Ex(o)plain: Subgroup-level analysis of exoplanet atmospheric parameters

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 August 2024, accepted 20 September 2024, date of publication 24 September 2024, date of current version 4 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3466919

APPLIED RESEARCH

Ex(o)plain: Subgroup-Level Analysis of Exoplanet Atmospheric Parameters

ALKIS KOUDOUNAS, (Graduate Student Member, IEEE), FLAVIO GIOBERGIA, (Member, IEEE), AND ELENA BARALIS, (Member, IEEE)
Department of Control and Computer Engineering, Politecnico di Torino, 10129 Turin, Italy
Corresponding author: Alkis Koudounas (alkis.koudounas@polito.it)

ABSTRACT Deep learning has been shown to be a valuable tool in astrophysics. In the field of exoplanetary science, deep learning-based approaches are being used extensively to automate the characterization of exoplanet atmospheres, reducing computational costs when compared to conventional methods. However, many atmospheric reconstruction models lack interpretability. We introduce *Ex(o)plain*, a model-agnostic framework to identify and describe the most meaningful traits that characterize exoplanet atmospheres. Our approach categorizes exoplanets into subgroups based on combinations of various metadata, such as surface gravity, planet radius, and star temperature. We analyze these subgroups to identify those for which the deep learning model performs better or worse than average. This provides useful insights into what is being effectively learned by these black box models and where they still struggle. We explore a practical case based on the synthetic observations generated for the upcoming Ariel mission. Experimental results demonstrate the effectiveness of adopting explanation techniques in revealing meaningful variations in reconstruction quality between individual models and their aggregated ensemble. We additionally show that ensemble approaches significantly outperform single learners. We leverage the same subgroup-based exploration techniques to assess the situations that are most beneficial for the ensemble. Our work provides a more nuanced understanding of deep learning results for exoplanet characterization, aiming to delineate feasible accuracy limits and enable more informed evaluations of these techniques' atmospheric reconstruction capabilities.

INDEX TERMS Deep learning, divergence, exoplanet atmospheric parameters, explainable AI, subgroup detection.

I. INTRODUCTION

Exoplanets orbiting other stars outside the solar system have transformed our understanding of planetary science. Characterizing atmospheric diversity across these worlds contextualizes planetary evolution within and beyond our solar system. Transmission spectroscopy is a technique that consists of measuring atmospheric absorption at different wavelengths when exoplanets transit in front of their star. This approach can reveal properties such as chemical composition, temperature, and the presence and characteristics of clouds by analyzing the observed spectra [1].

While transmission spectroscopy studies have been conducted for a few dozen exoplanets over the past two decades, collecting this observational data has been gradual. Upcoming space missions aim to broadly expand both the quantity and quality of transmission spectral measurements. For example, the James Webb Space Telescope [2] is expected to more than triple the volume of existing observations. Meanwhile, the Ariel space mission [3] has been designed to conduct an even larger-scale atmospheric survey characterizing over 1,000 exoplanet atmospheres through transit spectroscopy. These upcoming astrophysics attempts can be expected to achieve significant progress in the field by delivering significantly more transmission spectra to the scientific community for further analysis.

Atmospheric retrieval is the process of inferring exoplanet atmospheric properties from observational data, and it is challenging with low-resolution transmission spectra due to the complex, degenerate parameter spaces where multiple

The associate editor coordinating the review of this manuscript and approving it for publication was Binit Lukose.

combinations of temperature, chemical composition, and cloud characteristics can produce similar spectral signatures. This introduces degeneracies wherein multiple solutions could plausibly fit the observed spectral features [4]. To disentangle these degenerate cases, atmospheric retrieval involves determining a posterior probability distribution rather than identifying a single, definitive solution. The posterior distribution delineates the range of plausible atmospheric compositions that align with the observational data, considering measurement uncertainties. By outlining this spectrum of compatible solutions, the posterior offers a more comprehensive characterization of a planet's atmosphere than isolated point estimates. While traditional Bayesian techniques employing computationally demanding algorithms currently achieve top performance [5], they do not easily scale up to process the vast datasets expected from upcoming missions [6]. Machine learning (ML) and deep learning (DL) techniques have been employed to enhance retrieval efficiency, but many existing approaches still lack interpretability, that is, the ability to understand and explain how the models make their predictions [7]. Specifically, models may consistently underperform when characterizing certain types of exoplanetary atmospheres. However, such patterns tend not to emerge when reporting only average performance.

