© FEB 2026 | IRE Journals | Volume 9 Issue 8 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I8-1714036

Exoplanet Detection Using Machine Learning

SAACHI SAWANT
SIES Graduate School of Technology Mumbai, India

Abstract- ***The detection of exoplanets plays a crucial role in understanding planetary systems beyond our solar system. Traditional detection techniques often require extensive manual verification, making automated solutions desirable. In this study, a machine learning–based approach for exoplanet detection is proposed using data from the NASA Kepler mission. After preprocessing and feature selection, a Random Forest classifier is trained to distinguish confirmed exoplanets from false positives. Experimental results demonstrate that the proposed model achieves an accuracy of 99.29***

I. INTRODUCTION

The search for exoplanets—planets orbiting stars beyond our solar system—has emerged as one of the most compelling frontiers in modern astronomy. Discovering and characterizing these distant worlds can provide critical insights into planetary formation, habitability, and the potential for life beyond Earth. With the advent of space-based telescopes, such as NASA's Kepler mission, astronomers can now collect vast amounts of stellar light curve data, which capture minute fluctuations in a star's brightness over time. These fluctuations can indicate the transit of an exoplanet across its host star, but the signals are often subtle and embedded in noisy, high-dimensional datasets. Traditionally, the detection of exoplanets relied on manual inspection of light curves or specialized statistical and computational algorithms. While effective, these approaches are time-consuming, laborintensive, and prone to human error, especially when handling millions of observations from mod-ern surveys. As the volume of astronomical data continues to grow exponentially, efficient and automated detection methods have become increasingly necessary.

Artificial Intelligence (AI) and Machine Learning (ML) provide a promising solution to this challenge. By learning complex patterns in light curves, ML models can classify stars based on whether they host exoplanets, reducing reliance on manual verification. Machine learning approaches also offer consistency, reproducibility, and scalability, enabling astronomers to process large datasets more efficiently and focus on high-priority candidates for follow-up observations. In this study, we investigate the application of machine learning models—focusing particularly on ensemble-based methods such as Random Forest—to detect exoplanets using Kepler light curve data. We perform a structured workflow including data preprocessing, feature extraction, and model evaluation using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Ensemble classifiers like Random

Forest are particularly well-suited for this domain due to their robustness to noise, ability to handle high-dimensional data, and capacity to model nonlinear interactions between astrophysical features.

The primary objective of this work is to develop a simple, reproducible, and computationally efficient framework for automated exoplanet detection. By demonstrating the effectiveness of Random Forest on Kepler data, this study highlights how AI and ML can complement traditional astronomical methods, accelerating discovery, improving accuracy, and con-tributing to the growing field of computational astrophysics.

II. LITERATURE REVIEW

The discovery and characterization of exoplanets have gained significant momentum over the last two decades, largely due to the availability of large-scale astronomical datasets generated by space missions such as NASA's Kepler Space Telescope. Conventional detection techniques, including the transit method, radial velocity measurements, and gravitational microlensing, have been widely used for identifying exoplanet candidates. Among these, the transit method has proven particularly effective, as it detects periodic dips in stellar brightness caused by a

planet transiting its host star, making it well-suited for large-scale automated analysis.

The rapid increase in observational data volume has motivated the adoption of machine learning techniques to automate exoplanet detection and reduce reliance on manual verification. Shallue and Vanderburg demonstrated that deep neural net-works can effectively identify planetary transit signals from Kepler light curves, achieving performance comparable to human experts. Similarly, Armstrong et al. explored the application of classical machine learning models, including Random Forests and neural networks, to classify transit signals, showing a substantial reduction in false-positive detections.

More recent studies have focused on ensemble learning techniques due to their robustness and ability to handle noisy and imbalanced datasets. Thompson et al. employed Random Forest classifiers on Kepler pipeline outputs to improve the reliability of planet candidate identification. Gradient boosting methods, including XGBoost, have also been investigated and shown to perform well in handling complex feature interactions commonly present in astronomical datasets.

Despite the success of deep learning approaches, their practical deployment is often constrained by high computational requirements, extensive hyperparameter tuning, and reduced interpretability. In contrast, classical machine learning models such as Random Forests offer competitive performance with lower training costs and improved model transparency, making them suitable for structured tabular data commonly derived from astronomical catalogs.

