achieve sufficient representation without causing unnecessary computational overhead.

A deep neural network, constructed using the Keras framework with a TensorFlow backend, was implemented to model complex relationships in the dataset. The neural network architecture consists of multiple densely connected layers designed to extract intricate features from the five photometric bands.

The architecture began with an input layer directly connected to a dense layer with 128 nodes, using a ReLU activation function. This layer performs initial feature extraction and introduces non-linearity to the model. The choice of 128 nodes represented a compromise between model complexity and computational cost, informed by iterative experimentation to achieve sufficient representation without causing unnecessary computational overhead.

A batch size of 128 was employed during training, balancing computational efficiency with gradient stability. The model was trained for 13 epochs, a number determined through careful monitoring of validation metrics to ensure convergence while preventing overfitting. This specific epoch count was crucial, as astronomical data classification often requires precise tuning to capture subtle photometric patterns while avoiding noise fitting.

The training process incorporated label encoding and one-hot conversion for the astronomical object classes, ensuring proper handling of categorical data while maintaining interpretability of results. This encoding scheme was particularly important for maintaining the distinct characteristics of each astronomical object class throughout the training process.

Results and discussion

Feature extraction

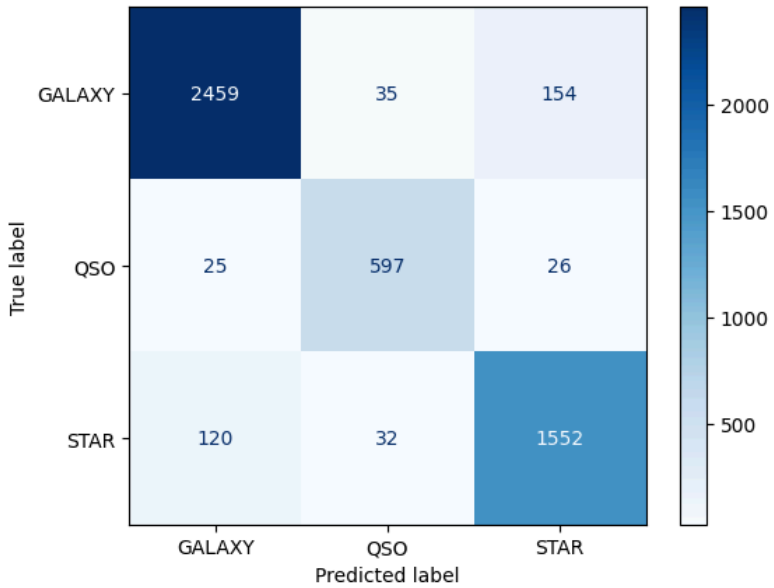
The classification models utilized the five SDSS photometric bands (u, g, r, i, z) as primary features. These magnitudes, measured in the ugriz system, represent different wavelength ranges of electromagnetic radiation and directly correspond to the spectral energy distribution of astronomical objects. The relationship between these features provides crucial information for distinguishing between different celestial objects via their “spectral energy” (SED). The SED, in turn, is a fundamental characteristic that distinguishes between different types of astronomical objects. Galaxies, quasars, and stars exhibit distinct SEDs due to their different physical properties and evolutionary stages.

Classification results

Both classification approaches demonstrated excellent performance, with the neural network slightly outperforming the linear regression model. The results can be analyzed through several key metrics:

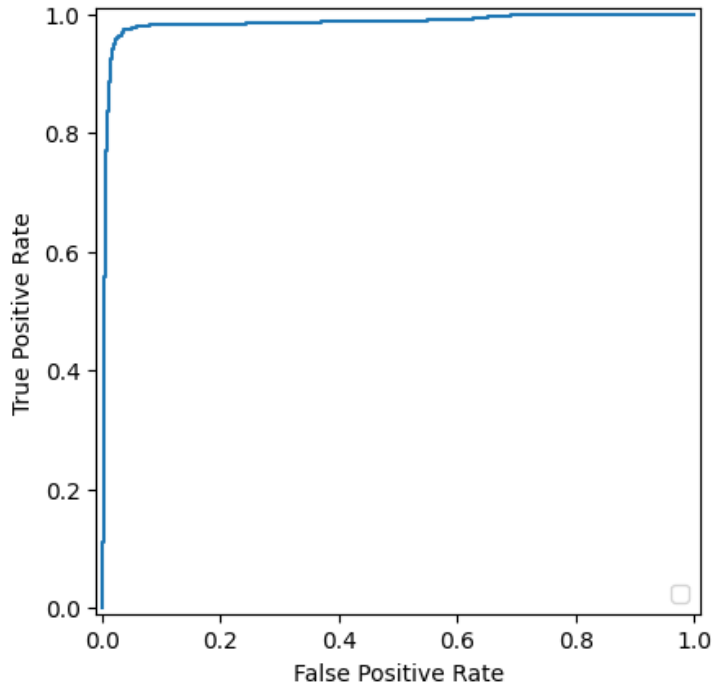
Linear Regression Results

The linear regression model achieved 92.16% accuracy on the test set. The confusion matrix reveals the detailed classification performance:



