HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 3 (2024)

Cooperated Supervised and Semi-Supervised Machine Learning for Identification of Exoplanet Habitability

Haoxuan Xu[1] and Yong Zhang[#]

[1]Dulwich International High School Suzhou, China
[#]Advisor

ABSTRACT

The study of planetary habitability has gained widespread attention, and most existing studies focused on analyzing the habitability of a single planet based on a single feature, which makes it difficult to process a large amount of planetary data quickly. In this paper, we propose a machine learning-based identification method for efficiently distinguishing the habitability of a batch of planets. Firstly, a planet dataset comprising 5476 unlabeled records from the NASA Exoplanet Archive and 63 labeled entries with habitability from the Habitable Worlds Catalog is collected. Following that, a binary particle swarm optimization approach is used to select the most relevant features according to the 63 labeled planets. To address the missing values in the NASA data, next a standardized median imputation technique is applied. Two distinct methods, namely K-means clustering and distance-based filtering, are developed to label a subset of uninhabitable exoplanets by integrating the unlabeled 5476 and labeled 63 data points. Finally, KNN classifier and a semi-supervised label spreading classifier are trained and cooperated, contributing to the accomplishment of the final classification task. The experimental outcomes demonstrate the viability and effectiveness of the proposed method.

Introduction

Owing to population growth, resource exploitation, industrial pollution, and various other factors, the Earth's environment has grown increasingly inhospitable, and resources have become progressively scarce. Therefore, the imperative for humanity to identify habitable exoplanets for future migration, habitation, and reproduction has reached a critical juncture. Since the 1990s (Wolszczan & Frail, 1992; Wolszczan, 1994), the discovery of exoplanets and the exploration of inner planets and satellites through spacecraft have yielded vital criteria for assessing habitability. Extensive geophysical comparisons have been conducted, revealing a substantial corpus of information to facilitate the comprehensive investigation into exoplanetary habitability.

NASA defined the primary criteria for habitability as " extended regions of liquid water (Dyches & Chou, 2015), conditions favorable for the assembly of complex organic molecules, and energy sources to sustain metabolism" (NASA, 2015). When assessing the habitability potential of an object, emphasis is placed on its bulk composition, orbital properties, atmosphere, and potential chemical interactions. Key characteristics of stellar objects include mass, luminosity, stable variability, and high metal abundances. Astro biological studies primarily concentrate on rocky, most Earth-like planets and satellites with Earth-like chemical potential.

NASA maintains a comprehensive database about over 5,000 exoplanets and sponsors astrobiology strategy, which specifically focuses on investigating the conditions for the existence of life and the quest for extraterrestrial life. Managed by NASA's Caltech, the Jet Propulsion Laboratory (JPL) has made significant discoveries of water in numerous unexpected locations both within and beyond our solar system through various

exploratory missions. As highlighted earlier, water is a fundamental essence for the existence of life. The Habitable Worlds Catalog (HWC) presented a habitability criterion as exoplanets up to 2.5 Earth radii or 10 Earth masses orbiting within the optimistic stellar habitable zone to be habitable. According to these criteria, they identified up to 63 potential habitable worlds out of over 5000 NASA-recorded exoplanets. Of these, 28 conservative samples include exoplanets up to 1.6 Earth radii or 3 Earth masses, or those that are more likely rocky. The other 35 optimistic samples consist of larger planets and might include super-Earths, ocean worlds, or mini-Neptunes, and therefore less likely to be rocky or support surface liquid water with a lower likelihood of habitable conditions.

Besides these organizations, many university research teams are also actively investigating planet habitability from various perspectives. For example, The Laboratory for Planet Habitability of Harvard University aims to understand the habitability of Earth and other celestial bodies, including habitable conditions on planets inside and outside the solar system, focusing on atmospheric and geological processes, as well as the origin and evolution of life. Planetary Science Laboratory studies planet habitability from the perspective of planetary science, including the composition and structure of exoplanetary atmospheres and potential signs of life. Massachusetts Institute of Technology Habitability Research Center concentrates on exploring various aspects of planetary habitability, covering multiple dimensions from planetary atmospheres to surface conditions. Their research spans habitability factors within and beyond the solar system.

A number of individuals have also made outstanding contributions in this area. Schulze-Makuch D. et al. (2011) suggested a two-tiered classification scheme of exoplanet habitability, i.e., the Earth Similarity Index (ESI) and the Planetary Habitability Index (PHI). The ESI considers available data as mass, radius, and temperature, and the PHI focuses on the presence of a stable substrate, available energy, appropriate chemistry, and the potential for holding a liquid solvent. They pointed out that PHI can be used to minimize the biased search for life and indicates the possibility of having lives under extreme conditions. Kaltenegger et al. (2010) discussed methods to analyze the spectrum of a planet and search for the biosphere, focusing on planets with masses within 10 Earth masses (MEarth) known as super-Earths that might be habitable. The research will most likely to generate a scope of various planets that will set planet formation, evolution, and our planet into an overall context. Meadows (2018) addressed that Proxima Centauri b is extremely important for observing the evolution and nature of terrestrial planets orbiting M dwarfs. To show its habitability, a self-consistent atmosphere for several evolutionary scenarios, including high-O2, high-CO2, and more Earth-like atmospheres, with both oxic and anoxic compositions being generated by 1-D coupled climate-photochemical models. Seager (2010) suggested that Hot Jupiters are the type of exoplanet most amenable to study at the current stage. Key points include the detection of molecular spectral features; observation of day-night temperature gradients; and constraints on vertical atmospheric structure. Through detection and analysis of atmospheric biosignatures, the final goal is to look for lives on other planets mainly focusing on transiting super Earth and direct image of true Earth analogues in the habitable zone. Vannah (2024) applied information theory to a variety of simulated exoplanet transmission spectra to create a diagnostic tool to search for signatures of potential life on Earth-analogue planets. The algorithm was tested on three epochs of evolution for Earth-like planets orbiting a range of host stars. Chen (2023) uses current data from the NASA Exoplanet Archive to research planet distribution and presence in habitable zones around Red Giant stars. They were also updating the power law relation between planet mass and stellar radius found in previous studies and providing more investigations specified in this topic. In an optimistically calculated habitable zone there are ten Red Giant-hosted exoplanets, five of them are in a more conservatively calculated habitable zone.

Much of the existing research has focused on assessing the habitability of individual or a few specific exoplanets. Studies of Vannah (2024) and Chen (2023) offer the advantage of relatively high accuracy and yields more detailed and comprehensive results regarding the likelihood of habitability and the potential existence of extraterrestrial life. However, it comes with the drawback of requiring an extensive amount of data, which is often not readily available or detailed enough. The current data on exoplanets may fall short of meeting

these requirements, leaving us reliant on the hope for more precise and detailed observational data on exoplanets in the future.

Furthermore, the one-by-one analysis employed in these studies has notable drawbacks. The calculation cycle is lengthy, the workload is substantial, and efficiency is compromised. As NASA and other organizations provide relevant research data, using machine learning based on these data to overcome the limitations, offering a more efficient and extensive way of selecting habitable exoplanets.

