JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Predicting Alien Worlds: A Data Science Approach to Exoplanet Detection and Habitability

Mrs. R.Suruthi Sutharsana[1], Pingili Sai Amulya[2], Keesari Sai Akhil[3], Dachepally Kiran Kumar [4], Merugu Venkat Sai [5] and Pamidi Kalyan Kumar[6]

[1] Assistant Professor, Department of CSE (Data Science), ACE Engineering College, Hyderabad, Telangana, India

[2,3,4,5,6] IV B.Tech. Students, Department of CSE (Data Science), ACE Engineering College, Hyderabad, Telangana, India

Abstract:

"Predicting Alien Worlds: A Data Science Approach to Exoplanet Detection and Habitability" navigates for the identification of exoplanets, applying data science techniques to detect and categorize their habitability. Analyzing data from missions like Kepler, the project seeks to understand diverse exoplanets, going beyond detection to assess habitability. This initiative integrates insights from astronomy, physics, and data science,pushing the boundaries of our cosmic knowledge. Ultimately, it enhances understanding of exoplanetary systems and lays groundwork for future research in the quest for extraterrestrial life, marking a significant step in unraveling the mysteries of distant worlds.

Key words: Transit Method, Radial Velocity, Direct Imaging, Gravitational Microlensing, Astrophilation, Spectroscopy, Scaling Techniques, Stellar Characteristics, Exoplanet Detection, Habitability Prediction, Astronomical Data Analysis, Data Science In Astronomy, Astrobiology, Predictive Modeling, Kepler Space Telescope, Stellar Flux, Equilibrium Temperature, Planetary Radius, Orbital Parameters, And Insolation Flux.

1.Introduction:

An exoplanet is a celestial body which has planetary properties and revolves around star outside solar system. This very celestial object has cultivated the possibility of alien world that holds the key to understanding the formation and evolution of planetary systems and potentially discovering life beyond Earth. The detection of exoplanets has primarily been accomplished through various methods, each contributing unique insights into their characteristics. Among these methods, the transit photometry technique, which involves monitoring the dimming of a star's light as a planet passes in front of it, has proven particularly effective. Other methods, such as radial velocity, direct imaging, and gravitational microlensing, complement transit photometry by providing additional data on exoplanetary masses, atmospheres, and orbits.

A critical aspect of exoplanet research is the identification of habitable planets those that have the potential to support life. This determination hinges on several factors, including the planet's location within the habitable zone of its star, where conditions might allow for liquid water to exist on its surface. Moreover, the planet's size, atmospheric composition, and geological activity are also crucial in assessing its habitability. Recent advancements in space-based telescopes, such as Kepler, K2, and TESS, have significantly expanded our ability to detect and study these distant worlds. By leveraging the high-precision photometric data from these missions, researchers can refine their models of planetary habitability and uncover new candidates for detailed study.

2.Literature Review:

Review of Existing Literature Related to the Problem Domain:

The existing literature pertaining to exoplanet detection and habitability assessment presents a diverse landscape of research endeavors, spanning various detection methodologies, habitability criteria, and analytical techniques. This corpus of work illustrates how scientific research in the topic has changed over time, driven by technological breakthroughs and multidisciplinary cooperation.Across numerous studies, authors have explored a multitude of detection methods, with particular emphasis on the transit and radial velocity techniques. Works by Seager (2010) and Mayor and Queloz (1995) underscore the effectiveness of these methods, leading to significant discoveries such as TOI 700 d and GJ 367 b. Moreover, direct imaging techniques, as demonstrated by Macintosh et al. (2015), hold promise for observing exoplanet surfaces, albeit with inherent technological challenges.

Habitability assessment remains a central theme within the literature, with researchers investigating concepts such as habitable zones and biosignatures. Pioneering works by Kasting (1993) lay the groundwork for understanding habitable zones, delineating planetary conditions conducive to life. Studies by Meadows et al. (2017), Des Marais et al. (2005), and Tsiganis et al. (2008) delve into biosignatures, elucidating potential indicators of life processes and providing frameworks for their detection. These contributions collectively enrich our understanding of habitability assessment and guide the search for potentially habitable exoplanets.