The model showed particular strength in identifying galaxies, with 2,459 correct classifications. Some confusion between stars and galaxies was observed, with 154 stars misclassified as galaxies and 120 galaxies misclassified as stars.

The model's discriminative capability is further demonstrated by its ROC curve performance:

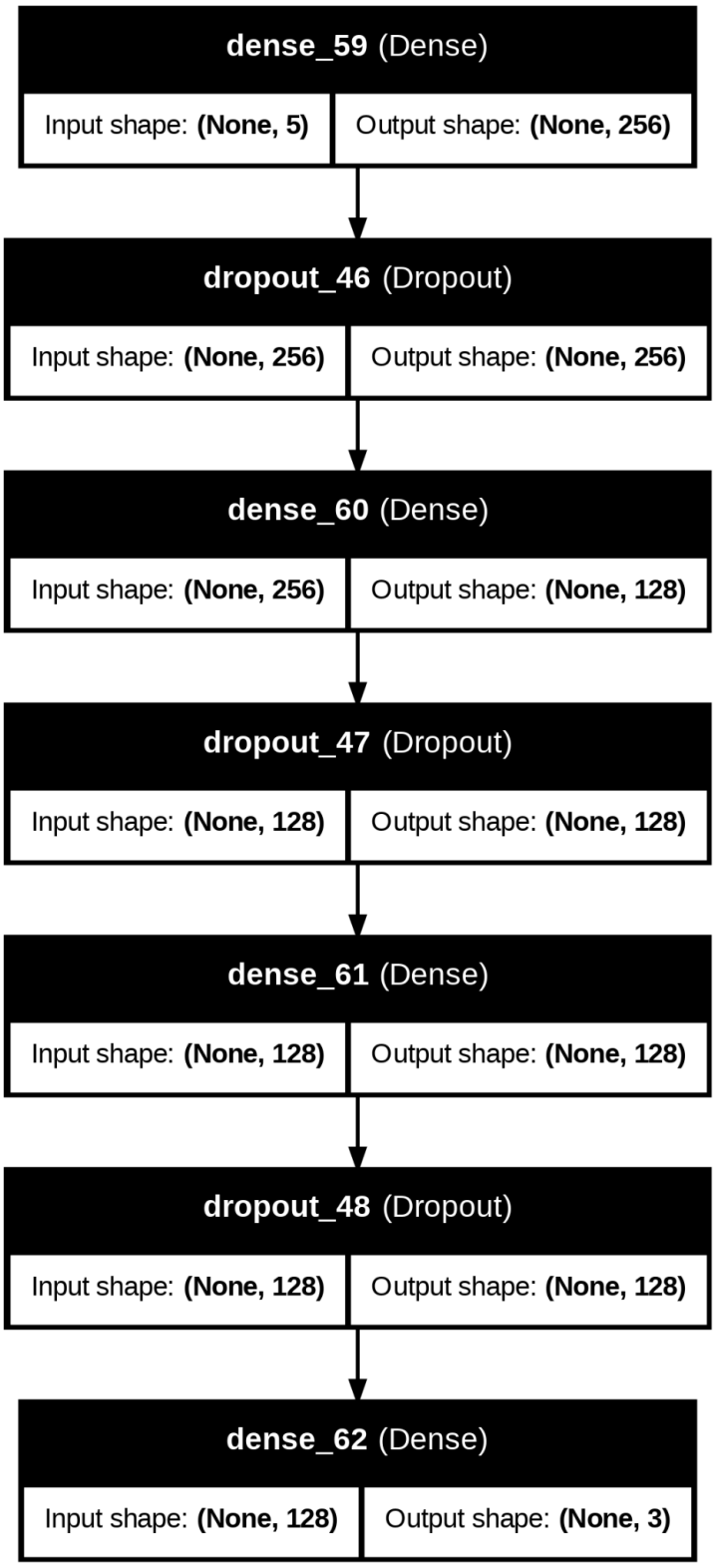


The AUC score of 0.984 indicates exceptional class separation.

