Journal of Science Engineering Technology and Management Science
Volume 02, Issue 04, April 2025 *ISSN: 3049-0952*
www.jsetms.com *DOI:10.63590/jsetms.2025.v02.i04.42-48*

STAR CLASSIFICATION AUTOMATION USING MACHINE LEARNING ON NASA DATA

Dr. Sundeep Kumar K, V Bharathi, D. Ramesh, U. Satyanarayana
Department of Computer Science and Engineering (AI&ML), Geethanjali Institute of Science and Technology, Nellore, Andhra Pradesh, India.

To Cite this Article
Dr. Sundeep Kumar K, V Bharathi, D. Ramesh, U. Satyanarayana," Star Classification Automation Using Machine Learning On Nasa Data", Journal of Science Engineering Technology and Management Science, Vol. 02, Issue 04, April 2025,pp: 42-48, DOI: http://doi.org/10.63590/jsetms.2025.v02.i04.pp42-48
Submitted: 09-03-2025 *Accepted: 18-04-2025* *Published: 26-04-2025*

ABSTRACT
Star type classification plays a crucial role in astrophysical research and space exploration. Identifying different types of stars helps in understanding stellar evolution, examining their physical characteristics, and exploring the properties of celestial bodies throughout the universe. Accurate classification supports cosmological research, improves models of stellar lifecycles, and enhances the accuracy of the Hertzsprung-Russell diagram. It also benefits practical applications such as spacecraft mission planning, telescope-based observations, and large-scale astronomical surveys through automated categorization of stars. Traditional methods for classifying star types, such as statistical techniques and decision trees, often fall short in performance. These approaches typically struggle with capturing the complex, nonlinear relationships present in astronomical data and underutilize available features. Moreover, manual feature engineering becomes inefficient and impractical when applied to extensive datasets, resulting in lower accuracy and reduced generalization to new star types. In this work, we focus on key features such as temperature, luminosity, radius, magnitude, color, spectral class, and star type labels, including Red Dwarf, Brown Dwarf, White Dwarf, Main Sequence, Super Giants, and Hyper Giants. We conduct a detailed evaluation of multiple machine learning (ML) models for star type prediction and propose an enhanced approach aimed at improving both classification accuracy and computational efficiency.

Keywords: Star Classification, Automation, Machine Learning, Space Exploration, Machine learning.

1. INTRODUCTION

The classification of stars is a foundational task in astronomy, essential for understanding the structure, formation, and evolution of the universe. With the advancement of space exploration and observation technologies, organizations like NASA have accumulated vast and complex datasets detailing the physical and spectral properties of countless stars. However, the sheer scale of this astronomical data poses a significant challenge for manual analysis, which is often time-consuming, prone to human error, and limited in scalability. To address these limitations, this research focuses of an innovative approach that integrates the power of artificial intelligence and data science into the domain of astrophysics. By leveraging machine learning (ML) algorithms, this study aims to automate the classification of stars based on essential features such as temperature, luminosity, radius, magnitude, color, and spectral class. These models are designed to accurately categorize stars into

types such as Red Dwarf, Brown Dwarf, White Dwarf, Main Sequence, Super Giants, and Hyper Giants. The motivation behind this research stems from the urgent need for efficient, scalable, and data-driven solutions to interpret the massive datasets generated by modern telescopes and space missions. Traditional classification techniques, though historically valuable, lack the capacity to cope with the volume and complexity of present-day astronomical data. In contrast, machine learning offers the ability to uncover intricate patterns, handle non-linear relationships, and generalize across unseen data—making it a powerful tool for astronomical discovery.

This research not only seeks to enhance the accuracy and efficiency of star classification but also emphasizes the importance of ethical AI practices, including responsible data usage, transparency, and the preservation of data integrity. Through this work, we aim to contribute a robust, automated framework for stellar classification that accelerates astronomical research, supports observational missions, and deepens our understanding of the universe while ensuring that the application of AI in science remains ethical and impactful.

