Stellar Object Classification Using Ensemble Machine Learning Techniques

Swapnil Yadav
MIT Art, Design & Technology University, Pune

Anirudha Agrawal
MIT Art, Design & Technology University, Pune

Atharva Arole
MIT Art, Design & Technology University, Pune

Abstract:
Modern astronomical surveys generate massive volumes of observational data, requiring automated systems for efficient analysis and classification. Large-scale sky surveys such as the Sloan Digital Sky Survey (SDSS) have cataloged millions of celestial objects, including stars, galaxies, and quasars, using multi-band photometric and spectroscopic measurements. Conventional classification approaches rely on manual inspection or rule-based statistical thresholds derived from color–magnitude diagrams; however, these methods are computationally intensive, time-consuming, and not scalable for contemporary big-data astronomy. Accurate stellar object classification is essential for studying galactic evolution, cosmological structure formation, and large-scale universe mapping. To address this challenge, this research introduces a machine learning–based stellar classification framework that leverages photometric attributes including right ascension, declination, redshift, and five-band magnitude measurements (u, g, r, i, z). The system performs data preprocessing, feature engineering using photometric color indices, and comparative evaluation of multiple supervised learning algorithms, including Random Forest, Support Vector Machine, Gradient Boosting, and XGBoost. Ensemble learning techniques are employed to enhance classification robustness and generalization performance. The dataset utilized in this study is derived from publicly available SDSS records , and experimental results demonstrate that boosting-based ensemble models achieve classification accuracy exceeding 97%, significantly outperforming traditional single-model approaches. The findings validate that integrating advanced machine learning algorithms into astronomical data pipelines enables scalable, accurate, and automated stellar object classification, supporting the transition of modern astronomy toward data-driven computational methodologies .

Keywords: Stellar Object Classification, Astronomical Data Mining, Sloan Digital Sky Survey (SDSS), Machine Learning, Ensemble Learning, Random Forest, Gradient Boosting, XGBoost, Photometric Features, Galaxy Classification, Quasar Detection, Supervised Learning, Data-Driven Astronomy.

1. INTRODUCTION

1.1 Background

The field of astronomy has undergone a major transformation with the emergence of large-scale digital sky surveys and high-resolution observational instruments. Modern astronomical projects continuously generate massive volumes of structured data describing celestial objects across multiple wavelength bands. Among these, the Sloan Digital Sky Survey (SDSS) has played a pivotal role in cataloging millions of stars, galaxies, and quasars through photometric and spectroscopic observations [1]. The availability of such large datasets has significantly advanced cosmological research, galactic evolution studies, and large-scale structure mapping of the universe.

Celestial objects observed in SDSS are characterized by attributes such as Right Ascension (RA), Declination (Dec), redshift, and five-band photometric magnitudes (u, g, r, i, z). These features provide essential information about an object's spatial location, brightness, and spectral properties.

However, due to overlapping photometric characteristics among stars, galaxies, and quasars, accurate classification remains a complex task. Traditional classification methods rely on manual inspection of spectral lines or rule-based statistical separation using color–magnitude diagrams [2][10]. While effective for small datasets, these approaches are computationally intensive and not scalable for modern astronomical databases containing millions of entries [2][5].

The rapid growth of astronomical data has led to the emergence of data-driven astronomy, where machine learning and statistical modeling techniques are applied to automate classification tasks [10][11]. Supervised learning algorithms such as Support Vector Machines, Decision Trees, and ensemble methods have demonstrated significant improvements in classification accuracy compared to classical threshold-based approaches [4][6]. In particular, ensemble learning techniques such as Random Forest and Gradient Boosting have shown strong generalization performance for high-dimensional astronomical datasets [6][8].

Despite these advancements, challenges remain due to high feature dimensionality, measurement noise, class imbalance, and nonlinear decision boundaries. Therefore, there is a growing need for robust machine learning frameworks capable of efficiently classifying stellar objects while maintaining scalability and accuracy.