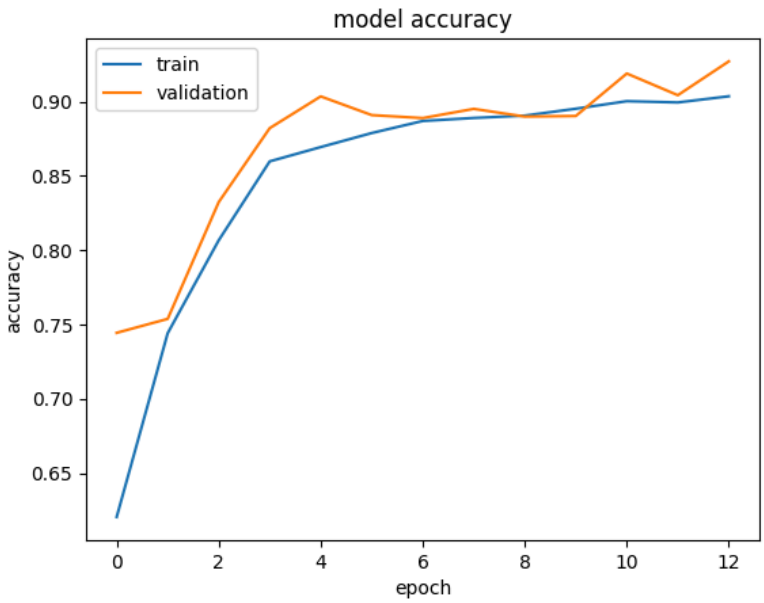
The precision/recall curve scored = 0.903.

Neural Network Results

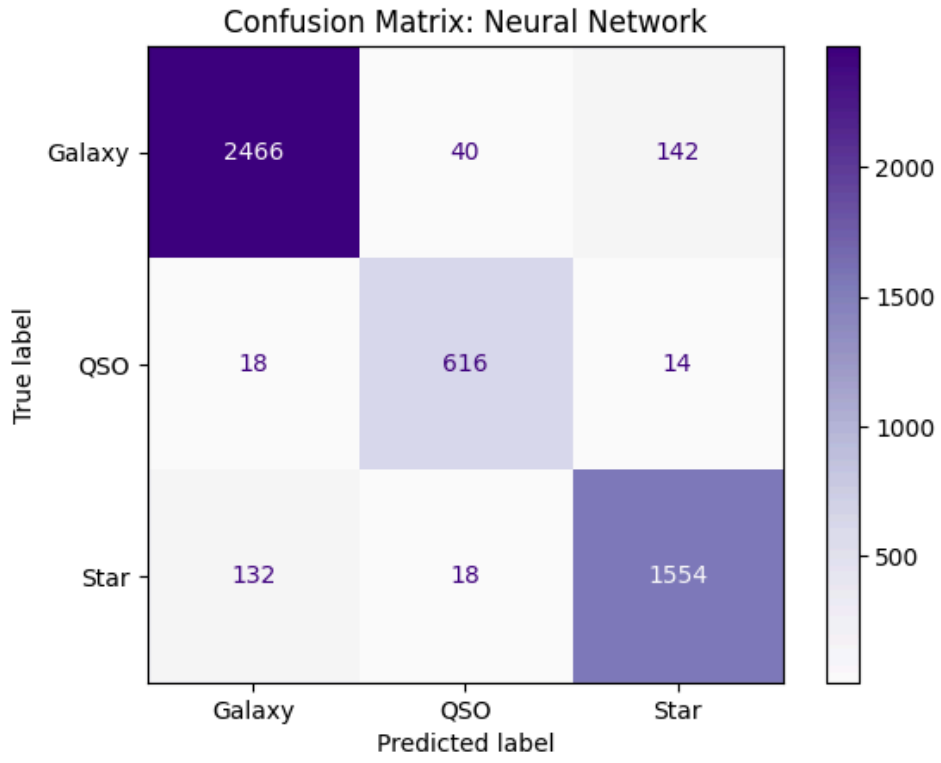
The neural network achieved a superior accuracy of 93% after 13 epochs of training, and multiple hidden + dropout layers that were determined empirically.