Machine learning approaches have found applications in diverse fields such as large language models (Deng, Xia, Peng, Yang & Zhang, 2023), computer vision (Sebe, 2005), speech recognition (Deng & Li, 2013), email filtering (Dada & Bassi,2019), agriculture (Yoosefzadeh-Najafabadi & Earl, 2020), and medicine (Deo, 2015). This is particularly significant in situations where the development of traditional algorithms for required tasks would be prohibitively costly. While machine learning has proven to be immensely beneficial in numerous fields, few research were carried out in the field of astrophysics. Nasios (2024) applied machine learning methods including Recurrent Neural Network, Convolutional Neural Network and simple Neural Networks to detect the potential habitability of ancient Mars by analyzing two mass spectrometry techniques, evolved gas analysis (EGA-MS) and gas chromatography (GC-MS). Gebhard (2024) presented a latent variable model based on a neural network that learns a pressure-temperature (PT) profile distribution. Each profile is represented by a low-dimensional vector that can be used to condition a decoder network that maps P to T which would be able to help Atmospheric retrievals and gain a deeper understanding of exoplanet atmospheres and their habitability. The impact of machine learning on astrophysics, particularly in the analysis of planetary habitability and the search for habitable exoplanets, has been relatively limited.

Motivated by these, we here propose a machine learning-based method for planetary habitability identification. As aforementioned, the HWC has a database of 63 potentially habitable exoplanets that are categorized into conservative samples and optimistic samples. Combined with the NASA Exoplanet Archive, there should also be tremendous unhabitable exoplanets. Accordingly, the identification on planetary habitability is a classification task for machine learning. Machine learning algorithms can be trained starting from a labeled training dataset, and then used for inference on any data alike.

The idea of our research is to first collect basic planetary data from NASA and HWC, and preprocess the missing data; then perform feature selection to construct the labeled dataset; followed by designing both supervised and semi-supervised classifiers to achieve the habitability identification. Specifically, unlabeled 5476 exoplanet data are downloaded from the NASA Exoplanet Archive together with 63 labeled data for habitable exoplanets from HWC. The missing data are processed by using mode imputation. A particle swarm optimization (PSO) is used to select the most relevant and important features. Two methods, i.e., K-means clustering and distance-based filtering, are designed to label a small number of unhabitable exoplanets by combining the unlabeled 5476 and labeled 63 data. KNN classifier and a semi-supervised label spreading classifier are trained to perform the final classification task. Experimental results are sufficiently demonstrated to show the feasibility of the proposed method.

The main contributions are as follows: (1) Machine learning-based framework is applied to identify the habitability of numerous exoplanets. (2) Unhabitable exoplanets are labeled by calculating the distance between the labeled and unlabeled data sets. (3) Classification results obtained by supervised and semi-supervised are cross-validated to show the feasibility of the proposed method.

Data Collection and Preprocessing

We collect two types of exoplanet data, i.e., labeled with habitability and unlabeled ones. A dataset comprising over 5,500 exoplanets, each characterized by more than 300 features, is sourced from the NASA Exoplanet Archive (NASA 2015). The data from NASA is authoritative and comprehensive, however, it lacks labels for

exoplanet habitability definitions. Therefore, this dataset is chosen as the original unlabeled dataset for classification. Additionally, data on 63 labeled habitable exoplanets, including information on planet type, mass, radius, etc., is obtained from HWC. However, HWC does not provide specific criteria and boundaries for assessing habitability concerning each feature. Instead, it offers two categories of estimates—optimistic and conservative—based only on radius and mass. It is therefore difficult to determine the exact features and criteria for reliable classification of unlabeled exoplanets. But it can be used as the training and testing data for machine learning.

The following two figures demonstrate the snapshot of the data from the NASA Exoplanet Archive and HWC.

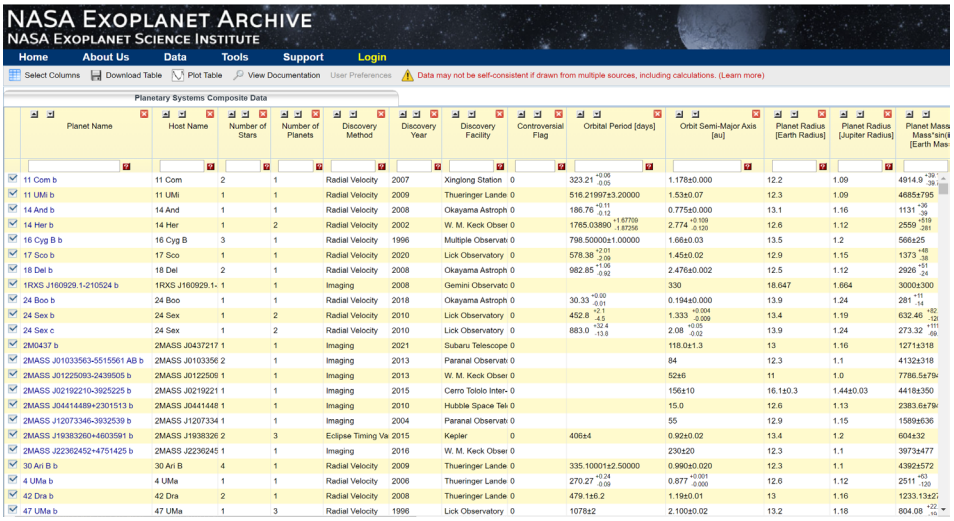

Figure 1. Unlabeled data from NASA Exoplanet Archive

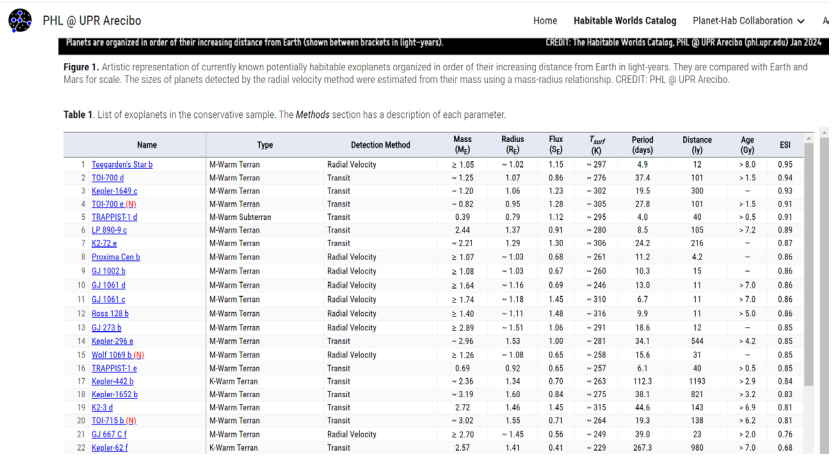

Figure 2. Labeled data from HWC

It is clear from Fig. 1 that there are a large number of null values in these datasets. Therefore, the standardized median filling is applied to the dataset and all non-numerical data was removed manually for successfully performing the following machine learning.

Framework

For easy understanding, the framework of our method is illustrated in Figure.3. It contains four main components, i.e., the binary bare bones PSO (BPSO) (Zhang, Gong, Hu & Zhang, 2015) based feature selection, K-means and labeled-unlabeled distance-based labeling, classification as supervised KNN and semi-supervised label spreading. The outputs of KNN and label spreading are compared for cross-validation and the final result is obtained. The details of these four components will be presented in the following sections.

