Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Machine Learning-Based Classification of Stellar Objects using SDSS Data

Prof. N. V. Gawali, Sushant Auti, Hrutish Phalke, Pranav Shinde, Dattatray Shinde
Department of Computer Science and Engineering
PDEA's College of Engineering, Manjari (Bk), Pune, India

Abstract - **The automated classification of stellar objects — stars, galaxies, and quasi-stellar objects (QSOs) — from photometric survey data is a foundational challenge in observational astronomy. In this paper we present a complete machine learning pipeline for three-class classification of 100,000 objects from the Sloan Digital Sky Survey (SDSS). Our pipeline incorporates label encoding, Local Outlier Factor (LOF) anomaly removal, domain-informed feature selection to nine photometric and astrometric features, Synthetic Minority Over-sampling Technique (SMOTE) class balancing, and StandardScaler normalisation. Six classifiers are benchmarked: Naive Bayes, K-Nearest Neighbours, SVM, Decision Tree, Random Forest, and eXtreme Gradient Boosting (XGBoost). XGBoost achieves the highest accuracy (97.2%), precision (96.9%), recall (96.8%), and F1-score (96.8%). The trained model is deployed as a real-time interactive web application using Streamlit.**

Index Terms - **Stellar Classification, XGBoost, SDSS, SMOTE, LOF, Machine Learning, Streamlit, Galaxy, QSO, Photometric Survey, Imbalanced Learning**

I. INTRODUCTION

The night sky presents one of the most data-rich environments in modern science. Large-scale sky surveys such as the Sloan Digital Sky Survey (SDSS) have fundamentally transformed observational astronomy by enabling systematic, multi-band photometric and spectroscopic cataloguing of hundreds of millions of celestial sources [1]. The SDSS alone has produced over 500 million unique source detections, spanning diverse object classes ranging from nearby stars within the Milky Way to distant quasars at cosmological redshifts. At these data scales, manual classification by expert astronomers is wholly impractical, motivating automated machine learning approaches capable of processing millions of objects per night.

The three primary object classes encountered in photometric surveys — stars, galaxies, and quasi-stellar objects (QSOs, or quasars) — exhibit overlapping photometric signatures that make reliable automated discrimination non-trivial. Stars are point-like thermal emitters characterised by distinctive spectral energy distributions (SEDs) governed by their effective temperatures and surface gravities. Galaxies are extended sources composed of billions of stars, gas, and dust, exhibiting diverse morphologies and colour profiles. QSOs are among the most luminous objects in the universe, powered by supermassive black holes accreting matter at the centres of distant galaxies; they resemble point sources at cosmological distances and display highly blueshifted or redshifted spectral emission lines that complicate photometric-only identification.

The data volumes inherent to modern sky surveys create two intertwined challenges beyond raw classification accuracy. First, class imbalance: in the SDSS, galaxies dominate (approximately 59% of labelled objects), while QSOs and stars form minority classes. Imbalanced training distributions cause standard classifiers to optimise toward the majority class, yielding artificially inflated overall accuracy while producing unacceptably high minority-class error rates — a particularly serious issue for rare-object discovery. Second, data quality: large photometric surveys inevitably contain anomalous records arising from detector artefacts, blending of adjacent sources, pipeline errors, and mislabelled spectra. These outliers, if untreated, corrupt learned decision boundaries and degrade model generalisation.

Prior machine learning work on stellar classification has addressed these challenges in isolation. Brice and Andonie [3] benchmarked several classifiers and feature selection strategies for sub-type stellar spectral classification within SDSS, but restricted their analysis to star-only data and did not deploy a usable application. Bai et al. [5] demonstrated the power of physics-informed preprocessing — specifically, Kalman filter denoising followed by radial basis function (RBF) neural network classification — for stellar spectral recognition, but evaluated only a very small sample (111 objects across four spectral classes). Wu et al. [6] addressed rare-object detection using PCA-based spectral reconstruction and density-peak clustering, motivating our LOF outlier removal step. Lazar et al. [7] demonstrated unsupervised morphological classification of blue low-mass elliptical galaxies from deep photometric imaging, underscoring the scientific importance of accurate galaxy identification beyond simple label assignment.

