Astronomy and Computing 23 (2018) 141–150

Contents lists available at ScienceDirect

Astronomy and Computing

journal homepage: www.elsevier.com/locate/ascom

Full length article

Theoretical validation of potential habitability via analytical and boosted tree methods: An optimistic study on recently discovered exoplanets

S. Saha [a], S. Basak [a], M. Safonova [c], K. Bora [b,*], S. Agrawal [a], P. Sarkar [a], J. Murthy [d]

[a] *Department of Computer Science and Engineering, PESIT-BSC, Bangalore, India*
[b] *Department of Information Science and Engineering, PESIT-BSC, Bangalore, India*
[c] *M. P. Birla Institute of Fundamental Research, Bangalore, India*
[d] *Indian Institute of Astrophysics, Bangalore, India*

ARTICLE INFO

Article history:
Received 18 September 2017
Accepted 15 March 2018
Available online 26 April 2018

ABSTRACT

Seven Earth-sized planets, known as the TRAPPIST-1 system, was discovered with great fanfare in the last week of February 2017. Three of these planets are in the habitable zone of their star, making them potentially habitable planets (PHPs) a mere 40 light years away. The discovery of the closest potentially habitable planet to us just a year before — Proxima b and a realization that Earth-type planets in circumstellar habitable zones are a common occurrence provides the impetus to the existing pursuit for life outside the Solar System. The search for life has two goals essentially: looking for planets with Earth-like conditions (Earth similarity) and looking for the possibility of life in some form (habitability). An index was recently developed, the Cobb–Douglas Habitability Score (CDHS), based on Cobb–Douglas habitability production function (CD-HPF), which computes the habitability score by using measured and estimated planetary parameters. As an initial set, radius, density, escape velocity and surface temperature of a planet were used. The proposed metric, with exponents accounting for metric elasticity, is endowed with analytical properties that ensure global optima and can be scaled to accommodate a finite number of input parameters. We show here that the model is elastic, and the conditions on elasticity to ensure global maxima can scale as the number of predictor parameters increase. K-NN (K-Nearest Neighbor) classification algorithm, embellished with probabilistic herding and thresholding restriction, utilizes CDHS scores and labels exoplanets into appropriate classes via feature-learning methods yielding granular clusters of habitability. The algorithm works on top of a decision-theoretical model using the power of convex optimization and machine learning. The goal is to characterize the recently discovered exoplanets into an "Earth League" and several other classes based on their CDHS values. A second approach, based on a novel feature-learning and tree-building method classifies the same planets without computing the CDHS of the planets and produces a similar outcome. For this, we use XGBoosted trees. The convergence of the outcome of the two different approaches indicates the strength of the proposed solution scheme and the likelihood of the potential habitability of the recently announced discoveries.

1. Introduction

With discoveries of exoplanets pouring in hundreds, it is becoming necessary to develop some sort of a quick screening tool – a ranking scale – for evaluating habitability perspectives for the follow-up targets. We have proposed a novel inductive approach, inspired by the Cobb–Douglas model from production economics, to verify theoretical conditions of global optima of the functional form to model and to compute the habitability score of exoplanets — the Cobb–Douglas Habitability Score (CDHS; Bora et al., 2016). The discovery of an exoplanet, Proxima b (Anglada-Escudé, 2016), orbiting the nearest star (Proxima Centauri), generated a lot of stir in the news (Witze, 2016) because it is located in the habitable zone and its mass is in the Earth's mass range: 1.27–3 $M_\oplus$, making it a potentially habitable planet (PHP) and an immediate destination for the Breakthrough Starshot initiative (Breakthrough Starshot, 2016). A few months after the announcement of Proxima b, another family of terrestrial-size exoplanets – the TRAPPIST-1 system – was discovered (Gillon et al., 2017).

* Corresponding author.
E-mail addresses: snehanshusaha@pes.edu (S. Saha), k_bora@pes.edu (K. Bora).

https://doi.org/10.1016/j.ascom.2018.03.003
2213-1337/© 2018 Elsevier B.V. All rights reserved.

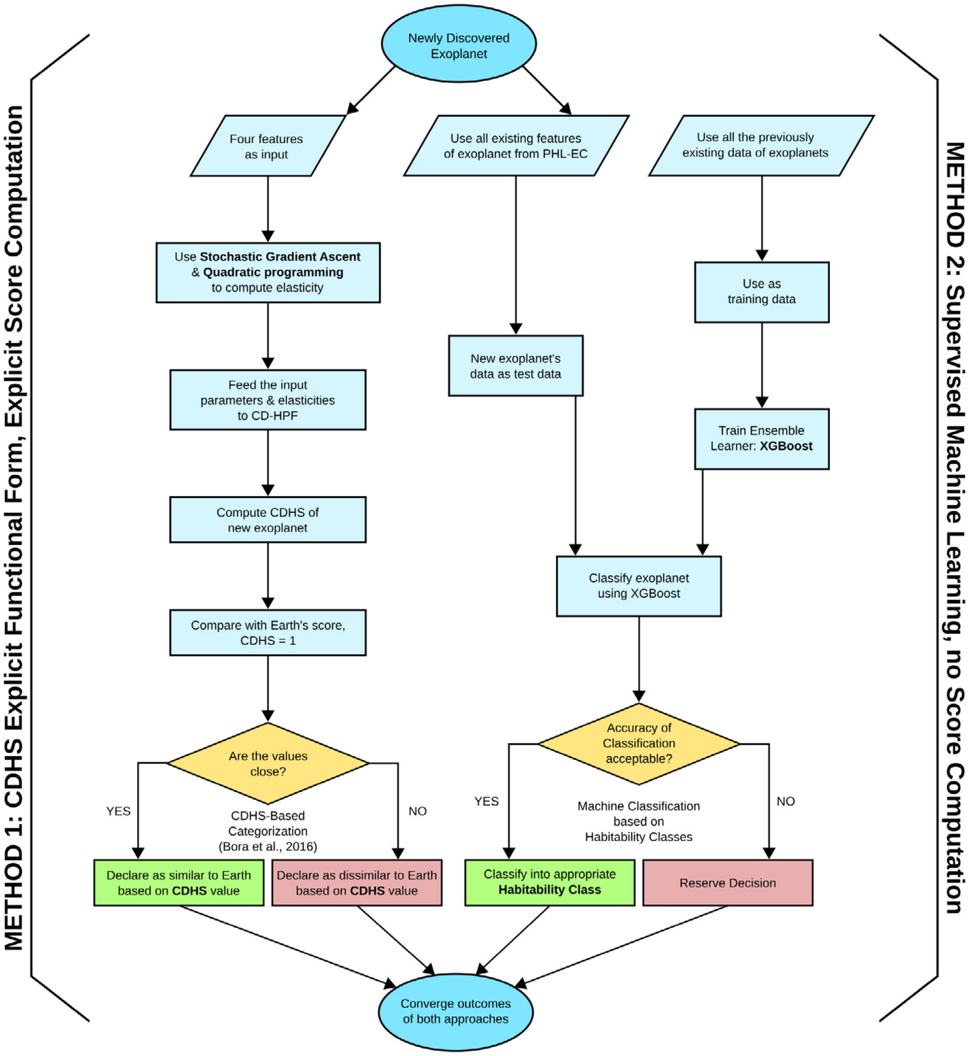
Newly Discovered Exoplanet
Four features as input
Use all existing features of exoplanet from PHL-EC
Use all the previously existing data of exoplanets
Use Stochastic Gradient Ascent & Quadratic programming to compute elasticity
Use as training data
New exoplanet's data as test data
Feed the input parameters & elasticities to CD-HPF
Train Ensemble Learner: XGBoost
Compute CDHS of new exoplanet
Classify exoplanet using XGBoost
Compare with Earth's score, CDHS = 1
Accuracy of Classification acceptable?
Are the values close?
YES
NO
CDHS-Based Categorization (Bora et al., 2016)
YES
Machine Classification based on Habitability Classes
NO
Declare as similar to Earth based on CDHS value
Declare as dissimilar to Earth based on CDHS value
Classify into appropriate Habitability Class
Reserve Decision
Converge outcomes of both approaches
METHOD 1: CDHS Explicit Functional Form, Explicit Score Computation
METHOD 2: Supervised Machine Learning, no Score Computation

Fig. 1. The convergence of two different approaches in the investigation of potential habitability. The outcome of the explicit scoring scheme i.e. Method 1 (based on CDHS) placed Proxima b in the "Earth League" (Earth-Similarity classification of exoplanets) and is empirically synonymous to being classified into a potentially habitable class of exoplanets via Method 2 (Not based on Earth-Similarity classification, rather supervised feature based machine classification, independent of CDHS). The convergence in outcome is remarkable but not accidental.

