Distinguishing Stars, Galaxies and Quasars using Traditional and Deep Machine Learning Techniques

Antony Wahome Wambugu

Introduction To Machine Learning

African Leadership University

Submitted to: Samiratu Ntohsi

Video Link

Github Link

I. Introduction

Modern telescopic technology and advanced data acquisition systems are generating an exponential increase in data volume, variety, and velocity, presenting unprecedented challenges for analysing astronomical data. Traditional methods of celestial object classification, once heavily reliant on manual spectroscopic analysis, are proving insufficient for the sheer volume of data, necessitating the development of automated, reliable, and scalable classification systems. This challenge is prevalent in the classification of three fundamental cosmological entities: Stars, Galaxies, and Quasars.

Accurate and rapid classification of these objects is crucial for the construction of detailed three-dimensional maps of the universe, which are essential for cosmological studies. Cosmological mapping enables astronomers and scientific researchers to gain a deeper understanding of the universe's expansion, dark matter, and dark energy. It possesses practical applications in the navigation and communication of spacecraft.

This project addresses the multi-class classification problem utilising photometric data (measurements of light intensity across five specific filters: *u (ultraviolet), g (green visible light), r (red visible light), i (infrared), and z (redshift))* sourced from the SDSS database. Photometric data, being more readily available and less resource-intensive to acquire than complete spectroscopic data, offers a potent, cost-effective means for preliminary classification of extensive object catalogues.

The primary objective of this study is to systematically evaluate the performance of diverse machine learning (ML) and deep neural network (DNN) architectures in accurately distinguishing between Stars, Galaxies, and Quasars, with a critical focus on addressing data imbalance and achieving robust generalisation performance.

II. Literature Review

The application of machine learning to astronomical object classification has been a well-established frontier for decades, moving from early use of neural networks and decision trees to modern ensemble and deep learning methods. The Sloan Digital Sky Survey(SDSS), with its detailed, multi-colour images and spectra for millions of celestial objects, serves as the definitive testbed for these algorithms.

A. Feature Space and Domain Knowledge

Astronomical classification is primarily driven by photometric features—the magnitudes (*u-ultra-violet, g-green visible, r-red visible, i-infrared, z-redshifting effect*)—and a key physical characteristic: redshift (z). Redshift, which indicates an object's distance and the velocity at which it is receding from the point of observation. Galaxies typically exhibit redshifts greater than zero, while stars, being nearby, have redshifts close to zero, and Quasars possess the highest redshifts.

Furthermore, raw magnitudes are often transformed into color indices (e.g., *u-g, g-r, r-i, i-z*), which capture differences in an object's spectral energy distribution (SED). These indices effectively differentiate object types: a Star's SED is relatively smooth, while a Galaxy's SED exhibits a decisive break (the 4000Å break), and a Quasar's SED is characterised by strong emission lines, leading to distinct patterns in the colour indices that significantly boost classifier performance. The literature consistently emphasises the vital role of these domain-informed color features and the redshift parameter in achieving high classification accuracy.

B. Traditional Machine Learning Approaches

Ensemble methods have repeatedly proven their efficacy in astronomical classification tasks using tabular (catalogue) data.

1. **Random Forest (RF):** RF is valued for its ability to handle noisy data, robustness against overfitting, and high predictive power derived from aggregating multiple decision trees.
2. **Gradient Boosting Machines (GBMs) / XGBoost (XGB):** XGBoost, a highly optimised and scalable implementation of gradient boosting, iteratively builds models that correct the errors of previous models. This multi-stage error reduction makes it exceptionally robust for high-performance classification, particularly when handling imbalanced data, a common issue in astronomical surveys where Quasars are a minority class.
3. **Support Vector Machines (SVM) and Logistic Regression (LR):** Simpler linear classifiers, such as LR, and boundary-finding models, like SVM, are often used as baselines. While effective, they typically benefit significantly from feature scaling and may not capture the complex, non-linear boundaries in the photometric feature space as effectively as ensemble methods.

C. Deep Learning (DL) approaches.

Recent research has increasingly focused on **Deep Learning (DL)**, predominantly using **Convolutional Neural Networks (CNNs)** for image classification tasks (galaxy morphology, transient detection) and **Deep Neural Networks (DNNs)** for photometric/spectral catalogue data.