In recent years, the literature has witnessed the emergence of innovative approaches to exoplanet research, leveraging advanced techniques such as machine learning and data science. Sarkar et al. (2023) propose a novel method for identifying potentially habitable exoplanets based on temperature anomalies, offering insights beyond traditional habitability zone models. Additionally, Wandel's exploration of extended habitability resulting from subglacial water oceans challenges conventional notions, opening new avenues for investigation. These innovative approaches underscore the dynamic nature of exoplanetary research and the continual quest to expand our knowledge of distant worlds.

The integration of machine learning and data science into exoplanet research marks a pivotal advancement in the field, as demonstrated by recent studies. Notably, papers such as those referenced as [3] and [4] showcase the efficacy of machine learning algorithms in habitability prediction and classification tasks, achieving notable accuracy rates through various supervised learning techniques. Additionally, [5] introduces innovative methods for habitability prediction, employing Gradient Boosted Regression Trees to attain remarkable accuracy levels.

However, alongside these advancements come notable challenges and avenues for future exploration. Despite the successes observed, issues such as imbalanced datasets and model optimization persist as significant hurdles. Works such as those referenced as [6], [7], and [9] delve into topics such as classification abilities, algorithm performance, and evaluation metrics, shedding light on areas ripe for improvement and indicating potential future research trajectories. As technology continues to evolve and methodologies advance, the pursuit of uncovering hidden worlds and assessing their habitability remains fueled by curiosity, innovation, and the collaborative efforts of interdisciplinary teams.

The area surrounding a star that is conducive to liquid water existing on a planet's surface, enhancing the possibility of habitability, is referred to as the habitable zone—also referred to as the Goldilocks zone.

Factors influencing habitability and the delineation of habitable zones include the luminosity and spectral type of the host star, the planet's distance from the star, its atmospheric composition, and its geophysical properties such as size and mass. The habitable zone is typically defined based on the balance between stellar flux and a planet's atmospheric conditions, ensuring that water neither freezes nor evaporates but remains in a liquid state on the planet's surface.

Research by Kasting (1993) and Kopparapu et al. (2013) has been instrumental in shaping our understanding of habitability and habitable zones, providing theoretical frameworks and models for assessing the potential habitability of exoplanetary systems. These works emphasize the importance of considering multiple factors in determining habitability and highlight the dynamic nature of habitable zones, which can vary depending on the properties of the host star and the planet itself.

This work explores exoplanets by leveraging advanced data science techniques to analyze and predict their existence and habitability. With an increasing number of exoplanets being confirmed, we employ classification algorithms to detect these distant worlds and assess their potential to support life. Specifically, we train and test binary classification algorithms for exoplanet detection, utilizing the Planetary Habitability Laboratory (PHL) catalog to assist in predicting the percentage likelihood of habitability. By integrating empirical data and theoretical models, this approach aims to enhance the accuracy and reliability of exoplanet detection and habitability assessments, contributing valuable insights to the field of exoplanetary research.

3.Existing Systems:

In exoplanet detection, classical methods have traditionally been employed, with gradient boosted trees serving as a prevalent technique for analyzing light curves. This process typically involves several stages, including data preprocessing, feature extraction, and model training using gradient boosted trees. For instance, in a recent study, researchers extracted a substantial number of features, totaling 789, from light curves to train a gradient boosting classifier using lightgbm. Within this framework, Transit Timing Variation events (TCEs) are scrutinized within light curves to filter out non-signals, with the Box Least Squares (BLS) algorithm commonly utilized to identify potential exoplanetary candidates.

However, there is a growing interest in exploring deep learning algorithms, particularly convolutional neural networks (CNNs), to enhance the vetting process of Kepler TCEs. By training models on human-classified Kepler TCEs, these CNNs demonstrate the ability to discern meaningful patterns through representation learning, offering a promising avenue for automating the detection process. Furthermore, researchers are exploring various neural network architectures tailored to the specific task of vetting transiting planet candidates, seeking to optimize performance and accuracy.

