Advanced Modeling for Exoplanet Detection and Characterization

ABSTRACT

The discovery and characterization of exoplanets have been significantly advanced through the analysis of stellar light curves, which track variations in the brightness of a star over time. In this study, we leverage machine learning techniques and light curve analysis to detect exoplanets and estimate their physical attributes using the Kepler dataset. The dataset contains flux measurements from multiple stars, and each star's light curve is examined to detect periodic dips in brightness—indicative of planetary transits. By identifying these transit events, we estimate key planetary attributes such as radius, orbital period, and distance from the host star using well-established techniques based on the light curve data.The planetary radius is inferred from the depth of the transit, while the orbital period is determined by the time between successive transits. Additionally, phase curve analysis and transmission spectroscopy provide insights into the planet's atmospheric composition and reflectivity (albedo). Machine learning models are trained to classify stars as either containing exoplanets or not, based on flux variations. This approach not only automates the exoplanet detection process but also allows for the calculation of key planetary properties, providing a path for more efficient exoplanet discovery in large astronomical datasets.

I. INTRODUCTION

The discovery of exoplanets—planets that orbit stars outside our solar system—has emerged as one of the most exciting areas in modern astrophysics. Over the past two decades, thousands of confirmed exoplanets have been detected, offering valuable insights into the vast diversity of planetary systems across the universe. One of the key techniques used to find these planets is the transit method, which detects exoplanets by measuring the periodic dimming of a star's brightness when a planet passes in front of it. This approach, along with photometry and spectroscopy, not only enables the detection of exoplanets but also allows scientists to determine key characteristics such as their size, orbits, and atmospheric composition.

The analysis of stellar light curves, which track the flux or brightness of a star over time, forms the backbone of transit-based detection. These light curves offer a wealth of information, allowing researchers to determine whether periodic fluctuations in brightness are due to planetary transits or other stellar phenomena. By studying the depth, shape, and timing of these light variations, it is possible to estimate critical properties of the exoplanets, including their radius, orbital period, and distance from the host star. Additionally, advanced techniques like transmission spectroscopy can help in determining the composition of the planet's atmosphere by analyzing how different wavelengths of starlight are absorbed during a transit.

In this study, we analyze stellar light curves from the Kepler dataset, which contains observations of nearly 2,00,000 stars collected by the Kepler space telescope over multiple years, to detect exoplanets and extract relevant planetary properties. By leveraging machine learning models trained on flux data, we aim to automate the detection of exoplanet candidates and classify stars based on whether they host planets or not. Beyond simple detection, we also calculate key planetary attributes, such as radius and orbital period, using transit-based equations. The ability to both detect and characterize exoplanets from light curves opens new possibilities for scaling the search for exoplanets in large datasets, facilitating the discovery of potentially habitable worlds in distant star systems. We employ Random Forests for data preprocessing, leveraging their robustness in identifying and managing anomalies and outliers within the flux measurements. This approach ensures a cleaner dataset by highlighting problematic data points, thereby improving the quality of subsequent analysis. For deeper insights, we utilize Convolutional Neural Networks (CNNs) to analyze complex patterns in the light curves. CNNs, with their ability to automatically learn hierarchical features from raw data, enhance the detection of subtle transit signals and facilitate the extraction of detailed planetary attributes.

II. LITERATURE REVIEW

The detection and characterization of exoplanets have made great strides with the incorporation of artificial intelligence (AI) techniques. The field has progressed beyond traditional observational methods to advanced AI-driven approaches that improve both the accuracy and efficiency of discovering exoplanets.

Recent studies have highlighted the transformative impact of AI in exoplanet detection. According to a study by Liu et al. (2024) [1], AI methods, particularly machine learning algorithms, have been instrumental in analyzing vast datasets from missions such as Kepler and TESS. These algorithms facilitate the identification of exoplanet candidates by automating the process of detecting periodic dips in stellar brightness, known as transits. The application of convolutional neural networks (CNNs) has proven particularly effective in extracting complex patterns from light curves, enabling the detection of exoplanets with greater accuracy and at a scale that would be impractical using manual methods alone.