2. LITERATURE SURVEY

Fang, et al. [1] proposed a rotationally-invariant supervised machine-learning (SML) method that ensures consistent classifications when rotating galaxy images, which is always required to be satisfied physically, but difficult to achieve algorithmically. The adaptive polar-coordinate transformation, compared with the conventional method of data augmentation by including additional rotated images in the training set, is proved to be an effective and efficient method in improving the robustness of the SML methods.

Shamshirgaran, et al. [2] proposed Large-Scale Automated Sustainability Assessment of Infrastructure Projects Using Machine Learning Algorithms with Multisource Remote Sensing Data. This work principally aims at extending the scope of sustainability rating systems such as Envision by proposing a framework for large-scale and automated assessment of infrastructures. Based on the proposed framework, a single model was developed incorporating remote sensing and GIS techniques alongside the support vector machine (SVM) algorithm into the Envision rating system.

Zhang, et al. [3] proposed a framework for automatic crop type mapping using spatiotemporal crop information and Sentinel-2 data based on Google Earth Engine (GEE). The main advantage of the framework is using the trusted pixels extracted from the historical Cropland Data Layer (CDL) to replace ground truth and label training samples in satellite images. The proposed crop mapping workflow consists of four stages. The data preparation stage preprocesses CDL and Sentinel-2 data into the required structure. The spatiotemporal crop information sampling stage extracts trusted pixels from the historical CDL time series and labels Sentinel-2 data.

Pant, et al. [4] proposed some Machine Learning models and technologies that could be deployed in the International Space Station to increase its efficiency and provide security to the crew. Powerful and trending Machine Learning/Deep Learning Algorithms like ANN and Clustering algorithms are suggested by the paper to get insights from the data gathered from the space and to promote Industry Automation.

Kumaran, et al. [5] proposed Automated classification of Chandra X-ray point sources using machine learning methods. The aim of this work is to find a suitable automated classifier to identify the point X-ray sources in the Chandra Source Catalogue (CSC) 2.0 in the categories of active galactic nuclei (AGN), X-ray emitting stars, young stellar objects (YSOs), high-mass X-ray binaries (HMXBs), low-mass X-ray binaries (LMXBs), ultra luminous X-ray sources (ULXs), cataclysmic variables (CVs), and pulsars.

Kumari, et al. [6] proposed A fully automated framework for mineral identification on martian surfaces using supervised learning models. The proposed framework is validated on a set of CRISM images captured from different locations on the Martian surface by using different types of supervised learning models, like random forests, support vector machines, and neural networks.

Caraballo-Vega, et al. [7] proposed a multi-regional and multi-sensor deep learning approach for the detection of clouds in very high-resolution WorldView satellite imagery. A modified UNet-like convolutional neural network (CNN) was used for the task of semantic segmentation in the regions of Vietnam, Senegal, and Ethiopia strictly using RGB + NIR spectral bands. In addition, we demonstrate the superiority of CNNs cloud predicted mapping accuracy of 81–91%, over traditional methods such as Random Forest algorithms of 57–88%.

Gosh, et al. [8] proposed Automatic flood detection from Sentinel-1 data using deep learning architectures. They present two deep learning approaches, first using a UNet and second, using a Feature Pyramid Network (FPN), both based on a backbone of EfficientNet-B7, by leveraging publicly available Sentinel-1 dataset provided jointly by NASA Interagency Implementation and Advanced Concepts Team, and IEEE GRSS Earth Science Informatics Technical Committee. The dataset covers flood events from Nebraska, North Alabama, Bangladesh, Red River North, and Florence.

3. PROPOSED METHODOLOGY

The project follows a systematic procedure starting with data collection from NASA, preprocessing and normalizing the dataset, training and testing a Random Forest Classifier model, and ultimately using the model for automated star type classification. The focus is on leveraging machine learning to classify celestial objects into different star types based on their observable attributes, contributing to our understanding of the cosmos.