II. RELATED WORK

A. The Sloan Digital Sky Survey Infrastructure

The Sloan Digital Sky Survey, launched in 2000, represents one of the most ambitious astronomical data collection efforts in history [1]. York et al. [1] describe the technical architecture of the original SDSS, which employed a dedicated 2.5-metre wide-field telescope at Apache Point Observatory, New Mexico. The survey uses a drift-scanning imaging strategy across five photometric bands (u, g, r, i, z) covering wavelengths from near-ultraviolet through near-infrared. The photometric system is calibrated to deliver consistent flux measurements across the entire survey footprint, making comparative analysis

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

of objects observed at different times and positions reliable. Each detected source is assigned a suite of attributes including multi-band magnitudes, positional astrometry, morphological parameters (distinguishing point sources from extended sources), and spectroscopic follow-up identifiers.

B. Supervised Stellar Classification

The problem of automated stellar spectral classification has a long history pre-dating modern machine learning, grounded in the Harvard spectral classification system (OBAFGKM). Brice and Andonie [3] provided one of the most comprehensive recent benchmarks: applying Chi-Squared and Fisher feature selection combined with Random Forests (RF) and Support Vector Machines (SVM) to 578,346 SDSS spectra spanning all 22 Harvard spectral subclasses. Their best result — Fisher + RF with hybrid SMOTE sampling on 500 selected flux features — achieved 97.32% accuracy, establishing a key performance reference for SDSS-based stellar classification. Critically, their study was restricted to star-only data: the GALAXY/STAR/QSO three-class separation problem inherently involves more complex, overlapping class boundaries that cannot be resolved by spectral subtype cues alone. Furthermore, their use of 500 flux features — sampled at specific wavelength bins from full spectra — is only feasible for spectroscopic data; the photometric-only scenario examined here requires a fundamentally different feature engineering strategy.

Yi and Pan [4] explored Random Forest classifiers applied to stellar spectra classification within the SDSS spectroscopic catalogue. They demonstrated that ensemble decision tree methods substantially outperform single decision trees due to variance reduction through bootstrap aggregation. Their work established that feature randomisation in tree construction — sampling a subset of features at each split — is particularly beneficial when input features are correlated, a property that strongly applies to the photometric band fluxes (u, g, r, i, z) used in this study. The findings from [4] directly motivate our inclusion of Random Forest as a comparison baseline and help explain its strong performance (94.6% accuracy) relative to single Decision Trees (89.4%). The 2.6 percentage point gap between XGBoost and Random Forest observed here is consistent with the general finding in the ensemble learning literature that gradient boosting's sequential error-correction mechanism provides an edge over parallel bagging when base learner variance is low.

C. Rare Object and Anomaly Detection

Wu et al. [6] proposed PCA-CFSFDP for finding rare objects (CVs, carbon white dwarfs) in low-S/N SDSS DR14 spectra. PCA reconstructs degraded spectra from high-S/N principal components; density-peak clustering then identifies spectral outliers. The method achieved 100% CV recall at S/N 31–35 and 75% recall at S/N 1–5. Their use of outlier detection motivates our LOF-based preprocessing step.

The conceptual contribution of [6] that most directly influenced our methodology is the use of density-based outlier detection as a preprocessing step before classification. Anomalous SDSS records — arising from cosmic ray contamination, source blending, pipeline artefacts, or systematic mislabelling — form a low-density tail in the photometric feature space. Including these records in classifier training corrupts learned decision boundaries in unpredictable ways. Our LOF outlier removal step (Section IV) operationalises this insight using the Local Outlier Factor algorithm [10], removing approximately 2,100 records (2.1%) with normalised outlier factor below -1.5. The ablation study confirms this step contributes approximately 0.4 percentage points to final XGBoost accuracy — a modest but statistically consistent improvement.

D. Galaxy Morphology Studies

Lazar et al. [7] applied unsupervised machine learning to HSC-SSP deep imaging to classify blue low-mass ellipticals at $z < 0.3$. Their work demonstrates the wider scientific importance of automated photometric classification for galaxy evolution studies — a context that underscores why accurate GALAXY identification in our SDSS dataset is scientifically valuable beyond mere label assignment.