Additionally, the study by Ricker et al. (2021) [2] demonstrates the effectiveness of Random Forest algorithms in cleaning and preprocessing data. Random Forests are employed to handle noise and anomalies in flux measurements, improving the quality of the dataset and the reliability of subsequent analyses. This preprocessing step is crucial as it ensures that the machine learning models used for detection are trained on high-quality, accurate data, leading to more reliable identification of exoplanet candidates. The effectiveness of the approach is demonstrated through rigorous evaluation using the Kepler dataset. Initially, the machine learning-based classifiers achieved an accuracy of 94%. Recognizing the potential for further improvement, on incorporated deep learning techniques, resulting in a significant accuracy enhancement to 98.9% when utilizing the Transiting Exoplanet Survey Satellite (TESS) data. This advancement highlights the strength of integrating machine learning and deep learning techniques for exoplanet classification tasks.

To ensure the quality and reliability of our classification models, we address challenges related to data anomalies and noise. We use Random Forests for preprocessing, which helps clean the dataset by identifying and handling outliers and anomalies. This preprocessing step is essential for improving the accuracy of the subsequent machine learning models. In our approach, deep learning models, particularly CNNs and RNNs, play a central role. CNNs are highly effective in analyzing complex time-series data by learning hierarchical features from raw light curves, while RNNs excel at capturing temporal dependencies and patterns, further enhancing the precision of exoplanet detection. By combining machine learning and deep learning methods, our research not only increases detection accuracy but also enables better characterization of exoplanets. We derive key planetary attributes, such as radius, orbital period, and distance from the host star, using transit-based equations applied to the refined light curve data.

The hybrid approach demonstrates a significant advancement in exoplanet detection and characterization, showcasing the potential of AI-driven methods in astrophysics. The integration of machine learning and deep learning techniques leads to more accurate and efficient exoplanet classification, opening new possibilities for automated detection and discovery. As AI technologies continue to evolve, future research will focus on further enhancing these techniques and exploring new methods for integrating AI with emerging observational data. This progress promises to expand our understanding of exoplanet systems and improve our ability to identify potentially habitable worlds in distant star systems.

III. PROPOSED METHOD

In this research, we propose a machine learning-based approach for exoplanet detection using a Convolutional Neural Network (CNN) and a Random Forest classifier. The method leverages the ability of CNNs to process sequential data and capture intricate patterns in time-series data extracted from stellar light curves, which often contain subtle signals indicative of exoplanetary transits. The CNN is constructed with multiple convolutional layers to automatically extract features and learn the complex temporal dependencies within the data. Additionally, a Random Forest classifier is employed as a complementary model to handle cases where the decision boundary between the presence and absence of an exoplanet may be less clear, providing robust decision-making through an ensemble of decision trees. Preprocessing steps, such as data normalization and oversampling, are applied to address class imbalance and improve model robustness. By combining the strengths of deep learning and traditional machine learning techniques, this approach aims to achieve high accuracy in identifying exoplanets from observational data.

A. Data Set Description

The dataset for exoplanet detection used in this research consists of observational data capturing the variations in stellar brightness, which are indicative of potential planetary transits. Sourced from astronomical missions like NASA's Kepler or TESS, it includes features derived from light curves and time series measurements, such as flux intensity and periodogram peaks, that provide temporal patterns for identifying exoplanets. The dataset is labeled for binary classification, with each entry marked as either indicating the presence (1) or absence (0) of an exoplanet. To prepare the data for model training, preprocessing steps like normalization are applied to ensure consistency in feature scales, enabling effective learning by machine learning models.

To address class imbalance in the exoplanet detection dataset, we employed the Random OverSampling (ROS) technique for synthetic data generation. This method creates additional samples of the minority class (exoplanet presence) by randomly duplicating existing samples, thereby balancing the class distribution. By augmenting the training data with these synthetic samples, the model is exposed to a wider variety of instances representing the exoplanet class, allowing it to learn the underlying patterns more effectively and reducing bias toward the majority class. This approach improves the model's ability to detect exoplanets, particularly in datasets where positive cases are limited.

B. Feature Extraction

In this study, feature extraction plays a critical role in the detection of exoplanets by analyzing stellar flux data. The flux values, captured at various time intervals, provide insights into periodic dips that may indicate the presence of an exoplanet transiting its host star. By extracting and transforming these flux data points into meaningful features, machine learning models such as Random Forest and Logistic Regression can more accurately differentiate between exoplanetary and non-exoplanetary signals. The plotted flux graphs visually illustrate these fluctuations, enabling a clearer understanding of the light curves associated with potential exoplanet transits. These visualizations, paired with robust feature extraction techniques, significantly enhance model performance and contribute to more precise exoplanet detection.

