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 KOI planet candidates to enhance classification accuracy and gain insights into feature importance for planet classification. The vast amount of data from KOI 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 KOI 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 KOI, have played a pivotal role, generating a wealth of data – the KOI 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 KOI 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.

1) Data Preprocessing: Data preprocessing is a critical step in the machine learning pipeline that involves transforming raw data into a format suitable for analysis and modeling. Missing values in our dataset are identified and addressed using imputation, ensuring dataset completeness and preserving valuable information for accurate analysis and modeling. Categorical variables are converted into a numerical format suitable for modeling using one-hot encoding. The dataset is divided into training and testing sets to evaluate the performance of the machine learning model. This allows for the model to be

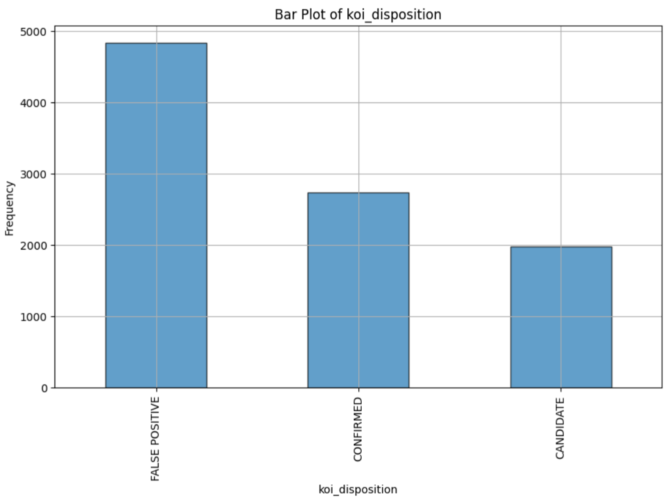

Fig. 1. The histogram of the target feature - koi_disposition

trained on one subset of the data and validated on another subset, ensuring that the model's performance generalizes well to unseen data.

2) Data separability: The classes in the dataset appear to be reasonably well separated based on the distribution of the target variable koi_disposition. There are three distinct class labels: "CONFIRMED", "CANDIDATE", and "FALSE POSITIVE". CONFIRMED: 2741 instances, CANDIDATE: 1984 instances, FALSE POSITIVE: 4839 instances. While there is a class imbalance with more instances of "FALSE POSITIVE" compared to the other classes. The histogram of the koi_disposition in Fig.1 shows the comparision.

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.

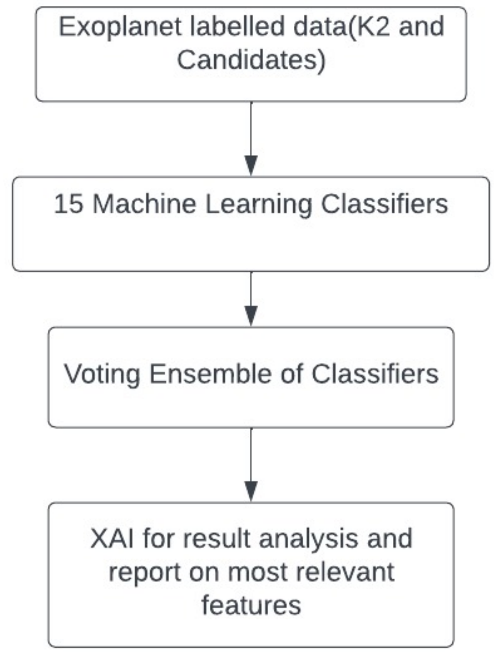

Fig. 2. The flowchart of the methodology.

IV. RESULTS AND DISCUSSIONS

A. KNN Classifier

With a smaller value of k, such as k=1, the model becomes more complex as it relies heavily on the nearest neighbor for classification. This can lead to high variance and overfitting. As k increases, the model's complexity decreases, and it tends to generalize better to unseen data. This is because the decision boundaries become smoother, and the classification is based on a larger number of neighbors. The trend suggests that as the value of k increases, the accuracy tends to decrease. This is evident from the decreasing accuracies: 0.778, 0.736, 0.770, 0.736, etc. This behavior is expected as larger values of k generally lead to simpler models with lower variance but potentially higher bias. Therefore, the highest accuracy is achieved with k=1 (0.778), indicating that a smaller neighborhood size is optimal for this particular dataset.

KNN classifier can be considered as an average classifier. We get around 77% accuracy, the accuracy changes significantly over 11 values of k, choosing the best possible value of k=3 we our maximum accuracy. Considering our data having 3 class labels, the classifier could have performed significantly better. KNN does not learn about the data and follows the prediction of the datapoints with the least distance with our testing point. It then finds which class label is closest to the class labels of the top k closest datapoints. It predicts according to which ever class label is predicted more and "assumes" that the new data point falls under that category rather than making an accurate classification by learning like

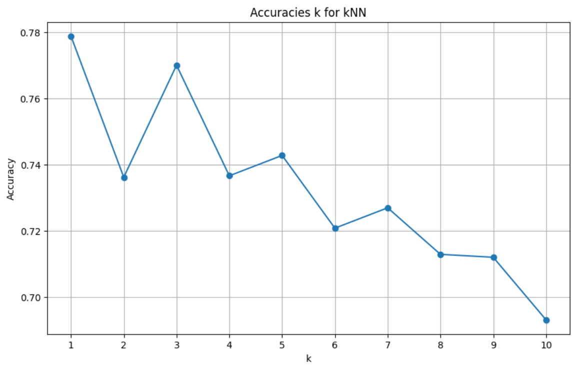

Fig. 3. The accuracies of the KNN classifiers from k = 1 to 11

other algorithms.

