Exploring Exoplanet Diversity: Ensemble Classification with XAI

Ishvarya G
Department of Computer Science and Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
bl.en.u4aie22118@bl.students.amrita.edu

Kavya Sree Kammari
Department of Computer Science and Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
bl.en.u4aie22121@bl.students.amrita.edu

Sharanya Vanraj Thambi
Department of Computer Science and Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
bl.en.u4aie22151@bl.students.amrita.edu

***Abstract*—This study explores the classification of exoplanets using machine learning (ML) algorithms and ensemble classifiers. It leverages data from the K2 planet candidates to enhance classification accuracy and gain insights into feature importance for planet classification. The vast amount of data from K2 presents both challenges and opportunities, requiring robust and informative feature selection. We investigate the performance of various ensemble classifiers, such as random forests and gradient boosting, aiming to achieve superior classification accuracy compared to single models.**

The comparative analysis will evaluate metrics like recall, precision, and F1-score. We will employ explainable AI (XAI) techniques to understand the significance of individual features in the ensemble models' predictions. This analysis will reveal which features are most crucial for accurate classification of K2 planets. This study represents a stepping stone towards developing even more sophisticated and reliable machine learning models for exoplanet classification, leveraging future datasets and incorporating additional features.

***Index Terms*—Exoplanets, Machine learning, Classification, explainable AI**

I. Introduction

For centuries, the existence of exoplanets, planets orbiting stars beyond our own solar system, has increased the scientific curiosity and fueled our deepest questions about the universe. With the advent of powerful telescopes and innovative techniques like transit photometry, the veil covering these distant worlds has begun to lift. NASA's Kepler mission and its successor, K2, have played a pivotal role, generating a wealth of data – the K2 planet candidate catalog containing with potential exoplanets awaiting confirmation and further characterization.

However, analyzing and interpreting this vast data poses significant challenges. Machine learning, a rapidly evolving field known for its ability to recognize patterns from complex datasets, is now making a few breakthroughs on exoplanet research. By using the power of algorithms, scientists can automate tasks, identify subtle patterns invisible to human analysis, and ultimately, accelerate the discovery and characterization of exoplanets.

This research leverages the power of ensemble classifiers, a machine learning technique that combines the predictions of multiple individual models to achieve superior accuracy and robustness compared to single models. By applying this approach to the K2 dataset, we aim to not only improve the classification accuracy of potential exoplanets but also gain deeper insights into the crucial features that differentiate them from other astronomical objects.

Moreover, we employ explainable AI (XAI) techniques to demystify the "black box" nature of ensemble classifiers. XAI methods shed light on the specific features that contribute most significantly to the model's predictions, providing valuable scientific validation and fostering trust in the results.

This research represents a significant step forward in leveraging machine learning to unravel the mysteries of exoplanets. By combining the power of ensemble classifiers with the interpretability of XAI, we aim to refine our understanding of these distant worlds and their characteristics.

II. Literature Survey

The findings underscore the potential of machine learning methodologies in assisting astronomers with the efficient and accurate verification of exoplanet candidates within extensive astronomical datasets [1]. By employing statistical and machine learning techniques, significant features can be identified, leading to enhanced classification accuracy and facilitating the discovery process. Future research endeavors could focus on further refining machine learning models by incorporating additional features and optimizing algorithms.

Utilizing a machine learning model, this study [2] has successfully increased the accuracy of classifying Kepler exoplanets from galaxy data. By harnessing computational methods, the research significantly improves the efficiency and reliability of exoplanet identification within the Kepler dataset. Future scope should focus on refining the machine learning

model by incorporating additional features and optimizing algorithms.

In exoplanet detection, a novel transit method utilizes light curve observations, detecting host star dimming[3]. To address dataset imbalance, Synthetic Minority Over-sampling Technique (SMOTE) is applied. Among six machine learning models, Majority Voting, integrating insights from four models, achieves a notable accuracy of 99.97%. Future research may explore alternative feature engineering and ensemble learning techniques to enhance accuracy and innovate exoplanet detection methodologies.

