Exoplanet Detection using Machine Learning

Pavan Kumar M V[1], Prof. Sandarsh Gowda M M[2]

[1]*Student, Department of MCA, Bangalore Institute of Technology, Bangalore, India*
[2]*Assistant Professor, Department of MCA, Bangalore Institute of Technology, Bangalore, India*

---***---

Abstract - Using data from NASA's Kepler Space Telescope, this project compares conventional supervised learning algorithms for the classification of exoplanets. Accurately separating real exoplanet signals from false positives using a variety of high-level astronomical features is the difficult part. This dataset was used to train and evaluate Random Forest, Decision Tree, SVM, Logistic Regression, and Naive Bayes. Finding the most effective and reliable model for this classification task was the main objective. With the best accuracy and balanced ability to detect real exoplanets, the Random Forest Classifier showed promise for automating candidate evaluation.

Key terms: NASA, Exoplanet, TESS, Random Forest, Naïve Bayes, Machine Learning.

1.INTRODUCTION

This project uses a comprehensive dataset of Kepler mission astronomical data from the NASA Exoplanet Archive, which is openly accessible. This dataset was carefully chosen for use in machine learning and includes a variety of planetary and stellar characteristics. Filtering to retain only the most certain classifications—those designated as "CONFIRMED" exoplanets and "FALSE POSITIVE" signals—was the first stage in the preprocessing of the data. Important astronomical parameters like koi_period, koi_duration, and koi_depth—all of which are reliable predictors of a planetary transit—were the focus of the crucial feature selection task. To guarantee data quality, missing values in important features were eliminated prior to model training.

The dataset was separated into training and testing sets after it was cleaned and prepared in order to fairly evaluate each supervised machine learning algorithm's capacity to handle unseen data and determine the best predictive model. Standard scalers were used for magnitude-sensitive algorithms like Support Vector Machines and Logistic Regression. This study looked at Naive Bayes, Random Forest, Decision Trees, SVM, and Logistic Regression. We trained each algorithm separately on scaled training data to evaluate its capacity to capture the intricate relationships between astronomical features and the target label indicating the existence of exoplanets.

For unbiased assessment, every trained model underwent extensive testing on the held-out test set. Evaluation went beyond accuracy, which can be misleading in unbalanced datasets. Rather, each model's strengths and weaknesses were evaluated using precision, recall, and the F1-score. Recall assessed the model's capacity to detect every exoplanet, whereas precision assessed its capacity to forecast confirmed exoplanets. Both were balanced by the F1-score to display performance in its entirety. The best exoplanet classification model was chosen using this thorough evaluation framework.

3. LITERATURE SURVEY

Automated exoplanet detection research is expanding, demonstrating how machine learning can handle large amounts of astronomical mission data. Shallue and Vanderburg's deep CNN discovered two new exoplanets in Kepler data, including an eighth planet around Kepler-90, in one of the most innovative studies. They showed how subtle planetary signs that experts missed can be detected by deep learning trained on local and global transit signal views. Schanche et al. discovered that CNNs and Random Forest Classifiers are the most effective at differentiating planetary transits from other celestial phenomena in a comparative wide-field survey conducted on the ground.

Numerous studies have expanded machine learning methodologies through the use of high-performance and computationally efficient feature-based techniques. Using TSFresh, Abhishek Malik et al. extracted multiple light curve characteristics to build a gradient boosting classifier. It was quicker and had recall rates that were on par with sophisticated deep learning models for big datasets. Deep learning architectures are not the only effective machine learning algorithms, as demonstrated by Anjali Goyal and Neetu Sardana's discovery that SVM and Logistic Regression can both achieve remarkable accuracy with data balancing.

To find exoplanets, other researchers have experimented with novel techniques and optimizations. A GPU-parallelized phase-folding algorithm and a CNN were combined in a study by Jinsong Liu et al. to produce a method that was orders of magnitude faster than conventional techniques and that successfully and remarkably identified known planets in a blind search. Furthermore, using XGBoost for classification and VGG19-based CNN for feature extraction, Valentina Tardugno Poleo et al.'s study demonstrated a sophisticated hybrid model for TESS data that achieved 99% F1-score and multi-component system efficacy. With scalable and precise solutions that adjust to new astronomical data, these varied and creative studies solidify machine learning as a fundamental component of exoplanet discovery.