In the analysis of the kNN classifier's performance, the evaluation focused on both training and testing accuracy metrics. The plot_train_test_accuracy function was employed to train the model with varying numbers of neighbors (k) and assess its accuracy on both the training and test datasets.

The results revealed an intriguing pattern. The training accuracy exhibited a notable increase from 0.676 to 1, signifying that the model achieved a perfect fit to the training data. However, such a high training accuracy may raise concerns about overfitting, suggesting that the model potentially memorized the intricacies of the training set.

Conversely, the testing accuracy experienced an increase from 0.59 to 0.70, indicating an enhancement in the model's ability to generalize to unseen data. Despite this improvement, the testing accuracy remained relatively modest, highlighting potential limitations in the model's generalization capacity. The discernible gap between training and testing accuracy is indicative of overfitting, where the model's high performance on the training set does not seamlessly translate to novel, unseen data.

To mitigate the overfitting concern, further model refinement is recommended. This may involve fine-tuning hyperparameters, such as the number of neighbors (k), or adopting advanced techniques like cross-validation to achieve a more balanced compromise between accurately capturing training data and ensuring robust generalization to new instances. Striking this equilibrium is essential for developing a kNN classifier that not only excels in learning from training data but also demonstrates resilience and efficacy in handling previously unseen patterns. overfitting tends to occur when the model becomes too complex, capturing noise and fluctuations in the training data rather than underlying patterns. Specifically, overfitting in kNN can happen when the number of neighbors (k) is too small.

In our dataset and more generally, if a kNN classifier has a small value of k, it tends to be more sensitive to local variations in the training data. With a low k, the model might fit the training data too closely, effectively memorizing the training instances, including noise and outliers. As a result, the classifier may struggle to generalize well to new, unseen data, leading to a situation where the model performs exceptionally well on the training set but poorly on the test set.

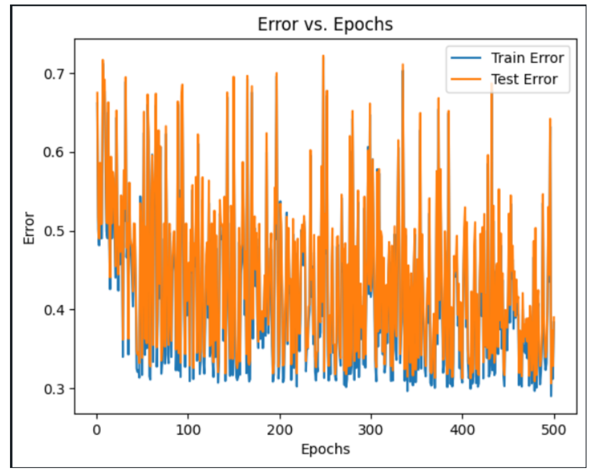

Fig. 4. The testing ad training errors of the multi layer perceptron

TABLE I
CLASSIFICATION REPORT OF CLASSIFIERS

Classifiers	**Precision**	**Recall**	**F1 Score**	**Accuracy**
Naive Bayes	0.57	0.53	0.47	51.54%
Decision Trees	0.88	0.88	0.88	91%
SVM	0.16	0.33	0.22	49%
CatBoost 1.0	1.0	1.0	1.0	100%
AdaBoost	1.0	1.0	1.0	100%
XGBoost	0.94	0.93	0.93	95%
Random Forest	0.93	0.91	0.92	94%

B. Multi Layer Perceptron

The multi-layer perceptron (MLP) model with two hidden layers, comprising 100 nodes in the first layer and 50 nodes in the second, has been trained on labeled data using label encoding. Throughout 500 epochs of training, the model achieved a training accuracy of 61.6% and a testing accuracy of 61%. The choice of Rectified Linear Unit (ReLU) activation function was utilized to introduce non-linearity into the model, enhancing its ability to learn complex patterns within the data. Despite the modest accuracy scores, the model's performance suggests it has successfully captured some underlying relationships within the dataset, although there may be room for further optimization to improve its predictive capabilities. The fig.4 shows the testing and training errors of the training process.

C. Multiple classifiers results

We employed a diverse set of classifiers to tackle our classification task. Naive Bayes yielded an accuracy of 51.54%. Decision Trees achieved an accuracy of 91%. Support Vector Machines (SVM) showed promising results with an accuracy of 49%. Adaboost and CatBoost give 100% accuracy each, while checking for overfitting K-fold cross-validation was performed along with SHAP to understand the contribution of each feature to the classification. K-fold performed for

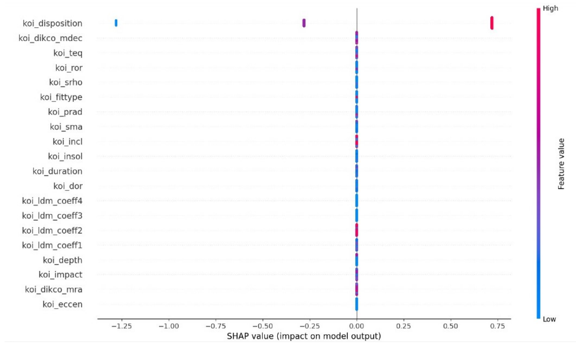

Fig. 5. The Shap values of AdaBoost Classifier

10 Folds, for the 9000 instance dataset gives 100% accuracy for all combinations of test and train set. SHAP distribution describes all the features to have a nearly equal contribution to the classification. Boosting algorithms tend to give high accuracy as they combine weak learners sequentially as well as identify important features. XGBoost demonstrated an accuracy of 95%.SVM with rbf kernel gives an accuracy of 49% this can be attributed to the absence of feature selection and the large number of features present in the dataset. Lastly, Random Forest exhibited strong performance with an accuracy of 94%. These results highlight the variability in performance across different classifiers, emphasizing the importance of exploring multiple approaches to achieve optimal accuracy.

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.