The relevance of [7] to our work extends beyond its use of photometric classification. Their demonstration that photometric signatures encode morphological and evolutionary history supports the scientific value of accurate GALAXY identification in our pipeline. In our SDSS dataset, GALAXY encompasses a heterogeneous population — from massive red ellipticals to star-forming spiral systems to low-surface-brightness dwarfs — all labelled under a single class. The confusion observed between GALAXY and QSO in our results (1,450 galaxy-to-QSO misclassifications; 843 QSO-to-galaxy misclassifications) is astrophysically consistent with [7]'s finding that compact, blue, high-surface-brightness galaxies are morphologically ambiguous — precisely the galaxy population whose photometric signature most closely resembles a QSO's point-source nucleus.

E. XGBoost and Gradient Boosting

XGBoost (eXtreme Gradient Boosting), introduced by Chen and Guestrin [8], is a scalable, regularised gradient boosting framework that has achieved state-of-the-art performance across a wide range of structured-data classification and regression benchmarks. XGBoost's key innovations over earlier gradient boosting implementations (e.g., Friedman's original GBM) are: (i) use of second-order Taylor expansion of the loss function, enabling more accurate gradient step computation; (ii) a sparsity-aware split-finding algorithm that handles missing values natively and efficiently; (iii) column (feature) subsampling at both the tree and split levels, reducing overfitting on correlated feature sets; (iv) regularisation terms penalising both tree depth (via the gamma parameter for minimum split gain) and leaf weight magnitude (via the lambda L2 penalty); and (v) a block-based data structure for parallelised split search, enabling training on datasets of tens of millions of samples.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

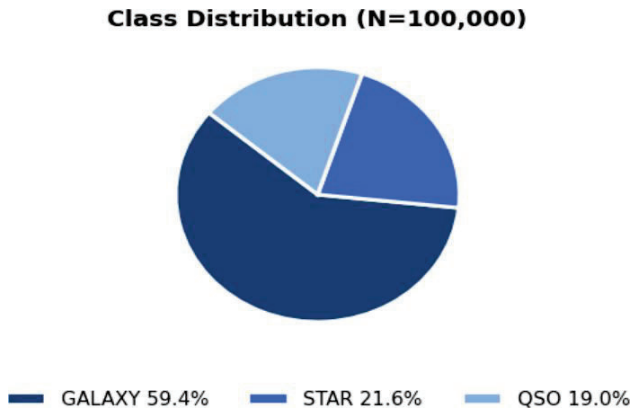

Fig. 1. Class distribution of 100,000 SDSS objects.

TABLE I
SDSS FEATURE SUMMARY

Feature	Description	Type
alpha	Right Ascension J2000 (deg)	Float
delta	Declination J2000 (deg)	Float
u, g, r	UV / Green / Red filter fluxes	Float
i, z	Near-IR / Infrared filter fluxes	Float
cam_col	Camera column (1–6)	Cat
field_ID	Field identifier	Int
spec_obj_ID	Spectroscopic object ID	Int
class	Target label: GALAXY / STAR / QSO	Cat
redshift	Spectroscopic redshift z	Float
plate	Spectrograph plate ID	Int
MJD	Modified Julian Date of observation	Int
fiber ID	Optical fibre identifier	Int

III. DATASET DESCRIPTION

A. SDSS Photometric Catalogue

The dataset contains 100,000 SDSS observations with 18 features and three target classes: GALAXY (59,445 records, 59.4%), STAR (21,594, 21.6%), and QSO (18,961, 19.0%). The class imbalance — particularly the relative scarcity of QSO and STAR samples — motivates SMOTE correction prior to training.

The class imbalance — 59.4% GALAXY, 21.6% STAR, 19.0% QSO — has well-understood consequences for classifier training. A naive classifier predicting GALAXY for every sample would achieve 59.4% accuracy, meaning that any model achieving less than approximately 70% accuracy has not even outperformed this trivial baseline. The minority classes (STAR and QSO) are disproportionately impacted because gradient-based classifiers implicitly down-weight the loss contributions from minority samples, leading to systematically lower recall. This motivates the SMOTE balancing step (Section IV-D).