With this in mind, we introduce *Ex(o)plain*, an interpretable framework for probabilistic atmospheric characterization. We aim to provide human-understandable insights into the patterns – or exoplanet categories – that *diverge* the most in terms of performance, posing difficulties for machine learning algorithms in building accurate predictive models. By identifying where models fall short, our approach seeks to help develop more robust and transparent solutions for atmospheric retrieval.

In addition to transmission spectra, exoplanets are often accompanied by supplementary planetary attribute data relating to factors like mass, surface gravity, and proximity to the host star. We refer to this contextual information as *planetary* metadata. Combinations of such metadata values characterize distinct data *subgroups* within the entire dataset. Typically, models are evaluated either globally on the entire evaluation set or within predefined, hand-picked subgroups.

We adopt efficient techniques to comprehensively compare model performance across all possible subgroups containing a minimum number of exoplanets without manually isolating some subgroups as particularly relevant. Naively assessing every possible combination of metadata criteria is impractical, given the combinatorial growth in possible subgroups. Instead, we adapt bias analysis methods, which identify and examine disparities in model performance across different data cohorts, to focus on "frequent" subgroups accounting for a significant and meaningful portion of the dataset [8]. The number of these statistically significant frequent subgroups does not scale in the same way as the total subgroup count. This allows the evaluation and comparison of models over relevant metadata configurations.

While being capable of analyzing a single model performance at the subgroup level, our approach also facilitates a meaningful comparison of different model performances across statistically significant frequent subgroups. This provides deeper insights beyond conventional evaluation methods reporting only average accuracy. Specifically, it reveals *gaps* (i.e., variations) in performance across metadata-defined cohorts that may not emerge at the global scale. Identifying where and how models succeed or struggle within distinct regions of the parameter space can guide the development of more robust solutions. This can help both determine the most promising models *for specific subsets* of exoplanets and guide the process of characterizing and addressing shortcomings in the existing models.

We propose an experimental section that uses synthetic observations simulated for the upcoming Ariel space telescope mission [9]. Through this data, we provide preliminary results demonstrating our proposed *Ex(o)plain* framework's capacity to identify and characterize patterns contributing to less accurate or degraded atmospheric reconstruction. We argue that interpretable machine learning presents a promising path toward unlocking scientific insights from the immense volumes of data expected from new astronomy efforts. By bringing transparency and explainability to complex models employed to analyze huge future exoplanet observations, our techniques aim to advance scientific understanding and develop robust approaches optimized for datasets of unprecedented scale.

The remainder of the paper is organized as follows: Section II focuses on examining relevant prior work, exploring studies on reconstructing exoplanet atmospheric parameters, and providing explanations and introspection into these modeling techniques. Section III reviews the problem under analysis, along with an exploration of the data and target task. Section IV introduces the proposed methodology's relevant aspects for identifying divergent subgroups. Section V reports the main experimental results obtained. Finally, Section VI draws conclusions and summarizes possible future directions. We release our code implementing these experiments in the project repository to enable reproducibility.[1]

II. RELATED WORKS

This section introduces the previous works of most relevance in the literature. In particular, it presents the main works regarding the reconstruction of atmospheric parameters via either traditional methods (e.g., Bayesian inference) or deep learning techniques. Next, the most relevant works in terms of the interpretability of exoplanetary deep learning techniques are presented.

A. RECONSTRUCTING EXOPLANET ATMOSPHERIC PARAMETERS

State-of-the-art atmospheric retrieval outcomes are currently achieved via traditional Bayesian inference employing

[1] https://github.com/koudounasalkis/Ex-o-plain

computationally intensive sampling algorithms [5]. However, when attempting to apply these resource-intensive methods at the massive scale of upcoming exoplanet survey datasets, such as those expected from Ariel, computational runtimes become impractical barriers [6].