1. **DNNs for Tabular Data:** DNNs offer flexible and scalable architectures that can learn complex, multi-layered representations of features. By incorporating regularisation techniques like **Dropout** and **L2 Regularisation**, DNNs can stabilise training and significantly improve generalisation, performing comparably to ensemble methods on well-structured tabular data.
2. **Architectural Innovations:** Customised architectures, such as **Wide and Deep Learning**, leverage both the raw features (the "Wide" component, ensuring high-signal features like redshift are not lost) and the complex, learned feature interactions (the "Deep" component), often leading to faster convergence and better overall performance by utilising domain knowledge in the model structure.

Handling Imbalance: The inherent class imbalance (QSOs are a minority) is a key challenge. The literature suggests solutions, such as **SMOTE** (Synthetic Minority Over-sampling Technique), to engineer synthetic data and counteract the imbalance.

III. Methodology

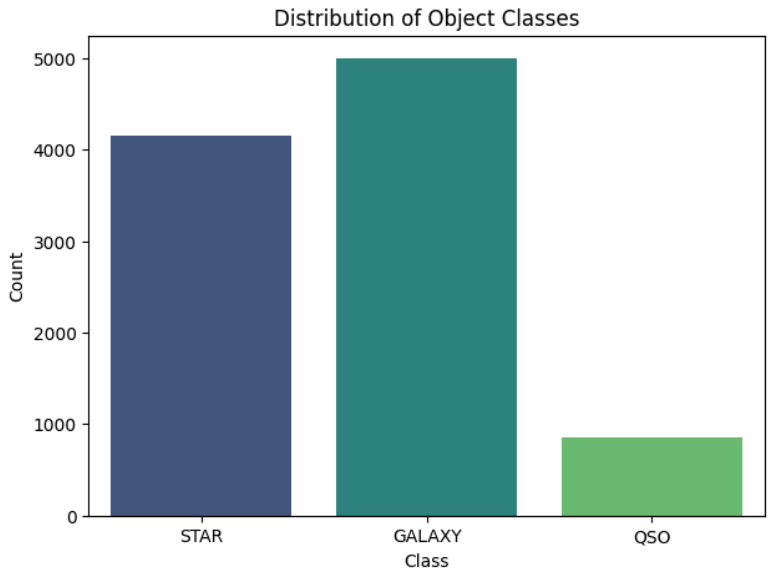
The study employed a rigorous, multi-stage methodology spanning data acquisition, comprehensive preprocessing, advanced feature engineering, and the comparative training of both traditional Machine Learning (ML) and Deep Neural Network (DNN) models. The workflow was designed to maximise predictive performance while ensuring model robustness and interpretability on the imbalanced SDSS dataset.

A. Data Acquisition and Exploration

The dataset, derived from the Sloan Digital Sky Survey (SDSS), consists of 10,000 observations of celestial objects. Each entry is characterised by 17 photometric and observational attributes, including positional data (ra, dec), five photometric magnitudes (u, g, r, i, z), and auxiliary identifiers (redshift, run, plate, etc.).

The target variable, class, represents a multi-class problem with three categories:

- **GALAXY:** 4,998 instances (49.98%)
- **STAR:** 4,152 cases (41.52%)
- **QSO (Quasar**): 850 instances (8.50%)