This work is motivated by testing the efficacy of the suggested model, CDHS, in determining the habitability score, the proximity to the "Earth-League", of the recently discovered exoplanets. Therefore, it was natural to test whether the model can also classify Proxima b and the planets in the TRAPPIST-1 system as potentially habitable by computing their habitability scores. This could indicate whether the model may be extended for a quick check of the potential habitability of newly discovered exoplanets in general. As we see from the results of our work, this is indeed the case with the TRAPPIST-1 planets.

The flowchart in Fig. 1 summarizes our new approach to the habitability investigation of exoplanets (on the example of Proxima b and TRAPPIST-1 system). This approach is based on the combination of two methods. The outcome of classification of exoplanets based on the CDHS (Method 1) is tallied with another machine classification method which discriminates samples (exoplanets) into classes based on the features/attributes of the samples (Method 2). The similar outcome from both approaches (the exoplanets being characterized for their potential habitability), is markedly different in structure and methodology, fortifies the growing advocacy of using machine learning in astronomy.

Our habitability score model considers four parameters/features, namely mass, radius, density and surface temperature of a planet extracted from the PHL-EC (Exoplanet Catalog hosted by the Planetary Habitability Laboratory (PHL)).[1] Though the catalog contains 68 observed and derived stellar and planetary parameters, we have considered only four for the CDHS model as a lot of parameters are redundant. We show that the CDHS model is scalable, i.e. capable of accommodating more parameters (see Section 4 on model scalability and Section 3 in supplementary file). Therefore, we may use a greater number of parameters in the future to compute the CDHS.

PHL classifies exoplanets into five categories based on their thermal characteristics: non-habitable, and potentially habitable: psychroplanet, mesoplanet, thermoplanet, and hypopsychroplanet. Proxima b and the TRAPPIST-1 system are amongst the recent additions to the catalog with recorded features. Here, we employ a non-metric classifier (Gradient Boosted Trees, XGBoost) to predict the class label of the recently discovered exoplanets. We compute the accuracy of our classification method, and aim to converge the result with the habitability score of the recently discovered exoplanets, which may suggest its proximity to the

[1] The latest updated (November 2017) dataset can be downloaded from the PHL website: http://phl.upr.edu/projects/habitable-exoplanets-catalog/data/database.

Table 1
Observed and calculated parameters of TRAPPIST-1 planets. Physical parameters are given in Earth units (EU), except for surface temperature T_s. ESI is the Earth Similarity Index used by the PHL (http://phl.upr.edu/projects/earth-similarity-index-esi) to indicate potential habitability. Earth's ESI and CDHS are both 1 and considered as the baseline. Computed habitability score and class of the planets belonging to planets system satisfy the threshold condition (Algorithm 3). The outcome of the machine learning algorithm fortifies the experimental findings (see Fig. 1 for visual illustration). Class label 6 implies most likely habitable exoplanets in a probabilistic sense (see "Earth-League" candidate selection algorithm in Section 4 of the supplementary file).

Planet	Mass	Radius	Mean insolation	Mean $T_s(K)$	CDHS_{DRS}	CDHS_{CRS}	Class	*P.ESI*
b	0.86	1.09	4.2	396.5	1.0318	1.0410	5	0.56
c	1.38	1.06	2.25	347.9	1.14084	1.1589	5	0.73
d	0.41	0.77	1.13	292.4	0.9642	0.8870	5	0.9
e	0.64	0.92	0.65	260.4	0.9722	0.9093	6	0.85
f	0.67	1.04	0.38	229.7	0.9803	0.9826	6	0.68
g	1.34	1.13	0.26	216.1	1.0951	1.1085	6	0.58
h	0.35	0.75	0.14	181.8	0.9511	0.8025	5	0.45

"Earth League". We call this an investigation in the optimistic determination of potential habitability. The hypothesis is the following: if a machine learning-based classification method classifies exoplanets by mining the features present in the PHL-EC (Method 2 in Fig. 1) and this process is independent of computing the explicit habitability score for recently developed exoplanets (*aka* Method 1 in Fig. 1), the habitability class indicated by learning from attributes (from the catalog) should match the outcome suggested by the CDHS. In other words, the class label of exoplanets predicted by the implicit method (Method 2) should correspond to the appropriate CDHS of those exoplanets. This is demonstrated in Table 7.

The second approach is based on XGBoost (Chen and Guestrin, 2016) — a statistical machine-learning classification method used for supervised learning problems where the training data with multiple features are used to predict a target variable. XGBoost is available as a toolkit which works by massively parallelizing gradient boosted trees (GBTs). We intended to test whether the outcomes of the two different approaches used to investigate the habitability of Proxima b and the TRAPPIST-1 planets, analytical and statistical, converge with a reasonable degree of confidence. In the first part, the CDHS of Proxima b and TRAPPIST-1 system are computed, and in the second part, considering these planets as a test case, the XGBoost model was trained and tested.

The paper is organized as follows. Sections 2, 3, 4, and 5 elaborate the theory and methods and discuss the implications on Proxima b and the planets in the TRAPPIST-1 system as a test case. In Section 6, we discuss our method, compare it with existing metrics, and discuss its highlights. In Section 7, we conclude, and state possibilities for improving our method as our future work.

2. Analytical approach via CDHS

We begin by discussing the key elements of the analytical approach. The parameters of Proxima b and TRAPPIST-1 System were extracted from the PHL-EC.

2.1. The data of proxima b and the TRAPPIST-1 system

The discovery of Proxima b was announced on 24th August 2016 (Anglada-Escudé, 2016). According to the PHL-EC, its radius is 1.12 EU, density is 0.9 EU, surface temperature is 262.1 K, and escape velocity is 1.06 EU. These attributes are close to those of the Earth; therefore, there are plausible reasons to believe that Proxima b may be a habitable planet. In the PHL-EC dataset, Proxima b is classified as a psychroplanet.

The discovery of the TRAPPIST-1 system has caught the attention of the entire astronomy community recently (Gillon et al., 2017). TRAPPIST-1 is an ultra-cool dwarf, detected by the 2MASS Sky Survey. All seven planets in the TRAPPIST-1 system are likely to be Earth-sized and rocky (Gillon et al., 2017), with the estimated low equilibrium temperatures — due to the exceptionally low stellar luminosity (1/1000th of the Sun), the insolation on the planets is equivalent to the insolation on the terrestrial group, thus allowing the possibility of liquid water on the surface. Three of the planets are within in the stellar habitable zone. Though all planets are most probably tidally locked with the parent star, water could still exist even on the innermost planets (see Table 2).

Table 2
Habitability score of Proxima b and Kepler-186 f, two most potentially habitable planets before the discovery of TRAPPIST-1 system. CDHS values of the exoplanets here and in Table 1 are significantly closer to the baseline score (Earth's score) compared to the ESI.