In this pursuit, novel techniques such as recurrence plots and image augmentation are being incorporated to improve pattern recognition amidst noise. By leveraging these innovative approaches, researchers aim to refine the detection of exoplanets within vast datasets, ultimately advancing our understanding of distant worlds beyond our solar system.

Habitability Assessment:

In habitability assessment, machine learning (ML) models play a crucial role in refining our understanding of potential life-supporting environments beyond our solar system. These models are trained on a diverse range of stellar and planetary parameters, including mass, luminosity, and orbital distance, to delineate the boundaries of a star's habitable zone more accurately. By leveraging ML algorithms, researchers can automate tasks such as signal detection and feature extraction, significantly expediting the identification process of promising exoplanet candidates for further investigation.

Moreover, deep learning algorithms are increasingly employed to analyze atmospheric data from exoplanets, aiming to detect potential biosignatures like methane or ozone. These biosignatures offer valuable clues about the presence of life-sustaining processes on distant worlds, providing insights into the potential habitability of exoplanetary environments. By delving into the intricate details of exoplanetary atmospheres, deep learning algorithms contribute to our ongoing quest to unravel the mysteries of extraterrestrial life.

In parallel, traditional habitability models continue to analyze fundamental stellar and planetary parameters, such as mass and luminosity, to predict the likelihood of habitable conditions across various types of worlds. These models serve to expand the scope of our search for habitable environments beyond our solar system, offering insights into the diversity of potential habitats within the cosmos. Through the integration of ML techniques and traditional modeling approaches, researchers strive to advance our understanding of habitability factors and identify targets for future exploration in the quest for life beyond Earth.

4.Proposed System

Problem Statement:

The primary objective of our project is to classify potential exoplanets and determine their likelihood of habitation. To accomplish this goal, we will select the best-performing model from our trained models, considering factors such as accuracy and efficiency. Our chosen model will be configured with a specific number of estimators and will utilize the Gini Impurity metric to evaluate the quality of the split.

In our assessment of the model's performance, we will go beyond binary classification metrics and incorporate a comprehensive set of evaluation measures. These include accuracy, which provides an overall indication of the model's correctness, and other metrics such as recall, F1 score, and precision, which offer insights into the model's ability to correctly identify positive instances and minimize false positives.

Additionally, we will utilize a confusion matrix to visualize the model's performance across different classes and assess its ability to distinguish between habitable and non-habitable exoplanets accurately. By incorporating these diverse metrics and evaluation techniques, we aim to provide a thorough and nuanced evaluation of the chosen model's effectiveness in predicting the habitability potential of the identified exoplanets.

5.Methodology:

5.1 Data Collection:

Data collection focused on utilizing NASA's Exoplanet Archive to obtain detailed information about exoplanets detected by the Kepler mission. The archive provided a comprehensive database, including parameters such as planet size, orbital period, and distance from the host star. These datasets were meticulously curated and validated by NASA, ensuring accuracy and reliability for subsequent analysis. By accessing this archive, I was able to compile and organize data critical for evaluating the detection methods and assessing the potential habitability of these exoplanets.

5.2 Data Ingestion:

The initial phase involved acquiring the necessary data from authoritative sources. Specifically, the Exoplanet Dataset and the Habitable Planets dataset were downloaded from the NASA Exoplanetary Archive. These datasets served as the foundation for subsequent analysis and model building.

5.3 Data Preprocessing:

To ensure data integrity and compatibility with the analytical tools, several preprocessing steps were undertaken:

1. Column Renaming: The column names in the datasets were renamed for improved clarity and interpretability, facilitating better understanding during analysis and documentation.

2. Handling Missing and Invalid Values: Instances with NaN (Not a Number), positive infinity (+inf), and negative infinity (-inf) values were identified and appropriately handled. These values were either imputed with suitable techniques or removed from the dataset, as deemed appropriate for training the model and conducting further analysis.