1.2 Problem Statement

Existing astronomical classification systems primarily depend on manual spectroscopic analysis or predefined statistical rules derived from photometric color indices [2]. Although spectroscopic classification provides high precision, it is resource-intensive and impractical for extremely large datasets generated by surveys such as SDSS [1]. Furthermore, rule-based photometric separation methods struggle when class boundaries overlap or when data distributions are nonlinear.

Current automated approaches often implement single machine learning models without extensive comparative evaluation or advanced feature engineering. As a result, classification performance may degrade when dealing with complex feature interactions and noisy observational measurements [5][10]. Additionally, some models are prone to overfitting, reducing their ability to generalize to unseen data.

Therefore, there is a need for a comprehensive stellar object classification framework [6][8][9] that integrates effective preprocessing, feature engineering, and ensemble learning techniques to improve predictive accuracy and robustness. The system must efficiently classify stars, galaxies, and quasars using photometric data while ensuring scalability for large astronomical datasets.

1.3 Objective

This paper aims to develop an ensemble-based machine learning framework for automated stellar object classification using SDSS photometric data. The key objectives are:

I.
To analyze and preprocess large-scale photometric datasets obtained from SDSS and prepare them for supervised classification [1][10].

II.
To implement and compare multiple machine learning algorithms, including Support Vector Machine, Decision Tree, Random Forest, Gradient Boosting, and XGBoost, for multi-class stellar classification [4][6][8][9].

III.
To apply feature engineering techniques such as photometric color index generation and correlation-based feature selection to enhance class separability and model performance [2][10].

IV.
To evaluate classification performance using comprehensive metrics and demonstrate the superiority of ensemble learning methods in achieving high accuracy and generalization capability [5][6][8].

2. LITERATURE SURVEY

Existing work related to stellar object classification and supporting computational technologies is categorized under the following subtitles:

2.1 Conventional Stellar Classification Methods

2.1.1 Problem Statement:

Traditional stellar classification methods rely on manual spectroscopic analysis and rule-based statistical separation using color–magnitude diagrams. Astronomers historically classified stars,

galaxies, and quasars by examining spectral lines and photometric color indices obtained from sky surveys such as the Sloan Digital Sky Survey (SDSS) [1].

Although spectroscopic analysis provides high precision, it is time-consuming, computationally expensive, and impractical for extremely large datasets. Photometric threshold-based separation methods are simpler but often fail when class distributions overlap or when nonlinear boundaries exist between object categories [2][10].

2.1.2 Loopholes:

- Manual inspection is not scalable for millions of observations [1].
- Rule-based color separation struggles with overlapping spectral distributions [2][10].
- High dependency on domain expertise and predefined thresholds.
- Limited adaptability to new large-scale survey data [5].

2.1.3 Proposed Solutions in Literature:

Several studies introduced statistical learning and pattern recognition approaches to automate classification [4][5]. Supervised machine learning models such as Support Vector Machines and Decision Trees have been applied to photometric datasets to improve classification efficiency. However, early implementations lacked extensive feature engineering and ensemble optimization, leaving scope for improvement in accuracy and robustness [6][8].

2.2 Astronomical Data Mining and Large-Scale Survey Platforms

2.2.1 Problem Statement:

Large astronomical databases such as SDSS generate structured photometric and spectroscopic data for millions of celestial objects [1]. These datasets are publicly accessible and support data-driven astronomy research. However, handling high-dimensional photometric data requires efficient computational frameworks.

2.2.2 Loopholes:

- High feature dimensionality leads to complex decision boundaries [5].
- Presence of noisy measurements due to atmospheric and instrumental variations [10].
- Class imbalance between stars, galaxies, and quasars.
- Limited preprocessing and feature engineering in some existing approaches.

2.2.3 Proposed Solutions in Literature:

Researchers have applied dimensionality reduction techniques and statistical feature analysis to improve separability [10]. Data normalization and correlation analysis are commonly used preprocessing strategies. However, integration of robust ensemble learning methods with structured feature engineering remains an area requiring further exploration.

2.3 Machine Learning-Based Stellar Classification

2.3.1 Problem Statement:

Supervised learning algorithms such as Support Vector Machines, Decision Trees, and K-Nearest Neighbors have been widely used for stellar object classification tasks [4][7]. These models learn decision boundaries from labeled photometric data and provide automated classification.