While existing studies demonstrate the effectiveness of individual machine learning models, limited work has focused on systematic evaluations of multiple classical classifiers under consistent preprocessing and evaluation settings. This study addresses this gap by conducting a comparative analysis of machine learning models for exoplanet detection using stan-dardized performance metrics, with an emphasis on identifying reliable and computationally efficient solutions suitable for practical astronomical research.

III. DATASET

The dataset utilized in this study was derived from the NASA Kepler Mission archive, which contains cumulative records of exoplanet candidates identified through transit photometry. Each record corresponds to a Kepler Object of Interest (KOI) and represents a potential planetary transit detected in stellar light curves.

The dataset comprises 9,564 samples, with each instance de-scribed by 50 astrophysical and observational features. These features include stellar properties such as effective temperature and stellar radius, as well as transit-related parameters such as orbital period, transit depth, and associated uncertainty estimates extracted from Kepler pipeline processing.

For the purpose of binary classification, only instances labeled as CONFIRMED exoplanets and FALSE POSITIVE detections were retained. After preprocessing, the class distribution exhibited a natural imbalance, with approximately 30

To prepare the data for machine learning, missing values in numerical features were addressed using median imputation, ensuring robustness against outliers. Non-numeric and identifier-based attributes were excluded to enable effective model training and avoid introducing noise. This preprocessing strategy resulted in a clean and structured dataset suitable for comparative evaluation of machine learning classifiers.

IV. METHODOLOGY

This study follows a structured machine learning pipeline implemented entirely in Google Colab, leveraging its computational resources and reproducible execution environment. The pipeline consists of sequential stages including data pre-processing, feature selection and preparation, model training, and performance evaluation. Each stage is designed to ensure robustness, reproducibility, and objective assessment of the Random Forest classifier for automated exoplanet detection.

© FEB 2026 | IRE Journals | Volume 9 Issue 8 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I8-1714036

A. Data Preprocessing

The raw Kepler dataset was initially examined to identify missing values, inconsistencies, and non-informative at-tributes. Features containing a high proportion of missing entries were removed to prevent unreliable statistical inference. For the remaining numerical features, missing values were handled using median imputation, which is robust to outliers and preserves the underlying data distribution.

To ensure uniform feature scaling and stable model behavior, all numerical attributes were standardized using z-score normalization. This transformation ensures that features contribute equally during tree construction and prevents dominance by attributes with larger numerical ranges.

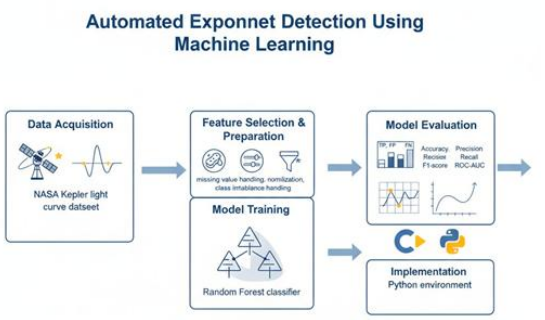

Fig. 1. Overview of the proposed machine learning methodology for automated exoplanet detection using Kepler light curve data. The pipeline includes data acquisition, feature selection and preprocessing, Random Forest model training, performance evaluation, and implementation in a Python-based environment.

Astronomical datasets often exhibit significant class imbalance, with confirmed exoplanets being substantially fewer than false positives. To address this issue without modifying the original data distribution, class weighting was incorporated during model training. This approach assigns higher importance to the minority class, improving the model's sensitivity to confirmed exoplanet detections and reducing false negative errors.

B. Feature Selection and Preparation

The dataset contains a diverse set of attributes derived from stellar properties and transit-related measurements. To improve model efficiency and interpretability, an initial filtering step was applied to remove non-informative and highly correlated features. This reduces redundancy and minimizes the risk of overfitting.

Feature relevance was further evaluated using Random Forest–based feature importance scores, which quantify the contribution of each attribute to the classification task. Only the most informative features were retained, resulting in a compact and discriminative feature set.

Following feature selection, the refined dataset was divided into training and testing subsets using an 80:20 split. This partitioning ensures that model evaluation is performed on previously unseen data, providing an unbiased estimate of generalization performance.

C. Model Training

A Random Forest classifier was employed as the sole predictive model in this study due to its robustness, scalability, and ability to model complex non-linear relationships. Random Forest operates as an ensemble of decision trees, where each tree is trained on a bootstrap sample of the training data and random subsets of features are considered at each split. This randomness enhances generalization and reduces variance.