Planet name	CDHS (DRS)	CDHS (CRS)	*P.ESI*
Kepler 186 f	1.075074	1.086295	0.61
Proxima Centauri b	1.08297	1.095255	0.87

It is worth mentioning that once we know one observable – the mass – other planetary parameters used in the ESI computation (radius, density and escape velocity) can be calculated based on certain assumptions. For example, the small mass of Proxima b suggests a rocky composition. However, since 1.27 EU is only a low limit on mass, it is still possible that its radius exceeds 1.5 – 1.6 EU, which would make Proxima b not rocky (Rogers, 2014). Since Proxima b mass is 1.27 EU, the radius is $R = M^{0.5} \equiv 1.12$ EU.[2] Accordingly, the escape velocity was calculated by $V_e = \sqrt{2GM/R} \equiv 1.065$ (EU), and the density by the usual $D = 3M/4\pi R^3 \equiv 0.904$ (EU) formula.

2.2. Cobb–Douglas habitability score (CDHS)

We have proposed the new model of the habitability score in Bora et al. (2016) using a convex optimization approach (Saha et al., 2016). In this model, the Cobb–Douglas function (Cobb and Douglas, 1928) is reformulated as the Cobb–Douglas Habitability Production Function (CD-HPF) to compute the habitability score of an exoplanet,

$$\mathbb{Y} = f\,(R, D, T_s, V_e) = K(R)^{\alpha} \cdot (D)^{\beta} \cdot (T_s)^{\gamma} \cdot (V_e)^{\delta}\,, \tag{1}$$

where the planetary parameters used are radius R, density D, surface temperature T_s, and escape velocity V_e. $\mathbb{Y}$ is the habitability score CDHS, and f is defined as CD-HPF. Elasticities α, β, γ and δ need to be estimated and the value of K is considered to be 1. The goal is to maximize the score, $\mathbb{Y}$, where the elasticity values of each parameter are subject to the condition $\alpha+\beta+\gamma+\delta < 1$. Note that the interior CDHS_i, denoted by $Y1$, is calculated using radius R and density D, while the surface CDHS_s, denoted by $Y2$, is calculated using surface temperature T_s and escape velocity V_e. The objective

[2] http://phl.upr.edu/library/notes/standardmass-radiusrelationforexoplanets Standard Mass–Radius Relation for Exoplanets, Abel Mendez, June 30, 2012.

is to find elasticity values that produce the optimal habitability score for the exoplanet, i.e. to find $Y_1 = \max_{\alpha,\beta} Y(R, D)$ such that, $\alpha > 0$, $\beta > 0$ and $\alpha + \beta \leq 1$. Similarly, we need to find $Y_2 = \max_{\gamma,\delta} Y(T, V_e)$ such that $\gamma > 0$, $\delta > 0$ and $\delta + \gamma \leq 1$. Elasticity values are obtained by a computationally fast Stochastic Gradient Ascent (SGA) algorithm described in Section 3.1. We calculate CDHS values based on the constraints known as returns to scale: Constant Returns to Scale (CRS) and Decreasing Returns to Scale (DRS) (Bora et al., 2016). Note that $\alpha + \beta < 1$ is the DRS condition for elasticity, which may be scaled to $\alpha_1 + \alpha_2 + \cdots + \alpha_n < 1$. Analogously, $\delta + \gamma < 1$ is the DRS condition for elasticity which may be scaled to $\delta_1 + \delta_2 + \cdots + \delta_n < 1$.

The analysis of CDHS for Proxima b and the TRAPPIST-1 system will help to explore how this method can be effectively used for newly discovered planets, as these planets are considered to be Earth-like. The eventual characterization of any exoplanet is accomplished by using the proximity of CDHS of that planet the CDHS value of the Earth, with additional constraints imposed on the algorithm termed as "probabilistic herding". The algorithm works by taking a set of values in the neighborhood of 1 (CDHS of Earth). A threshold of 1 implies that a CDHS value between 1 and 2 is acceptable for membership in the "Earth-League", pending fulfillment of further conditions. For example, the CDHS of the most potentially habitable planet before Proxima b, Kepler-186 f, is 1.086 (the closest to the Earth's value). While another PHP, GJ-163 c, has the farthest score (1.754) from 1, and may not be even a rocky planet as its radius can be between 1.8 and 2.4 EU, which is not good for a rocky composition theory (see e.g. Rogers, 2014).

2.3. CDHS calculation using radius, density, escape velocity and surface temperature

Using the values of the parameters from the PHL-EC, we calculated CDHS scores for the CRS and DRS cases, and obtained optimal elasticities. The CDHS values in CRS and DRS cases are shown in Table 1. The degree/extent of closeness is explained in Bora et al. (2016) in great detail.

2.4. CDHS calculation using stellar flux and radius

Following the simplified expression of the ESI in terms of only planetary radius and incident flux,[3] we repeated the CDHS computation using only radius and stellar flux for Proxima b (1.12 EU and 0.700522 EU, respectively). From the scaled down version of Eq. (1), we obtain CDHS$_{CRS}$ and CDHS$_{DRS}$ as 1.083 and 1.095, respectively. These values confirm the robustness of the method used to compute CDHS and validate the claim that Proxima b falls into the "Earth-League" category.

2.5. CDHS calculation using stellar flux and mass

The habitability score requires the use of available physical parameters, such as radius, or mass, and temperature, and the number of parameters is not extremely restrictive. As long as we have the measure of the interior similarity – the extent to which a planet has a rocky interior, and exterior similarity – the location in the HZ or the favorable range of surface temperatures, we can reduce (or increase) the number of parameters. Since radius is calculated from an observable parameter — mass, we decided to use the mass directly in the calculation, obtaining CDHS$_{DRS}$ as 1.168 and CDHS$_{CRS}$ as 1.196 for Proxima b. The CDHS achieved using radius and stellar flux (previous subsection) and the CDHS achieved using mass and stellar flux have the same values.

Table 3
CDHS values calculated for different parameters.

Parameters used	CDHS$_{CRS}$	CDHS$_{DRS}$
R, D, T_s, V_e	1.083	1.095
Stellar Flux, R	1.196	1.168
Stellar Flux, M	1.196	1.167

Does this imply that stellar flux and planet mass are enough to compute the habitability score as defined by our model? It cannot be confirmed until enough number of clean data samples are obtained containing the four parameters used in the CDHS formulation. We plan to perform a full-scale dimensionality analysis as future work.

The values of CDHS using different methods are summarized in Table 3.

The nicety in the result, i.e. little difference in the values of CDHS (in Table 3), is due to the flexibility of the functional form in the model proposed in Ginde et al. (2016), and the computation of the elasticities by the Stochastic Gradient Ascent method described in the next section. Using this method led to the fast convergence of the elasticities. Proxima b passed the scrutiny and is classified as a member of the "Earth League".

3. Elasticity computation: stochastic gradient ascent (SGA)

A library function *fmincon* was used in Bora et al. (2016) to compute the elasticity values. Here, we have implemented a more efficient algorithm to perform the same task. This was done for two reasons: to be able to break free from the in-built library functions, and to devise a sensitive method which would mitigate the oscillatory nature of Newton-like methods around the local minima/maxima. Although theoretically sound, algorithmic implementations of most of these methods face convergence issues in real time due to the oscillatory nature. We have employed a modified version of the descent, an SGA algorithm, to calculate the optimum CDHS and the elasticities for mass, radius, density and escape velocity (Eq. (1) in Section 2.2). As opposed to the conventional Gradient Ascent/Descent method, where the gradient is computed only once, stochastic version recomputes the gradient for each iteration and updates the elasticity values. Theoretical convergence, guaranteed otherwise in the conventional method, is sometimes slow to achieve though. Stochastic variant of the method speeds up the convergence, justifying its use in the context of the problem (the size of data, i.e. the number of discovered exoplanets, is increasing every day).

Output elasticity (α, β, γ or δ) of Cobb–Douglas habitability function is the accentual change in the output in response to a change in the levels any of the inputs. Accuracy in elasticity values is crucial in deciding the right combination for the optimal CDHS, where different approaches are analyzed before arriving at a final decision.