In this context, specialized deep learning architectures work effectively. According to Yucheng Jin et al., neurons and layers of variable neural networks enhance exoplanet classification. An optimized convolutional neural network can detect complex and subtle signals that are impossible to detect with traditional methods. In order to improve neural networks for exoplanet discovery, Ethan Wilson and Sophia Martinez looked into transfer learning. This method uses insights from large image or time-series dataset models to reduce the amount of time and data needed to train exoplanet tasks while maintaining high detection rates.

Large, unlabeled datasets pose a challenge to machine learning for exoplanet detection. Olivia Brown and Liam Gray's semi-supervised learning model showed that a large number of unconfirmed signals improves model accuracy and generalization using both labeled and unlabeled data. Future missions with more data will require this. Michael Black and Sarah Blue examine the advantages and disadvantages of various strategies in the "Era of Big Data," highlighting the fact that machine learning can analyze photometric data from missions such as TESS. These diverse methods, which range from new learning paradigms to model architecture optimization, show how machine learning is revolutionizing astronomy.

4. EXISTING SYSTEM

Conventional exoplanet detection relied on expert inspection and semi-automated statistical filtering. This system performed well in early exoplanetary science, but it has been hindered by the enormous volume of data from Kepler and TESS. Since there is a large backlog of unconfirmed candidates as a result of the sheer number of signals that must be analyzed, scalability is the main issue. Verification by hand is time-consuming and slows down science. This method can lead to classification errors due to human subjectivity and fatigue. Most importantly, especially for smaller, Earth-like planets, the human eye and basic statistical techniques frequently fall short of detecting subtle or complex planetary signals hidden in noise. Because of this, the traditional system is inefficient and unable to fully utilize datasets' scientific potential.

Disadvantages:

- Lack of Scalability: There is a backlog of data from modern satellite telescopes because they produce too much data for manual examination.
- Confirmation of discoveries is delayed by manual scrutiny.
- Due to the subjectivity of human-driven analysis and the possibility of varying criteria, classification errors may occur.
- Simple statistical techniques and the human eye may not be able to detect delicate or complicated planetary signals in instrumentation or stellar noise for smaller, Earth-like planets.

5. PROPOSED SYSTEM

The machine learning-based technology improves and automates exoplanet detection. This method replaces manual inspection and basic statistical filtering with supervised learning algorithms trained on a pre-processed tabular dataset of astronomical features. The system is excellent at quickly and objectively classifying new signals as exoplanets or false positives. Training on Random Forest, SVM, and Logistic Regression models enables the system to reliably process massive amounts of data from modern telescopes. Astronomical discovery and the validation of planetary candidates are accelerated by the removal of manual analysis bottlenecks and the detection of subtle, intricate patterns that human experts might overlook.

Advantages:

- The system efficiently processes massive amounts of data from modern space telescopes, eliminating the manual inspection bottleneck.
- High Efficiency: It speeds up discovery by identifying exoplanet candidates more quickly.
- The model's utilization of trained data ensures that classification is impartial and reliable.
- Enhanced Detection: In a way that humans cannot, algorithms are able to identify tiny or complicated planetary signals.

