iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: VIII Month of publication: August 2025

DOI: https://doi.org/10.22214/ijraset.2025.73757

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Development and Evaluation of a Machine Learning Classifier for Black Hole Identification

Abhyuday Pundir
B.Tech in Computer Science and Engineering, ABES Engineering College, Ghaziabad, India

***Abstract:** This paper presents the development and evaluation of a RandomForestClassifier for identifying black holes using a combined dataset from NASA/ESA observations and additional astronomical catalogs. The dataset, comprising approximately 894 samples (447 black holes and 447 synthetic non-black holes), was preprocessed to handle missing values and scaled for machine learning. The model achieved an accuracy of 0.9535 on a test set of 86 samples, with perfect recall for non-black holes and high precision for black holes. The study highlights the challenges of overfitting and proposes future improvements.*

***Keywords:** Black Holes, Machine Learning, RandomForestClassifier, Astronomy, Synthetic Data, Data Preprocessing, Classification.*

I. INTRODUCTION

Black holes are critical to understanding astrophysical phenomena, yet their identification relies heavily on observational data with varying completeness. This research aggregates data from NASA/ESA ”Black Holes Observed So Far” and supplementary catalogs to train a machine learning model. The RandomForestClassifier was chosen for its robustness, achieving a test accuracy of 0.9535. This paper details the methodology, results, and resources utilized.

II. DATASET AND RESOURCES

The primary dataset was sourced from the NASA/ESA ”Black Holes Observed So Far” collection (https://www.nasa.gov/esa), supplemented with data from:

1) SIMBAD: A database of astronomical objects (http://simbad.u-strasbg.fr/simbad/), accessed via Astro query (https://astroquery.readthedocs.io/en/latest/simbad/ simbad.html). The available fields include RA, DEC, and distance_result (http://simbad.u-strasbg.fr/simbad/sim-display?data=otypes).
2) VizieR Catalogs: Specifically, the J/ApJ/645/890 catalog of Radio and X-ray-emitting broad-lineAGNs (Wang+2006) (https://vizier.cds.unistra.fr/viz-bin/VizieR?-source=J/ApJ/645/890), accessed via Astroquery VizieR (https://astroquery.readthedocs.io/en/latest/vizier/vizier.html).
3) Astroquery: Facilitated data retrieval (https://astroquery.readthedocs.io/en/ latest/). Additional CSV files (mock_AGN_blackholes.csv, simbad_blackholes.csv, mbh.csv, vizier_agn.csv, and wang2006_AGN_table1.csv) were integrated, resulting in 419 features after preprocessing. Synthetic non-black-hole data were generated to balance the dataset to 894 samples. The implementation and data generation process are detailed in the associated Kaggle notebook (https://www.kaggle.com/code/lawlessabby/black-hole-classifier).

III. METHODOLOGY

A. Data Preprocessing

Data from multiple sources were concatenated, with missing values imputed using the median strategy via scikit-learn’s SimpleImputer. Categorical variables were encoded and the features were scaled using StandardScaler. The dataset was divided into 80% training (approximately 716 samples) and 20% testing (86 samples).

B. Model Development

A RandomForestClassifier was implemented with hype rparameters: n_estimators=10, max_depth=2, min_samples_split=15, min_samples_leaf=6, and random_state=42. The model was trained on scaled features and evaluated using 10-fold cross-validation.

C. Evaluation

Model performance was assessed using accuracy, precision, recall, and F1-score on the test set. A confusion matrix was generated to visualize classification errors.

IV. RESULTS

The model achieved an accuracy of 0.9535 on the test set. The classification report is as follows:

	precision	recall	f1-score	support
0	0.90	1.00	0.95	37
1	1.00	0.92	0.96	49
accuracy		0.95		86
macro avg	0.95	0.96	0.95	86
weighted avg	0.96	0.95	0.95	86