5.4 Exploratory Data Analysis (EDA)

The dataset is collected from NASA Exoplanet Archive which consists of 9564 instances and 49 attributes. Fig demonstrates the status of exoplanets mainly categorized to confirmed, candidate and false positive.

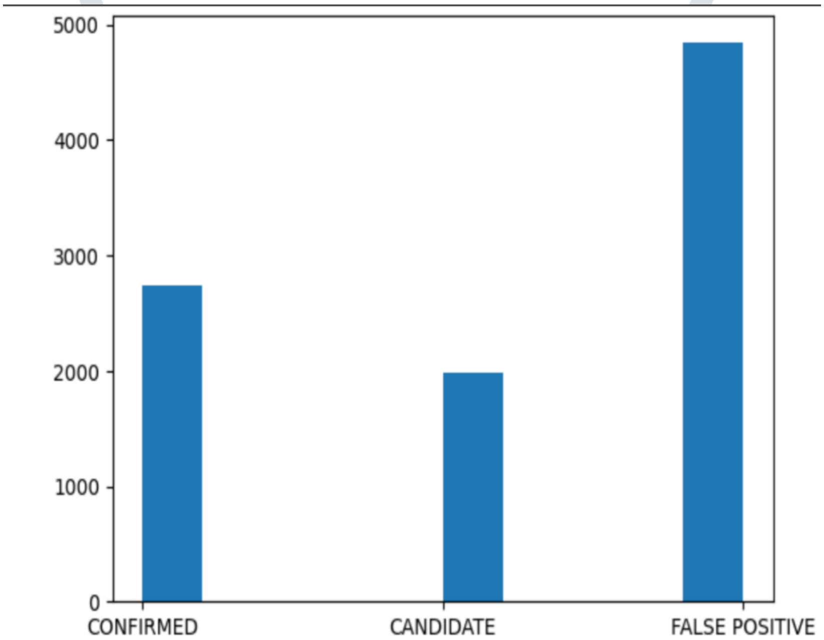

Fig 1: Categories of Exoplanets in dataset

1.Redundant Feature Identification: Redundant or highly correlated features were identified, as their inclusion could lead to multicollinearity issues and negatively impact model performance.

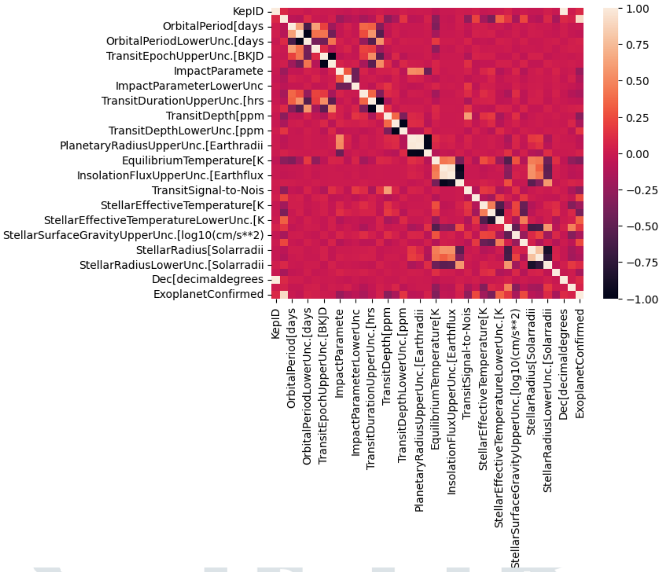

Fig 2: Correlation analysis of data

2. Feature Selection:

Feature Selection for Exoplanet Detection:

After preprocessing the data, utilizing feature importance in training models for exoplanet detection allows us to identify and prioritize the most influential factors affecting the model's predictions. This approach enhances the model's accuracy and efficiency by focusing on the critical attributes that significantly impact exoplanet detection. By understanding the importance of each feature, we can refine the model to improve its performance and reliability in identifying potential exoplanets.