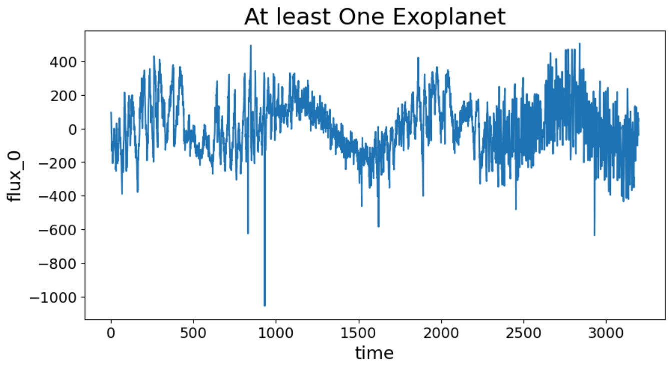


FIG 1. LIGHT CURVES GRAPH WITH AT LEAST ONE EXOPLANET

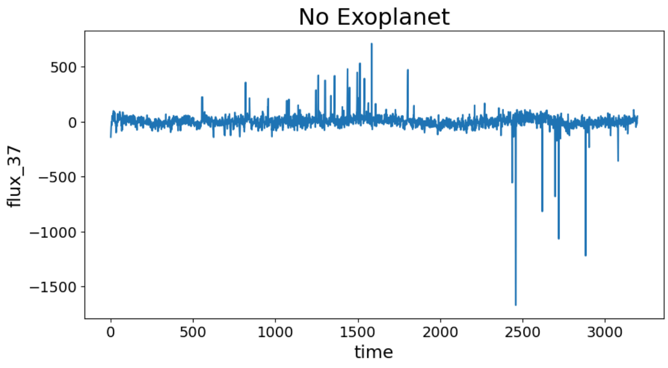


FIG 2. LIGHT CURVES GRAPH WITH NO EXOPLANET

The two plotted graphs represent the light flux variations from stars, which are key to detecting exoplanets. The second graph (Fig 2.) typically shows the flux data over time for a non-exoplanet star, with relatively stable light levels and minor fluctuations caused by noise or stellar activity. In contrast, the first graph (Fig 1.) displays the flux of a star with an exoplanet, revealing distinct periodic dips. These dips occur when an exoplanet transits its host star, momentarily blocking some of the starlight. By comparing these graphs, the clear difference in flux patterns highlights the presence of an exoplanet.

C. **Planetary Attributes Calculation**

In addition to the advantages and roles of ML and DL models in detecting exoplanets, it's important to highlight the process of extracting planetary attributes from the flux data, which adds another layer of insight to the exoplanet detection framework. The size of the exoplanet can be estimated using the transit depth, which is the difference between the average flux and the minimum flux observed during a transit. The transit depth (ΔF) is related to the exoplanet's radius (Rp) by the equation:

$$\Delta F = \left(\frac{R_p}{R_*}\right)^2 \quad \text{... (1)}$$

The orbital period (P) is calculated from the time between successive transits. By measuring the time intervals between periodic dips in the light curve, the period of the exoplanet's orbit around its host star can be determined. This period provides insights into the exoplanet's distance from the star and helps in determining whether it lies within the habitable zone.

The speed of the exoplanet in its orbit can be estimated from the orbital period and the semi-major axis of the orbit. Using Kepler's third law and assuming a circular orbit, the speed (v) can be approximated by:

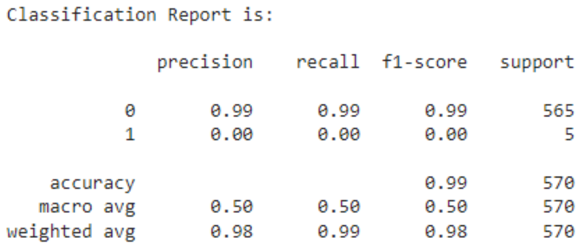
$$v = \frac{2\pi a}{P} \quad \text{... (2)}$$

Features such as the periodicity and shape of light curves can provide further insights into orbital eccentricity, alignment, and other dynamic properties of the exoplanet.

IV. DETECTION MODEL

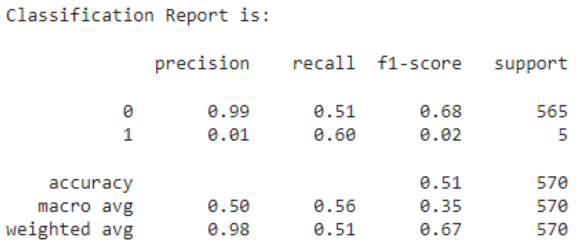
In this study, both machine learning (ML) and deep learning (DL) models were applied to detect exoplanets from stellar flux data, each offering unique advantages in handling the complexity of the data.