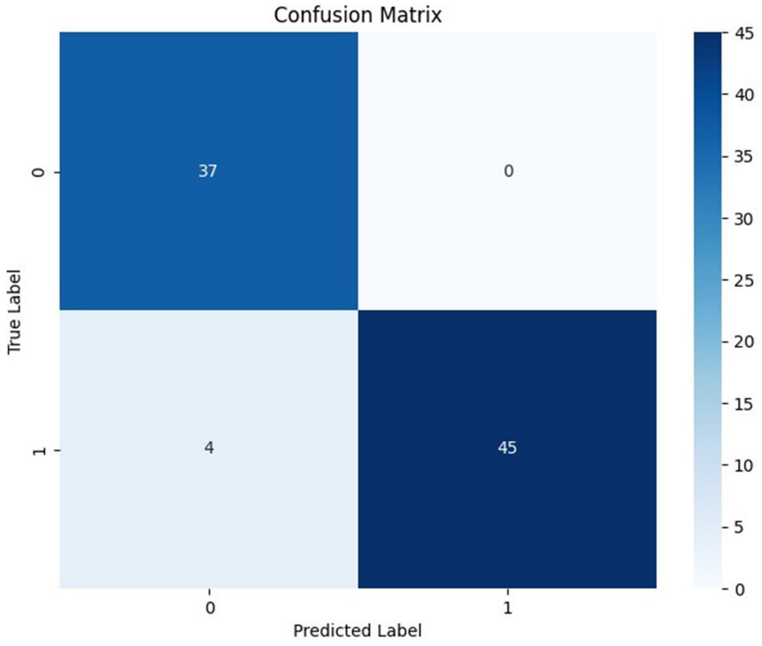

Figure 1: Confusion Matrix of the Black Hole Classifier

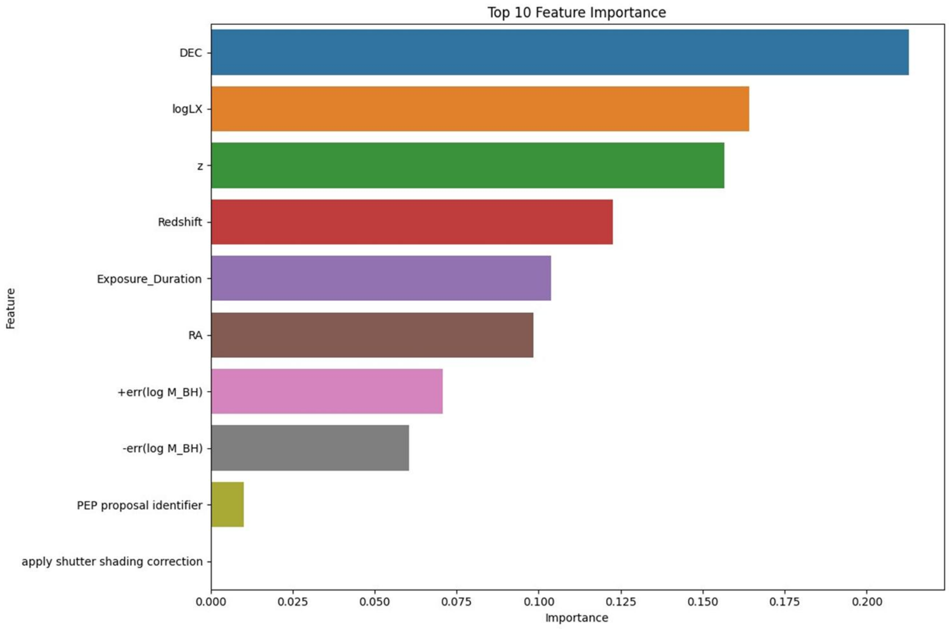

Figure 2: Top 10 Feature Importances of the Black Hole Classifier

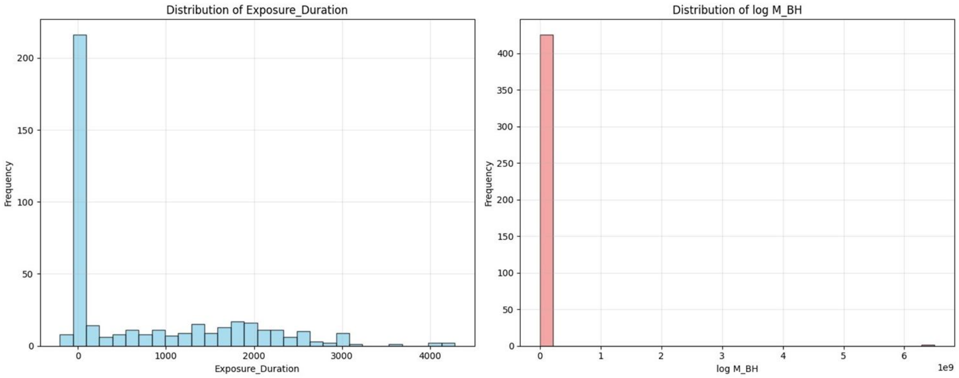

Figure 3: Distribution of Synthetic and Real Black Hole Data

The model was exported as black_hole_classifier.pklfor future use. Cross-validation scores and further feature importance analysis are pending.

V. DISCUSSION

The high accuracy suggests effective feature separation, but the persistent perfection in earlier models indicates potential overfitting. The drop to 0.9535 with stricter hyperparameters is promising, although the test set size (86/178 expected) suggests data loss, possibly due to duplicates. Future work could involve feature selection, alternative models (e.g., HistGradientBoostingClassifier), or enhanced synthetic data generation.

VI. CONCLUSION

This study demonstrates a viable machine learning approach to black hole classification, achieving 95.35% accuracy. The methodology and resources provide a foundation for further astrophysical research, with opportunities to refine the model and dataset.