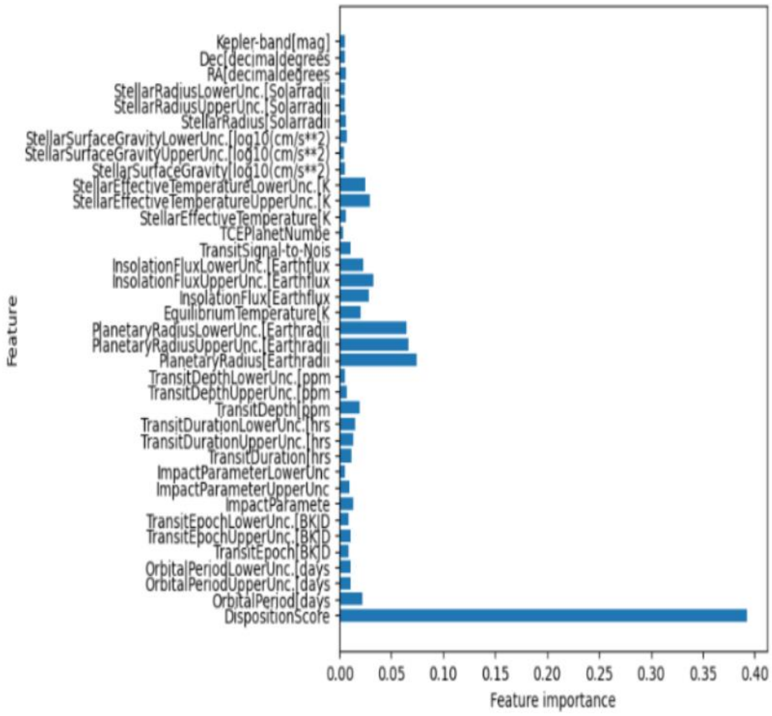

Fig 3: Feature Importance for Exoplanets

Feature selection for Habitable planets:

In the context of feature selection for habitability prediction, our model incorporates several key criteria derived from the PHL catalog to assess the potential habitability of exoplanets. These criteria include planetary radius, orbital semi-major axis, stellar surface gravity, equilibrium temperature, planetary density, radiative flux, eccentricity, and obliquity. The planetary radius is considered habitable if it falls between 0.8 and 1.5 times the radius of Earth. For the orbital semi-major axis, the exoplanet must lie within the habitable zone, which is defined as a range from 0.1 to 5 astronomical units (AU) from the host star, adjusted for the star's effective temperature.

Stellar surface gravity is included as a criterion, with suitable values ranging from 4.0 to 4.9 in log10(cm/s²). The equilibrium temperature of the exoplanet is another critical factor, with a habitable range set between 273 K and 373 K. The model also evaluates planetary density, differentiating between rocky planets and larger planets by assigning higher habitability scores to planets with radii up to 1.5 Earth radii, and lower scores to those with radii between 1.5 and 2.5 Earth radii. Radiative flux, or insolation, is calculated based on the Stefan-Boltzmann law, considering the host star's effective temperature and the exoplanet's orbital radius and stellar radius.

The exoplanet is deemed habitable if its insolation flux lies within an appropriate range. Additionally, the model takes into account the eccentricity of the exoplanet's orbit, with lower eccentricities (≤ 0.2) being favorable for habitability. Finally, obliquity, or axial tilt, is included, with the model favoring exoplanets that have an axial tilt within 45 degrees from the perpendicular to their orbital plane. These carefully selected features provide a comprehensive framework for assessing the habitability of exoplanets, leveraging empirical data and theoretical models to enhance the accuracy and reliability of habitability predictions in exoplanet research.

5.5 Model Building:

To address the classification problem, multiple suitable algorithms were implemented based on their relevance, computational efficiency, and ability to handle the dataset. The train test split of data is 70:30.

5.6 Model Testing:

Key metrics used in this evaluation included accuracy, precision, recall, and F1-score. Accuracy provided an overall measure of how often the model correctly predicted exoplanet detection. Precision indicated the proportion of true positive identifications among all positive predictions, reflecting the model's exactness. Recall measured the model's ability to identify all relevant instances, indicating its sensitivity. The F1-score, a harmonic mean of precision and recall, offered a balanced metric that accounts for both false positives and false negatives. These metrics collectively ensured a comprehensive assessment of the model's effectiveness in classifying and detecting exoplanets within the dataset.