2.3.2 Loopholes:

- Single classifiers are sensitive to noise and overfitting [5].
- Performance may degrade in high-dimensional feature space.
- Limited generalization when training data distribution shifts.
- Insufficient comparison across multiple algorithms in some studies.

2.3.3 Proposed Solutions in Literature:

Ensemble learning techniques such as Random Forest and Gradient Boosting have been proposed to improve classification stability and accuracy [6][8]. Boosting-based models iteratively reduce classification errors and enhance predictive power. Although these approaches demonstrate strong performance, comprehensive comparative frameworks integrating multiple ensemble models are still limited in literature.

2.4 Deep Learning Approaches in Astronomy

2.4.1 Problem Statement:

Recent research explores deep learning models for galaxy morphology classification and image-based object detection [9][12]. Convolutional Neural

Networks (CNNs) are applied directly to astronomical image data.

2.4.2 Loopholes:

- Require high computational resources and GPU infrastructure.
- Large labeled image datasets are necessary for training.
- Less interpretable compared to traditional ensemble models.
- Not always optimal for structured tabular photometric data.

2.4.3 Proposed Solutions in Literature:

Deep learning frameworks show promising results in image-based classification tasks [12][13]. However, for structured SDSS photometric datasets, ensemble machine learning techniques remain computationally efficient and highly accurate alternatives.

2.5 Overcoming the Loopholes [1][2][4]

This paper addresses the loopholes mentioned above in Table 1:

Loopholes in Existing Work	Our Proposed Solution (Stellar Classification Framework)
Manual spectroscopic classification is not scalable	Automated machine learning–based classification using SDSS photometric data
Rule-based threshold methods fail for overlapping classes	Ensemble learning models capture nonlinear decision boundaries
Single classifiers prone to overfitting	Random Forest, Gradient Boosting, and XGBoost improve robustness
Limited feature engineering	Color index generation and correlation-based feature selection
Lack of comprehensive model comparison	Comparative evaluation of multiple supervised and ensemble algorithms
Poor generalization in noisy datasets	Cross-validation and boosting techniques enhance stability

Table 1: Loopholes in Existing Stellar Classification Approaches and Proposed Solutions

3. METHODOLOGY

3.1 Overview

To overcome the limitations of traditional stellar classification methods that rely on manual inspection or rule-based statistical thresholds [2], the proposed Stellar Object Classification framework utilizes supervised machine learning techniques for automated multi-class classification. The system leverages photometric data obtained from the Sloan Digital Sky Survey (SDSS) [1], applies preprocessing and feature engineering techniques, and evaluates multiple classification models including ensemble learning approaches [6][8][9].

The proposed framework consists of structured data preparation, feature enhancement using photometric color indices, supervised model training, and performance evaluation using cross-validation. Ensemble models such as Random Forest and Gradient Boosting are employed to improve classification robustness and generalization performance [6][8].

The proposed methodology consists of four key components:

1. Data Acquisition & Environment Modeling

○ The astronomical dataset obtained from SDSS [1] is treated as the classification environment.
○ Each celestial object contains attributes such as Right Ascension, Declination, redshift, and photometric magnitudes (u, g, r, i, z).
○ The dataset includes labeled categories: Star, Galaxy, and Quasar.
○ The environment represents high-dimensional photometric feature space where each object is mapped to its corresponding class label.

2. Data Preprocessing & Feature Engineering

○ Raw photometric measurements are cleaned to remove duplicates, inconsistencies, and missing values.
○ Feature scaling is applied to normalize magnitude values and prevent dominance of high-range attributes [5].
○ Photometric color indices (differences between adjacent bands) are generated to enhance class separability [10].
○ Correlation analysis is performed to identify redundant features and reduce multicollinearity.

Since astronomical observations may contain measurement noise, preprocessing ensures improved stability of machine learning models.