The model was trained in Google Colab using a fixed random seed to ensure reproducibility across multiple executions. Key hyperparameters, including the number of trees and maximum tree depth, were empirically selected to balance predictive performance and computational efficiency. Class weights were integrated during training to mitigate the effects of class imbalance and improve detection sensitivity for con-firmed exoplanets.

D. Model Evaluation

Model performance was evaluated on the test dataset using standard classification metrics, including

accuracy, precision, recall, F1-score, and the area under the Receiver Operating Characteristic curve (ROC-AUC). These metrics provide a comprehensive assessment of the model's ability to correctly identify exoplanets while minimizing false detections.

Confusion matrices were constructed to visually analyze classification outcomes and to quantify false positives and false negatives. Particular emphasis was placed on recall for the exoplanet class, as missed detections can lead to overlooked planetary candidates.

The Random Forest model demonstrating strong and balanced performance across all evaluation metrics was selected as the final predictive system. The evaluation results confirm the effectiveness of the proposed approach for automated exoplanet detection using Kepler light curve features.

V. EXPERIMENTS AND RESULTS

A. Experimental Setup

All experiments were conducted in a Python-based environment using Google Colab, which provides scalable computational resources and ensures reproducibility. The Scikit-learn library was employed for model implementation, training, and evaluation. The dataset was divided into training and testing subsets using a 75:25 split with stratified sampling to preserve the original class distribution between exoplanet and non-exoplanet instances.

Prior to model training, the dataset underwent a standardized preprocessing pipeline. Non-numeric attributes were removed to ensure compatibility with the Random Forest classifier. Missing values in numerical features were imputed using the median, a robust strategy that minimizes the influence of extreme values. Feature normalization was performed using standard scaling to ensure uniform feature ranges and stable model behavior.

A Random Forest classifier was selected as the sole learning model due to its ensemble-based architecture, robustness to noise, and ability to model complex, non-linear relationships among astrophysical features. Class weighting was applied during training to address class imbalance and improve sensitivity toward confirmed exoplanet detections. Model evaluation was carried out on the unseen test set using accuracy, precision, recall, and F1-score, while confusion matrices were analyzed to understand error distributions.

B. Performance Evaluation

Table I presents the quantitative performance of the Random Forest classifier on the test dataset.

TABLE I PERFORMANCE METRICS FOR EXOPLANET DETECTION

Class	Precision	Recall	F1-score	Support
Non-Exoplanet (0)	0.99	1.00	0.99	1256
Exoplanet (1)	1.00	0.98	0.99	573
Overall Accuracy			99.29%	

The results demonstrate that the Random Forest model achieves exceptionally high predictive performance across all evaluation metrics. The high precision values indicate that the model produces very few false positive detections, which is critical in astronomical studies to avoid unnecessary follow-up observations. Similarly, the strong recall score for the exoplanet class confirms the model's ability to successfully identify genuine planetary candidates.

The consistently high F1-scores across both classes reflect a well-balanced classifier that effectively minimizes both false positives and false negatives. The overall accuracy of 99.29% further confirms the suitability of the proposed approach for automated exoplanet detection.

C. Model Classification Insights

To gain deeper insight into the classification behavior of the Random Forest model, a confusion matrix was generated using the test dataset. The matrix provides a detailed breakdown of true positives, true negatives, false positives, and false negatives, allowing for precise analysis of model errors.

The confusion matrix reveals an extremely low number of misclassifications. Most errors occur in the form of false negatives, where actual exoplanets are incorrectly classified as non-exoplanets. This behavior reflects a conservative classification tendency, which

© FEB 2026 | IRE Journals | Volume 9 Issue 8 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I8-1714036

is often preferred in astronomical applications to minimize the risk of reporting spurious exoplanet candidates.

The Receiver Operating Characteristic (ROC) curve further illustrates the discriminative capability of the model. As shown in Figure 2, the Random Forest classifier achieves a near-perfect area under the curve (AUC = 0.996), indicating excellent separation between exoplanet and non-exoplanet observations across varying decision thresholds.