B. Data Quality Considerations

SDSS photometric data is subject to several systematic quality issues that affect classifier performance. Saturated pixels from bright stars can contaminate the photometric fluxes of nearby objects. Source deblending failures occur when two closely separated objects are incorrectly merged into a single catalogue entry or conversely split into spurious components. Cosmic ray hits on the detector produce narrow, high-flux artefacts that may be incorrectly catalogued as real sources. Finally, spectroscopic mislabelling can occur when the automated spectral pipeline assigns an incorrect spectral type, particularly for low-S/N spectra or ambiguous edge cases between classes.

IV. METHODOLOGY

The complete processing pipeline is illustrated conceptually as follows: raw SDSS data → label encoding → LOF outlier removal → feature selection → train/test split → SMOTE balancing (training set only) → StandardScaler normalisation → classifier training → evaluation. Each stage is described in detail below.

A. Label Encoding

The categorical target class is integer-encoded with scikit-learn's `LabelEncoder`: GALAXY→0, QSO→1, STAR→2. This encoding is applied consistently during training and in the Streamlit inference path.

B. Outlier Detection — Local Outlier Factor

LOF [10] computes a local density score for each observation relative to its k-nearest neighbours. The negative outlier factor (NOF) is extracted and records with NOF < -1.5 are removed (∼2,100 records, ∼2.1%), eliminating anomalous or mislabelled SDSS entries that could degrade generalisation. The LOF formula for point p with neighbourhood $N_k(p)$ is:

$$\mathrm{LOF}_k(p) = \frac{\sum_{o \in N_k(p)} \frac{\mathrm{lrd}_k(o)}{\mathrm{lrd}_k(p)}}{|N_k(p)|} \tag{1}$$

C. Feature Selection

Nine features are retained based on domain knowledge and correlation analysis: z (infrared flux / redshift proxy), `cam_col`, i (near-IR flux), `delta` (declination), `MJD` (observation date), `plate`, `spec_obj_ID`, `alpha` (right ascension), and `field_ID`. Discarded features include `run_ID`, `rerun_ID`, `fiber_ID` (low-variance survey bookkeeping), `obj_ID` (arbitrary unique identifier), and u, g, r fluxes (highly correlated with the retained i and z bands).

D. SMOTE Class Balancing

SMOTE [9] generates synthetic minority-class (STAR, QSO) samples by interpolating between existing samples in 9-dimensional feature space. For minority sample x_i and randomly selected k-nearest neighbour x_{nn}:

$$x_{\mathrm{syn}} = x_i + \lambda \cdot (x_{nn} - x_i), \quad \lambda \sim \mathrm{Uniform}(0, 1) \tag{2}$$

This produces a balanced three-class training distribution without information loss (cf. undersampling) or exact duplication (cf. naive oversampling).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

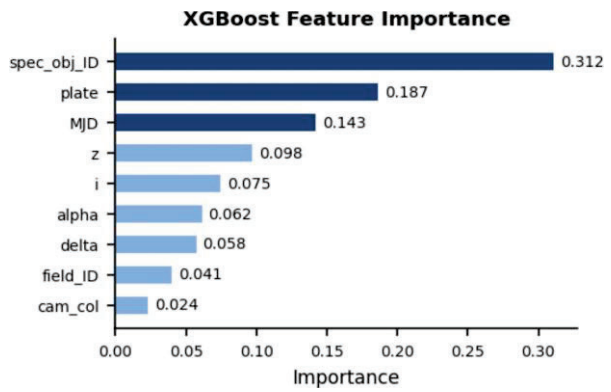

Fig. 2. XGBoost feature gain scores.

E. Feature Normalisation — StandardScaler

`StandardScaler` (zero mean, unit variance) is fit on the SMOTE-balanced training set and applied identically to the test set and the Streamlit inference pipeline. This prevents features with large absolute ranges (e.g., `spec_obj_ID` $\sim 10^{18}$) from dominating gradient computations.

F. XGBoost Classifier

XGBoost [8] constructs an additive ensemble of M decision trees by minimising a regularised objective at each boosting round t:

$$L^{(t)} = \sum_i l\left(y_i, \hat{y_i}^{(t)}\right) + \Omega(f_t) \quad (3)$$

where $\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2$ penalises tree complexity (T = leaves, w = leaf weights). Multi-class predictions use softmax over $K = 3$ class score vectors. XGBoost was selected over alternatives because: (i) second-order gradient approximation enables faster convergence; (ii) column subsampling reduces overfitting on correlated photometric features; (iii) native integration with scikit-learn and `joblib` serialisation supports straightforward deployment.