---

3. Model Training & Classification

○ The processed dataset is divided into training and testing subsets.
○ Multiple supervised learning algorithms are implemented, including:

- Decision Tree
- Support Vector Machine
- Random Forest
- Gradient Boosting
- XGBoost

○ Ensemble models are prioritized because they reduce overfitting and improve predictive stability [6][8][9].
○ Cross-validation is applied to ensure reliable generalization performance across unseen data.

Instead of relying on single-model predictions, ensemble methods aggregate multiple weak learners to produce a stronger classification outcome.

---

4. Evaluation & Decision Making

○ Model performance is evaluated using Accuracy, Precision, Recall, F1-score, and Confusion Matrix.
○ Feature importance analysis is performed to determine which photometric attributes contribute most significantly to classification.
○ The best-performing model is selected as the final Stellar Classification Engine.
○ If multiple models show similar accuracy, the model with better generalization stability and lower variance is preferred [6][8].

3.3 Algorithmic Workflow

1. Initialize the dataset environment using SDSS photometric records [1].
2. Import labeled data containing stars, galaxies, and quasars.
3. Perform preprocessing:
   a. Remove missing and duplicate entries.
   b. Normalize feature values.
   c. Generate color index features.
4. Split the dataset into training and testing subsets.
5. For each classification model:
   a. Train the model using training data.
   b. Predict object categories on testing data.
   c. Compute performance metrics (accuracy, precision, recall, F1-score).
6. Compare performance across all models and identify the classifier with highest predictive accuracy and stability.
7. Deploy the selected ensemble model as the final stellar object classification system.

3.4 Advantages Over Existing Approaches

● Automates stellar classification instead of relying on manual spectroscopic inspection [2].
● Utilizes ensemble learning to handle nonlinear decision boundaries and noisy data [6][8][9].
● Enhances feature representation using photometric color indices [10].
● Reduces overfitting through cross-validation and boosting techniques [5].
● Provides scalable classification suitable for millions of astronomical observations [1].

Table 2: Comparison of Classification Algorithms

Algorithm	Type	Strengths	Limitations	Suitability for Stellar Classification
Decision Tree	Supervised	Easy to interpret; handles nonlinear splits	Prone to overfitting	Useful baseline model but unstable alone
Support Vector Machine	Supervised	Effective in high-dimensional space; strong boundary separation	Computationally expensive for large datasets	Suitable for moderate-sized photometric datasets
Random Forest	Ensemble (Bagging)	Reduces variance; robust to noise; high accuracy	Less interpretable than single tree	Highly suitable for stellar classification
Gradient Boosting	Ensemble (Boosting)	High predictive accuracy; corrects previous errors	Requires careful tuning	Very effective for complex feature interactions
XGBoost	Optimized Boosting	Regularization; scalable; high performance	Hyperparameter sensitive	Excellent for large SDSS datasets
K-Nearest Neighbors	Instance-based	Simple; no training phase	Sensitive to scaling and high dimensions	Limited scalability for large astronomical data

4. DATA COLLECTION

The effectiveness of the proposed Stellar Object Classification system depends on the quality, completeness, and reliability of astronomical survey data. Since machine learning models heavily rely on structured labeled datasets, accurate photometric and spectroscopic measurements are essential for building a robust classification framework.

To develop the stellar classification model, two primary types of datasets were collected:

4.1 Astronomical Photometric Data

● The primary dataset was obtained from the Sloan Digital Sky Survey (SDSS), which provides publicly accessible astronomical data [1].

● The dataset contains labeled observations of celestial objects categorized as Star, Galaxy, and Quasar. Each record includes photometric magnitude measurements in five wavelength bands (u, g, r, i, z), along with spatial coordinates such as Right Ascension (RA) and Declination (Dec), and redshift values.

● Photometric magnitudes capture the brightness of celestial objects across different spectral bands, which helps differentiate stellar bodies based on their emission characteristics [10].

● The dataset also includes additional metadata such as observation identifiers and instrument details, which were filtered during pre-processing to retain only relevant classification features.

The collected SDSS photometric dataset forms the foundation of the classification system, enabling supervised learning algorithms to identify patterns that distinguish stars, galaxies, and quasars.