3.1. Computing elasticities via gradient ascent

Gradient Ascent is an optimization algorithm used for finding the local maximum of a function. Given a scalar function $F(x)$, gradient ascent finds the $\max_x F(x)$ by following the slope of the function. This algorithm selects initial values for the parameter x and iterates to find the new values of x which maximizes $F(x)$ (here CDHS). Maximum of a function $F(x)$ is computed by iterating through the following step,

$$x_{n+1} \leftarrow x_n + \chi \frac{\partial F}{\partial x}, \tag{2}$$

where x_n is an initial value of x, x_{n+1} the new value of x, $\frac{\partial F}{\partial x}$ is the slope of function $Y = F(x)$ and χ denotes the step size, which is a

[3] PHL: http://phl.upr.edu/projects/earth-similarity-index-esi.

random number greater than 0 and lesser than a permissible upper bound. The algorithm iteratively makes small jumps (descent or ascent algorithms are trained to make small jumps in the direction of the new update) towards an optima. Stochastic variant thus mitigates the oscillating nature of the global optima — a frequent malaise in the conventional Gradient Ascent/Descent and Newton-like methods, such as *fmincon* used in Bora et al. (2016). At this point of time, without further evidence of recorded/measured parameters, it may not be prudent to scale up the CD-HPF model by including more parameters other than the ones which we are currently using. But if it ever becomes a necessity (to utilize more than the four parameters), the algorithm will come in handy and multiple optimal elasticity values may be computed fairly easily.

3.2. Computing elasticities via constrained optimization

Let the assumed parametric form be $\log(y) = \log(K)+\alpha\log(S)+\beta\log(P)$.[4] Consider a set of data points,

$$\begin{array}{ccccccc} \ln(y_1) & = & K' & + & \alpha S'_1 & + & \beta P'_1 \\ \vdots & & \vdots & & \vdots & & \vdots \\ \ln(y_N) & = & K' & + & \alpha S'_N & + & \beta P'_N \end{array} \tag{3}$$

where $K' = \log(K)$, $S'_i = \log(S'_i)$ and $P'_i = \log(P'_i)$. If $N > 3$, this is an over-determined system, where one possibility to solve it is to apply a least squares method. Additionally, if there are constraints on the variables (the parameters to be solved for), this can be posed as a constrained optimization problem. These two cases are discussed below.

No constraints: This is an ordinary least squares solution. The system is in the form $y = Ax$, where

$$x = \begin{bmatrix} K' & \alpha & \beta \end{bmatrix}^T, \quad y = \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ y_N \end{bmatrix}, \tag{4}$$

and

$$A = \begin{bmatrix} 1 & S'_1 & P'_1 \\ & \cdots & \\ 1 & S'_N & P'_N \end{bmatrix}. \tag{5}$$

The least squares solution for x is the solution that minimizes

$$(y - Ax)^T(y - Ax). \tag{6}$$

It is well known that the least squares solution to Eq. (4) is the solution to the system $A^T y = A^T Ax$, i.e. $x = (A^T A)^{-1}A^T y$. In *MATLAB*, the least squares solution to the overdetermined system $y = Ax$ can be obtained by $x = A \setminus y$. Table 4 presents the results of least squares (no constraints) obtained for the elasticity values after performing the least square fitting, while Table 5 displays the results obtained for the elasticity values after performing the constrained least square fitting; in Table 6, the values of CRS and DRS from quadratic programming for Proxima b have been enunciated.

Constraints on parameters: This results in a constrained optimization problem. The objective function to be minimized (maximized) is still the same, namely,

$$(y - Ax)^T(y - Ax). \tag{7}$$

This is a quadratic form in x. If the constraints are linear in x, then the resulting constrained optimization problem is a quadratic program (QP). A standard form of a QP is

$$\max x^T Hx + f^T x, \tag{8}$$

Table 4
Elasticity values for IRS, CRS and DRS cases after performing the least square test (no constraints): elasticities α and β satisfy the theorem $\alpha+\beta < 1, \alpha+\beta = 1$, and $\alpha+\beta > 1$ for DRS, CRS and IRS, respectively, and match the values reported previously in Bora et al. (2016).

	IRS	CRS	DRS
α	1.799998	0.900000	0.799998
β	0.100001	0.100000	0.099999

Table 5
Elasticity values for IRS, CRS and DRS cases after performing the least square test (with constraints): elasticity values α and β satisfy the theorem $\alpha+\beta < 1, \alpha+\beta = 1$, and $\alpha+\beta > 1$ for DRS, CRS and IRS, respectively, and match the values reported previously (Bora et al., 2016).

	IRS	CRS	DRS
α	1.799998	0.900000	0.799998
β	0.100001	0.100000	0.099999

such that

$Cx \le b$; Inequality constraint

$C_{\mathrm{eq}}x = b_{\mathrm{eq}}$; Equality constraint.

Suppose the constraints are $\alpha, \beta > 0$ and $\alpha+\beta \le 1$. The QP can be written as (neglecting the constant term $y^T y$)

$$\max x^T(A^T A)x - 2y^T Ax, \tag{9}$$

such that

$$\begin{cases} \alpha > 0, \\ \beta > 0, \\ \alpha+\beta \le 1. \end{cases} \tag{10}$$

For the standard form as given in Eq. (8), Eqs. (9) and (10) can be represented by rewriting the objective function as follows:

$$x^T Hx + f^T x, \tag{11}$$

where

$$H = A^T A \text{ and } f = -2A^T y. \tag{12}$$

The inequality constraints can be specified as

$$C = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 1 \end{bmatrix}, \text{ and } b = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \tag{13}$$

In *MATLAB*, the QP can be solved using the function *quadprog*. The results in Table 5 were obtained by conducting quadratic programming.

Using active set: We have conducted the experiment using the *active learning* technique. This framework is best suited in our case as it can be applied to different performance targets and all types of classifications. The traditional active-set method is divided into two steps, focusing on feasibility and optimality, in that order. Instead of an *ad-hoc* start, active set methods bank on a good *initiator* estimate of the optimal active set. This is well suited for a sequence of quadratic programs to be solved, which is what our constrained optimization problem needs. Active set gave best results out of all the three algorithms which suffices our argument.

By solving the described constrained QP, we find that the results satisfy the condition $\alpha+\beta \le 1$ for the DRS case, $\alpha+\beta = 1$ for the CRS case, and $\alpha+\beta \ge 1$ for the IRS case. Elasticities α, β and K from both computations are very close, supporting our choice of α, β and K. Identical results are observed for elasticities γ and δ for the surface CDHS. Obtained elasticity values used for computing CDHS are $\alpha = 0.8$ and $\beta = 0.1$ for DRS, and $\alpha = 0.9$ and $\beta = 0.1$ for CRS cases, respectively. The algorithms are repeated to compute γ and

[4] This is a logarithmic transformation of the standard CDHS model (which has the exponential form).

Table 6
Results of quadratic programming by using the active-set learning. Exact match with SGA results and Method 1, which satisfy the conditions of CRS, DRS and IRS i.e. elasticity values α and β satisfy the theorem $\alpha+\beta=1; \alpha+\beta<1; \alpha+\beta>1$ and match the values reported in Bora et al. (2016).

	CRS	DRS
K	1	1
α	0.9000	0.8000
β	0.1000	0.1000

δ, and a convex combination of interior and surface CDHS is used to calculate the CDHS of Proxima b. The entire process, including classification of all exoplanets post-habitability score computation, is summarized in the supplementary file.