Machine learning and deep learning techniques have increasingly been leveraged within exoplanetary research to address this challenge of scaling probabilistic modeling to big observational future resources. Applications have ranged from pre-processing tasks like data detrending [10], [11] debris removal [12], [13] and planet detection and characterization [14], [15], [16]. The aim has been to advance retrieval efficiency while maintaining characterization performance on anticipated huge volumes of spectral observations. Various ML and DL techniques, including random forests, convolutional neural networks (CNNs), and generative adversarial networks (GANs), have been utilized to enhance the efficiency and reduce the computational demands of atmospheric retrieval. However, such approaches frequently rely on approximations that can impact the precision of posterior probability estimates. A recent study [17] proposed a multimodal 1D-CNN architecture that integrates spectral measurements with supplementary stellar and planetary attribute data. The authors demonstrated that this technique delivers satisfactory outcomes when assuming normality in the posterior distribution while also revealing computationally efficient performance.

B. INTERPRETABILITY OF DL EXOPLANETARY MODELS

Recent works have developed techniques for automatically identifying data subgroups, i.e., subsets of the data characterized by a set of metadata attribute values, exhibiting problematic predictive behaviors in structured datasets [8], [18], [19]. Our proposal draws inspiration from DivExplorer [8], a method to detect diverging behaviors that explores frequent subgroups accounting for a meaningful portion of the dataset. DivExplorer enables the identification of subgroups where models perform well or struggle. Other heuristic-driven subgroup discovery approaches do not provide straightforward model performance comparisons [18], [19]. Instead, DivExplorer supports subgroup-wise performance comparisons over statistically significant data subgroups (i.e., subgroups with support above a given relevance threshold). As demonstrated in prior works [20], [21], this distinguishes DivExplorer as uniquely capable of enabling the in-depth subgroup-level model analyses we introduce here for exoplanetary retrieval applications. Our work builds upon DivExplorer's approach by introducing a framework for systematically benchmarking models within the problematic patterns that surface from exoplanet metadata-defined subgroups.

To the best of our knowledge, the only existing work that interprets deep learning models for exoplanet atmospheric retrievals is the research from [22]. Their analysis primarily quantified how predictions of one parameter varied given changes in other parameters. Like their technique, our method is agnostic to specific models. In contrast, we advance interpretable model evaluation by assessing performance at the subgroup level while considering *all* recurring subgroups defined by combined metadata criteria. Moreover, we explore the influence on predictions from both individual and combined stellar/planetary attribute metadata, offering richer insight compared to [22], which focused solely on conditioning parameter predictions. Therefore, we introduce a more comprehensive framework for identifying and characterizing attributional effects and performance disparities in exoplanet atmospheric characterization models.

III. EXOPLANETS ATMOSPHERE RECONSTRUCTION

Exoplanets are mainly discovered using techniques like measuring the radial velocity of a star or observing changes in its brightness when a planet passes in front of it (transit) [23]. When an exoplanet transits, it causes a slight yet detectable decrease in the star's brightness as viewed from Earth. This *dip* is affected by the planet's atmosphere, which absorbs specific wavelengths of light. Analyzing this dip at different wavelengths (also known as *transit spectrum*) helps astronomers learn about the atmosphere's composition and properties. However, challenges like observational noise and limited wavelength coverage can make gathering accurate information and interpreting the data harder.

Atmospheric retrieval aims to deduce parameters that yield the optimal mathematical fit to an observed transit spectrum, employing a forward radiative transfer model and optimization techniques. This is often characterized as an ''inverse problem'' [24] and aims to determine the posterior probability distribution of possible atmospheric parameter solutions given the acquired data. Within a Bayesian statistical framework, the posterior represents the probability of diverse atmospheric composition models aligning with the observed spectrum, providing a probabilistic description of the atmosphere knowledge acquired after accounting for prior information and new observational evidence.