4.2 Derived Feature Dataset (Color Indices & Engineered Attributes)

● Since raw photometric magnitudes may not always provide optimal class separation, derived features were generated in the form of photometric color indices.

● Color indices were computed by subtracting adjacent magnitude bands (e.g., u–g, g–r, r–i, i–z), which represent spectral slope differences and improve discriminative power [2][10].

● These derived features help reduce ambiguity between overlapping classes and enhance model learning capability.

5.1 System Overview

The architecture is composed of four main modules:

1. Data Collection Module:

This module collects structured astronomical data from SDSS [1], including photometric magnitudes (u, g, r, i, z), spatial coordinates (RA, Dec), and redshift values.

The module ensures that only labeled observations (Star, Galaxy, Quasar) are retained for supervised learning. Raw data files are imported and stored in a structured format suitable for preprocessing and analysis.

● Correlation analysis was conducted to identify redundant or highly correlated attributes. This ensured that the final feature set maintained meaningful information without introducing multicollinearity [5].

Because astronomical datasets may contain measurement inconsistencies or noise, preprocessing and feature engineering steps were essential to ensure high data quality before training machine learning models.

Since SDSS observations may include minor measurement uncertainties due to atmospheric effects and instrument sensitivity variations, ensemble learning techniques—particularly Random Forest and Gradient Boosting—were utilized to handle noise and improve classification stability [6][8].

By integrating high-quality photometric data with engineered spectral features, the system ensures that every celestial object record contains sufficient discriminative information for accurate classification.

5. PROPOSED ARCHITECTURE

The proposed Stellar Object Classification architecture integrates large-scale astronomical survey data with advanced ensemble machine learning techniques to automate the classification of celestial objects [1][6][8]. The system is designed to efficiently process high-dimensional photometric data and accurately categorize objects into stars, galaxies, and quasars. The architecture ensures scalability, robustness, and adaptability for large astronomical datasets such as those provided by the Sloan Digital Sky Survey (SDSS).

2. Pre-processing and Feature Engineering Module:

This module cleans and pre-processes astronomical data by removing missing entries, duplicate observations, and irrelevant metadata [5].

Feature scaling is applied to normalize magnitude values and prevent bias in distance-based models.

To improve class separability, photometric color indices are generated by computing differences between adjacent magnitude bands (u–g, g–r, r–i, i–z) [2][10].

Correlation analysis is performed to eliminate redundant attributes and reduce multicollinearity.

This module ensures that the dataset is optimized for machine learning model training.

3. Model Training and Classification Module:

Multiple classification algorithms are implemented, including Decision Tree, Support Vector Machine, Random Forest, Gradient Boosting, and XGBoost [4][6][8][9].

The dataset is divided into training and testing subsets. Cross-validation techniques are applied to ensure generalization and prevent overfitting.

Ensemble learning methods such as Random Forest and Gradient Boosting are emphasized because they combine multiple weak learners to enhance classification robustness and predictive accuracy [6][8].

The trained models generate predictions for unseen celestial objects and classify them into Star, Galaxy, or Quasar categories.

4. Evaluation and Decision Module:

This module compares performance across all implemented models using evaluation metrics such as Accuracy, Precision, Recall, and F1-score [5].

Feature importance analysis is performed to identify which photometric attributes contribute most significantly to classification performance.

The model demonstrating the highest predictive accuracy and stability is selected as the final Stellar Classification Engine.

If multiple models achieve comparable performance, preference is given to the model with lower variance and better generalization capability [6][8].

5.2 Workflow Description

1. The system imports photometric data from SDSS [1].
2. Data pre-processing is performed, including cleaning, normalization, and feature engineering [2][5].
3. The processed dataset is split into training and testing subsets.
4. Multiple machine learning classification models are trained and validated [4][6][8][9].
5. Model predictions are evaluated using classification performance metrics.
6. The best-performing ensemble model is selected and deployed for automated stellar object classification.

5.3 Advantages of the Proposed Architecture

- Automates stellar classification instead of relying on manual spectroscopic analysis [2].
- Utilizes ensemble learning techniques to improve classification accuracy and robustness [6][8][9].
- Enhances feature representation through photometric color index generation [10].
- Reduces overfitting and improves generalization using cross-validation strategies [5].
- Scalable for millions of astronomical observations generated by modern sky surveys [1].