4. Model scalability

In Bora et al. (2016), we have shown the theoretical guarantee regarding the conditions on elasticity. However, the scalability of the model (scalability of the theoretical guarantee) depends on the fact that the conditions of global maxima continue to hold even if the number of input parameters increase. In addition, the theoretical guarantee in some cases (Saha et al., 2016) tends to relax when an arbitrary parameter is added to the model. This happens due to curvature violation of the functional form. In other words, if eccentricity (say) is added as one of the input parameters along with surface temperature, density, radius and mass, there needs to be a mathematical guarantee that the conditions on elasticity should scale in the same fashion. This has been illustrated previously via computer simulation. However, there needs to be a theoretical result fortifying the intuitive understanding of the proposed model and the scoring scheme — the CDHS. We define the theorem which lays the foundation for model scalability in the event any parameter is added to the existing model, already accommodating an arbitrary number of parameters. If the conditions of elasticity for a global maxima hold for a fixed set of input parameters (say, n), it will continue to hold when the number of parameters is increased by 1 (say, $n+1$). This is an inductive approach, non-traditional but powerful! The proof (given in Section 3 of the supplementary file) is based on the principle of mathematical induction.

Theorem. *If the global maxima for CDHS, i.e.*

$$\log(Y)=\frac{1}{1-\sum_{i=1}^{n}\alpha_i}\log\left\{k\prod_{i=1}^{n}\left(\frac{x_i p}{w_i}\right)^{\alpha_i}\right\} \tag{14}$$

holds, then the same condition for the global maxima will continue to hold if an additional input parameter is inserted in the habitability function CD-HPF, i.e., if

$$\log(Y_{\text{new}})=\frac{1}{1-\sum_{i=1}^{n+1}\alpha_i}\log\left\{k\prod_{i=1}^{n+1}\left(\frac{x_i p}{w_i}\right)^{\alpha_i}\right\} \tag{15}$$

holds as well. Further, it follows that the elasticity condition for DRS for $n+1$ parameters is true, i.e. $1-\sum_{i=1}^{m+1}\alpha_i>0$, if the elasticity condition for DRS for n parameters, i.e. $1-\sum_{i=1}^{m}\alpha_i>0$, holds.[5]

CDHS leads to a classification scheme (Algorithm 3 in Section 4 of the supplementary file) and depends on computing the habitability score of discovered exoplanets. However, the classification problem does not have to rely on having numerical values of the response variable of samples under classification. Instead, the hidden relationship between samples may be discovered by the construction of the decision rules connecting the feature values of the samples.

5. Classification via non-functional form: XGBoost, a feature-based learning and classification method

Here we illustrate a method by which the high habitability score of Proxima b may be predicted by using class labels and features from the PHL-EC. The method XGBoost (eXtreme Gradient Boosting) is a non-metric classifier, and a fairly recent addition to the suite of machine learning algorithms (Chen and Guestrin, 2016). Non-metric classifiers are applied in scenarios where there are no definitive notions of similarity between feature vectors.

A typical machine-learning problem processes input data and combines that with the learning algorithm to produce a model as output. Learning implies recognizing complex patterns and making intelligent decisions based on data. The machine comes up with its own prediction rule, based on which a previously unobserved sample would be classified as a certain type, meso or psychroplanets for example, with a reasonable accuracy.

In order to appropriately apply a method (including preprocessing and classification), a thorough study of the nature of the data should be done; this includes understanding the number of samples in each class and the separability of the data. Depending on the nature of the data, appropriate preprocessing and post processing (if needed) methods should be determined along with the right kind of classifier for the task.

5.1. Understanding the data for classification

The PHL-EC dataset contains more than 3500 samples and is growing steadily: from 1904 samples in November 2015 to 3635 samples at the time of writing.[6] We have considered 51 features of the data for classifying, and have eliminated the ones that are unimportant for classification, such as the name of the parent star (*S.Name*) and the name of the planet (*P.Name*). In the dataset PHL-EC, planets are already segregated into five classes based on their surface thermal properties:

(1) *Non-Habitable*: planets that do not have thermal properties required to sustain life.
(2) *Mesoplanet*: planets with a mean global surface temperature between 0 °C and 50 °C — a necessary condition for complex terrestrial life. These are generally referred to as Earth-like planets.
(3) *Psychroplanet*: planets with mean global surface temperature between −15 °C and +10 °C — somewhat colder than the optimal temperature for the sustenance of terrestrial life.
(4) *Thermoplanet*: planets with the temperature in the range of 50 °C – 100 °C – warmer than the temperature range suited for most terrestrial life.
(5) *Hypopsychroplanets*: planets with temperature below −50 °C. These planets are too cold for the survival of most terrestrial life.

Out of these, the classes of hypopsychroplanet and thermoplanet have too few samples (only two planets each) and hence are not useful for the analysis. The classification was performed on remaining three classes: psychroplanet, mesoplanet and non-habitable.

A planet having characteristics suitable for inhabitation is still a rare occurrence; naturally, most of the samples in the dataset

[5] The habitability score CDHS is computed using four parameters: R, D, T_s and V_e. If a new parameter from the PHL-EC needs to be added to the CD-HPF, it is important to know if the conditions of global maxima for habitability still hold. The above theorem validates our superposition conclusively.

[6] These numbers vary over time. We need these samples to train the classifier, so that it can classify new additions, such as e.g. Proxima b or TRAPPIST-1 planets.

belong to the class of non-habitable planets (3690 out of 3782 at the time of writing this paper). From a data analytic point of view, this is a *data bias* and can lead to overfitting, i.e., when a classifier becomes overly complex and extremely sensitive to the nuances in the data. Overfitting is a problem that needs to be dealt with carefully and not be overlooked as an administrative task. In a dataset such as the PHL-EC, where the number of samples belonging to one class is over a thousand times the total number of samples belonging to all the other classes, just reporting the numeric accuracy obtained by directly feeding the data to train a classifier would be an incorrect methodology.

5.2. Classification of data

As a first step, data from PHL-EC was pre-processed (the authors have tried to tackle the missing values by taking mean for continuous-valued attribute, and mode for categorical attributes). Certain attributes from the database, namely *P.NameKepler*, *S.NameHD, S.NameHid, S.Constellation, S.Type, P.SPH, P.InteriorESI, P.SurfaceESI, P.Disc.Method, P.Disc.Year, P.MaxMass, P.MinMass*, *P.Inclination* and *P.Habmoon* were removed as these attributes do not contribute to the nature of classification of habitability of a planet. Though individual ESI values (and planetary mass) do contribute to habitability determination, because the dataset directly provides total value of *P.ESI* — the global ESI of the planet, these features were neglected. Following this, classification algorithms were applied on the processed dataset, where in total 49 features were used.

To counter the potential problems due to the dominance by a single class, we used *artificially balanced* datasets by considering random samples from the classes of non-habitable and mesoplanets with the number of samples belonging mesoplanets being equal to the number of samples in the psychroplanet class (as it has the least number of samples) and the number of non-habitable planets being an appropriate fraction of the total number of non-habitable samples (between 5% and 10%). Then this balanced dataset was used as the training set. This cycle of balancing the dataset artificially, dividing it, training and testing the classifier was performed multiple times, and the mean accuracy of all the trials was considered to be representative of the potential of a classifier. By artificial balancing, the reported accuracies are also more reliable than without balancing. In the aggregate exercise of classification, 500 iterations of training–testing were performed.

We have applied a powerful ensemble classification for the task described above. *Boosting* refers to the method of combining the results from a set of *weak learners* to produce a *strong* prediction. Generally, a weak learner's performance is close to a random guess. A weak learner divides the job of a single predictor across many weak predictor functions, and optimally combines the votes from all smaller predictors. This helps in enhancing the overall prediction accuracy.

XGBoost is a tool developed by utilizing these boosting principles (Chen and Guestrin, 2016). XGBoost combines a large number of regression trees with a small learning rate. Regression can be used to model classifiers: here the word *regression* may refer to logistic or soft-max regression for the task of classification. XGBoost uses an ensemble of decision trees. We describe the detailed working principle in Section 4 (XGBoost: An Exploration of Machine Learning based Classification) of the supplementary file.