A. DATA AND METRIC

Our experiments utilize spectral data from the Ariel Big Challenge Database [9]. For data generation, the authors used A LFNOOR [25], a tool designed to enhance the forward model and atmospheric retrieval capabilities of TauREx 3 [26] for large populations of exoplanet atmospheres. This tool automates telescope simulations and large-scale atmospheric retrievals, allowing the authors to produce for the ESA-Ariel mission 105,887 simulated forward observations and 26,109 standardized retrieval outputs. This dataset thus provides simulated observations of the light absorption patterns of exoplanet atmospheres across a range of wavelengths. The patterns in which different gases absorb wavelengths of light are unique, allowing the spectral data to function effectively as a proxy for atmospheric composition. This dataset, containing 2972 confirmed and 2928 candidate exoplanets, thus a total of 5900 unique objects, aims to

provide a realistic sample of what data obtained from the upcoming Ariel mission will produce. The task is to predict the combined probability distributions of six fundamental atmospheric properties - temperature and the logarithmic abundances of five gases (water, carbon dioxide, carbon monoxide, methane, and ammonia gas) - based solely on the observed spectrum for a given exoplanet. These predictions take the form of posterior distributions.

A commonly adopted choice to evaluate the performance of predicted atmospheric parameter distributions is the Two-Sample Kolmogorov-Smirnov (K-S) statistical test [27], [28]. The K-S test determines whether two data samples originate from the same underlying continuous probability distribution. The test is conducted with the null hypothesis assuming that the two samples originate from the same distribution. The test returns a score in the [0,1] range, where lower numbers signify greater similarity between the samples being compared, and a score of zero represents identical distributions. Specifically, we calculate the K-S statistic for each exoplanet individually, then average these values across a separate holdout test set to evaluate the overall performance of our models on new observational data.

IV. METHODOLOGY

This section outlines the key concepts that form the basis of our proposed framework, *Ex(o)plain*. We aim to design a model-agnostic methodology for identifying and characterizing the predominant trends and patterns that underlie degraded predictive capability when reconstructing exoplanet atmospheric properties. To begin, we introduce an existing deep learning-based technique previously applied in the literature to infer atmospheric parameters from observational data (Subsection Section §IV-A). While deep learning is only one possible approach, understanding this method provides a concrete context for the problem we aim to analyze. We then investigate how our framework practically explores the model's predictions to identify and characterize patterns (in terms of recurring subgroups) of mispredictions (Subsection Section §IV-B). By extracting these subgroups, our framework facilitates systematically exploring where and why reconstruction degrades. Instead of concentrating on a specific model or technique, our framework examines prediction errors at a broader level, allowing us to extract extensive insights applicable across diverse models.

A. ATMOSPHERIC PARAMETERS ESTIMATION

We analyze the model architecture presented in [17] to demonstrate our proposed explanation technique. As shown in Figure 1, this 1-dimensional convolutional neural network (1D-CNN) approach leverages both the exoplanet's spectral observations and any available auxiliary planetary metadata to infer the target atmospheric parameters.

The target available is provided as samples from the ground truth distribution. To predict such a distribution, we assume it follows a multivariate Gaussian distribution with full covariance. This simplifying assumption allows for a simple parametrization of the distribution and enables the modeling of interactions across dimensions. The spectral data is processed through a series of convolutional layers, allowing the model to learn informative features directly from this input modality. Meanwhile, for the auxiliary data, an inverse transformation is applied, and the polynomials up to degree 2 are extracted. Starting from the original 8 features, their procedure allows to obtain a total of 152 transformed auxiliary features, which are then fed straight into a feedforward network to encode this contextual information separately. The outputs from these two streams are then combined to produce the final atmospheric parameter prediction. An L1 loss function is used to minimize the difference between predicted and ground truth distributions, ensuring better convergence properties compared to KL divergence. More details can be found in [17].

We adopt an ensemble modeling scheme to help address potential variance issues of the approach proposed in [17] and boost prediction stability. Specifically, we train multiple instances of the base architecture on randomly sampled subsets of the training data. We average the mean and covariance predictions across all learners at inference time. While this method treats each model as equally weighted regardless of confidence, we find that even modest ensemble sizes of around 5-10 learners significantly outperform relying on a single model alone based on the empirical results obtained.