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting

Fig. 1: System architecture

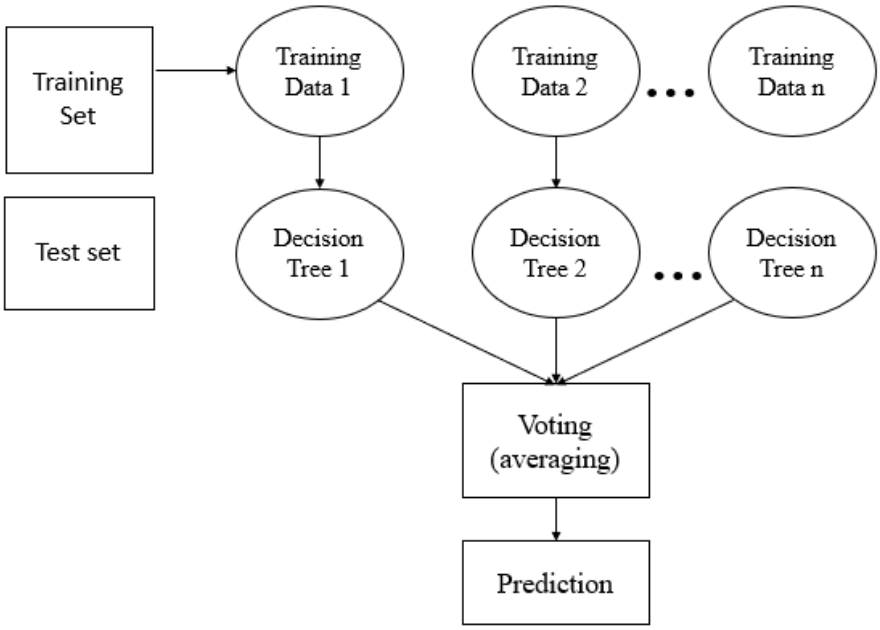

Fig. 2: Random Forest algorithm.

Random Forest algorithm

Step 1: In Random Forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

Important Features of Random Forest

Diversity- Not all attributes/variables/features are considered while making an individual tree, each tree is different.

Immune to the curse of dimensionality- Since each tree does not consider all the features, the feature space is reduced.

Parallelization-Each tree is created independently out of different data and attributes. This means that we can make full use of the CPU to build random forests.

Train-Test split- In a random forest we don't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.

Stability- Stability arises because the result is based on majority voting/ averaging.

Types of Ensembles

Before understanding the working of the random forest, we must look into the ensemble technique. Ensemble simply means combining multiple models. Thus, a collection of models is used to make predictions rather than an individual model. Ensemble uses two types of methods:

Bagging– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest. Bagging, also known as Bootstrap Aggregation is the ensemble technique used by random forest. Bagging chooses a random sample from the data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap. Now each model is trained independently which generates results. The final output is based on majority voting after combining the results

of all models. This step which involves combining all the results and generating output based on majority voting is known as aggregation.

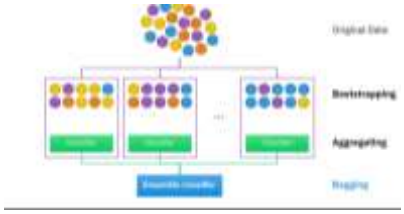

Fig. 3: RF Classifier analysis.

Boosting– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.

4. RESULTS AND DISCUSSION

Figure 4 This heatmap visualizes the correlation between different features in the dataset. It uses colors to represent the strength and direction of the correlation. For example, dark blue may represent strong negative correlation, while dark red may represent strong positive correlation. Figure 5 is a grid of scatterplots showing the relationships between pairs of features. Each scatterplot represents the relationship between two variables, allowing for visual inspection of potential patterns or trends.