The pursuit of finding the appropriate classification method for any classification problem requires a lot of experimentation and analysis of the nature of the data. We performed a convex hull test to understand the nature of the data and found that the data is not linearly separable. Hence, classifiers like SVM (Support Vector Machines), K-NN (*k* Nearest Neighbors) and LDA (Linear Discriminant Analysis) are not expected to perform well. All these classifiers were tried as candidates for classification; as expected from the convex hull test, they did not perform well. This motivated our choice of Tree-based classifiers and, more specifically, of the XGBoost.

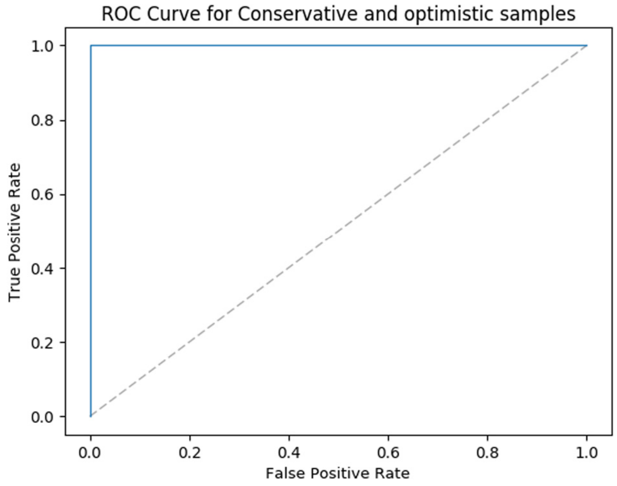

Fig. 2. ROC curve (blue line) of the final outcome of the machine classification of Proxima b and the TRAPPIST-1 system using XGBoost. The dotted line is the ROC for the case when the performance of the binary classifier is the same as a perfectly random guess. A good numeric representation of ROC curves is the percentage area of the coordinates which falls under the curve (AUC). Here, the AUC is 100%, which indicates a perfect performance of the classifier.

5.3. Classification accuracy of XGBoost

XGBoost was used to classify Proxima b and the TRAPPIST-1 system as test samples. The ROC (Receiver Operating Characteristic) curve obtained for this classification is shown in Fig. 2. Each point on the ROC plot represents a sensitivity/specificity pair which corresponds to a particular decision threshold. Sensitivity, or recall, is the proportion of positive tuples that are accurately identified, and specificity is the proportion of negative tuples that are correctly identified. A test with non-overlapping classes has an ROC plot that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore, the closer the ROC plot is to the upper left corner, the higher the overall accuracy of the test (Zweig and Campbell, 1993).

The overall classification accuracy of XGBoost was 100%, i.e. there were no false positives or false negatives in its classification. The method of classification was to select Proxima b as the training set and the remaining samples in the catalog as the test set (subject to artificial balancing by under-sampling the non-habitable class).

However, in cases where a dataset exhibits a data bias towards one class, the F-score test statistic is more representative than the accuracy of a classifier. It is used to analyze whether a classifier is able to achieve both high precision and high recall simultaneously (for details on precision, recall, F-score, ROC curves, etc., see Peres et al., 2015, Van Rijsbergen, 1979). The values for precision and recall were calculated to gauge the goodness of the classifier. This was done but considering the psychroplanet and mesoplanet classes as the positive classes one at a time. With respect to the psychroplanet class, the calculated F-score was 0.94, precision was 0.95, and recall was 0.93. With respect to the class of mesoplanets, the F-score was 0.95, precision was 0.93, and recall was 0.97. Using the XGBoost classifier, the class-wise accuracy of all the test samples was estimated to be 100%. We can thus say that XGBoost performs the classification remarkably well!

Table 7
Summary of results of both methods: samples which are labeled as Class 6, an indicator of potential habitability are also predicted as habitable; the outcome of both approaches matches. For example, TRAPPIST 1 d, labeled as psychroplanet by Method 2 with 100% accuracy, is also in Class 6 ("Earth-League", according to Method 1).

Exoplanet	Method 1: Explicit score calculation			Method 2: Classification using XGBoost	
	$CDHS_{DRS}$	$CDHS_{CRS}$	Class category	Classification accuracy (%)	Predicted class
Kepler 186 f	1.075074	1.086295	6	–	–
Proxima b	1.08297	1.095255	6	100.0	Psychroplanet
TRAPPIST-1 b	1.0318	1.0410	5	89.6	Non-habitable
TRAPPIST-1 c	1.14084	1.1589	5	88.4	Non-habitable
TRAPPIST-1 d	0.9642	0.8870	5	100.0	Mesoplanet
TRAPPIST-1 e	0.9722	0.9093	6	100.0	Psychroplanet
TRAPPIST-1 f	0.9803	0.9826	6	99.7	Psychroplanet
TRAPPIST-1 g	1.0951	1.1085	6	82.3	Psychroplanet
TRAPPIST-1 h	0.9511	0.8025	5	95.1	Non-habitable

5.4. Classification without surface temperature

We also present the results of machine classification based on only the following planetary parameters that can be observed. This experiment itself is divided into two parts: the first, where the parameters *P.Min Mass, P.Mass, P.Radius, P.SFlux Min, P.SFlux Min, P.SFlux Mean, P.SFlux Max* are used, and the second, where only *P.Min Mass, P.Mass* and *P.Radius* are used. The purpose of this exercise is to see how a classification scheme, based on surface temperature (which is a calculated variable), is reflected only by the variables which can be observed. The experiment is performed using the same general experimental pipeline which we have described in Section 5.2; the only change is that any of the columns in the dataset other than those aforementioned were removed.

In this experimental setup, just like the previous one, the overall results appropriately classify the planets into the classes as defined by the PHL-EC and reflect the properties of the habitability class. The results of this experiment are shown in Table 8.

6. Discussion

The theoretical premise of the CDHS is different from those of existing metrics such as Earth-Similarity Index (ESI) and Planetary Habitability Index (PHI) (Schulze-Makuch et al., 2011). Prior to the CDHS, to the best of our knowledge, a foundation for a habitability indicator based on optimization theory was not made. The rationale for this is to incorporate tradeoffs that may exist between various planetary parameters to create favorable conditions for the existence of life in a form similar to that on Earth.

Based on the CDHS values, Proxima b and TRAPPIST-1 e, f, and g fall in the "Earth-League" class: the difference between their CDHS and the Earth's CDHS is within the acceptable threshold of 1. The classification model, XGBoost, used in this work to classify the exoplanets also resulted in very high accuracies (Tables 7 and 8), which provides evidence of the strength of the model to automatically label and classify newly discovered exoplanets, such as Proxima b or TRAPPIST-1 planets in this case.

Our approach is emphatically exhibiting the validation of the potential habitability of Proxima b, Kepler-186 f, and TRAPPIST-1 planets, matching with the PHL findings. The robustness of the formula and the solid theory behind the formulation are validated by the proximity of the scores computed for different cases.