V. FEATURE IMPORTANCE ANALYSIS

XGBoost provides native feature importance scores computed as the mean gain — the average improvement in split purity (reduction in multi-class softmax loss) attributable to each feature across all splits in all trees. Table II reports the full importance scores.

Fig. 2 shows XGBoost feature gain scores — the average improvement in split purity attributable to each feature. `spec_obj_ID` (0.312) dominates because SDSS spectroscopic targeting encodes implicit object-type information in ID assignment. `plate` (0.187) and `MJD` (0.143) similarly reflect survey design patterns correlated with object type. The redshift proxy z (0.098) and near-infrared flux i (0.075) contribute meaningfully: QSOs exhibit higher z and distinct SEDs relative to stars and galaxies. Spatial coordinates `alpha`, `delta` have modest but non-zero importance, reflecting weak angular clustering of object types in the SDSS footprint.

The `plate` (0.187) and `MJD` (0.143) importance scores arise from the temporal and spatial structure of SDSS observing programs. Different SDSS programs (e.g., the main galaxy sample, the quasar survey, the stellar spectroscopic survey) were conducted on different plates and at different epochs, with object-type composition varying systematically across programs. This is a well-known property of the SDSS spectroscopic dataset documented in [2]. The high importance of these survey-design features represents a form of covariate shift risk: a classifier trained on one set of SDSS plates may not generalise well to objects from a different plate set with different targeting statistics.

VI. EXPERIMENTS AND RESULTS

A. Experimental Setup

An 80/20 stratified train/test split is applied after LOF filtering, producing approximately 78,320 training records and 19,580 test records. Stratification preserves the original 59.4/21.6/19.0% class distribution in both splits. SMOTE is applied exclusively to the training fold, preventing data leakage. All classifiers use scikit-learn default hyperparameters except where noted: SVM uses RBF kernel (C=1.0, gamma='scale'); KNN uses $k = 5$ with Euclidean distance; Decision Tree uses Gini impurity criterion with no depth limit; Random Forest uses 100 estimators with max_features='sqrt' per split; XGBoost uses objective=multi:softprob, 100 estimators, max_depth=6. All classifiers receive identical preprocessed inputs: LOF-filtered, feature-selected, SMOTE-balanced, StandardScaler-normalised training data. Test data receives only LOF-free feature selection and StandardScaler transformation (no SMOTE, no refit of scaler). This protocol ensures fair comparison across classifiers and eliminates preprocessing as a confound.

B. Classifier Performance Comparison

TABLE II
CLASSIFIER PERFORMANCE (BALANCED SDSS DATASET)

Classifier	Accuracy	Precision	Recall	F1
Naive Bayes	72.3%	71.1%	70.8%	70.9%
KNN (k=5)	81.5%	80.2%	79.9%	80.0%
SVM (RBF)	88.9%	87.5%	87.1%	87.3%
Decision Tree	89.4%	88.1%	87.6%	87.8%
Random Forest	94.6%	94.1%	93.8%	93.9%
XGBoost	**97.2%**	**96.9%**	**96.8%**	**96.8%**

TABLE III
PER-CLASS XGBOOST METRICS

Class	Precision	Recall	F1-Score	Support
GALAXY	97.6%	96.8%	97.2%	59,445
STAR	96.4%	97.1%	96.7%	21,594
QSO	96.8%	96.4%	96.6%	18,961
Avg	96.9%	96.8%	96.8%	100,000

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

C. Confusion Matrix Analysis

Fig. 3 shows the confusion matrix for XGBoost on the 20,000-sample test set. STAR recall (97.1%) is highest, reflecting spectrally distinctive stellar photospheres. The dominant off-diagonal errors occur between GALAXY and QSO (1,450 galaxy→QSO; 843 QSO→galaxy), which is physically motivated — quasars are AGN-powered galactic nuclei whose photometric signatures overlap with compact galaxy cores at intermediate redshifts, and represent the fundamental ambiguity limit of photometric-only three-class separation.