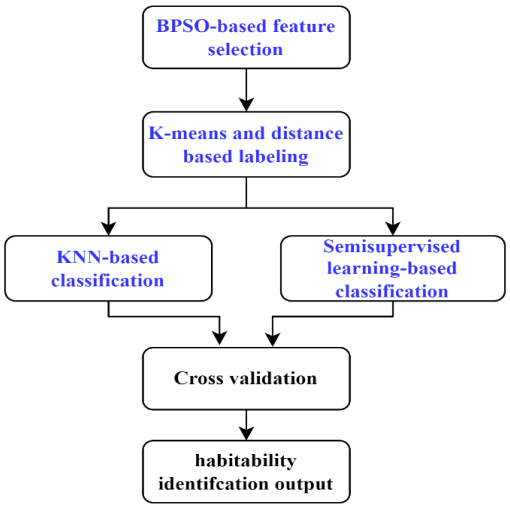

Figure 3. Framework

Methodology

BPSO-Based Feature Selection

For the purpose of direct habitability analysis and identification on the NASA data, it's necessary to find the most relevant features from the original table that affect the habitability of exoplanets. Therefore, we here first manually assign the labels of the 63 habitable planets to the same ones in NASA's data by matching their names. Subsequently, the most relevant features need to be selected from the original table, which consists of 315 features. To achieve this, the BPSO algorithm (Zhang et al., 2015) together with the KNN classifier is employed for feature selection since it has been successfully applied to many practical problems.

As it was addressed in Reference (Zhang et al., 2015), the BPSO algorithm involves two main steps. In Step 1, the swarm particles representing different feature selection modes are initially generated, and each particle sets its position as its personal best. Iterative steps are then executed to identify the optimal feature subset according to the classification precision until the maximum iteration limit (Tmax) is reached. During each iteration, particles update their personal best (Pbest) and global best (Gbest). Based on these two best values, each particle updates its feature mode for effectively exploring promising solutions. Detailed description please refer to the original research in (Zhang et al.).

The original features of NASA data are 315 dimensions with only 21 left after removing the textual ones and those with a large number of missing values. Feature encoding is imperative when applying the BPSO. Corresponding to the 21 features to be selected, 21 binary bits are randomly generated as a particle. The meaning

of the binary bits is shown in Figure.4. If the value of the *n*-th bit is 1, indicating that the *n*-th feature is selected, otherwise, the *n*-th feature is ignored in the following KNN-based classification task.

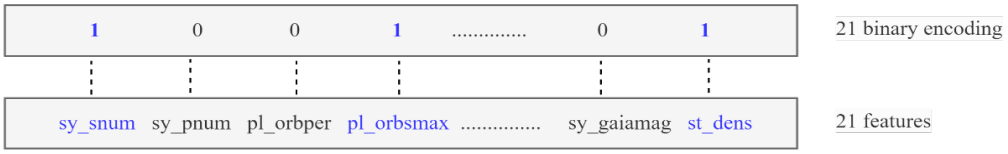

Figure 4. Feature Encoding

With each individual encoding, KNN with *K*=2 is conducted, and the classification error is calculated as the fitness function to evaluate the importance of the features selected. The smallest error means the individual is competitive and the corresponding selected features are more important. According to the fitness, the particles iteratively update by following the equations (Zhang et al., 2015):

Equation 1: Update principles of particles

$$x_{i,j}(t+1) = N\left(\frac{Pbest_{i,j}(t) + Gbest_{i,j}(t)}{2}, \left|Pbest_{i,j}(t) - Gbest_{i,j}(t)\right|\right)$$

Where $x_{i,j}(t+1)$ means the *j*-th value of the *i*-th particle at *t*+1 generation. *N*() indicates the Gaussian distribution with the mean $\frac{Pbest_{i,j}(t)+Gbest_{i,j}(t)}{2}$ and the variance $\left|Pbest_{i,j}(t) - Gbest_{i,j}(t)\right|$.

The most pertinent features obtained through this feature selection method encompass stellar density (st_dens), right ascension (RA), stellar surface gravity logarithm (st_logg), and stellar mass (st_mass).

Labeling and Classification

HWC has provided labels for 63 habitable planets, categorized into optimistic and conservative estimations. However, a substantial number of habitable planets in the NASA data remain unlabeled and are challenging to determine. To address this, the third category is compulsory for classifying these data. Since it is anticipated that the number of planets falling into the third category (uninhabitable exoplanets) will exceed the combined total of the other two, approximately 400 labels have been assigned to the third category for proper categorization.

Here two ways of determining the third-class labels are proposed, i.e., Using K-means clustering to find the class containing the minimum habitable planets, or calculating the distance and selecting the planets with the maximum values. K-means clustering is an unsupervised machine learning algorithm used for partitioning a dataset into distinct groups or clusters based on similarities among data points. The objective is to minimize the variance within each cluster. The algorithm works iteratively by assigning data points to the nearest cluster centroid and then updating the centroids based on the mean of the assigned points. This process continues until convergence, resulting in well-defined clusters. Here 5476 data from NASA, containing 63 labeled and the rest unlabeled, are clustered by using the four features filtered by the BPSO algorithm. NASA classifies exoplanets into four categories as Gas giant, Neptunian, super-Earth and terrestrial, so we cluster all planets here into four classes using the above features. Thus, the value of K, which here determines the number of clusters, is set to 4. The result is shown in Figure.5. Next, for each major category the labeled data is counted and the category containing the least labeled data is found as the uninhabitable set (with 664 planets).

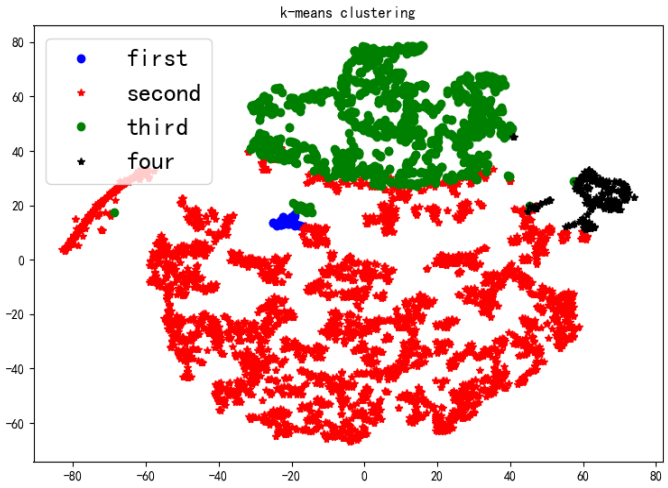

Figure 5. Demonstration of K-means based uninhabitable planets selection

Another way to find the labels of the third category is by calculating the Euclidean distance between each labeled and unlabeled data points, sorting these distances from the largest to the smallest, and then selecting the top *M* unlabeled data points as the third category of uninhabitable planets. For example, *M*=200 which ended up being 408 due to the large amount of data and the fact that some of the distances may correspond to more than one planet.

After obtaining the third class of labels, we propose two classification approaches, the first is to train a KNN classifier and the other is to use a semi-supervised label spreading algorithm. KNN is selected for its suitability in scenarios where the habitability decision boundaries are not distinctly defined, and the distribution of planet data is nonlinear. The algorithm operates on the principle that similar data points tend to share the same class. In this application, 70% of the habitable planets are employed as the training dataset, with the remaining 30% reserved for testing to evaluate the algorithm's accuracy. The selection of K, representing the number of neighbors, plays a pivotal role in the algorithm's performance. A small K may introduce sensitivity to noise, while a large K may result in over-smoothing. By carefully balancing the model's responsiveness to neighboring data points, the KNN algorithm aims to provide reliable predictions for habitability classifications.