It needs to be noted that the authors perceive habitability as a probabilistic measure or a measure with varying degrees of certainty. Therefore, the construction of different classes of habitability Classes 1 to 6 was contemplated, corresponding to measures as *most Earth-like* as Class 6, to *least Earth-like* as Class 1 (note that this scheme of classification is not identical to the automated machine classification for planetary habitability, which is enunciated by the second approach). As a further illustration, Classes 6 and 5 seem to represent the identical patterns in habitability, but they do not! Class 6 – the "Earth-League" – is different from Class 5 in the sense that it satisfies the additional conditions of thresholding and probabilistic herding and, therefore, ranks higher on the values of CDHS. This is in stark contrast to the binary definition of exoplanets being "habitable or non-habitable", and a deterministic perception of the problem itself. The approach therefore required classification methods that are part of machine learning techniques and convex optimization — a sub-domain strongly coupled with machine learning. CD-HPF and CDHS are used to determine the equivalence to habitability and the maximum habitability score of all exoplanets with confirmed surface temperatures in the PHL-EC (*confirmed surface temperatures alone do not imply habitability and therefore the machine classification problem of habitable exoplanets is non-trivial*). Global maxima are calculated theoretically and algorithmically for each exoplanet, exploiting intrinsic concavity of CD-HPF and ensuring no curvature violation. Computed scores are fed to the attribute enhanced K-NN algorithm — a novel classification method, used to classify the planets into different classes to determine how similar an exoplanet is to Earth. The authors would like to emphasize that, by using classical K-NN algorithm and not exploiting the probability of habitability criteria, the results obtained were good, having 12 confirmed potentially habitable exoplanets in the "Earth League". We have created web pages (https://habitabilitypes.wordpress.com/ and https://astrirg.org/) for this project to host all relevant data and results: sets, figures, animation video and a graphical abstract. The web pages also contain the catalog of all confirmed exoplanets with class annotations and computed habitability scores. The catalog is built with the intention of further use in designing statistical experiments for the analysis of the correlation between habitability and the abundance of elements (this work is briefly outlined in Safonova et al., 2016). It is a very important observation that our algorithm and method give rise to a score metric, CDHS, which is structurally similar to the Planetary Habitability Index (PHI) (Schulze-Makuch et al., 2011) as a corollary in the CRS case (when the elasticities are assumed to be equal to each other).

The two methods that we have explored in our work essentially try to answer two questions collectively: "*Is this planet potentially habitable?*" and "*How potentially habitable is this planet?*". By performing classification, we can affirm if a planet is expected to be potentially habitable or not, and by computing the CDHS, we are basically assigning a number to every planet which reflects its potential habitability. By doing this, in the future, we can gain deeper insights to planet's characteristics, understand what range of scores of the CD-HPF implies which classes of habitability, approximate unobserved attributes of a planet, etc. As the volume of data in the PHL-EC catalog increases with time, a robust automated method must be in place to analyze the data quickly and in an efficient way. An automated system primarily serves two purposes. The first is that it reduces human error in computation. The second, it eradicates the subjectivity that arises when different people try to classify or judge any data sample (here, a planet). One

Table 8
Results of machine classification based on a smaller set of planetary parameters as features for the classifier represented as classification accuracy (%). In the headers of the second and third columns, we mention the training parameters used for the experiments expressed in Earth Units (EU). The features in the third column are independent of surface temperature, so is the classification.

Exoplanet	*P.Min Mass, P.Mass, P.Radius, P.SFlux Min, P.SFlux Min, P.SFlux Mean, P.SFlux Max* (are expressed in %)	*P.Min Mass, P.Mass and P.Radius – these features are independent of S. Temp* (results are expressed in %)
Proxima Cen b	73.8	73.0
TRAPPIST-1 b	100.0	100.0
TRAPPIST-1 c	100.0	100.0
TRAPPIST-1 d	100.0	97.3
TRAPPIST-1 e	79.5	99.8
TRAPPIST-1 f	100.0	99.8
TRAPPIST-1 h	100.0	100.0

researcher's appraisal of a data sample might not be the same as that of another when evaluated based on general characteristics. However, when an algorithm is used for this, the results will be the same, regardless of which computer the system is deployed on. Hence, the implication of a system like this is the standardization of classification and of multiple ways of evaluating the potential habitability of an exoplanet.

To sum up, CD-HPF and CDHS turn out to be self-contained metrics for habitability. We would like to pose the following questions in this context: *How do the two approaches coincide/converge?* And, *what is the implication in the overall scientific context?*

Context is critical in solving a problem as complex as determining the habitability of discovered exoplanets. We mention with great regard the advances and contributions made by Dirk Schulze-Makuch, Abel Mendez, and other researches working in this field (Schulze-Makuch et al., 2011). In contrast to the other existing habitability metrics, our approach is entirely data-driven and inspired by machine learning and optimization.

This system may be extended in the future to analyze how the CDHS correlate with the classes of habitability. As there are multiple classes, it would be interesting to see if the CDHS falls into certain ranges for each class. The convergence of the score, however, gives rise to the following questions:

- Are only stellar flux/surface temperature and radius/mass enough to construct a reliable habitability score via machine learning?
- Should a full-scale dimensionality reduction technique be employed (completely data-driven approach) in the future, to analyze the context and validate such a claim?

The concept of developing a classifier based on our growing knowledge of exoplanets is intriguing. There is no reason why such an approach should not work other than to think that of the large number of possible habitable exoplanets. We have parameters based on only one example that is known to be habitable and in that regard assume that all non-Earth like exoplanets are non-habitable. Our definition of habitability may need to be refined as we find more truly habitable planets.

We have made use of stochastic gradient ascent and MATLAB's *fmincon* function (in the past) to find local maxima. Evolutionary algorithms may also be used to track dynamic functions of the type that allow for the oscillation that are instead mitigated with SGA. Additionally, we make use of 49 features through XGBoost. The results suggest that the use of newly discovered exoplanets for testing and remaining samples in the catalog for training performed well with XGBoost (AUC = 1.0) which is surprisingly good. It is possible that the neural network may also work well. It would be interesting to try a *fuzzy* approach on this problem where planets have membership in all class labels but just to differing degrees. Given the sparsity of our knowledge about planets, their features and habitability, a fuzzy approach may be worth exploring, and a comparative analysis with traditional classification approaches may be documented.

CDHS is the first of its kind where a sound theoretical basis has been provided regarding the scalability of the model to incorporate other parameters (such as flux, eccentricity, or orbital velocity). It is proven to guarantee global optima under countably many additional parameters. Moreover, if distributions in parameter data differ markedly, other metrics without a mathematical proof of scalability may not be appropriate. In a nutshell, a theoretical foundation of our metric is a cornerstone of the novelty of our work.

7. Conclusions and future work

Rapid discoveries of exoplanets notwithstanding, it is unrealistic and premature to predict how Earth-like are the conditions on any planet on the basis of the scant data available presently. The best case scenario is to adduce a list of optimistic targets for future detailed missions. This manuscript achieves that goal by combining physical observations, mathematical rigor, and ML techniques. Our approach might pay rich dividends considering the encouraging observations recently reported (Bourrier et al., 2017), where Space Telescope Imaging Spectrograph (STIS) has been used to study the amount of ultraviolet radiation received by the TRAPPIST-1 planets. It was inferred that three planets within the HZ, TRAPPIST-1 e, f and g, may still possess abundant amount of water on their surfaces, indicating potential habitability. We have predicted the same using ML and sophisticated modeling this paper and the supplementary file. From the CD-HPF (Method 1, Fig. 1), we found the habitability scores of TRAPPIST-1 e, f, and g to be close enough to that of Earth (within two decimal places, Table 7). TRAPPIST-1 e and f are also classified with high accuracies by the boosted-tree learning (Table 7). This is definitely encouraging, especially in the view of the most recent series of papers reporting the potential habitability of these planets (Barr et al., 2017; Grimm et al., 2018; Wolf, 2017; Papaloizou et al., 2017).

The mathematical and computational premise of our approach leaves some margin for future work. Here we address three main technological challenges which we will consider taking up in the future.

Replacing missing values: Sometimes, there is a missing data in the catalog. The unknown surface temperatures (or other parameters) can be estimated using various statistical, or *rule-based* models (Agrawal et al., 0000; Agrawal and Srikant, 1994). Future work may include incorporating more input parameters to the Cobb–Douglas function coupled with tweaking the attribute-enhanced K-NN algorithm by checking an additional condition.

Additionally, parameters such as orbital period, stellar flux, and distance of the planet from host star may be equally important to determine the habitability. It is pertinent to check for the dominant parameters that contribute more towards the habitability score. This can be accomplished by computing percentage contributions to the response variable — the habitability score. We would like to conclude by stressing on the efficacy of the method of using a few of the parameters rather than sweeping through a host of properties listed in the catalogs, effectively reducing the dimensionality of the problem.