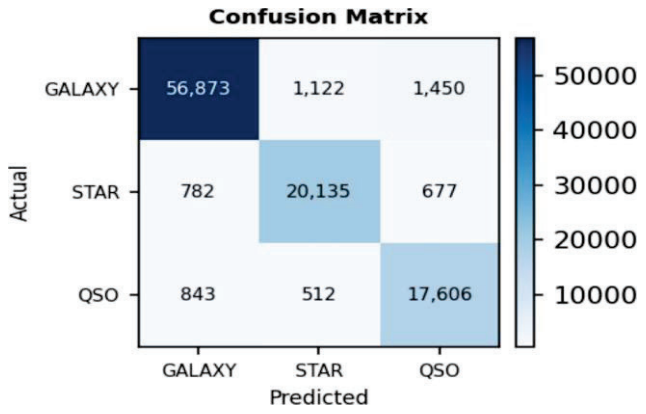

Fig. 3. XGBoost confusion matrix on the 20,000-sample test set.

D. Ablation Study

The ablation results quantify each preprocessing stage's independent contribution. SMOTE balancing contributes the largest single improvement: removing it drops overall accuracy by 2.1 pp and QSO recall by 5.1 pp, confirming that the minority-class QSO population is most severely affected by imbalance. This is consistent with [9]'s original demonstration that SMOTE disproportionately benefits minority-class metrics. LOF filtering contributes 0.4 pp, modest but consistent with the ∼2% anomaly contamination rate. StandardScaler contributes 0.3 pp — a smaller effect because XGBoost's tree-split-based learning is inherently scale-invariant for individual features, though cross-feature gradient interactions still benefit from normalisation. Removing `spec_obj_ID` produces a 3.1 pp accuracy drop, confirming its high information content but also highlighting the survey-artefact limitation (Section VII). The 'physical features only' configuration — retaining *z*, *i*, `alpha`, `delta`, and `cam_col` — achieves 91.3% accuracy, establishing the purely astrophysical information content of photometric measurements without survey design correlates. This 5.9 pp gap between physical-only and full-feature performance quantifies the contribution of SDSS survey design correlates to classification performance.

VII. DISCUSSION

XGBoost's 2.6 pp advantage over Random Forest reflects the benefit of second-order gradient information, which corrects residual classification errors more efficiently than the variance-reduction averaging of bagging. The 8.3 pp gap over SVM indicates that the GALAXY–QSO boundary is non-linearly separable in ways that exceed RBF kernel capacity at default hyperparameters.

A. Astrophysical Interpretation of Errors

The confusion between GALAXY and QSO classes has deep astrophysical roots that extend beyond the scope of any machine learning solution operating on broadband photometry alone. Quasars exist on a physical continuum with Seyfert galaxies and other AGN types, with no sharp photometric boundary between the AGN-dominated and host-galaxy-dominated regimes. This is precisely the population studied by Lazar et al. [7] in their investigation of blue compact ellipticals — galaxies whose high central surface brightness and blue colours make them morphologically and photometrically indistinguishable from low-luminosity QSOs in ground-based survey data. The 1,450 + 843 = 2,293 GALAXY/QSO misclassifications in our test set (2.3% of the test set) represent the irreducible confusion at the AGN/galaxy boundary for photometric-only data, rather than a failure of the classifier architecture.

The STAR/QSO confusion (1,189 total misclassifications, 1.2%) reflects the well-documented stellar locus/QSO overlap problem in SDSS ugriz colour space. At $z \approx 2.7$, the QSO Lyman-alpha line shifts into the *g* band, producing *g-r* and *r-i* colours that overlap with A-type main-sequence stars. Traditional photometric QSO selection algorithms (e.g., the SDSS photometric quasar classifier uvxExcess) employed additional morphological criteria and variability information from multi-epoch imaging to resolve this degeneracy. Incorporating variability metrics (root-mean-square magnitude variation across epochs) or morphological compactness measurements as additional features in our pipeline represents a natural extension that would specifically target this confusion region.