The semi-supervised label spreading algorithm is used as the other method of classification. The algorithm also starts with 70% of the labeled data as training set and 30% as testing one. It first trains a KNN model to generate initial predictions. Subsequently, the algorithm introduces unlabeled data into the KNN model, incorporating them into the learning process. During this phase, the algorithm extracts valuable insights from the labeled instances, utilizing this information to either make predictions or assign pseudo-labels to the unlabeled instances, i.e., an operation commonly referred to as label propagation. To enhance its predictive capabilities, the model undergoes an iterative refinement process, continuously updating both the initially labeled and pseudo-labeled instances. This iterative approach is designed to enhance the model's comprehension of underlying patterns within the data. Following the iterative refinement process, the performance of the final model is evaluated by using a separate validation or test set, providing a comprehensive assessment. The flowchart of the supervised KNN together with the semi-supervised label spreading identification method is shown in Figure.6.

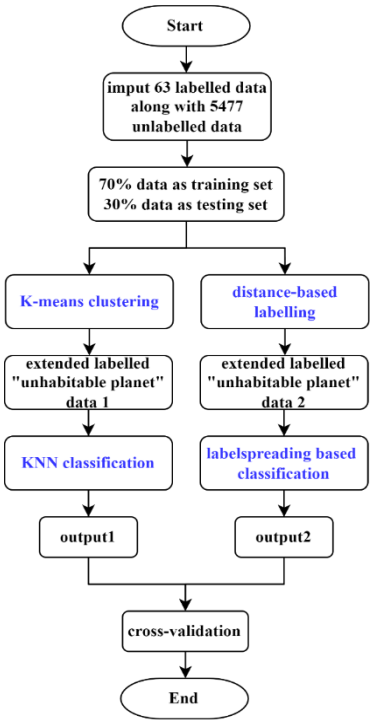

Figure 6. Flowchart of KNN and semi-supervised label spreading identification

Experiment

The algorithm consists of four main points: feature selection, data preprocessing, determination of the third category of labels, i.e., uninhabitable planet labels, and classification-based identification of planet habitability. We carry out the following six-group experiments to illustrate the performance of each component and to obtain a relatively accurate and realistic result.

Group 1: Effectiveness of feature selection: 63 labeled data are used to select key features by using BPSO, and KNN is applied to the classification. Result with higher accuracy means selected features are more competitive.

Group 2: Methods of data preprocessing: two methods as Mode+Normalization (MN) and Standardization+Median (SM) for null values filling will be compared based on the results of KNN and SVC.

Group 3: Different ways of determining the third label: one is to use K-means clustering to find the class with the least number of known habitable planets as the uninhabitable label, and the other is a distance-based method, which calculates the distance between each labeled 63 planetary data and those unlabeled 5481 ones, then use a certain number of the furthest points as the uninhabitable label.

Group 4: Key parameters setting: the important parameters as K for the KNN classifier and threshold of distance-based third label determination are experimentally determined by trial and error.

Group 5: Performance of semi-supervised and supervised classifications by comparing their results and accuracy and the ratio between training and testing data sets.

Group 6: Cross-validation of the results given by the two classification methods.

Results of Feature Selection

63 labeled exoplanet data with optimistic habitable as label "1" and conservative habitable as label "2" are used to perform the BPSO-based feature selection. KNN is used as the classifier and the accuracy is calculated to compare the result of BPSO and that of all features involved. When 21 features are all used for classification, the accuracy is 0.730. With BPSO, the accuracy increased to 0.857 with 4 features as density (st_dens), right ascension (RA), stellar surface gravity logarithm (st_logg), and stellar mass (st_mass). According to the selected features, further data preprocessing for the unlabeled 5476 exoplanets is carried out.

Methods of Data Preprocessing

As mentioned before, there are a lot of null values in the original data, which need to be filled in to facilitate the subsequent operation of the program and the experiment. Therefore, we propose two ways to fill the null values. One is to fill the null value using mode filling and then normalize the data (MN). The other is to first standardize the original data and then fill the null using median (SM). To experimentally show the effectiveness of these two methods, the classification results of KNN and SVC will be compared. In this experiment, the K-means algorithm with K being 4 is here used to simply determine the third label, i.e., uninhabitable planets. The classification indices as precision (P), recall (R), f1-score (f1) and Accuracy (A) calculated by the following Equations on testing data will be compared here. A higher value means better performance of the compared methods. The average values of the four indices are demonstrated in Figure.7, and the detailed results on each class (denoted as "1", "2", and "3") are listed in Table 1 and Table 2.

Equation 2: Classification precision (P):

$$P = \frac{True\ positive}{True\ positive + False\ positive}$$

Equation 3: Classification recall (R):

$$R = \frac{True\ positive}{True\ positive + False\ negative}$$

Equation 4: Classification f1-score (f1):

$$f1 = 2 \times \frac{R}{P + R}$$

Equation 5: Classification Accuracy (A):

$$A = \frac{\#\ of\ correct\ Predictions}{Total\ \#\ of\ predictions}$$

Where True positive means actual positive samples are correctly predicted as positive ones, and False positive is actually positive, but falsely predicted as negative ones. False negative just represents actual positive samples are falsely predicted as positive ones, and True negative means negative samples are correctly predicted as negative ones.

Figure 7 illustrates the average outcomes obtained through various data processing techniques. Upon analysis, it becomes apparent that the Precision value and Accuracy score are nearly indistinguishable for the two distinct methods. However, it is evident that the Recall value and f-1score of SM significantly surpass those of MN for both SVC and KNN classifications. Upon closer examination of the detailed results, it becomes

apparent that aside from the precision of identifying label 2, SM either equals or outperforms MN in all other aspects of performance. Consequently, we have opted for SM as our preferred method of data preprocessing.

Table 2 further presents the outcomes of two classification methods following two distinct data preprocessing approaches. However, the obtained results seem unrealistic, as the number of exoplanets belonging to label 2 (conservative habitable) is notably larger than that of label 3 (uninhabitable). In this case, we believe that it is because of the way of determining label 3. Therefore, we proposed another way of determining uninhabitable planets, namely distance-based labeling. The effectiveness of this labeling method will be explained in the following subsection.