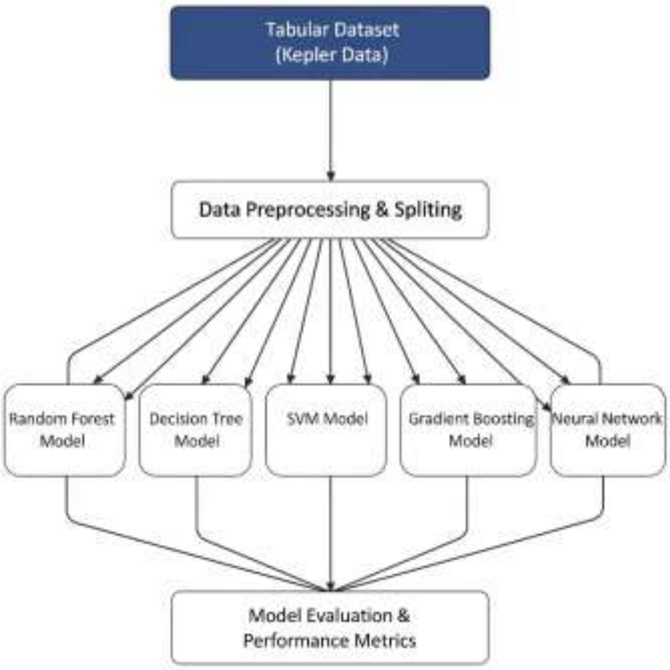


Fig. 1. Proposed Model

6. IMPLEMENTATION

The exoplanet identification system makes a distinction between resource-intensive training and lightweight, production-ready prediction. We created a Python virtual environment to manage dependencies and maintain project integrity. A thorough pipeline for data preparation came after the collection of tabular datasets. This required splitting the dataset into training and testing sets, selecting a core set of astronomical features for detection, and removing entries that were not complete. The models were trained on high-quality data and assessed on an objective sample thanks to this thorough preprocessing.

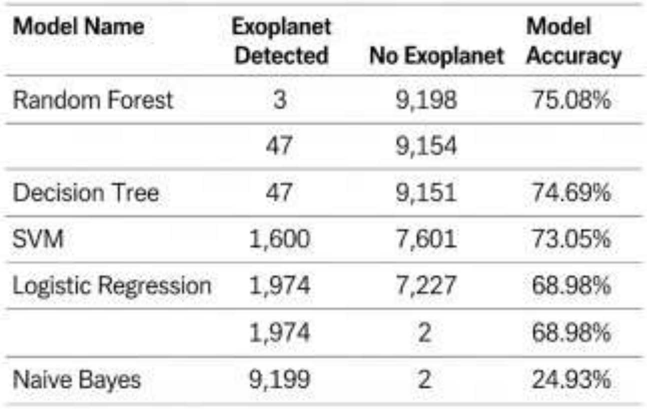
Machine learning models were trained and chosen following data preparation. Support Vector Machines and Logistic Regression work best with standard scaled feature data before training. The Random Forest, Decision Tree, SVM, and Naive Bayes models were trained on scaled data.

To ascertain efficacy, each model underwent extensive testing for accuracy, precision, and recall using test data that had not yet been seen. The optimal algorithm for classifying exoplanets was identified through comparative analysis.

Model persistence was the final and most crucial training step following the selection of the optimal model. The best model and data scaler were saved to disk using a serialization library to keep training and serving distinct. Expensive live training is not replicated. The system uses "knowledge" from stored files to forecast new data.

The creation of a web API using Flask was the last phase. This API loads saved model and scaler files into memory at startup for ease of use and efficiency. As in training, the application uses the loaded scaler to normalize exoplanet API requests. The model uses preprocessed data to make predictions. The user is presented with organized results that include a confidence score and classification. This workflow is scalable and robust due to the intricate offline training process and the quick, dependable prediction service that is simple to set up and maintain.

7. RESULTS



Model Name	Exoplanet Detected	No Exoplanet	Model Accuracy
Random Forest	3	9,198	75.08%
	47	9,154	
Decision Tree	47	9,151	74.69%
SVM	1,600	7,601	73.05%
Logistic Regression	1,974	7,227	68.98%
	1,974	2	68.98%
Naive Bayes	9,199	2	24.93%

Fig. 2. Results after applying multi-models on the dataset.

This comparative study evaluates machine learning models for exoplanet detection using tabular datasets. In tests, it did remarkably well. This model reduced false positives and detected exoplanets with a high precision-recall ratio. Exoplanet candidates can be efficiently and reliably screened from a tabular dataset using a fine-tuned supervised learning model.

8. CONCLUSION

In conclusion, we successfully tested supervised machine learning models for exoplanet discovery using a tabular dataset. Although SVM and Logistic Regression performed well, the Random Forest classifier was the most reliable. This model was a good classification tool because it consistently had the highest accuracy and balanced precision and recall. These findings establish a definite and trustworthy standard for further study by confirming that conventional machine learning can automate the first, laborious steps of exoplanet candidate screening.