We expect the models characterizing the ensembles to display varying behaviors for different subgroups of the data (as defined in the next section): this is because each learner of the ensemble has been trained on a different sample of the original training set, and has been initialized differently.

B. SUBGROUPS IDENTIFICATION

In our framework, we characterize data subgroups using itemsets, which are collections of *attribute-value* pairs that describe aspects of the data.

For the models under consideration, we define the *divergence* of a data subgroup as the difference between a model's performance, i.e., its K-S score, on that subgroup and its overall performance on the entire dataset.

We typically observe different performance for different subgroups of data for a variety of reasons. Typically, the divergence in performance occurs because some subgroups are better represented in the training data (i.e., they are more commonly observed) than others.

Similarly, the *subgroup gain* refers to the difference in performance between two models for a specific subgroup.

When analyzing the exoplanet spectral data from Ariel [9], we leverage the accompanying planetary metadata, which includes eight interpretable attributes of star and planet properties: star distance, stellar mass, radius and temperature, planet mass, orbital period, semi-major axis, and surface gravity. Since this metadata is continuous, we discretize each attribute into ten bins of equal frequency for our initial analysis. This decision strikes a nice balance between

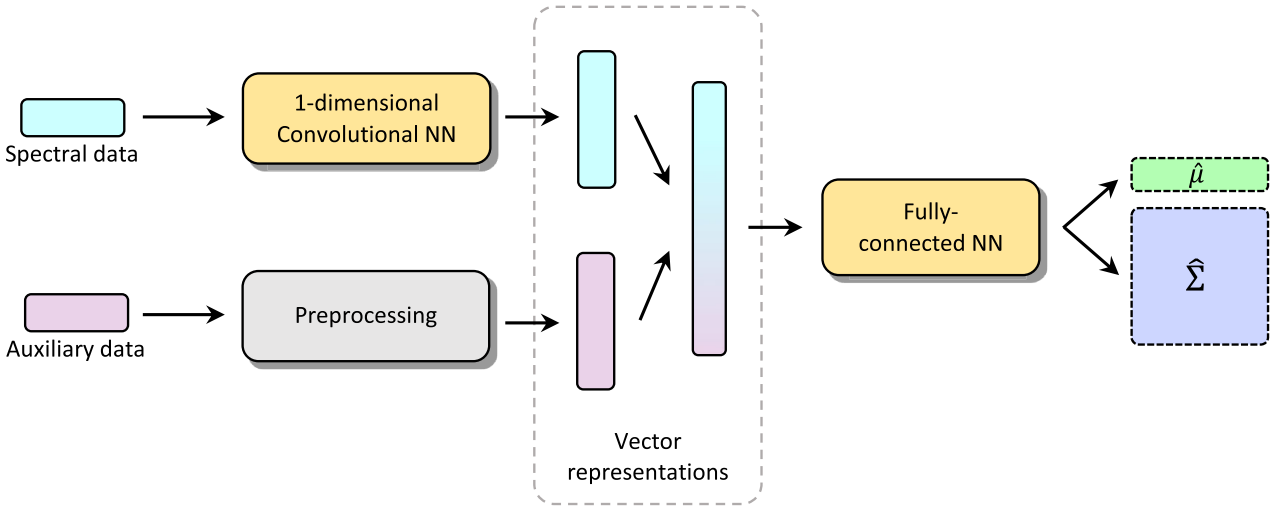


FIGURE 1. High-level architecture of the methodolgoy proposed in [17]. Spectral data is processed through a CNN, whereas the auxiliary data is preprocessed and concatenated to the CNN's output. A fully-connected neural network produces a mean vector and a covariance matrix.

subgroup granularity and interpretability. This level of discretization provides sufficient detail to capture meaningful variations in the data while ensuring that the resulting subgroups remain interpretable and manageable for analysis. This balance can be adjusted if needed based on specific case-by-case considerations. We can systematically identify recurrent patterns corresponding to high divergence and subgroup gain values between models by representing the data as items of attribute-value pairs. This provides insight into how specific regions of the attribute space influence the performance of reconstructions.