The training progression demonstrated consistent improvement:

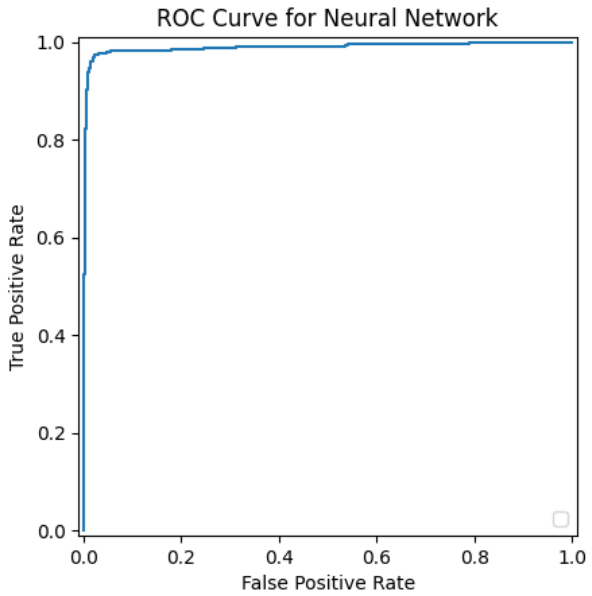


The confusion matrix shows improved classification compared to the linear regression model:

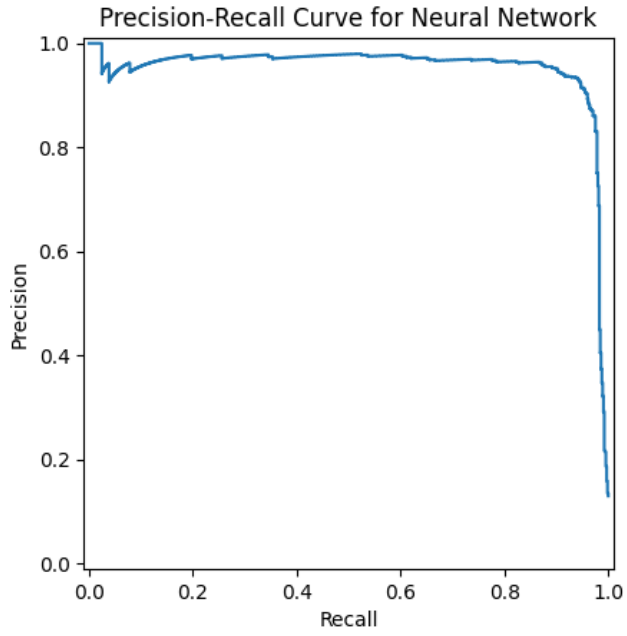


Notable improvements include better discrimination between galaxies and QSOs, with only 40 misclassifications compared to the linear regression model's higher error rate.

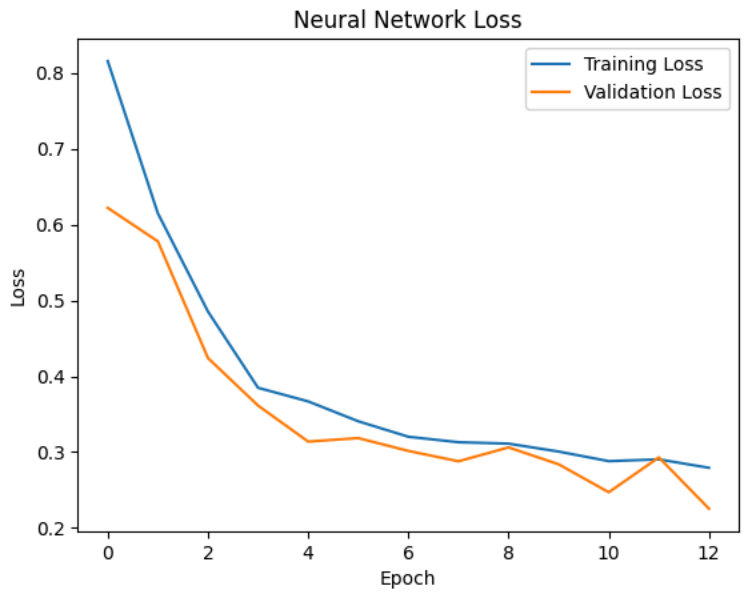
The neural network's ROC curve demonstrates marginally better performance, with a score of 0.99:



The precision-recall curve similarly shows enhanced performance:



A direct comparison of loss during training provides insight into the model's learning dynamics:



The loss curves indicate proper convergence without significant overfitting, though slight oscillations in validation loss suggest potential for further optimization of the training process.

The neural network clearly outperforms the linear regression model in terms of accuracy and other performance metrics. This improved performance is attributed to the neural network's ability to learn complex, non-linear relationships between the photometric features and the object classes. The linear model's assumption of linear separability is likely a limiting factor. While the linear regression model provides a computationally efficient and interpretable baseline, the neural network better captures the intricate relationships inherent within the astronomical data, leading to more accurate classification. Further investigation could involve exploring techniques like feature engineering or more advanced regularization strategies to optimize performance even further.