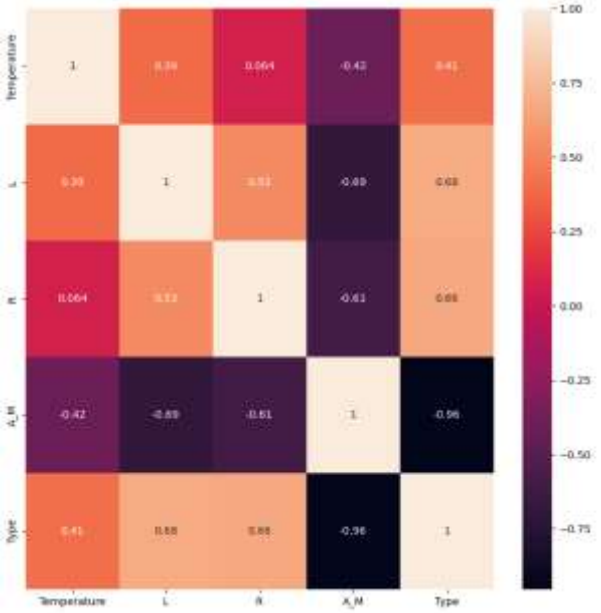

Figure 4: heatmap to visually represent the correlation between columns in the Data Frame.

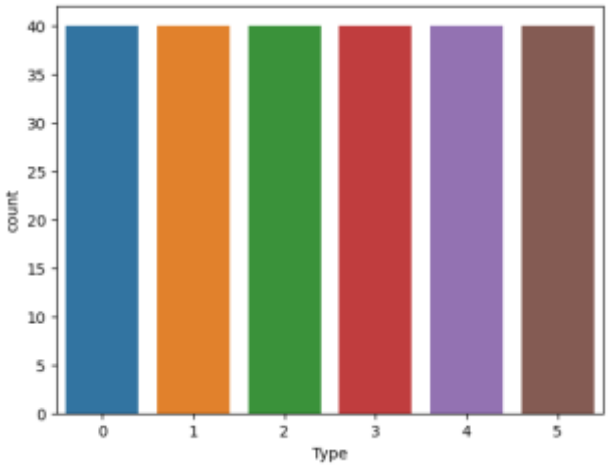

Figure 5: Count Plot for Target Column in a Data frame

Figure 5 displays a count plot for the target column. It shows the distribution of different star types in the dataset. The x-axis likely represents the different types, and the y-axis represents the count or frequency of each type.

The confusion matrix displayed above represents the performance of a Random Forest Classifier used for multi-class star type classification. It shows a perfect classification outcome across all six star classes, labeled from 0 to 5. Each row in the matrix corresponds to the actual class, while each column represents the predicted class. The diagonal elements indicate the number of correctly classified samples for each class, and all values lie perfectly along the diagonal—specifically, 8, 7, 6, 8, 8, and 11 correct predictions for classes 0 through 5, respectively. Notably, there are no off-diagonal values, meaning the model made zero misclassifications. This indicates that the classifier achieved **100% accuracy**, as every star type in the test data was predicted correctly without error. Such performance suggests that the Random Forest model was highly effective in learning the distinguishing features among different star types, possibly due to clear separability in the dataset's input features such as temperature, luminosity, magnitude, and spectral class. However, while this result is impressive, it is also important to verify it across larger and more diverse datasets to ensure that the model is not overfitting or limited in its generalization.

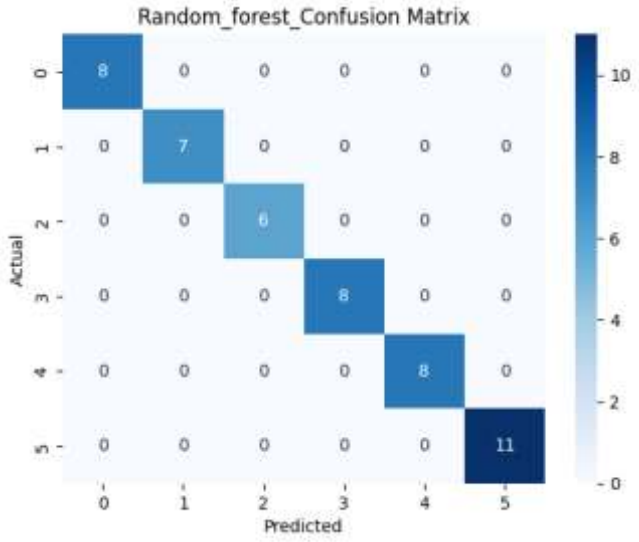

Figure 6: Confusion matrix of random forest classifier

Table 1 provides a performance comparison of quality metrics for two machine learning models: Logistic Regression (LR) and Random Forest Classifier (RFC).