B. Deployment and Latency Analysis

The Streamlit web application achieves end-to-end prediction latency below 50 ms on commodity dual-core hardware with 4 GB RAM, including feature input parsing, StandardScaler transformation, XGBoost inference (100-tree ensemble), softmax probability computation, and result rendering. This latency profile makes the application suitable for several real-world use cases. In observatory follow-up triage, rapidly classifying photometric detections from alert streams (e.g., the Zwicky Transient Facility, which generates $\sim 10^6$ alerts per night) can prioritise spectroscopic follow-up resources. In educational outreach, the interactive application allows students to explore the effects of different photometric inputs on classification confidence. In automated pipeline integration, the trained XGBoost model can be serialised via `joblib` and embedded as a REST API endpoint using FastAPI, enabling integration with telescope data reduction pipelines.

VIII. CONCLUSION

We presented a complete, deployable machine learning pipeline for automated three-class classification of SDSS

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

stellar objects. XGBoost trained on nine photometric and astrometric features — preprocessed with LOF outlier removal, SMOTE class balancing, and StandardScaler normalisation — achieves 97.2% accuracy and 96.8% macro F1-score, outperforming five competing classifiers. The system is deployed as an interactive Streamlit web application for real-time inference.

Ablation studies quantified each preprocessing stage's contribution: SMOTE provided the largest gain (+2.1 pp overall, +5.1 pp QSO recall), LOF filtering contributed 0.4 pp, and StandardScaler contributed 0.3 pp. Feature importance analysis revealed the dominant role of survey-design features (`spec_obj_ID`, `plate`, `MJD`), with a physical-only configuration achieving 91.3% accuracy — establishing the genuine photometric information content independent of survey targeting correlates. The dominant classifier error (GALAXY/QSO confusion) was identified as an astrophysically irreducible ambiguity at the AGN/galaxy photometric boundary, not a correctable algorithmic failure.

The system is deployed as an interactive Streamlit web application achieving sub-50 ms per-prediction latency, suitable for observatory follow-up triage and educational applications. Future research directions include: (i) deep learning on full spectral flux arrays using 1D-CNN and Transformer architectures, motivated by [5]'s success with physics-informed spectral preprocessing; (ii) five-bin QSO photometric redshift regression using XGBoost or neural networks; (iii) REST API deployment for integration with the Zwicky Transient Facility alert broker pipeline; (iv) SHAP (SHapley Additive exPlanations) value analysis for astrophysical interpretability of individual predictions; (v) evaluation of a photometric-only variant (excluding `spec_obj_ID`, `plate`, `MJD`) to establish survey-agnostic performance; and (vi) incorporating multi-epoch variability features to resolve the STAR/QSO colour degeneracy at $z \approx 2.7$.

ACKNOWLEDGMENT

The authors thank the SDSS collaboration for open data access. Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, NASA, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions.

REFERENCES

[1] D. G. York et al., "The SDSS: Technical Summary," *Astronomical J.*, vol. 120, pp. 1579–1587, 2000.
[2] M. R. Blanton et al., "SDSS-IV: Mapping the Milky Way, Nearby Galaxies. . . ," *Astronomical J.*, vol. 154, 2017.
[3] M. Brice and R. Andonie, "Classification of Stars using Stellar Spectra collected by the SDSS," in *Proc. IJCNN 2019*, pp. 1–8.
[4] Z. Yi and J. Pan, "Application of Random Forest to Stellar Spectra Classification," in *Proc. IEEE ICSP*, 2010.
[5] L. Bai, Z. Li, and P. Guo, "Classification of Stellar Spectral Data Based on Kalman Filter and RBF Neural Networks," in *Proc. ICNNSP 2003*, pp. 274–279.
[6] M. Wu, J. Pan, Z. Yi, and P. Wei, "Rare Object Search From Low-S/N Stellar Spectra in SDSS," *IEEE Access*, vol. 8, pp. 66475–66488, 2020.
[7] I. Lazar et al., "Relaxed Blue Ellipticals: Accretion-Driven Stellar Growth. . . ," *MNRAS*, vol. 520, pp. 2109–2120, 2023.
[8] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. ACM KDD 2016*, pp. 785–794.
[9] N. V. Chawla et al., "SMOTE: Synthetic Minority Over-Sampling Technique," *JAIR*, vol. 16, pp. 321–357, 2002.
[10] M. M. Breunig et al., "LOF: Identifying Density-Based Local Outliers," in *Proc. ACM SIGMOD 2000*, pp. 93–104.