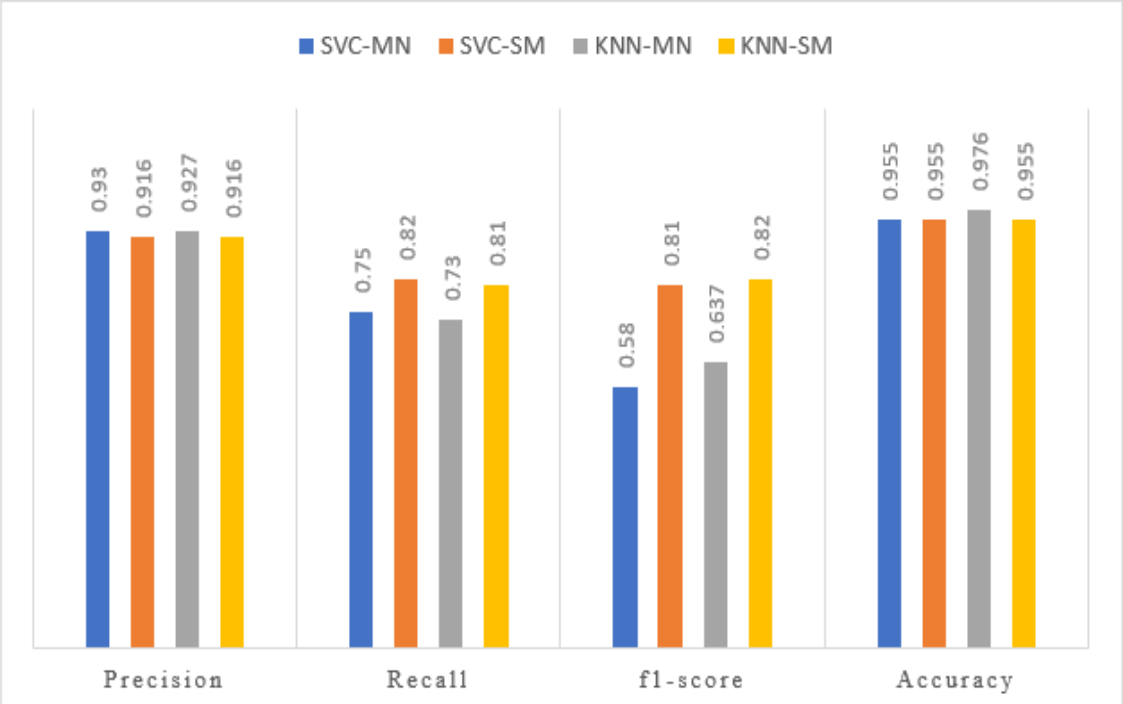

Figure 7. Average Results of Compared Data Preprocessing.

Table 1. Detailed results of data preprocessing

performance / Method of filling		precision			recall			f1-score		
		1	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
KNN	**MN**	1.00	**0.79**	0.99	0.25	0.94	1.00	0.40	0.86	0.99
	SM	1.00	0.75	**1.00**	**0.43**	**1.00**	1.00	**0.60**	0.86	**1.00**
SVC	**MN**	1.00	**0.83**	0.96	0.12	0.62	1.00	0.22	0.71	0.98
	SM	1.00	0.75	**1.00**	**0.43**	**1.00**	1.00	**0.60**	**0.86**	**1.00**

Table 2. Classification results

Number / Method of filling		1	2	3
KNN	MN	351	4230	895
	SM	67	4969	211
SVC	MN	11	3641	1824
	SM	51	4964	232

Different Ways of Determining the Uninhabitable Planets

Based on the results of Experiment 5.2, we here adopt the SM as the data preprocessing method due to its superior performance. Our current labeling method, which is K-means based labeling, has yielded results shown in Figure. 8. Through K-means clustering, both labeled and unlabeled data are grouped into four clusters, with label "3" being assigned to the cluster containing the least amount of labeled data. In this instance, utilizing SM as the data preprocessing method resulted in a total of 232 instances labeled as label "3".

To maintain consistency with the K-means based method, the distance-based method has been employed to select 40 furthest distances from the 63 labeled data points, totaling 232 in all, unlabeled data points as label "3", which mirrors the approach taken by the K-means method.

The KNN with K being 5 is also used here to compare the performance of these two methods of selecting uninhabitable planets, and the values of Precision, Recall, and f1-score are recorded in Table 3.

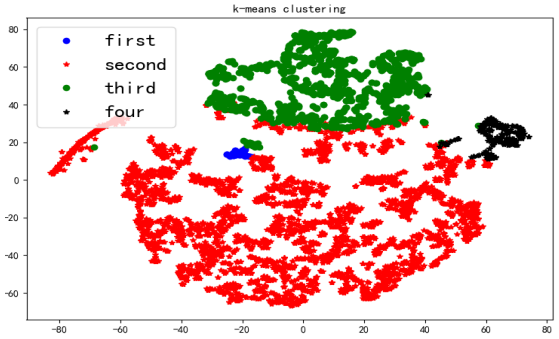

Figure 8. Results of K-means clustering

Table 3. Results of different methods for labeling uninhabitable planets

perfor-mance / Method of labeling	precision			recall			f1-score		
	1	2	3	1	2	3	1	2	3
K-means-based	1.00	0.75	1.00	0.43	1.00	1.00	0.60	0.86	1.00
Distance-based	0.80	0.56	0.89	**0.57**	0.42	0.96	**0.67**	0.48	0.92

From Table 3, it is evident that although the K-means based method is superior to the distance-based method overall, the Recall and f1-score are higher for the distance-based method for label "1". This suggests

that the distance-based method may offer better overall performance for label "1" classification, potentially enabling the identification of more habitable exoplanets. The classification results obtained by KNN indicate 80 optimistic habitable, 408 conservative habitable, and 4758 uninhabitable planets. Despite these scores, the results obtained through this approach are more realistic compared to the previous one.

However, to enhance the overall accuracy, precision, recall, and f1-score, a more reliable and convincing classification method is required. Therefore, label spreading semi-supervised classification has been cooperated, as it demonstrates better performance compared to KNN in this case, particularly with a small amount of labeled data and a large amount of unlabeled data.

Key Parameters Setting

After SM data preprocessing and distance-based labeling, the important parameters such as *K* for the KNN classifier and threshold of distance-based third label determination are experimentally searched to possibly obtain the optimal values.

First, the value of *K* for the KNN classifier is selected from 1 to 21 since odd number is often preferred for this algorithm to avoid Tie-ballot problems in classification. The precision, recall, f1-score, accuracy, and number of each category (labeled as "1", "2", and "3") are given in Table 4. For intuitive comparison, the values of P, R, and f1 for label "1" and label "2" along with A are shown in Figure. 9. The final classification results for only label "1" and label "2" are illustrated by Figure. 10. We here only consider label "1" and label "2" because the focus of the project is to look for habitable planets.

Table 4. Different K values and classification results

	1				**2**				**3**				
K	**P**	**R**	**f1**	**N**	**P**	**R**	**f1**	**N**	**P**	**R**	**f1**	**N**	**A**
1	0.67	0.57	0.62	175	0.40	0.33	0.36	666	0.89	0.93	0.91	4405	0.818
3	0.50	0.57	0.53	252	0.44	0.33	0.38	479	0.90	0.93	0.91	4515	0.818
5	0.80	**0.57**	**0.67**	80	**0.56**	**0.42**	**0.48**	408	**0.89**	0.96	0.92	4758	**0.852**
7	**1.00**	0.43	0.60	36	0.50	0.33	0.40	356	0.87	**0.97**	0.92	4854	0.841
9	1.00	0.43	0.60	43	0.57	0.33	0.42	326	0.87	0.99	0.93	4877	0.852
11	1.00	0.43	0.60	36	0.75	0.25	0.38	236	0.85	1.00	0.92	4974	0.852
13	1.00	0.43	0.60	34	0.67	0.17	0.27	127	0.84	1.00	0.91	5085	0.841
15	1.00	0.43	0.60	27	1.00	0.08	0.15	53	0.82	1.00	0.90	5166	0.830
17	1.00	0.43	0.60	29	0.00	0.00	0.00	28	0.81	1.00	0.90	5189	0.818
19	1.00	0.43	0.60	30	0.00	0.00	0.00	8	0.81	1.00	0.90	5280	0.818
21	1.00	0.43	0.60	26	0.00	0.00	0.00	0	0.81	1.00	0.90	5220	0.818