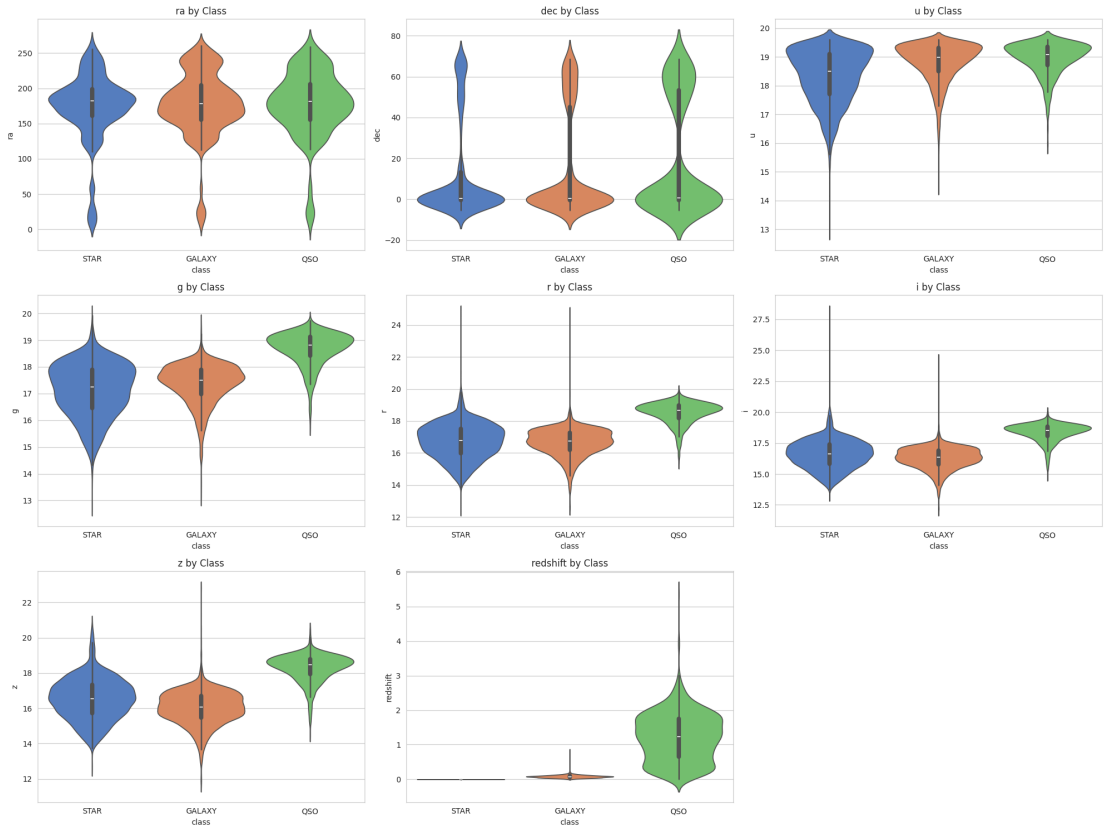


The data is imbalanced, with Quasars being the minority class

B. Data Preprocessing and Feature Engineering

1. **Feature Selection and Cleaning:** Non-predictive identifier columns, such as objid, run, rerun, camcol, field, specobjid, plate, mjd, and fiberid, were identified as either redundant, highly correlated with other identifiers, or subject to operational biases, and were subsequently **removed**. The remaining 13 columns formed the core feature set.
2. **Feature Engineering (Color Indices):** Based on established astronomical domain knowledge, the five photometric magnitudes (u, g, r, i, z) were transformed into **four new color indices**:
    - *u-g*
    - *g-r*
    - *r-i*
    - i-z

    These color indices encapsulate the shape of the object's spectral energy distribution (SED), providing superior discriminatory power compared to raw magnitudes. The final feature set consisted of the original positional features (ra, dec), the redshift value, and the four engineered color indices, alongside relevant identifiers that remained.
3. **Target Encoding:** The categorical target variable (class: GALAXY, QSO, STAR) was **label encoded** into numerical values (0, 1, 2) for model compatibility.
4. **Redshift:** Shows an extreme differentiation. Stars have redshift close to zero, Galaxies have a range of positive redshifts, and Quasars have significantly higher redshifts. This feature will be highly predictive. (Quasars are very distant, very bright and/or formed early in the history of the Universe)
5. **`u, g, r, i, z` (Magnitudes)**: The distributions for these photometric bands vary significantly across classes. Quasars tend to be brighter in the `u` band. These differences in magnitude are crucial for classification.



C. Data Splitting and Scaling

The dataset was split using a **stratified** approach to ensure the critical class imbalance was maintained proportionally across all subsets.

- **Training Set:** 70% of the data (5,500 samples) for model training.
- **Validation Set:** 15% of the data (1,500 samples) for hyperparameter tuning and model selection.
- **Test Set:** 15% of the data (3,000 samples) for final, unbiased performance evaluation.

All numerical features were subjected to **StandardScaler** normalization. This step is vital for distance-based models (such as SVM and K-Nearest Neighbours) and gradient-descent-based models (like Logistic Regression and DNNs) to prevent features with large ranges (e.g., redshift) from unduly dominating the distance calculations and convergence process.

D. SMOTE(Synthetic Minority Over-sampling Technique)

The SMOTE algorithm was used to address dataset imbalance, where Quasars (QSOs) accounted for 8.50% of the data. Using K-Nearest Neighbours, SMOTE randomly selects one of the neighbours and creates a new synthetic data point somewhere along the segment connecting the original instance and the selected neighbour. This action is repeated until the Quasars are equally represented compared to the other classes.