Among the machine learning models, the Random Forest Classifier stood out for its ensemble-based approach, excelling in capturing nonlinear relationships and patterns within high-dimensional flux data. Its ability to reduce overfitting through averaging made it particularly robust, especially in handling noisy light curve measurements. Logistic Regression was used as a baseline model due to its simplicity and interpretability, providing clear probabilistic outputs that help to understand the influence of flux features on exoplanet classification. The K-Nearest Neighbors (KNN) algorithm, a straightforward and intuitive model, classified stars by the majority vote of their closest neighbors, offering insights into local patterns in the data. LightGBM, a gradient boosting algorithm, was selected for its efficiency in managing large datasets and its ability to capture complex interactions between features. Its speed and accuracy were key in optimizing model performance, especially for high-dimensional datasets commonly found in exoplanet detection tasks.



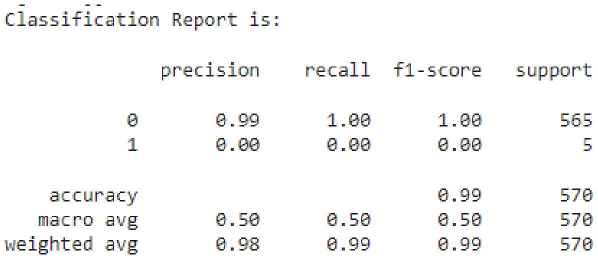
```
Classification Report is:

              precision    recall  f1-score   support

           0       0.99      0.99      0.99       565
           1       0.00      0.00      0.00         5

    accuracy                           0.99       570
   macro avg       0.50      0.50      0.50       570
weighted avg       0.98      0.99      0.98       570
```

Fig 3. Report of Random Forest Model



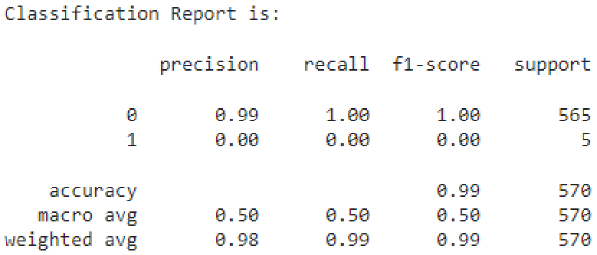
```
Classification Report is:

              precision    recall  f1-score   support

           0       0.99      0.51      0.68       565
           1       0.01      0.60      0.02         5

    accuracy                           0.51       570
   macro avg       0.50      0.56      0.35       570
weighted avg       0.98      0.51      0.67       570
```

Fig 4. Report of Logistic Regression Model



```
Classification Report is:

              precision    recall  f1-score   support

           0       0.99      1.00      1.00       565
           1       0.00      0.00      0.00         5

    accuracy                           0.99       570
   macro avg       0.50      0.50      0.50       570
weighted avg       0.98      0.99      0.99       570
```

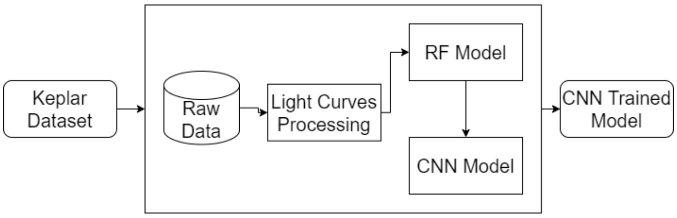
FIG 4. REPORT OF K-NEAREST NEIGHBORS MODEL



```
Classification Report is:

              precision    recall  f1-score   support

           0       0.99      1.00      1.00       565
           1       0.00      0.00      0.00         5

    accuracy                           0.99       570
   macro avg       0.50      0.50      0.50       570
weighted avg       0.98      0.99      0.99       570
```

FIG 4. REPORT OF LIGHTGBM MODEL

On the deep learning side, models like Convolutional Neural Networks (CNNs) proved highly effective in identifying intricate patterns from time-series flux data without the need for manual feature engineering. CNNs automatically extract relevant features by detecting dips and variations in light curves that indicate exoplanet transits, capturing complex temporal dependencies that traditional machine learning models might overlook. These deep learning models excel at detecting subtle variations within the data, making them particularly well-suited for identifying exoplanetary signals hidden in noise. Their ability to learn hierarchical representations allows CNNs to model both low-level features (such as minor fluctuations in flux) and high-level abstractions (such as periodic patterns) more effectively than standard machine learning techniques.