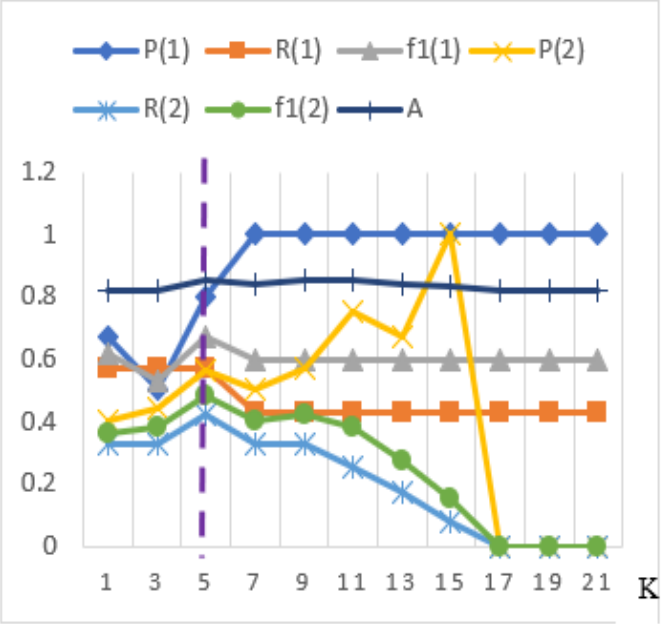

Figure 9. Trend of P, R, f1, and A for label "1" and "2"

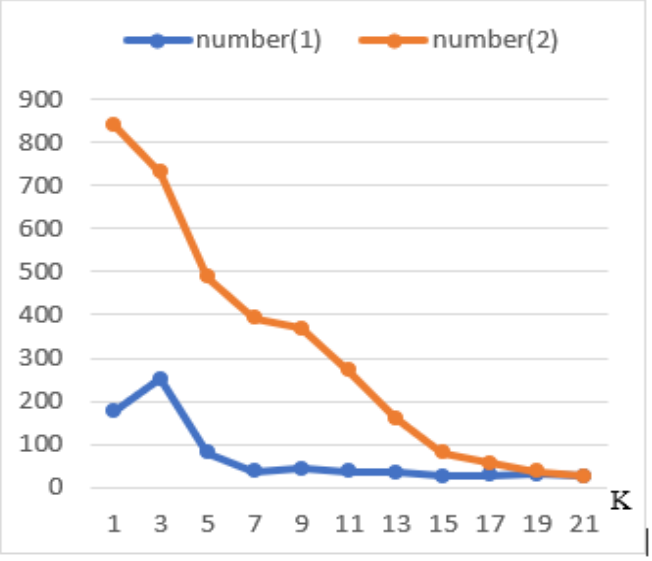

Figure 10. Variations on the number of label "1" and "2".

Based on Figure. 9, it is evident that several optimal values are observed when the *K* is set to 5. Although the P-value of label "1" shows improvement when the *K* value is increased to 7, the majority of the values do not exhibit the same level of superiority as those obtained when k is set to 5. Hence, after thorough consideration, we ultimately opted for *K*=5 to ensure the validity and accuracy of the KNN model.

Then values of the threshold for selecting uninhabitable planets with distance-based method are further explored. The classification accuracy of KNN with *K* being 5 is plotted in Figure.11 when the threshold varied from 40 to 250 with 10 spans.

From the figure, it is evident that both the P-values of label "1" and label "2" exhibit continuous fluctuations within the range of values from 40 to 210. They both reach their peak values at 210 and decrease rapidly as the range increases. Similarly, the accuracy shows minimal variation but also peaks at 210 before experiencing a slight decrease. Consequently, we decide to select a distance threshold of 210. At this threshold, there are 571 exoplanets categorized as the third type of labeled samples. The overall accuracy of classification is 0.926 under this condition.

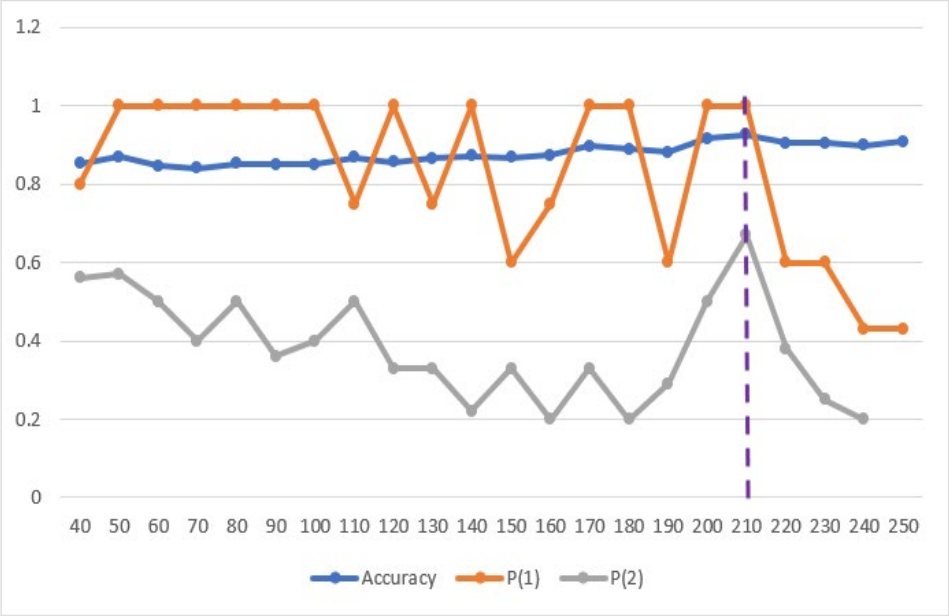

Figure 11. Accuracy and *P* varied along with the threshold of distance

Performance of Semi-Supervised Label Spreading Classification

On the basis of former experiments, given the constraint of having 571 labels of type "3", we conduct an experiment similar to the previous study to determine the optimal value of *K* in the KNN kernel function. Following identical procedures, we observed that all indicators remained consistent when the value of *K* ranged between 19 and 31. To reduce computational complexity while ensuring the authenticity of classification, we here set the midpoint value of 25 as the optimal *K* value.

Figure.12 shows the A and P of label "1" and label "2" when the training set separation varied from 0.3 to 0.9. From the figure, it is clear that as the size of the training set gradually increases, the overall accuracy as well as the P and A metrics exhibit a gradual upward trend. This trend continues until reaching a peak around 0.7. Subsequently, there is a gradual decline in performance, and by the time it reaches 0.9, the label spreading indexes attain their highest values.