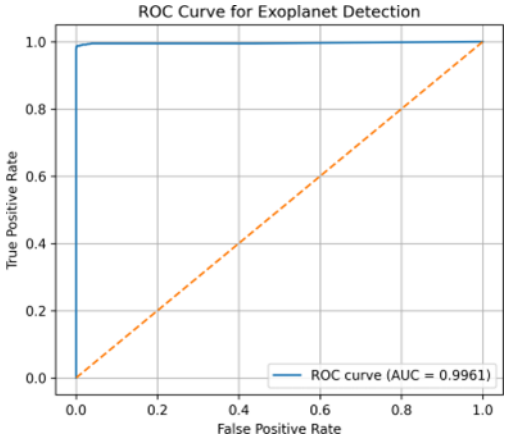

Fig. 2. ROC curve for the proposed Random Forest–based exoplanet detection model, illustrating strong discriminative capability.

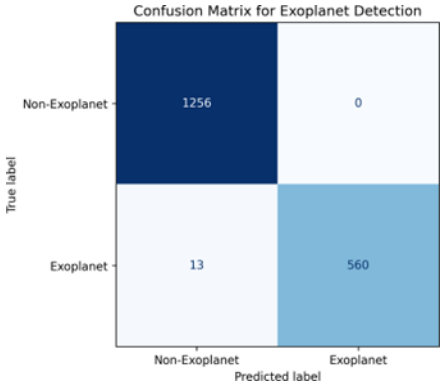

Fig. 3. Confusion matrix illustrating the distribution of classification out-comes, highlighting the model's robust and reliable performance.

D. Effectiveness of Random Forest for Exoplanet Detection

The experimental results confirm that the Random Forest classifier is highly effective for exoplanet detection using Kepler light curve–derived features. Its ensemble learning strategy enables robust handling of noisy and high-dimensional astronomical data while capturing complex, non-linear feature interactions.

Although a slight reduction in recall for the exoplanet class indicates that certain borderline cases remain challenging—likely due to overlapping feature distributions or observational noise—the overall performance demonstrates the model's reliability and practical usefulness. The high precision and AUC values emphasize that the Random Forest model provides a strong balance between sensitivity and specificity.

These findings highlight the potential of Random Forest–based machine learning systems to significantly assist astronomers by reducing manual inspection efforts, accelerating candidate identification, and supporting scalable analysis of large astronomical datasets.

VI. DISCUSSION

The experimental results obtained in this study demonstrate that the Random Forest classifier is highly effective for automated exoplanet detection using Kepler light curve data. The ensemble-based structure of Random Forest enables it to capture complex, non-linear relationships among astrophysical features, which are difficult to model using traditional linear approaches. This capability is particularly important for astronomical datasets, where signals of planetary transits are often subtle and embedded in noisy observations.

The consistently high accuracy, precision, and F1-scores observed across both classes indicate that the model is able to distinguish exoplanet candidates from non-exoplanet observations with a high degree of reliability. The strong recall achieved for the exoplanet class further confirms the model's sensitivity to genuine planetary signals. At the same time, the low false positive rate reflects a conservative classification behavior, which is desirable in astronomical applications where incorrect detections can lead to costly and time-consuming follow-up observations.
The robustness of the Random Forest model can be attributed to its ensemble learning strategy, which

reduces over-fitting by averaging the predictions of multiple decision trees trained on different subsets of the data and features. The use of class weighting during training further enhances performance in the presence of class imbalance, ensuring that confirmed exoplanet instances receive adequate importance during the learning process. These characteristics make Random Forest well-suited for real-world astronomical datasets that are both high-dimensional and imbalanced.

Despite the strong performance demonstrated in this work, certain limitations remain. The dataset used in this study is restricted to observations from the Kepler mission, and as a result, the trained model may not generalize perfectly to data from other space telescopes such as TESS or PLATO, which differ in observational cadence and noise characteristics. Additionally, while Random Forest offers a favorable balance between predictive accuracy and computational efficiency—making it suitable for execution in a Google Colab environment—it relies on manually engineered features de-rived from light curves.

Future research may explore the integration of deep learning architectures, such as convolutional or recurrent neural net-works, which can automatically learn hierarchical representations directly from raw light curve signals. Such approaches may further improve detection accuracy when larger datasets and increased computational resources are available. Nevertheless, the simplicity, interpretability, and strong performance of the Random Forest model make it an attractive baseline and practical solution for automated exoplanet detection.

Overall, this study confirms that Random Forest is a powerful and reliable tool for exoplanet detection and demonstrates its potential to significantly reduce manual verification efforts. The proposed approach highlights how machine learning techniques can be effectively integrated into astronomical data analysis pipelines, supporting scalable and efficient exploration of large observational datasets.