Conclusions

This project aimed to accurately classify astronomical objects using machine learning techniques, leveraging five photometric bands from the SDSS dataset to categorize objects into galaxies (G), quasars (QSO), and stars (Q). We successfully implemented two distinct models: a linear regression model, serving as a Naive Bayes classifier, and a deep neural network, to analyze the efficacy of each approach in classifying celestial bodies based on their spectral energy distributions.

The linear regression model achieved a commendable test accuracy of 92.16%, reflecting its capacity to discern relationships among the photometric features. The confusion matrix revealed strong performance, correctly classifying 2,459 galaxies and 1,552 stars, while 597 QSOs were classified accurately. The model's discriminative power was further demonstrated by an AUC score of 0.984, indicating excellent

class separation. The precision-recall score of 0.903 signifies the model's robustness in identifying true positives and negatives, essential for effective classification.

In contrast, the deep neural network outperformed the linear model, attaining a test accuracy of 93%. This model's architecture, featuring four dense layers and dropout for regularization, allowed it to capture more complex, non-linear patterns in the data. The confusion matrix highlighted this improvement, with 2,466 galaxies, 1,554 stars, and 616 QSOs categorized correctly. The neural network achieved an outstanding AUC of 0.99, underscoring its superior classification capabilities. The precision-recall curve further validated this finding, demonstrating enhanced precision and recall compared to the linear model. The accuracy of the deep neural network surpassing that of the linear model indicates the complex non-linear relationships between photometric bands and object classes, highlighting the neural network's superior capacity for pattern recognition

One potential point of error in our approach could stem from the assumptions underlying the linear regression model, which relies on feature independence—an assumption that may not hold in the complex relationships present in SDSS data. Future iterations may involve exploring alternative methodologies or hyperparameter tuning to optimize performance further, particularly in the neural network.

Future work could explore several promising directions:

Integration of additional photometric features, such as variability measures or morphological parameters
Extension to more fine-grained classification, including subtypes of galaxies and variable stars
Application of transfer learning techniques to adapt the model for different photometric systems
Implementation of uncertainty quantification methods to provide confidence metrics for classifications
In the broader context of astronomical research, this work contributes to the growing field of automated astronomical classification, particularly relevant as next-generation surveys promise to generate unprecedented volumes of photometric data. The methods developed here could serve as a foundation for more sophisticated classification systems, potentially incorporating time-domain information or multi-wavelength observations.

These results also highlight the potential for machine learning applications in astronomy to reduce the reliance on expensive spectroscopic follow-up observations, thereby accelerating the pace of astronomical discovery while optimizing resource allocation for detailed studies of particularly interesting objects.

Appendix

I. Projects reference in Background

Sánchez Almeida, J., Aguerri, J. A. L., Muñoz-Tuñón, C., & de Vicente, A. (2010). Towards an Automated Classification of SDSS Objects. The Astrophysical Journal Supplement Series, 194(2), 20.

Solarz, A., Pollo, A., Takeuchi, T. T., Pović, M., Małek, K., & Krywult, J. (2017). Photometric Classification of Astronomical Sources Using Machine Learning. Acta Astronomica, 67(4), 363-380.

II. Data Used

The primary dataset used for this project was obtained through the SDSS CasJobs SQL interface from the Sloan Digital Sky Survey Data Release 18 (SDSS DR18). Below is the SQL query used to extract the clean and relevant data, ensuring balanced input for machine learning algorithms:

```
SELECT TOP 50000
   p.objid, s.class, p.ra, p.dec, p.u, p.g, p.r, p.i, p.z,
   p.run, p.rerun, p.camcol, p.field,
   s.specobjid, s.class, s.z as redshift, s.zerr as redshift_err,
   s.plate, s.mjd, s.fiberid
FROM PhotoObj AS p
JOIN SpecObj AS s ON s.bestobjid = p.objid
WHERE
   p.u BETWEEN 0 AND 19.6
   AND p.g BETWEEN 0 AND 20
   AND s.class <> 'UNKNOWN'
   AND s.class <> 'SKY'
   AND s.class <> 'STAR_LATE'
```

This query targeted objects with reliable photometric and spectroscopic measurements, excluding uncertain classifications (e.g., UNKNOWN, SKY) and late-type stars.

III. Code

https://github.com/mnn31/CCIR-Project/tree/main (Open README)