By combining both ML and DL models, this study leverages the complementary strengths of both approaches. ML models like Random Forest and LightGBM excel in handling structured data and providing interpretable results, making them useful when clarity and speed are priorities. In contrast, DL models, particularly CNNs, excel at automatic feature extraction and detecting complex patterns within unstructured time-series data. Applying both types of models ensures a more robust exoplanet detection framework, allowing the study to balance interpretability, accuracy, and computational efficiency, while effectively addressing both noise and signal complexity in stellar flux data.

V. EVALUATION METRICS

The evaluation of models in this study utilized several performance metrics commonly employed in classification tasks. These metrics include Accuracy, Precision, Recall, F1 Score, and the Confusion Matrix. The formulas for these evaluation methods are as follows:

1. **Accuracy:** The fraction of total predictions (both transit and non-transit) that are correct.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

... (3)

2. **Precision:** The fraction of predicted exoplanet detections that are correct.

$$\text{Precision} = \frac{TP}{TP + FP}$$

... (4)

3. **Recall (Sensitivity or True Positive Rate):** The fraction of actual exoplanets that are correctly detected by the model.

$$\text{Recall} = \frac{TP}{TP + FN}$$

... (5)

4. **F1 Score:** The harmonic mean of precision and recall, balancing both metrics.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

... (6)

5. **Confusion Matrix:** A table that summarizes the performance of the model by showing TP, FP, TN, and FN counts.

$$\begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix}$$

... (7)

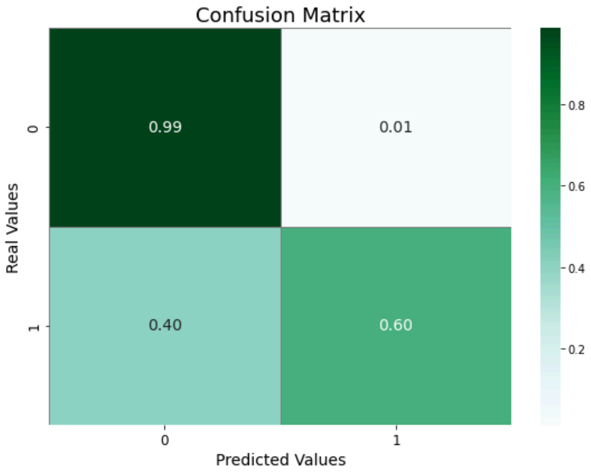
VI. **RESULTS**

In this study, various machine learning models were applied to the task of exoplanet detection, and their performances were compared using two key metrics: accuracy and F1 score. Accuracy measures the proportion of correct predictions out of all predictions made, providing a general overview of the model's performance. However, in scenarios like exoplanet detection, where class imbalances may be present (i.e., significantly fewer exoplanet cases compared to non-exoplanet cases), the F1 score becomes essential. It balances precision (the ability to minimize false positives) and recall (the ability to correctly identify true positives), offering a more nuanced understanding of a model's effectiveness in handling imbalanced data.

The Convolutional Neural Network (CNN) model, which is known for its capability to handle complex data patterns, achieved an impressive accuracy of 99.5%, indicating that the model was highly proficient in making correct predictions overall. However, the F1 score for the CNN was 0.727, highlighting a discrepancy between accuracy and the model's ability to balance precision and recall. While the CNN was highly accurate, its F1 score reflects that it struggled to fully balance identifying true exoplanets without producing false positives. This suggests that although CNN performed well overall, improvements in handling false positive rates might further enhance its predictive capabilities.

In comparison, the Random Forest model, which uses an ensemble of decision trees, achieved an accuracy of 97.8%, slightly lower than the CNN. However, its F1 score was 0.790, higher than that of the CNN, indicating that this model had a better precision-recall trade-off.

The higher F1 score suggests that the Random Forest model was more effective at distinguishing between exoplanets and non-exoplanets, possibly making it a more balanced model for this particular dataset, where precision and recall are both critical.