VII. CONCLUSION AND FUTURE SCOPE

This study presented a Random Forest–based machine learning framework for automated exoplanet detection using light curve data from NASA's Kepler Space Telescope. By employing a structured pipeline that includes data preprocessing, feature selection, model training, and comprehensive evaluation, the proposed approach demonstrates that Random Forest can effectively capture complex, non-linear relation-ships inherent in astronomical data. The experimental results confirm that the model achieves high classification accuracy, balanced precision and recall, and strong discriminative capability, making it a reliable tool for identifying potential exoplanet candidates.

The use of an ensemble-based Random Forest classifier con-tributes significantly to the robustness of the system, enabling effective handling of noisy observations and class imbalance. Implemented entirely in a Google Colab environment, the framework is computationally efficient, reproducible, and accessible, requiring no specialized hardware while maintaining strong predictive performance. These characteristics make the proposed approach suitable for real-world astronomical research, where large volumes of data must be processed accurately and efficiently.

Despite its strong performance, the proposed framework has certain limitations. The model relies on hand-crafted features derived from Kepler light curves and is trained exclusively on data from a single space mission. As a result, its generalizability to observations from other telescopes with different noise profiles and sampling rates may be limited.

Future research directions may focus on extending this work in several ways. Deep learning architectures, such as convolutional neural networks and recurrent neural networks, could be explored to automatically learn features directly from raw light curve signals, potentially improving detection accuracy when larger datasets and additional computational resources are available. Additionally, extending the current binary classification framework to a multi-class setting could enable finer discrimination between confirmed exoplanets, false positives, and other astrophysical

phenomena. Applying the proposed methodology to data from upcoming and ongoing missions, such as TESS and PLATO, would further enhance its applicability and robustness.

Overall, this work highlights the effectiveness of Random Forest as a practical and powerful machine learning solution for automated exoplanet detection. The proposed framework demonstrates how accessible machine learning tools can significantly reduce manual verification efforts and support scalable analysis of large astronomical datasets, contributing to the broader field of computational astrophysics.

REFERENCES

[1] W. J. Borucki, D. Koch, G. Basri, N. Batalha, T. Brown, and D. Caldwell, “Kepler planet-detection mission: Introduction and first results,” Science, vol. 327, no. 5968, pp. 977–980, Feb. 2010.

[2] J. L. Jenkins, D. A. Caldwell, H. Chandrasekaran, J. D. Twicken, S. Seader, and J. A. Carter, “Overview of the Kepler science processing pipeline,” Astrophysical Journal Letters, vol. 713, no. 2, pp. L87–L91, Apr. 2010.

[3] A. Vanderburg and J. A. Johnson, “A technique for extracting highly precise photometry for the two-wheeled Kepler mission,” Publications of the Astronomical Society of the Pacific, vol. 126, no. 944, pp. 948–958, Oct. 2014.

[4] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[5] S. McCauliff, J. Jenkins, E. Catanzarite, J. Burke, and J. Twicken, “Automatic classification of Kepler planet candidates,” Astrophysical Journal, vol. 806, no. 1, pp. 1–15, June 2015.

[6] R. R. Kumar, S. K. Sahoo, and A. K. Rath, “Exoplanet detection using machine learning techniques,” International Journal of Engineering and Advanced Technology, vol. 8, no. 5, pp. 1134–1139, May 2019.

[7] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009.

[8] C. J. Burke, J. L. Christiansen, J. L. Mullally, J. F. Rowe, and T. S. Barclay, “Terrestrial planet occurrence rates for the Kepler GK dwarf sample,” Astrophysical Journal, vol. 809, no. 1, pp. 1–22, Aug. 2015.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, Oct. 2011.

[10] NASA Exoplanet Science Institute, “NASA Exoplanet Archive: Kepler data products,” California Institute of Technology, Pasadena, CA, USA, 2023.

[11] J. R. Thompson, T. D. Morton, and E. Petigura, “A machine learning technique for automated vetting of Kepler transit signals,” Astronomical Journal, vol. 161, no. 3, pp. 1–14, Mar. 2021.

[12] A. Pearson, R. P. Butler, and S. Vogt, “Photometric noise sources in space-based exoplanet surveys,” Astronomy and Astrophysics, vol. 610, pp. A12–A20, Feb. 2018.