This architecture enables a shift from manual and rule-based classification toward fully automated, data-driven stellar object classification, supporting the advancement of computational astronomy and large-scale astrophysical research.

6. RESULTS & DISCUSSION

The proposed Stellar Object Classification framework was evaluated using labelled photometric data obtained from the Sloan Digital Sky Survey (SDSS) [1]. Experiments were conducted to compare the performance of traditional single-model classifiers against ensemble-based learning approaches such as Random Forest, Gradient Boosting, and XG Boost [6][8][9].

The objective of the evaluation was to measure classification accuracy, generalization capability, and robustness across multiple celestial object categories (Star, Galaxy, Quasar).

6.1 Classification Accuracy Comparison

Table 3 summarizes the comparative performance of traditional classification algorithms and ensemble learning models.

Classification Strategy	**Accuracy (%)**	**Improvement Over Baseline (%)**
Decision Tree (Baseline)	91.2	–
Support Vector Machine	93.4	2.2%
Random Forest	96.1	5.4%
Gradient Boosting	97.8	7.2%
XGBoost (Proposed System)	97.5	6.9%

Observations:

- Traditional Decision Tree models misclassified certain quasars as stars due to overlapping photometric properties.
- Ensemble models significantly reduced misclassification rates across all three categories.

Table 3. Stellar Classification Performance Comparison

The ensemble-based models consistently outperformed traditional single classifiers. Random Forest improved stability by aggregating multiple decision trees [6], while boosting algorithms further enhanced predictive performance by iteratively correcting classification errors [8][9].

These results demonstrate that ensemble learning significantly enhances classification accuracy compared to standalone decision-tree models.

6.2 Class-wise Performance Analysis

Figure 2 and Figure 3 illustrate confusion matrix comparisons between the baseline Decision Tree model and the proposed Gradient Boosting model.

Fig. 2 Actual Result 1 (Decision Tree Confusion Matrix Representation)

Fig. 3 Actual Result 2 (Gradient Boosting Confusion Matrix Representation)

- Galaxy classification achieved near-perfect precision due to distinctive redshift characteristics [10].
- Boosting algorithms improved minority class detection and reduced false positives [8][9].

The results confirm that ensemble learning improves inter-class boundary separation and handles nonlinear feature interactions effectively.

6.3 Feature Importance Analysis

Feature importance scores extracted from the Random Forest and Gradient Boosting models revealed the following key insights:

1. Photometric color indices (u–g and g–r) were the most influential features.
2. Redshift played a crucial role in distinguishing galaxies and quasars.
3. Spatial coordinates (RA, Dec) had comparatively lower predictive impact.

These findings align with astronomical domain knowledge, confirming that spectral color differences are strong discriminators in stellar classification [2][10].

The inclusion of engineered color indices significantly improved classification stability and separability.

6.4 Model Stability and Generalization

Cross-validation results indicated:

- Ensemble models showed lower variance compared to single decision trees.
- Boosting techniques maintained consistent accuracy across different data splits [6][8].
- Overfitting was minimized through aggregation and regularization mechanisms.

The results confirm that ensemble learning frameworks are well-suited for high-dimensional astronomical datasets containing noisy measurements.

6.5 Discussion

- **Effectiveness:** The proposed ensemble-based classification framework achieved accuracy exceeding 97%, demonstrating strong predictive capability for large-scale astronomical data [6][8][9].

- **Scalability:** The system can efficiently process large SDSS datasets and can be extended to future sky surveys generating millions of observations [1].

- **Robustness:** Boosting algorithms effectively handled overlapping class distributions and noisy photometric measurements.

- **Limitations:** Performance may vary when applied to datasets from different astronomical instruments due to calibration differences or varying observational conditions [10].

Future work includes exploring hybrid deep learning–ensemble architectures and integrating image-based classification techniques alongside structured photometric data to further enhance classification performance.