1) ITEMS AND ITEMSETS

Let D represent our dataset, $\mathcal{A}$ its set of metadata attributes, and $\mathcal{I}$ its set of *items*. Each item is defined as $a = v$, where a is an attribute $\in \mathcal{A}$ and v indicates the corresponding value. Since v represents a range of values, it can be written as an interval $[x, y)$. As such, an alternative representation for the item is $a \in [x, y)$. For instance, if we consider attributes like *planet surface gravity* and *planet distance*, examples of items could be $\{planet\_surface\_gravity \in [13.10, 18.38)\ ms^{-2}\}$ and $\{star\_radius \in [1.62, 6.30)\ R_{\odot}\}$.

The *subgroup* associated with a specific item refers to the portion of the data that meets the criteria defined by that item. For each attribute, the subgroups delineated by items should partition the dataset distinctly without overlapping and should collectively include all potential values of the attribute. For instance, considering the *"star radius"* attribute, the specified ranges should be non-overlapping, and together they should cover all feasible radii within a predefined range.

An *item* enables slicing or selecting a subset of data based on a single attribute. Moreover, we can conduct multi-attribute slicing by employing *itemsets*, which are collections of zero or more items. Each item within an itemset corresponds to a unique attribute. For example, an itemset could be represented as follows: $\{planet\_mass \in [0.01, 0.02)\ M_J, star\_radius \in [1.62, 6.30)\ R_{\odot}\}$.

The *support* of an itemset I is defined as the fraction of the dataset that satisfies the criteria specified by I. In other words, it is the ratio of the subgroup size meeting the criteria of I to the total size of the dataset. For example, an itemset with a support of 0.01 indicates that it is present in 1% of the dataset. The empty itemset, which represents the entire dataset, has a support of 1.

2) SUBGROUP DIVERGENCE AND GAIN

We first define a statistic measure f to quantify the performance in a specific modeling task. As previously mentioned, we will use the Kolmogorov-Smirnov (K-S) statistic as our metric f to evaluate predictions of atmospheric parameters.

We formally represent the performance of a given model M on any data subgroup, which we refer to as an itemset I. We denote $f(I, M)$ to indicate the value of the performance measure f when model M is applied only to the examples contained within the subgroup I. We introduce the concept of *divergence* to capture differences in how well a model behaves with specific subsets of the data compared to its overall performance. Specifically, the divergence related to a subgroup I for model M represents the difference between the performance on subgroup I alone $f(I, M)$ and the performance when evaluated over the entire dataset D as a whole $f(D, M)$. In formulas:

$$div_f(I, M) = f(I, M) - f(\emptyset, M)\ . \tag{1}$$

By quantifying the divergence, we aim to systematically identify any subgroups where a model's predictive capabilities degrade noticeably or improve significantly relative to its average behavior. As described in [8] and [20], this technique

TABLE 1. Summary of dataset characteristics, including planetary and stellar auxiliary information and the corresponding unites of measure (UoM), number of retrieved subgroups with DivExplorer (on the test set), and average time in seconds (across ten runs) of DivExplorer subgroup exploration.

Stellar Metadata [UoM]		*Planetary Metadata [UoM]*		*# Subgroups*	*Computation time (s)*
distance	$[pc]$	mass	$[M_J]$	63	0.20±0.07
mass	$[M_\odot]$	orbital period	$[days]$		
radius	$[R_\odot]$	semi-major axis	[AU]		
temperature	$[K]$	surface gravity	$[ms^{-2}]$		

provides insights into how attribute combinations that define specific data cohorts influence reconstruction quality.

We also establish the concept of *subgroup gain* when transitioning from model M_1 to model M_2 for a specific subgroup I. This gain refers to the improvement in performance achieved on the itemset I when model M_1 is substituted with model M_2:

$$gain_f(I, M_1, M_2) = f(I, M_2) - f(I, M_1)\ . \tag{2}$$

To identify meaningful itemsets with significant divergence or gain values in our data, we leverage the DivExplorer [8] tool. DivExplorer efficiently extracts all itemsets exceeding a predefined support threshold while calculating the associated divergence values. The support threshold parameter plays a crucial role in focusing our analysis. By setting a minimum level of data representation for itemsets, e.g., by ensuring they cover at least 1% of the dataset, we effectively filter out anomalies and retain patterns that are sufficiently significant to be considered operationally meaningful. DivExplorer also offers a redundancy pruning of the identified subgroups so that if subgroups S and $S \cup \{j\}$ have very close divergence, only S may be returned, the item j not having a significant effect. The following analysis utilizes a support threshold of 0.01 and a redundancy threshold of 0.03 to ensure that the retrieved subgroups are not entirely overlapped.