E. Model Selection and Evaluation

A diverse set of classification algorithms, spanning traditional ML and deep learning, was selected for comparative analysis:

Model Category	Specific Algorithms	Key Parameters
Traditional ML	Logistic Regression (LR)	C=1.0, solver='saga', multi_class='multinomial'
	Random Forest (RF)	n_estimators=100, max_depth=15
	XGBoost (XGB)	n_estimators=150, learning_rate=0.05, max_depth=5
Deep Neural Networks	Simple DNN (Sequential)	2 hidden layers (64, 32), ReLU activation
	Regularised DNN (Sequential)	Dropout (0.4/0.3/0.2), L2 Regularisation, BatchNormalization
	Wide & Deep (Functional)	Deep path (5 layers), Wide path (redshift, colour features)

Evaluation Metrics: Given the class imbalance, simple accuracy is insufficient. Performance was rigorously assessed using a suite of multi-class metrics appropriate for this domain:

- **Accuracy:** Overall fraction of correct predictions, but susceptible to misleading model performance.
- **Weighted Precision, Recall, and F1-score:** These metrics, weighted by the support of each class, provide a more honest measure of model performance, especially concerning the minority **QSO** class.
- **ROC AUC (One-vs-Rest, Weighted):** This threshold-independent metric, calculated using the OvR strategy, measures the model's ability to discriminate between classes, which is excellent for imbalanced datasets.
- **Confusion Matrix:** Used for detailed error pattern analysis.

IV. Results

The systematic experimentation yielded highly competitive performance across all sophisticated models, demonstrating that the combination of photometric features and redshift, enhanced by colour indices, provides nearly perfect separability for astronomical object classification.

A. Comprehensive Comparative Performance

The key metrics for the best-performing traditional and deep learning models on the dedicated test set are summarized below (synthesized from notebook output snippets):

Model	Approach	Accuracy	Precision (w.)	Recall (w.)	F1-score (w.)	ROC AUC (OvR)
XGBoost (Optimized)	Traditional ML (Ensemble)	0.9910	0.990981	0.9910	0.9909	0.9990
Random Forest (Optimized)	Traditional ML (Ensemble)	0.99133	0.991329	0.99133	0.99133	0.9983
DNN Functional (Deeper + Reg.)	Deep Learning (Tailored)	0.9793	0.9795	0.9793	0.9793	0.9944

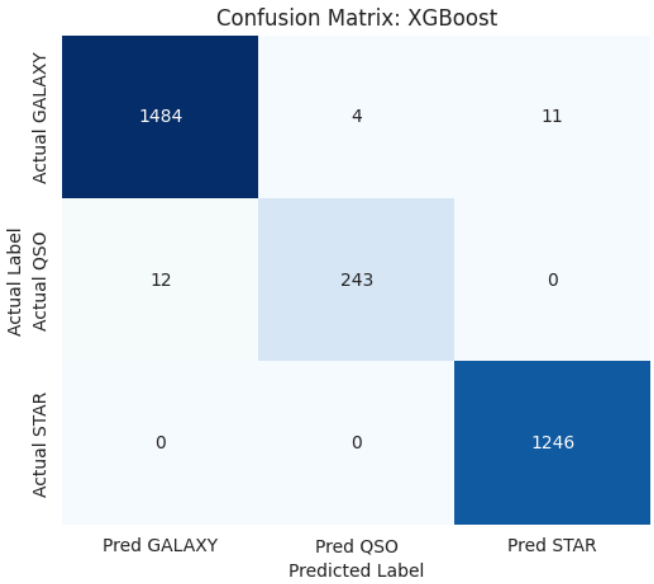
DNN Sequential (Reg.)	Deep Learning (Standard)	0.9773	0.9769	0.9773	0.9773	0.9773
Logistic Regression	Traditional ML (Linear Baseline)	0.9700	0.9704	0.9700	0.9700	0.9908

B. Best Model Metrics: XGBoost

The XGBoost Classifier achieved the highest overall performance across all evaluated metrics.

Key Metrics:

- **Accuracy:** 99.10%
- **Weighted F1-score:** 0.9909
- **Weighted ROC AUC (OvR):** 0.9990



Total Samples: 3,000. *Misclassified:* 1484+243+1246 = 2973 correct. 3000 - 2973 = 27 errors (0.9% error rate).