7. CONCLUSION

This study presents a comprehensive machine learning framework for automated Stellar Object Classification, designed to accurately categorize celestial objects into stars, galaxies, and quasars using large-scale photometric datasets obtained from the Sloan Digital Sky Survey (SDSS) [1]. Unlike traditional classification methods that rely on manual spectroscopic inspection or predefined statistical thresholds [2], the proposed system leverages ensemble learning techniques to adaptively learn complex patterns from high-dimensional astronomical data [6][8][9].

The implemented framework integrates data preprocessing, photometric color index generation, supervised model training, and comparative evaluation of multiple classification algorithms. Ensemble methods such as Random Forest, Gradient Boosting, and XGBoost demonstrated superior performance in terms of classification accuracy, robustness, and generalization capability when compared to conventional single-model approaches [6][8][9].

Experimental results confirm that boosting-based ensemble models achieve accuracy exceeding 97%, significantly reducing misclassification among overlapping object categories. The findings validate the effectiveness of data-driven astronomical classification systems and emphasize the importance of feature engineering and ensemble learning in modern computational astrophysics [5][10].

Overall, this research highlights the potential of machine learning–based approaches to transform stellar classification from manual and rule-based methodologies to scalable, automated, and highly accurate systems. The proposed framework supports the advancement of data-driven astronomy and can be extended to future large-scale sky surveys, enabling efficient analysis of ever-growing astronomical datasets.

8. FUTURE SCOPE

The Stellar Object Classification system offers several opportunities for further research and practical deployment in large-scale astronomical surveys. Future enhancements may include:

1. Integrating real-time sky survey streams from next-generation astronomical missions such as the Vera C. Rubin Observatory and space-based telescopes to test the model under continuously updating observational conditions [3][4].
2. Expanding the classification framework to support multi-class categorization beyond stars, galaxies, and quasars, including rare objects such as pulsars, brown dwarfs, and active galactic nuclei.

3. Incorporating deep learning architectures such as Convolutional Neural Networks (CNNs) to directly analyze raw astronomical images instead of relying only on tabular photometric features [8][9].
4. Combining ensemble learning with automated hyperparameter optimization techniques to further improve prediction accuracy and robustness in highly imbalanced datasets.
5. Enhancing explainability by integrating interpretable AI techniques, enabling astronomers to better understand feature importance and model decision-making processes.
6. Deploying the system as a scalable cloud based analytical platform capable of handling petabyte-scale datasets generated by modern sky surveys.

9. REFERENCES

[1] Sloan Digital Sky Survey (SDSS). [Online]. Available: https://www.sdss.org. Accessed: Nov. 2025.

[2] Zeraatgari, F. Z., et al., "Photometric classification of stars, galaxies, and quasars using machine learning," *arXiv preprint arXiv:2311.02951*, 2023.

[3] Zhang, Y., et al., "Stellar classification with vision transformer and SDSS photometric images," *arXiv preprint*, 2024.

[4] He, H., et al., "Stellar-ViT: Vision transformer for stellar classification," *Universe*, vol. 10, no. 5, pp. 214, 2024.

[5] Li, X., et al., "Active learning for stellar spectral classification," *arXiv preprint arXiv:2406.18366*, 2024.

[6] Kumar, A., et al., "Galaxy classification using machine learning models," *arXiv preprint arXiv:2603.24435*, 2026.

[7] Chen, Y., et al., "Machine learning for predicting galaxy properties," *arXiv preprint arXiv:2405.15566*, 2024.

[8] Smith, J., et al., "Unsupervised machine learning for galaxy evolution," *arXiv preprint arXiv:2404.09958*, 2024.

[9] Wang, L., et al., "CNN-based star-galaxy classification using SDSS data," *arXiv preprint arXiv:2404.01049*, 2024.

[10] Patel, R., et al., "Stellar life cycle prediction using machine learning," *Preprints.org*, 2025.

[11] Rahman, M., et al., "Automated classification of celestial objects using machine learning," *International Journal of Pervasive Computing and Communications*, 2025.

[12] Garcia, P., et al., "Random forest for astronomical spectral classification," *Algorithms*, vol. 16, no. 6, pp. 293, 2023.