Table 1 provides an overview of relevant dataset characteristics and results from applying DivExplorer to our scenario. It includes details on the type of planetary and stellar metadata available as auxiliary information, the number of unique *frequent* subgroups identified by DivExplorer, i.e., those subgroups exceeding the support threshold, and the average runtime in seconds to perform the exhaustive subgroup exploration process across ten experimental runs.

3) SHAPLEY VALUES FOR LOCAL AND GLOBAL CONTRIBUTION

After having identified subgroups exhibiting significant divergence or gain through DivExplorer, we also aim to understand each attribute-value pair's, or item's, contribution to those metrics. Building upon principles from [8], we introduce the concept of an item's contribution to the overall metric value $g(I)$ of a subgroup I, where g represents either divergence or gain.

To quantify these individual item contributions, we adopt concepts from game theory by leveraging the *Shapley values*. In a team I of players, the *Shapley value* of $i \in I$ represents the value added by i, and is measured as the average of $g(J[:i^+]) - J[:i^-]$ over all permutations J of I, where $J[:i^+]$ is the prefix of J up to and including i, and $J[:i^-]$ is the prefix of J up to and excluding i. For a given subgroup I exhibiting divergence/gain, each constituent item i is assigned a Shapley contribution value $s_g(i, I)$ that captures the notion of how much i contributed to the divergence or gain of S. Notably, the sum of all item Shapley values within I will precisely equal the overall divergence or gain metric, i.e., $\sum_{i \in I} s_g(i, I) = g(I)$.

In addition, we examine an item i's *global* Shapley value, denoted $\mathcal{S}_g(i)$, which captures its average impact when included in all compatible subgroups. This provides a complementary view of how impactful an attribute-value pair, i.e., an item i, tends to be across the entire dataset, on average.

Collectively, the contributions of Shapley-based item analyses and global values enable a systematic exploration of the factors influencing non-negligible variations or improvements within subgroups.

V. EXPERIMENTAL RESULTS

We evaluate the effectiveness of our proposed *Ex(o)plain* framework through several analyses. First, we demonstrate its advantages in identifying granular and human-understandable challenging subgroups compared to traditional clustering techniques (§V-A). Second, we show its ability to uncover and characterize factors contributing to erroneous predictions for the ensemble model (§V-B) and one individual weak learner (§V-C). Third, we conduct a comparative subgroup-level assessment of the performance of different models to identify their differing areas of success and failure. We specifically compare at the subgroup level the ensemble vs. an individual learner (§V-D) and two weak learners (§V-E). Finally, we examine the overall influence of metadata attributes on model performance using global Shapley values (§V-F).

Experiments were run on a machine equipped with Intel® Core™ i9-10980XE CPU, 2 × Nvidia® RTX A6000 GPU, 128 GB of RAM running Ubuntu 22.04 LTS. We provide detailed information about the model used for the evaluation and the training procedure in the official project repository.