In predicting exoplanet habitability, a thorough study on the PHL_EXOPLANET_CATALOG.csv dataset, featuring 112 variables, addressed data imbalance through oversampling across 5640 records[4]. Preceding model application, essential steps including cleaning, preprocessing, resampling, and feature selection were executed. Machine learning models LR, SVM, XG Boost, and K-fold cross-validation were employed, with stratified K-fold cross-validation proving most effective at 95.2% accuracy. Future work emphasizes integrating explainable ML for model interpretability and testing on updated NASA datasets for broader applicability.

The classic approach of gathering data then pre-processing it followed by feature selection is done then the data is trained and evaluated[5]. The dataset is from NASA Exoplanet Science Institute. Redundant features were removed and outliers were dropped after detection. The obtained dataset is balanced and a train test validation ratio of 75:15:10 was taken. Dimensionality reduction was done using RFE and PCA. The approach taken is using PCA followed by Fully Connected Model and Random Forest Model each then RFE with Fully Connected Model and Random Forest Model each. Accuracy increases with number of hidden layers, 98-99% was the accuracy obtained. Different combinations of models can be used and outliers can be removed manually as a future scope.

Efficiency comparison of Machine Learning algorithms for detection of exoplanets[6]. The proposed model first uses the available flux dataset of stars that exoplanets orbit, which is trained using the KNN algorithm, and the accuracy is recorded. Then, the original data is balanced by using SMOTE to generate more data to create a balanced dataset. This data is trained using three algorithms: KNN, Logistic Regression, and Decision Trees, and the results are recorded. SMOTE oversamples the data by synthesising more data by observing the available data so that the minority class is now equal to the majority class. The accuracy report is generated before and after applying SMOTE. KNN yields the best accuracy while Logistic Regression has the least. Different ML algorithms can be studied, and their combinations as well, to increase accuracy.

The ThetaRay algorithm is used[7] to find anomalies in the light curves of exoplanets. This algorithm employs various mathematical and AI/ML methods to detect abnormalities. Subsequently, semi-supervised processing takes place, where some labels of Kepler TCE are provided, but no TESS data is available. A new dataframe is generated by an augmented algorithm, and this dataframe is used by unsupervised learning algorithms. The three unsupervised learning algorithms used are geometric-based NY, algebraic-based LU, and neural network-based AE. The reduction in false positives is plotted. In the future, ThetaRay can be optimized further, and false positives can be further reduced.

In response to the increasing complexity of spacecraft missions, this paper introduces a machine learning-based anomaly detection system for spacecraft telemetry data. Leveraging NASA datasets from SMAP and MSL missions, the proposed approach outperforms traditional methods and expands anomaly detection capabilities[8]. Key contributions include addressing five anomaly types through comprehensive feature extraction, comparing three machine learning methods, and providing an explainability analysis. Acknowledging limitations, the paper suggests future research directions, including sensitivity to noise exploration, consideration of more complex models, expansion of anomaly types, real-time implementation, and integration with human-in-the-loop validation. This work signifies a crucial step in enhancing anomaly detection in spacecraft telemetry, laying the groundwork for advanced applications in evolving space missions.

PIML explores the techniques of physics-informed machine learning (PIML) in the realms of astronomy and cosmology, extending beyond conventional classification and prediction tasks [9]. The future scope is paving the way for future advancements in these fields, includes further refinement of algorithms for complex simulations, especially explaining the interpretability of the ML black boxes and the integration of XAI and PIML.

The analysis reveals valuable insights into the suitability of different machine learning algorithms for multi-touch attribution modeling [10]. By assessing their performance metrics such as accuracy, precision, and recall, it becomes apparent which algorithms excel in accurately attributing conversions across multiple touchpoints. Future research directions may involve investigating novel machine learning algorithms specifically tailored for multi-touch attribution modeling.

III. Methodology

A. Dataset

The K2 dataset, stemming from NASA's K2 mission, contains a vast array of astronomical data capturing stars, galaxies, and exoplanets across different parts of the sky. It encompasses various features like photometric measurements, light curves, and celestial coordinates, providing invaluable insights into the universe.

we will simplify the preprocessing of the K2 dataset by focusing on removing null values, addressing error ranges, and implementing labeling techniques. These steps will ensure the data's quality and usability for efficient analysis and exploration in astronomical research.