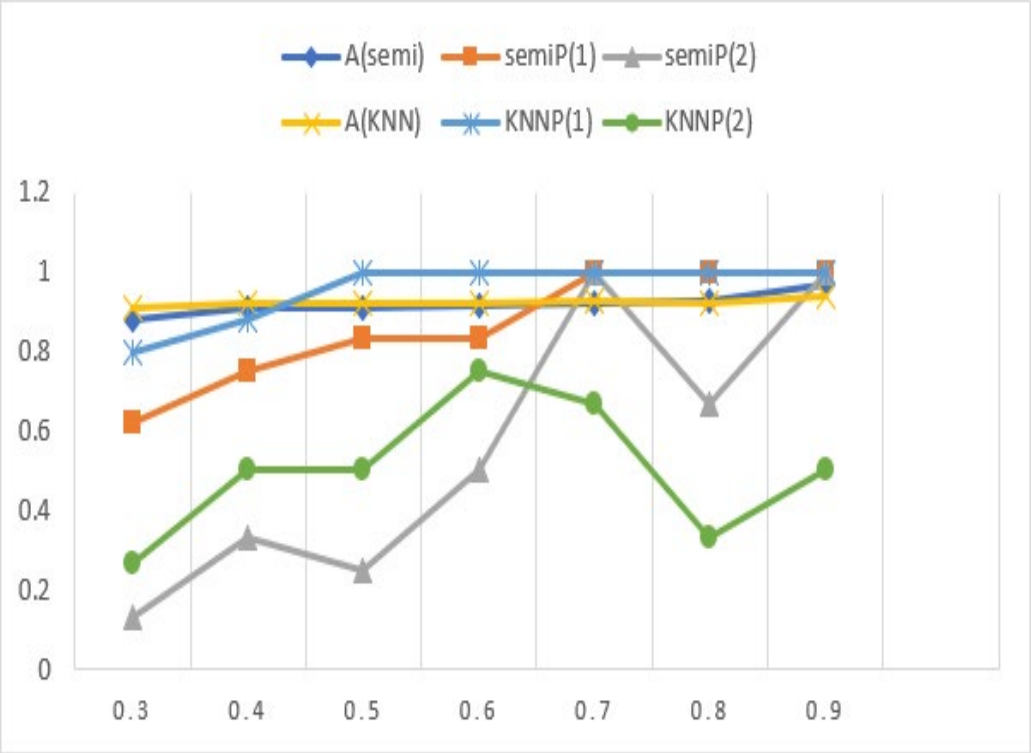

Figure 12. Trend of A, P for label "1" and label "2" for label spreading and KNN

Final Classification Results with Cross-Validation

To enhance the accuracy of our results, particularly for label "1", which represents a conservative estimate of inhabitable planets, we conduct a cross-check of the labels obtained from two distinct classification methods: KNN and label spreading. Two scenarios with 90% and 70% testing data are considered for KNN and label spreading respectively, and the results together with the name of the obtained inhabitable planets are shown in the following figure. The "red" name means the same planets identified under these four scenarios, and the "blue" ones are those only achieved under 70% testing condition. * and ** respectively represent the different inhabitable planets identified by semi-supervised label spreading and KNN under 90% and 70% testing.

It can be seen that 11 planets are identified as inhabitable ones among 5476 NASA data under these four scenario, 18 ones are obtained when the testing samples are set as 70%. When the testing ratio is 90%, 34 and 47 planets are recognized as inhabitable exoplanets by label spreading and KNN with 11 identical ones, i.e., 32.35% for semi-supervised label spreading and 23.40% for KNN. Similarly, when the testing ratio is 70%, the results of semi-supervised label spreading and KNN are 39 and 54 with 18 identical inhabitable planets. The successful ratio for label spreading and KNN are 46.15% and 33.33%. Furthermore, comparing the results obtained by label spreading under 90% and 70% conditions, the ratio of getting the same planets is 31/(31+3+8)=73.81%, and that of the KNN is 37/(37+10+16)=58.73%. Clearly, the label spreading can get more reliable results of inhabitable planets identification.

	semi		KNN		semi		KNN		Name of Identical inhabitable planets under four scenario	Name of inhabitable planets only for SS&KNN under 70% testing
1	90% (34)		90% (47)		70% (39)		70% (54)			
2										
3	GJ 1148 c		GJ 3634 b		GJ 1148 c		COCONUTS-2 b	**	GJ 3634 b	K2-319 b
4	GJ 3634 b		GJ 367 b		GJ 3634 b		GJ 1148 c	**	GJ 367 b	Kepler-304 b
5	GJ 367 b		K2-16 c	**	GJ 367 b		GJ 3634 b		Kepler-1257 b	Kepler-305 b
6	HIP 91258 b	*	Kepler-1018 b		K2-319 b	*	GJ 367 b		Kepler-1532 b	Kepler-1881 b
7	Kepler-1128 b		Kepler-1028 b	**	Kepler-1128 b		K2-319 b	**	Kepler-1907 b	Kepler-2001 b
8	Kepler-1139 b		Kepler-1108 b		Kepler-1139 b		K2-43 b	**	Kepler-193 b	Kepler-282 b
9	Kepler-1257 b		Kepler-1162 b		Kepler-1257 b		Kepler-1018 b		Kepler-1946 b	GJ1148C
10	Kepler-1332 b		Kepler-118 b		Kepler-1332 b		Kepler-1108 b		Kepler-261 b	
11	Kepler-1532 b		Kepler-1197 b	**	Kepler-1468 b	*	Kepler-1162 b		Kepler-398 b	
12	Kepler-1710 b	*	Kepler-1257 b		Kepler-1532 b		Kepler-118 b		Kepler-667 b	
13	Kepler-1737 b	*	Kepler-1322 c		Kepler-1740 b	*	Kepler-1257 b		Kepler-960 b	
14	Kepler-1881 b		Kepler-1355 b		Kepler-1881 b		Kepler-1322 c			
15	Kepler-19 b		Kepler-1513 b		Kepler-19 b		Kepler-1355 b			
16	Kepler-1907 b		Kepler-153 b	**	Kepler-1907 b		Kepler-1458 b	**		
17	Kepler-193 b		Kepler-1532 b		Kepler-193 b		Kepler-1513 b			
18	Kepler-1946 b		Kepler-1612 b		Kepler-1946 b		Kepler-1532 b			
19	Kepler-1954 b		Kepler-1642 b	**	Kepler-1954 b		Kepler-1612 b			
20	Kepler-2001 b		Kepler-1723 b		Kepler-1960 b	*	Kepler-1710 b	**		
21	Kepler-254 b		Kepler-1860 b	**	Kepler-2001 b		Kepler-1723 b			
22	Kepler-260 b		Kepler-1887 b		Kepler-254 b		Kepler-1740 b	**		
23	Kepler-260 c		Kepler-1907 b		Kepler-260 b		Kepler-1881 b	**		
24	Kepler-261 b		Kepler-192 b		Kepler-260 c		Kepler-1887 b			
25	Kepler-269 b		Kepler-193 b		Kepler-261 b		Kepler-1907 b			
26	Kepler-269 c		Kepler-1946 b		Kepler-269 b		Kepler-192 b			
27	Kepler-282 b		Kepler-222 b	**	Kepler-269 c		Kepler-193 b			
28	Kepler-304 b		Kepler-261 b		Kepler-282 b		Kepler-1946 b			
29	Kepler-310 b		Kepler-261 c		Kepler-304 b		Kepler-2001 b	**		
30	Kepler-398 b		Kepler-297 c		Kepler-305 b	*	Kepler-261 b			
31	Kepler-542 b		Kepler-346 b		Kepler-310 b		Kepler-261 c			
32	Kepler-553 b		Kepler-375 b		Kepler-398 b		Kepler-282 b	**		
33	Kepler-570 b		Kepler-398 b		Kepler-542 b		Kepler-297 b	**		
34	Kepler-667 b		Kepler-48 c		Kepler-553 b		Kepler-297 c			
35	Kepler-75 b		Kepler-63 b		Kepler-570 b		Kepler-304 b	**		
36	Kepler-960 b		Kepler-667 b		Kepler-667 b		Kepler-305 b	**		
37			Kepler-671 b		Kepler-75 b		Kepler-346 b			
38			Kepler-706 b		Kepler-77 b	*	Kepler-375 b			
39			Kepler-728 b		Kepler-960 b		Kepler-398 b			
40			Kepler-751 b		KOI-351 b	*	Kepler-48 c			
41			Kepler-770 b		SR 12 AB c	*	Kepler-63 b			
42			Kepler-939 b				Kepler-667 b			
43			Kepler-949 b				Kepler-671 b			
44			Kepler-960 b				Kepler-688 b	**		
45			Kepler-97 c				Kepler-706 b			
46			Kepler-977 b				Kepler-728 b			
47			Kepler-985 b	**			Kepler-751 b			
48			KOI-351 b	**			Kepler-770 b			
49			LSPM J2116+0234 b	**			Kepler-939 b			
50							Kepler-949 b			
51							Kepler-960 b			
52							Kepler-97 c			
53							Kepler-977 b			
54							LP 791-18 b	**		
55							LTT 1445 A b	**		