Model uncertainty: The uncertainty in the model is related to the inefficiency of the estimators, a technicality that we have not encountered in our model. However, from a theoretical perspective, it is a pertinent point to discuss. We have assumed the value of k as 1 for simplicity (in the general formulation of the model, there exists a multiplicative proportionality constant k). However, k in our model formulation may be estimated from data by using sophisticated fitting models and constrained optimization techniques. Once k is suitably estimated, elasticity may then be predicted/fitted accordingly. The CDHS model can scale up to any number of inputs in theory. However, the increase in the numbers of inputs leads to exponential increase in the complexity of this model. This increase in complexity may cause curvature violation of the Cobb–Douglas model, which would cause erroneous elasticity coefficients estimations. In order to mitigate this, we may use stochastic frontier analysis (Goswami et al., 2018) to estimate the values of elastic coefficients and proportionality constant k. This is a theoretical possibility and may happen when we augment our model to include more input parameters. A stochastic frontier CDHS for two input parameters, say K, L, has the production frontier in the form:

$$y = f(K, L)TE, \tag{16}$$

where TE is the technical inefficiency, the ratio of observed output to maximum possible output. If $TE = 1$, maximum output is achieved. This production frontier is deterministic as the entire deviation from maximum feasible output is attributed to technical inefficiency. It does not consider random *shocks*, which is not beyond control of production function. To address the random shocks, the production frontier function can be redefined as follows:

$$y = f(K, L)TE \exp(v), \tag{17}$$

where v is the stochastic variable which defines the random shocks, uncertainty, etc.

Measurement uncertainty: If there exists measurement uncertainties in any of the observables and/or input features used in our model, it can be shown that CDHS continues to have a global optima under perturbed (stochastic measurement error) input values. Estimation techniques witness observed choices deviate from optimal ones due to two factors: failure to optimize, i.e. inefficiency, and random noise. Stochastic variant of our model incorporates these factors seamlessly. It is one of the best technique to model input behavior, produces individual estimate scores that have greater accuracy. The basic idea lies in the introduction of an additive error term consisting of noise and an inefficiency term. Thus, the method can help to identify the predictor variables which need corrective measures. Hence, the model may produce efficiency estimates or efficiency scores. These estimates may then be used to identify the predictor variables which need intervention and corrective measures. It is important to note that the efficiency score varies as it is dependent on input characteristics. This relationship can be expressed in terms of a function of single dependent variable (output) with one or more explanatory variables (inputs). Another random variable may be included which represents noise. We plan to extend current work by exploiting Cobb–Douglas Frontier (Goswami et al., 2018) to address the issues of measurement uncertainties.

Acknowledgments

We would like to thank the Science and Engineering Research Board (SERB)-Department of Science and Technology (DST), Government of India for supporting our research. The project reference number is SERB-EMR/2016/005687.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at doi:10.1016/j.ascom.2018.03.003.

References

Agrawal, R., Imielinski, T., Swami, A., 0000. Mining association rules between sets of items in large databases. In: Proc. 1993 ACM SIGMOD International Conference on Management of Data, Washington DC (USA), pp. 207–216.

Agrawal, R., Srikant, R., 1994. Fast algorithms for mining association rules in large databases. In: Bocca, Jorge B., Jarke, Matthias, Zaniolo, Carlo (Eds.), Proc. 20th International Conference on Very Large Data Bases (VLDB 94). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 487–499.

Anglada-Escudé, G., et al., 2016. A terrestrial planet candidate in a temperate orbit around Proxima Centauri. Nature 536, 437–440. http://arxiv.org/abs/1609.03449.

Barr, A.C., Dobos, V., Kiss, L.L., 2017. Interior Structures and Tidal Heating in the TRAPPIST-1 Planets, http://arxiv.org/abs/1712.05641. doi:10.1051/0004-6361/201731992.

Bora, K., Saha, S., Agrawal, S., Safonova, M., Routh, S., Narasimhamurthy, A., 2016. CD-HPF: new habitability score via data analytic modeling. Astron. Comput. 17, 129–143. http://arxiv.org/abs/1604.01722.

Bourrier, V., de Wit, J., Bolmont, E., Stamenkovic, V., Wheatley, P.J., Burgasser, A.J., Delrez, L., Demory, B.-O., Ehrenreich, D., Gillon, M., Jehin, E., Leconte, J., Lederer, S.M., Lewis, N., Triaud, A.H.M.J., Van Grootel, V., 2017. Temporal evolution of the high-energy irradiation and water content of TRAPPIST-1 Exoplanets. Astron. J. 154, 121. doi:10.3847/1538-3881/aa859c. http://arxiv.org/abs/1708.09484.

2016. Breakthrough Starshot. A Russian billionaire has a crazy plan to reach a nearby planet that might harbor life . http://www.businessinsider.in, Retrieved 12.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). ACM, New York, NY, USA, pp. 785-794. doi:10.1145/2939672.2939785. http://arxiv.org/abs/1603.02754.

Cobb, C.W., Douglas, P.H., 1928. A theory of production. Amer. Econ. Rev. 18 (Supplement), 139.

Gillon, M., Triaud, A.H.M.J., Demory, B.-O., et al., 2017. Seven temperate terrestrial planets around the nearby ultracool dwarf star TRAPPIST-1. Nature 542, 456–460.

Ginde, G., Saha, S., Mathur, A., Venkatagiri, S., Vadakkepat, S., Narasimhamurthy, A., Daya Sagar, B.S., 2016. ScientoBASE: A framework and model for computing scholastic indicators of non-local influence of journals via native data acquisition algorithms. J. Scientometrics 107 (1), 1–51. http://arxiv.org/abs/1605.01821.

Goswami, B., Sarkar, J., Saha, S., Kar, S., 2018. CD-SFA: Stochastic frontier analysis approach to revenue modeling in cloud data centers. Int. J. Comput. Netw. Distrib. Syst. http://arxiv.org/abs/1610.00624. (In Press).

Grimm, S.L., Demory, B.-O., Gillon, M., et al., 2018. The nature of the TRAPPIST-1 exoplanets. http://arxiv.org/abs/1802.01377. doi:10.1051/0004-6361/201732233.

Papaloizou, J.C.B., Szuszkiewicz, E., Terquem, C., 2017. The TRAPPIST-1 system: Orbital evolution, tidal dissipation, formation and habitability. http://arxiv.org/abs/1711.07932.

Peres, D.J., Iuppa, C., Cavallaro, L., Cancelliere, A., Foti, E., 2015. Significant wave height record extension by neural networks and reanalysis wind data, Vol. 94, pp. 128-140, doi:10.1016/j.ocemod.2015.08.002.

Rogers, L.A., 2014. Most 1.6 earth-radius planets are not rocky. Ap. J. 801 (1), 41. http://arxiv.org/abs/1407.4457.

Safonova, M., Murthy, J., Shchekinov, Y.A., 2016. Age aspects of habitability. Int. J. Astrobiol. 15, 93. doi:10.1017/S1473550415000208. http://arxiv.org/abs/1404.0641.

Saha, S., Sarkar, J., Dwivedi, A., Dwivedi, N., Anand, M.N., Roy, R., 2016. A novel revenue optimization model to address the operation and maintenance cost of a data center. J. Cloud Comput. 5 (1), 1–23.

Schulze-Makuch, D., Méndez, A., Fairén, A.G., et al., 2011. A two-tiered approach to assessing the habitability of exoplanets. Astrobiology 11, 1041.

Van Rijsbergen, C.J., 1979. Information Retrieval, ISBN:9780408709293.

Witze, A., 2016. Earth-sized planet around nearby star is astronomy dream come true. Nature 536, 381–382.

Wolf, E.T., 2017. Assessing the habitability of the TRAPPIST-1 system using a 3D climate model. Astrophys. J. 839, L1. http://arxiv.org/abs/1703.05815.

Zweig, M.H., Campbell, G., 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin. Chem. 39 (4), 561–577.