Models displayed in Fig are used for model building and and evaluation results are sorted based on accuracy of model.

Table 1: Evaluation Metrics for Models trained

Model Name	Accuracy	Precision	Recall	F1_score
Random Forest	0.960282	0.971178	0.952088	0.961538
Decision Tree	0.931454	0.934809	0.933661	0.934235
KNN	0.797566	0.776667	0.858722	0.815636
Logistic Regression	0.796284	0.771335	0.866093	0.815972
SVM	0.710442	0.647154	0.977887	0.778865

6.Results and Analysis:

In the results and analysis phase, the performance of the Random Forest algorithm in both exoplanet detection and prediction of habitability emerged as highly robust, achieving an impressive accuracy rate of approximately 95 percent. This notable accuracy underscores the efficacy of Random Forest in effectively classifying exoplanetary systems and assessing their potential habitability.

During exoplanet detection, the Random Forest model demonstrated remarkable precision in identifying true positive instances of exoplanets, minimizing false positive detections and thereby enhancing the reliability of the classification process. The algorithm's ability to effectively handle large datasets and complex feature spaces contributed to its robust performance, enabling the identification of subtle patterns indicative of exoplanetary transits or other detection signals.

Moreover, in the prediction of habitability, the Random Forest model exhibited a high degree of accuracy in assessing the suitability of exoplanets for life based on predefined habitability criteria. By leveraging a multitude of input features such as stellar parameters, planetary characteristics, and environmental factors, the algorithm provided reliable predictions of whether a given exoplanet falls within the habitable zone and thus has the potential to harbor life as we know it.

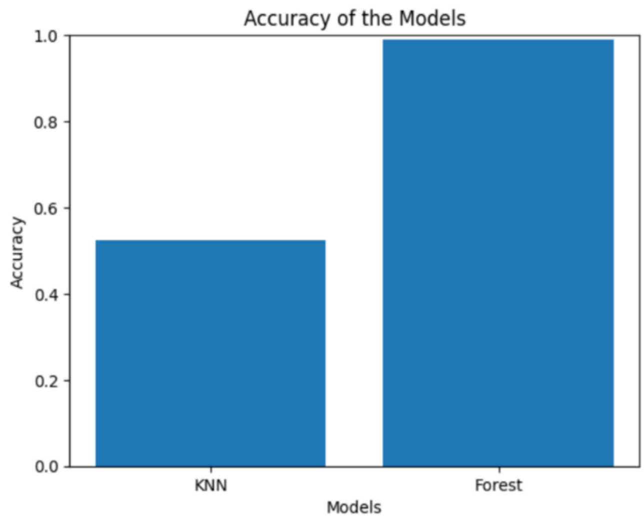

Fig 4: Performace of Models for Habitability Prediction

These results highlight the versatility and efficacy of the Random Forest algorithm in addressing diverse challenges within exoplanet research, from detection to habitability assessment. To further ensure the robustness and generalizability of the model's performance, k-fold cross-validation was implemented following the initial training and testing phases. By dividing the dataset into k subsets, or folds, and iteratively using one fold for validation while training on the remaining k-1 folds, this technique provided a comprehensive assessment of the model's accuracy, precision, recall, and F1-score across different subsets of data. This process is crucial for mitigating the risk of overfitting, where a model performs well on training data but fails to generalize to new, unseen data.

The importance of k-fold cross-validation lies in its ability to offer a more reliable estimate of the model's performance. In the context of exoplanet detection and habitability prediction, where datasets can be complex and varied, k-fold cross-validation ensures that the model's evaluation is not biased by a particular split of the data. This method provides a detailed understanding of how the model performs across different portions of the dataset, leading to more robust and trustworthy results. The insights gained from k-fold cross-validation guide further refinement of the model, ensuring it maintains high accuracy and generalizability in real-world applications, ultimately enhancing the reliability of exoplanetary research.