Figure. 13 Identified inhabitable exoplanets with label spreading and KNN

Conclusion

Planetary habitability identification stands as a prominent area within contemporary astrophysical research, often characterized by its time-intensive nature. In this study, leveraging 63 labeled samples from the Habitable Exoplanet Catalog (HWC), we proposed a rapid habitability classification and identification approach employing K-Nearest Neighbors (KNN) and semi-supervised classification on NASA's current dataset comprising 5476 unlabeled data points. Initially, a Binary Particle Swarm Optimization (BPSO) strategy was utilized to extract four key features, namely density (st_dens), right ascension (RA), stellar surface gravity logarithm (st_logg), and stellar mass (st_mass), from a pool of 21 planetary features. Subsequently, a standardized median filling strategy was employed to process missing data, facilitating data preprocessing. Furthermore, an uninhabitable planet selection strategy, leveraging K-means clustering and distance measures, was devised to categorize samples into habitable, more habitable, and uninhabitable sets. Utilizing this sample set, KNN and label spreading semi-supervised classifiers were trained, and the classification results were cross validated to ascertain the final livable planet outcomes. Through extensive comparative experiments, core algorithm parameters were refined, affirming the effectiveness of the proposed feature selection, numerical preprocessing, distance-

based sample set construction, as well as supervised and semi-supervised classification methods. Ultimately, nearly 20 habitable planets were identified through cross-validation, underscoring the efficacy of the proposed methodology.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

Chen, R. E., Jiang, J. H., Rosen, P. E., Fahy, K. A., & Chen, Y. (2023). Exoplanets around Red Giants: Distribution and Habitability. *Galaxies*, 11, 112. https://doi.org/10.3390/galaxies11060112

Dada, E. G., Bassi, J. S., Chiroma, H., Abdulhamid, S. i. M., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon*, 5(6), e01802. https://doi.org/https://doi.org/10.1016/j.heliyon.2019.e01802

Deng, L., & Li, X. (2013). Machine learning paradigms for speech recognition: An overview. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(5), 1060-1089.

Deng, Y., Xia, C. S., Peng, H., Yang, C., & Zhang, L. (2023). *Large Language Models Are Zero-Shot Fuzzers: Fuzzing Deep-Learning Libraries via Large Language* Models Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis, Seattle, WA, USA. https://doi.org/10.1145/3597926.3598067

Deo, R. C. (2015). Machine Learning in Medicine. *Circulation*, 132(20), 1920-1930. https://doi.org/doi:10.1161/CIRCULATIONAHA.115.001593

Dyches, Preston; Chou, Felcia (7 April 2015). "The Solar System and Beyond is Awash in Water". *NASA*. Retrieved 8 April 2015.

Gebhard, T. D., Angerhausen, D., Konrad, B. S., Alei, E., Quanz, S. P., & Schölkopf, B. (2024). Parameterizing pressure–temperature profiles of exoplanet atmospheres with neural networks. *Astronomy & Astrophysics*, 681, A3.

Kaltenegger, L., Selsis, F., Fridlund, M., Lammer, H., Beichman, C., Danchi, W., Eiroa, C., Henning, T., Herbst, T., Léger, A., Liseau, R., Lunine, J., Paresce, F., Penny, A., Quirrenbach, A., Röttgering, H., Schneider, J., Stam, D., Tinetti, G., & White, G. J. (2010). Deciphering Spectral Fingerprints of Habitable Exoplanets. *Astrobiology*, 10(1), 89-102. https://doi.org/10.1089/ast.2009.0381

Meadows, V. S., Arney, G. N., Schwieterman, E. W., Lustig-Yaeger, J., Lincowski, A. P., Robinson, T., Domagal-Goldman, S. D., Deitrick, R., Barnes, R. K., Fleming, D. P., Luger, R., Driscoll, P. E., Quinn, T. R., & Crisp, D. (2018). The Habitability of Proxima Centauri b: Environmental States and Observational Discriminants. *Astrobiology*, 18(2), 133-189. https://doi.org/10.1089/ast.2016.1589

NASA Astrobiology Strategy 2015, https://astrobiology.nasa.gov/about/astrobiology-strategy/

Nasios, I. (2024). Analyze mass spectrometry data with artificial intelligence to assist the understanding of past habitability of Mars and provide insights for future missions. *Icarus*, 408, 115824.

Schulze-Makuch, D., Méndez, A., Fairén, A. G., von Paris, P., Turse, C., Boyer, G., Davila, A. F., António, M. R. d. S., Catling, D., & Irwin, L. N. (2011). A Two-Tiered Approach to Assessing the Habitability of Exoplanets. *Astrobiology*, *11*(10), 1041-1052. https://doi.org/10.1089/ast.2010.0592
Seager, S., & Deming, D. (2010). Exoplanet Atmospheres. *Annual Review of* Astronomy *and Astrophysics*, 48(1), 631-672. https://doi.org/10.1146/annurev-astro-081309-130837

Sebe, N. (2005). Machine *learning in computer vision (Vol. 29)*. Springer Science & Business Media.

Vannah, S., Gleiser, M., & Kaltenegger, L. (2024). An information theory approach to identifying signs of life on transiting planets. Monthly *Notices of the Royal Astronomical Society*, 528, L4-L9. https://doi.org/10.1093/mnrasl/slad156

Wolszczan, A., & Frail, D. A. (1992). A planetary system around the millisecond pulsar PSR1257 + 12. *Nature*, 355(6356), 145-147. https://doi.org/10.1038/355145a0

Wolszczan, A. (1994). Confirmation of Earth-Mass Planets Orbiting the Millisecond Pulsar PSR B1257 + 12. *Science*, 264(5158), 538-542. https://doi.org/doi:10.1126/science.264.5158.538

Yoosefzadeh-Najafabadi, M., Earl, H. J., Tulpan, D., Sulik, J., & Eskandari, M. (2020). Application of Machine Learning Algorithms in Plant Breeding: Predicting Yield From Hyperspectral Reflectance in Soybean. *Front Plant Science*, 11, 624273. https://doi.org/10.3389/fpls.2020.624273
Zhang, Y., Gong, D., Hu, Y., & Zhang, W. (2015). Feature selection algorithm based on bare bones particle swarm optimization. *Neurocomputing*, 148(1), 150-157. https://doi.org/10.1016/j.neucom.2012.09.049