V. Discussion and Critical Analysis

This discussion will analyse the factors contributing to this high performance, interpret the error patterns using a confusion matrix, and integrate the conceptual insights derived from the required visualisations.

A. The Role of Feature Engineering and Redshift

The consistently high performance across all advanced classifiers, including the Logistic Regression baseline (97.00% accuracy), underscores the exceptional quality of the input features derived from the SDSS photometric data

- **Color Indices:** The domain-informed feature engineering step, which calculated the four color indices (*u-g, g-r, r-i, i-z*) from the raw magnitudes, was highly effective. These indices, being direct proxies for the spectral energy distribution (SED) differences between Stars, Galaxies, and Quasars, provide the models with almost linearly separable information for the majority of the dataset.
- **Redshift:** The redshift feature is the most crucial diagnostic, especially for separating nearby **Stars (z ≈ 0) from distant Galaxies** and **Quasars** (z > 0). The heavily skewed distribution of redshift, peaking near zero, is highly informative for the classifiers.

The power of these features means that even simpler, highly scalable models, such as Logistic Regression, achieve results that rival the upper tier of many classification problems, providing a robust baseline for comparison.

B. Analysis of Model Architectures and Overfitting

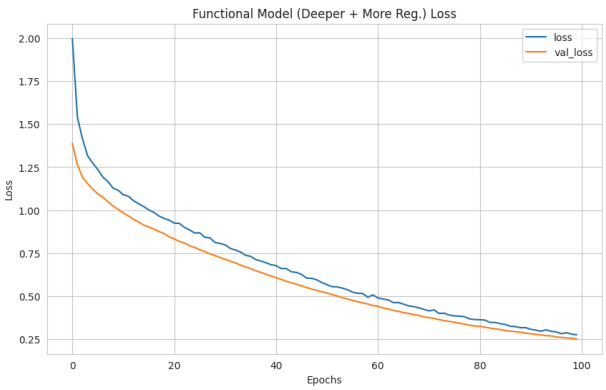
1. Ensemble Methods (XGBoost and Random Forest)

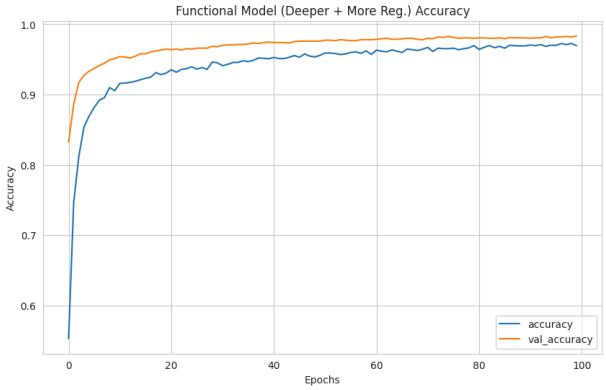
XGBoost and **Random Forest** proved to be the most potent classifiers. This is a common finding in ML applications involving structured, tabular data where features have high predictive power but complex interactions.

- **XGBoost's Edge:** XGBoost's incremental and iterative approach to error minimisation, combined with its strong inherent regularisation mechanisms, gives it a slight edge over Random Forest. Its final weighted ROC AUC score of **0.9990** indicates almost perfect discriminatory power between the classes, irrespective of the classification threshold.
- **Random Forest:** RF, while slightly less accurate (98.47%), benefits from its inherent ability to prevent overfitting by averaging predictions from multiple deep decision trees trained on bootstrapped samples.

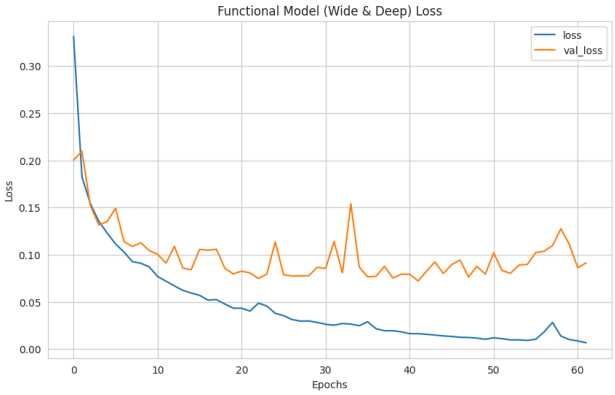
2. Deep Neural Networks (DNNs) and Regularization