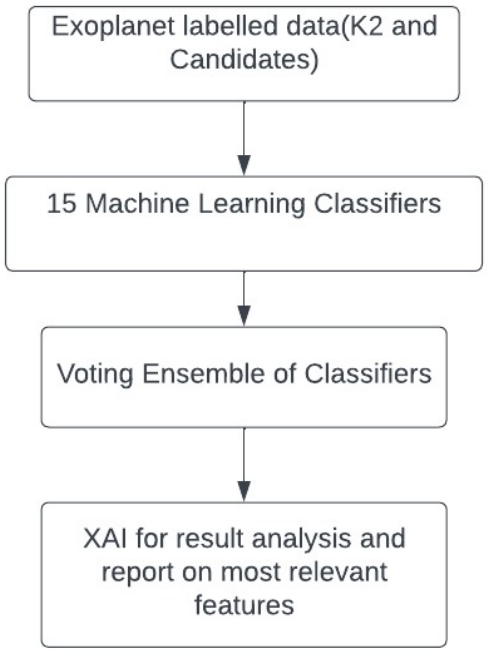

Fig. 1. The flowchart of the methodology.

B. Ensemble Learning

An ensemble classifier combines multiple individual classifiers to improve prediction accuracy and robustness by aggregating their outputs. In the future, our ensemble approach will yield good accuracy by combining the strengths of multiple algorithms. This strategy ensures robustness and reliability in our predictions, leading to more accurate and dependable results.

C. XAI

We're employing eXplainable Artificial Intelligence (XAI) techniques to shed light on why certain results are obtained. This involves breaking down the decision-making process of our algorithms, making it easier to understand and interpret. For each algorithm that yields the highest accuracy within our ensemble, we're implementing XAI methods. This approach not only helps us identify the most effective algorithms but also provides insights into their inner workings, ultimately enhancing the transparency and trustworthiness of our models.

REFERENCES

[1] Karim, Abdul, Jamal Uddin, and Md Mahmudul Hasan Riyad. "Identifying Important Features For Exoplanet Detection: A Machine Learning Approach." GPH-International Journal of Applied Science 7, no. 01 (2024): 01-17.

[2] Nayak, Sasmita Kumari. "Classification of Kepler exoplanet searching from galaxy using machine learning model." (2024).

[3] Rakesh, G. Venkata Sai, M. Jahnavi Bhuvana Chandrika, Ch Venkata Rami Reddy, and Muvva Suneetha. "Exoplanet Detection Using Feature Engineering with Ensemble Learning." In 2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN), pp. 116-122. IEEE, 2023.

[4] Raminaidu, Ch, V. Priyadarshini, Ch Ravi Swaroop, and R. Shiva Shankar. "Building Accurate Machine Learning Models for Predicting the Habitability of Exo-Planets." In 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT), pp. 961-967. IEEE, 2023.

[5] Virmani, Chaitanya, Ria Singhla, Priyanka Gupta, and Hardeo Kumar Thakur. "Identifying Exoplanet Candidates with Machine Learning." In Advances in Signal Processing, Embedded Systems and IoT: Proceedings of Seventh ICMEET-2022, pp. 333-343. Singapore: Springer Nature Singapore, 2023.

[6] Herur, Aahish Nagesh, Raquib Tajmohamed, and J. Godwin Ponsam. "Exploring Exoplanets using kNN, Logistic Regression and Decision Trees." In 2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), pp. 1-7. IEEE, 2022.

[7] Ofman, Leon, Amir Averbuch, Adi Shliselberg, Idan Benaun, David Segev, and Aron Rissman. "Automated identification of transiting exoplanet candidates in NASA Transiting Exoplanets Survey Satellite (TESS) data with machine learning methods." New Astronomy 91 (2022): 101693.

[8] Cuéllar, Sara, Matilde Santos, Fernando Alonso, Ernesto Fabregas, and Gonzalo Farias. "Explainable anomaly detection in spacecraft telemetry." Engineering Applications of Artificial Intelligence 133 (2024): 108083.

[9] Meskhidze, Helen. "Beyond Classification and Prediction: The Promise of Physics-Informed Machine Learning in Astronomy and Cosmology." (2024).

[10] Pattanayak, Satyabrata, Peeta Basa Pati, and Tripty Singh. "Performance Analysis of Machine Learning Algorithms on Multi-Touch Attribution Model." In 2022 3rd International Conference for Emerging Technology (INCET), pp. 1-7. IEEE, 2022.