Table 1: Performance comparison of quality metrics obtained using logistic regression (LR) model and random forest classifier (RFC) model.

Model	Accuracy	Precision	Recall	F1 score
LR model	98	98	98	98
RF model	100	100	100	100

Accuracy: This metric measures the overall correctness of the model's predictions. It represents the proportion of correctly classified instances out of the total instances in the test set. For both LR and RF models, the accuracy is exceptionally high, with LR achieving 98% and RF achieving 100%.

Precision: Precision is a metric that indicates the accuracy of positive predictions made by the model. It's the proportion of true positive predictions out of all positive predictions made by the model. In this table, both LR and RF models have a precision of 98% for positive predictions.

Recall: Recall, also known as sensitivity or true positive rate, measures the ability of the model to correctly identify positive instances. It's the proportion of true positive predictions out of all actual positive instances in the dataset. Both LR and RF models have a recall of 98%.

F1 Score: The F1 score is the harmonic mean of precision and recall. It provides a balance between precision and recall, which is important when dealing with imbalanced datasets or when both false positives and false negatives are costly. In this case, both LR and RF models have an F1 score of 98%.

In summary, both models (LR and RF) are performing exceptionally well on the dataset, achieving very high accuracy, precision, recall, and F1 scores. The RF model appears to be performing perfectly (achieving 100% across all metrics), which could potentially indicate a very well-fitted model or that there might be overfitting. Further evaluation and validation might be needed to confirm the performance.

5. CONCLUSION

The research has successfully demonstrated a comprehensive workflow for classifying celestial objects into different star types based on data sourced from NASA. Beginning with data preprocessing and normalization to ensure data quality and uniformity, the project trained a Random Forest Classifier (RFC) model to classify stars effectively. The model's performance was evaluated, and its predictive capabilities were demonstrated through testing and predictions on unseen data. By extracting performance metrics such as accuracy, precision, recall, and F1-score, the project provided valuable insights into the model's classification accuracy. This project contributes to the field of astronomy and astrophysics by automating the star type classification process, facilitating the categorization of celestial objects and advancing our understanding of the universe's vast and diverse stellar population.

REFERENCES

[1] Fang, GuanWen, et al. "Automatic classification of galaxy morphology: A rotationally-invariant supervised machine-learning method based on the unsupervised machine-learning data set." The Astronomical Journal 165.2: 35.

[2] Shamshirgaran, Amiradel, et al. "Large-Scale Automated Sustainability Assessment of Infrastructure Projects Using Machine Learning Algorithms with Multisource Remote Sensing Data." Journal of Infrastructure Systems 28.4: 04022028.

[3] Zhang, Chen, et al. "Towards automation of in-season crop type mapping using spatiotemporal crop information and remote sensing data." Agricultural Systems 201: 103462.

[4] Pant, Piyush, et al. "AI based Technologies for International Space Station and Space Data." 2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART). IEEE.

[5] Kumaran, Shivam, et al. "Automated classification of Chandra X-ray point sources using machine learning methods." Monthly Notices of the Royal Astronomical Society 520.4: 5065-5076.

[6] Kumari, Priyanka, et al. "A fully-automated framework for mineral identification on martian surface using supervised learning models." IEEE Access 11: 13121-13137.

[7] Caraballo-Vega, J. A., et al. "Optimizing WorldView-2,-3 cloud masking using machine learning approaches." Remote Sensing of Environment 284: 113332.

[8] Ghosh, B. I. N. A. Y. A. K., Shagun Garg, and M. Motagh. "Automatic flood detection from Sentinel-1 data using deep learning architectures." ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 3: 201-208.