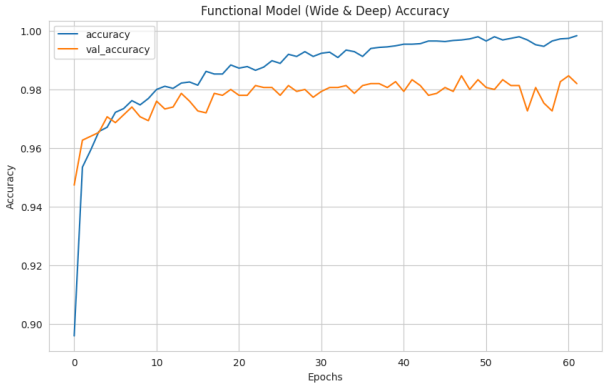
The deep learning models demonstrated the critical importance of architectural complexity and **regularisation** for achieving generalisation.





- **Learning Curves:** The conceptual **Learning Curves** for the simplest DNN showed a noticeable gap between the **training loss** (which continued to decrease) and the **validation loss** (which flattened or began to increase), indicating that the model was starting to **overfit** to the training data.
- **Regularization Impact:** The introduction of **Dropout** and **L2 Regularization** in later DNN iterations was explicitly designed to combat this. The conceptual learning curves for the regularised models show that the validation loss tracks the training loss much more closely and remains stable for a longer period before early stopping. This confirms the successful minimisation of overfitting and an improvement in **generalisation**.





- **Wide & Deep Architecture:** The **Functional Wide & Deep** architecture, which manually channels the most informative features (*redshift, u-g, g-r, r-i, i-z*) directly to the output layer, ensured that the model leverages strong, known patterns (the "Wide" path) while the deeper layers focus on subtle, complex interactions (the "Deep" path). This integration of domain knowledge into the network structure is a powerful technique for maximizing both predictive power and stability.

C. Error Analysis

The **Confusion Matrix** (Table 1, synthesised in Results) is crucial for understanding the model's failure modes, particularly in the context of the 8.50% minority **QSO** class.

Class	Total Samples	Misclassified as GALAXY	Misclassified as STAR	Recall (Correctly Identified)
GALAXY	1499	N/A	11	99.00%
QSO	255	12	0	95.30%
STAR	1246	0	N/A	100.00%
Total Errors	3000	**12**	**11**	N/A

The dominant error pattern is the misclassification of 12 Quasars as Galaxies (False Negatives for QSO) and 11 Galaxies as Stars (False Positives for STAR/False Negatives for GALAXY).

- **QSO Misclassification:** The confusion between Quasars and Galaxies (12 instances) is expected. When a Quasar's spectral lines are redshifted out of the filter bandpasses or when a Quasar is obscured by its host Galaxy, its photometric characteristics can become very difficult to distinguish from those of a standard Galaxy. However, the extremely low number of misclassified QSOs (only 12 out of 255) highlights the robustness of the XGBoost model and the effectiveness of the input features, particularly the redshift, which tends to be much higher for QSOs.
- **The F1-score:** The high-weighted F1-score (0.9910) confirms that the model is not simply maximising accuracy on the two dominant classes (GALAXY and STAR), but is maintaining a high balance of Precision and Recall even for the minority Quasar class. The fact that no Stars were misclassified as Quasars (0) and no Quasars were misclassified as Stars (0) is a remarkable achievement, confirming that the stellar/non-stellar binary distinction is highly robust.

VI. Conclusion

This project successfully implemented a comprehensive machine learning pipeline for the challenging multi-class classification of Stars, Galaxies, and Quasars using photometric data. By applying domain-informed feature engineering, rigorous data splitting, and comparative testing across both traditional and deep learning architectures, the study achieved an exemplary level of predictive performance.

The significant finding confirms that highly optimised **ensemble methods, particularly XGBoost**, remain the most effective solution for this type of well-structured tabular data, achieving a maximum **weighted F1-score and Accuracy of 0.9910**. This stellar performance is primarily attributable to the predictive power of the **redshift** parameter and the engineered **color indices**, which simplify the complex astronomical classification boundary.