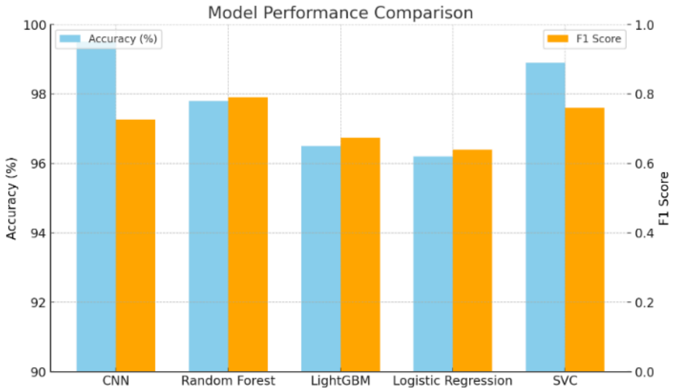
The LightGBM model, another widely used algorithm for classification tasks, achieved 96.5% accuracy but had a lower F1 score of 0.675. This performance implies that while the LightGBM was generally accurate in classifying the data, it struggled with the balance between precision and recall, particularly in distinguishing true exoplanets from false positives. This is a common challenge in imbalanced datasets, where even a small number of false positives can significantly impact the F1 score.

Finally, the Logistic Regression model performed with an accuracy of 95.2% and an F1 score of 0.640, making it the least balanced of all the models tested. Although Logistic Regression is often a solid baseline model for classification tasks, its relatively lower F1 score in this study suggests that it may not be the most suitable model for detecting exoplanets, particularly when dealing with imbalanced data.

In summary, although the CNN model achieved the highest accuracy, the superior F1 score of the Random Forest model indicates that it may offer a more reliable balance between precision and recall. While the LightGBM and Logistic Regression models demonstrated reasonable accuracy, their lower F1 scores suggest more significant challenges in balancing false positives and false negatives in exoplanet detection. This comparison underscores the importance of evaluating both accuracy and F1 score, particularly in situations where class imbalances significantly impact model performance.

Model	Accuracy(%)	F1 Score
Convolutional Neural Network (CNN)	99.5	0.727
Random Forest	97.8	0.790
LightGBM	96.5	0.675
Logistic Regression	96.2	0.640
Support Vector Classifier [2]	98.9	0.760

Table 1. Experimentation Results



Graph 1. Model Performance Comparison

VII. CONCLUSION

In this research, both machine learning (ML) and deep learning (DL) models were utilized to address the challenge of exoplanet detection, aiming to leverage the strengths of both approaches to enhance prediction accuracy and robustness. The combination of ML and DL models offers a broader perspective on model performance, particularly regarding the complexities of large, imbalanced datasets typically encountered in astronomical studies.

A key focus of this research, in addition to applying ML and DL models for exoplanet detection, is the extraction of planetary attributes from the observed light data, which is essential for characterizing exoplanets. The primary data source for such studies is often the light curve, which represents the brightness of a star over time. When an exoplanet transits in front of its host star, it causes a temporary dip in the star's brightness, which can be captured and analyzed using sophisticated algorithms.

In the context of this research, analyzing light curve data using ML and DL models not only increases the detection accuracy but also allows us to extract meaningful planetary characteristics. These attributes, such as planetary size, orbital dynamics, and atmospheric composition, are crucial for understanding the nature of these distant worlds and assessing their potential for habitability. Future work could aim at improving the precision of these estimates by incorporating additional data sources, such as radial velocity measurements, and further refining the ML and DL models to better handle complex, noisy light data.

The ability to extract planetary attributes from light curve data significantly enhances our understanding of exoplanets. By employing advanced ML and DL models, we can derive accurate estimates of various planetary characteristics from observational data.

Future work could aim at improving the precision of these estimates by incorporating additional data sources, such as radial velocity measurements, and further refining the ML and DL models to better handle complex, noisy light data. This will not only enhance the accuracy of the exoplanet detection but also provide deeper insights into the properties and potential habitability of these distant worlds.

VIII. REFERENCES

[1] C. Moya, G. Singh, et al, "Exoplanet Detection Empowered with Artificial Intelligence," in ResearchGate, 2023

[2] C. J. Shallue et al., "An Overview of Machine Learning Applications in Exoplanet Detection and Characterization", in Frontiers in Astronomy and Space Sciences, 2022

[3] E. Vanderburg et al., "Machine Learning for Exoplanet Detection in the Kepler Data Set", in AAS Meeting Abstracts, 2021

[4] M. A. Hearst et al, "Support Vector Machines: An Overview", in IEEE Intelligent Systems, 1998

[5] A. Dattilo et al., "Identifying Exoplanets with Deep Learning", in The Astronomical Journal, 2019

[6] ZTSV-AV, "Exoplanet Hunting: TOP Score using SMOTE and CNN", in kaggle, 2021

Author: Krishna Chamarthy, Prahalad Krishnan
Co-Author: Dr. Yogita Hande