7.CONCLUSION:

In conclusion, our data-driven approach has provided valuable insights into exoplanet detection and habitability assessment. We've identified distinct detection patterns, established habitability indicators, and uncovered the diversity of exoplanetary systems. Our findings have implications for future missions, enhancing detection sensitivity, optimizing observational strategies, and streamlining validation processes. Moreover, our research contributes to a comprehensive understanding of alien worlds and their potential to support life. Looking ahead, addressing limitations in data quality and model uncertainty will be crucial, along with exploring complex interactions and integrating multi-wavelength observations. Ultimately, the study of exoplanets not only expands our cosmic perspective but also fuels humanity's quest for extraterrestrial life, inspiring future generations to explore the mysteries of the cosmos.

8.FUTURE WORK:

In the future, advancements in observational technologies, such as the JWST and the Roman Space Telescope, will enable detailed studies of exoplanet atmospheres and surfaces. This will aid in the search for biosignatures, crucial indicators of potential life expanding the dataset to include K2 and TESS data, which can significantly enhance the detection and analysis of exoplanets and their habitability. By integrating K2 data, with its improved photometric precision and extended field of view, we can refine our detection algorithms and identify a broader range of exoplanetary systems. Additionally, leveraging TESS data, known for its all-sky survey capabilities and focus on the brightest stars, allows us to gather more comprehensive and high-fidelity measurements of exoplanetary properties. This integration will not only increase the discovery rate of exoplanets but also provide deeper insights into their atmospheric compositions, orbital dynamics, and potential habitability, thereby advancing our understanding of exoplanetary science and contributing to the broader quest for finding life beyond Earth.

9.REFERENCES

1. Howell, S. B., Sobeck, C., Haas, M., Still, M., & Barclay, T. (2014). The K2 Mission: Characterization and Early Results. Publications of the Astronomical Society of the Pacific, 126(938), 398. K2 Mission
2. Ricker, G. R., Winn, J. N., Vanderspek, R., Latham, D. W., Bakos, G. Á., Bean, J. L., & Villasenor, J. (2015). Transiting Exoplanet Survey Satellite (TESS). Journal of Astronomical Telescopes, Instruments, and Systems, 1(1), 014003. [TESS]
3. National Aeronautics and Space Administration, “Kepler and K2”, https://www.nasa.gov/mission_pages/kepler/main/index.html
4. PHL HEC (Habitable Exoplanets Catalog), http://phl.upr.edu/projects/habitableexoplanets-catalog/data/database.
5. Saha, S., Bora, K., Agrawal, S., Routh, S., & Narasimhamurthy, A.(2015). “ASTROMLSKIT: A New Statistical Machine Learning Toolkit:A Platform for Data Analytics in Astronomy,” (No. arXiv: 1504.07865)

6. Hora, K. (2018). “Classifying Exoplanets as Potentially Habitable Using Machine Learning,” in ICT Based Innovations, pp. 203-212, Springer, Singapore.
7. Zhu Weijun, and Wang Xin. “Predicting the Habitability of Exoplanets based on GBRT Algorithm.”
8. Saha, S., Basak, S., Safonova, M., Bora, K., Agrawal, S., Sarkar, P., & Murthy, J. (2018). “Theoretical validation of potential habitability via analytical and boosted tree methods: An optimistic study on recently discovered exoplanets. Astronomy and Computing”, vol. 23, 141-150.
9. Agrawal, S., Basak, S., Saha, S., Rosario-Franco, M., Routh, S., Bora, K., & Theophilus, A. J. (2015), “A Comparative Study in Classification Methods of Exoplanets: Machine Learning Exploration via Mining and Automatic Labeling of the Habitability Catalog”.
10. Maxwell, A. E., Warner, T. A., & Fang, F. (2018), “Implementation of machine learning classification in remote sensing: An applied review”, International Journal of Remote Sensing, 39(9), 2784-2817.
11. Saito, T., & Rehmsmeier, M, “The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets”. PloS one, 10(3), 2015.