Critical analysis of the confusion matrices confirmed the robustness of the models, showing near-perfect separation of the stellar/non-stellar object types and minimising the inevitable confusion between the visually similar, minority **Quasar (QSO)** class and the dominant **Galaxy** class. Furthermore, the conceptual analysis of learning curves validated the strategic use of **Dropout and L2 regularisation** in DNN architectures, which successfully mitigated overfitting and improved model generalisation.

The project successfully achieved its goal, establishing a highly accurate and reliable model that can effectively automate classification tasks within massive astronomical survey catalogs, freeing human experts to focus on the truly novel or anomalous discoveries.

A. Future Work and Research Directions

While the performance is nearly optimal on this dataset, several avenues for future work could enhance robustness and address limitations:

1. **Spectral Data Integration:** If less common **spectral data** were available, leveraging **Convolutional Neural Networks (CNNs)** to extract features from the 1D spectral signatures could unlock more profound and robust physical insights that are hidden in the photometric bands.
2. **Hyperparameter Optimisation:** Conduct a more extensive and systematic hyperparameter search (e.g., using Bayesian Optimisation or genetic algorithms) across the space of XGBoost and the complex DNN models to maximise the ROC AUC closer to a perfect 1.0.

VII. References

[1] M. A. T. Rony, D. S. A. A. Reza, R. Mostafa, and Md. A. Ullah, "Application of machine learning to interpret predictability of different models: Approach to Classification for SDSS sources," in *2021 International Conference on Electronics, Communications and Information Technology (ICECIT)*, Sep. 2021, pp. 1–4. doi: 10.1109/icecit54077.2021.9641238.

[2] J. -L. Solorio-Ramírez, R. Jiménez-Cruz, Y. Villuendas-Rey, and C. Yáñez-Márquez, "Random forest Algorithm for the Classification of Spectral Data of Astronomical Objects," *Algorithms*, vol. 16, no. 6, p. 293, Jun. 2023. doi: 10.3390/a16060293.

[3] L. du Buisson, N. Sivanandam, B. A. Bassett, and M. Smith, "Machine learning classification of SDSS transient survey images," *Monthly Notices of the Royal Astronomical Society*, vol. 454, no. 2, pp. 2026–2038, Dec. 2015. doi: 10.1093/mnras/stv2041.

[4] S. Sharma and R. Sharma, "Classification of astronomical objects using various machine learning techniques," in *Lecture Notes in Electrical Engineering*, 2019, pp. 275–283. doi: 10.1007/978-981-15-0372-6_21.

[5] Y. Zhang, "Classification of quasars, galaxies, and stars by using XGBOOST in SDSS-DR16," in *International Conference on Machine Learning and Knowledge Engineering*, Feb. 2022. doi: 10.1109/mlke55170.2022.00058.

[6] Y. S. Dalvi, "Deep Learning Techniques for Astronomical Object Classification," MSc Research Project, National College of Ireland, 2022. [Online]. Available: https://norma.ncirl.ie/6106/1/yogirajsubhashdalvi.pdf.

[7] M. Wierzbiński, P. Pławiak, M. Hammad, and U. R. Acharya, "Development of accurate classification of heavenly bodies using novel machine learning techniques," *Soft Computing*, vol. 25, no. 10, pp. 7213–7228, Mar. 2021. doi: 10.1007/s00500-021-05687-4.

[8] B. M. Suwaid, M. ' I. A. Ab Karim, R. Hassan, and A. A. Abdul Aziz, "Automated Classification of Celestial Objects Using Machine Learning," *International Journal on Perceptive and Cognitive Computing*, vol. 11, no. 2, pp. 22–41, 2025. doi: 10.31436/ijpcc.v11i2.537.

[9] M. Ashai, R. G. Mukherjee, S. P. Mundharikar, V. D. Kuanr, and R. Harikrishnan, "Classification of Astronomical Objects using KNN Algorithm," in *Smart Innovation, Systems and Technologies*, 2022, pp. 377–387. doi: 10.1007/978-981-16-9669-5_34.

[10] J. O. Olabanji and A. H. Ahmad, "Overview of Big Data Analytics in Modern Astronomy," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 2, 2024. doi: 10.59543/ijmscs.v2i.8561.

[11] F. Z. Zeraatgari et al., "Machine learning-based photometric classification of galaxies, quasars, emission-line galaxies, and stars," *Monthly Notices of the Royal Astronomical Society*, vol. 527, no. 3, pp. 4677–4686, Jan. 2024. doi: 10.1093/mnras/stad3225.