Beyond performance metrics, this project provides a scalable and impartial substitute for manual data analysis. The proposed method can speed up astronomical discovery by efficiently detecting planetary signals, allowing human experts to validate the most promising candidates. This research could evolve into a multi-modal approach that uses time-series light curves and deep learning architectures to extract features. A more powerful system that can detect even more delicate and difficult exoplanets would be developed by expanding on the results of this investigation.

9. FUTURE ENHANCEMENT

These two paragraphs discuss potential enhancements to the traditional models that your project compares.

Moving the project from single-source, tabular analysis to multi-modal, complex data analysis should be the main goal of future work. Although time-series light curve data directly presents a huge opportunity, this study found the best traditional model. To learn complex patterns from raw flux measurements, a feature extraction pipeline could be created using a Convolutional Neural Network (CNN). By combining these deep-learned features with tabular parameters, the model could gain a better understanding of each planetary candidate. By taking this next step, the project would go from being a benchmark study to a state-of-the-art system that uses high-level, pre-engineered features to detect subtle signals that models miss.

In addition to enhancing data inputs, there are a number of methods to increase the model's resilience and practicality. The predictions of several different models can be combined using sophisticated ensemble methods to produce a classification system that is more robust and dependable. From a practical standpoint, the project might develop into an app that is ready for production. Docker would be used to containerize the system in order to guarantee cloud deployment and reproducibility. These enhancements would accelerate the discovery and validation of exoplanets by making the project more technically sophisticated and scalable for the scientific community.

10. REFERENCES

[1] Shallue, Christopher J., and Andrew Vanderburg. "Identifying Exoplanets with Deep Learning: A Five Planet Resonant Chain around Kepler-80 and an Eighth Planet around Kepler-90."

[2] Schanche, N., A. Collier Cameron, and G. Hébrard, et al. "Machine-learning Approaches to Exoplanet Transit Detection and Candidate Validation in Wide-field Ground-based Surveys."

[3] Malik, Abhishek, Benjamin P. Moster, and Christian Obermeier. "Exoplanet Detection using Machine Learning."

[4] Liu, Jinsong, Jingying Tang, and Songhu Wang, Jianfeng Liu. "GPU phase folding and deep learning method for detecting exoplanet transits."

[5] Tardugno Poleo, Valentina, and Nora Eisner, et al. "Detection of Exoplanets in Transit Light Curves with Conditional Flow Matching and XGBoost."

[6] Goyal, Anjali, and Neetu Sardana. "Detection of Exoplanets Using Classical Machine Learning Techniques."

[7] Wang, Zefang, Xiaokai Zhang, and Kejian Wang. "Identifying Light-curve Signals with a Deep-learning-based Object Detection Algorithm."

[8] Armstrong, David J, Jevgenij Gamper, and Theodoros Damoulas, et al. "Exoplanet validation with machine learning: 50 new validated Kepler planets."

[9] Jin, Yucheng, Lanyi Yang, and Chia-En Chiang. "Training a convolutional neural network for exoplanet classification with transit photometry data."

[10] [Authors not available]. "Classifying Kepler light curves for 12 000 A and F stars using supervised feature-based machine learning."

[11] Garvin, Emily O., Markus J. Bonse, and Jean Hayoz, et al. "Machine Learning for Exoplanet Detection in High-Contrast Spectroscopy: Revealing Exoplanets by Leveraging Hidden Molecular Signatures in Cross-Correlated Spectra with Convolutional Neural Networks."

[12] Prithivraj, G., and Alka Kumari. "Identification and Classification of Exoplanets Using Machine Learning Techniques."

[13] Nath-Ranga, R., O. Absil, V. Christiaens, and E. O. Garvin. "Machine Learning for Exoplanet Detection in High-Contrast Spectroscopy: Combining Cross-Correlation Maps and Deep Learning on Medium-Resolution Integral-Field Spectra."

[14] Agrawal, Vaibhav, and Rahul Kumar Singh. "Deep Learning based Exoplanet Detection using TESS Data."

[15] Thompson, Aaron, and Jane Smith, et al. "A Semi-Supervised Approach to Exoplanet Candidate Validation from Photometric Surveys."