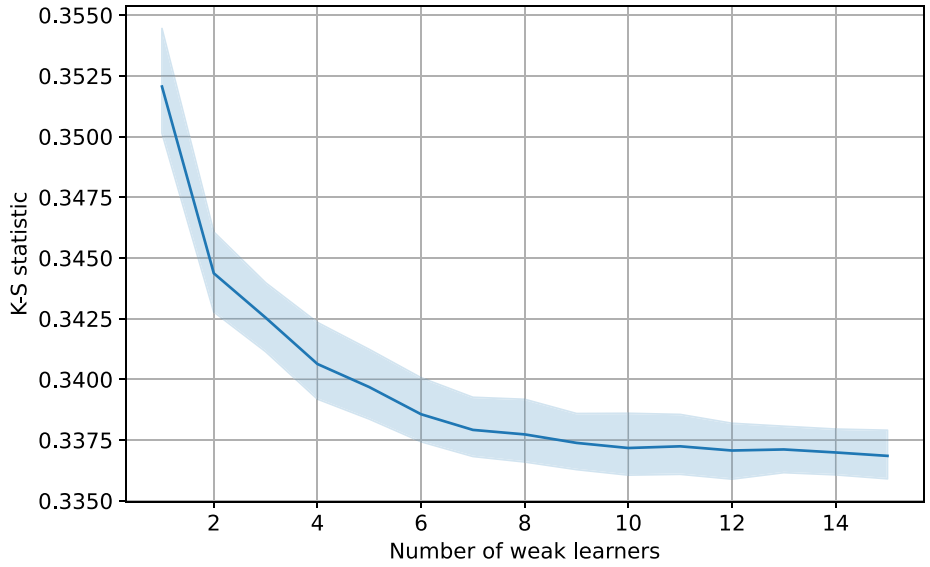


FIGURE 2. K-S statistic computed as the number of estimators in an ensemble increases (lower is better). The estimated parameters are obtained as the average across all learners. Error bars are obtained as the 95% confidence interval when running the experiment 10 times, with different initializations and train/test splits.

TABLE 2. Top-3 highest negatively divergent subgroups on performance for the *ensemble*. The *score* column denotes the K-S statistical test, Δ_{score} the divergence w.r.t. the performance on the whole set, while *t* indicates the Welch's t-test score.

Subgroup	*score*	Δ_{score}	t
{*planet_orbital_period* $\in$ *[34.76, 731.94) days*, *planet_surface_gravity* $\in$ *[13.10, 18.38)* ms^{-2}}	0.413	0.076	3.276
{*planet_semi_major_axis* $\in$ *[0.01, 0.04) AU*, *planet_surface_gravity* $\in$ *[4.36, 5.58)* ms^{-2}}	0.412	0.075	2.808
{*planet_semi_major_axis* $\in$ *[0.21, 1.50) AU*, *planet_surface_gravity* $\in$ *[13.10, 18.38)* ms^{-2}}	0.409	0.072	3.256

To begin, we demonstrate that employing an ensemble of models yields a notable boost in predictive accuracy w.r.t. individual classifiers, as quantified via K-S tests across multiple trials. Figure 2 illustrates this aspect, showing the balance between computational demands and predictive gains as the number of models in the ensemble increases. We find an ensemble size of approximately 5 achieves a reasonable trade-off between the computational cost required and the quality of the obtained performance. The KS improvement of approximately 0.02 when using an ensemble is statistically significant, as evidenced by the confidence intervals shown in the plot (Figure 2). Further, ensembling allows for a comparative study of individual models versus the ensemble, providing a nuanced understanding of performance enhancements and trade-offs at the subgroup level. With this in mind, the subsequent experimental analyses thus focus on studying a 5-model ensemble compared to a single baseline model (§V-D) or two weak learners together (§V-E).

A. COMPARISON WITH CLUSTERING

To assess the effectiveness of our approach compared to classical unsupervised clustering methods like K-Means, we investigate more granular subgroup definitions and compare the two approaches as such granularity changes. We extract, for both techniques, a number of subgroups ranging from 60 to 220. This range of values provides a reasonable number of subgroups that can be manually analyzed by a domain expert in a reasonable amount of time. For K-Means, we directly vary the number of extracted subgroups K. For DivExplorer, we vary the minimum support threshold so as to extract the desired number of subgroups. We consider, for each set of extracted subgroups, the average and the maximum divergence across all subgroups. This information is indicative of the usefulness of the extracted result since we aim to identify the most diverging subgroups.

Figure 3 highlights the differences between the two approaches. DivExplorer identifies more divergent results (both in terms of average, as well as maximum divergence) when the number of extracted subgroups is low. K-Means, instead, produces better representations (only in terms of maximum divergence identified) when a large number of subgroups is extracted. This is an interesting behavior that is worth noting. However, it should be pointed out that keeping the number of extracted subgroups low is generally

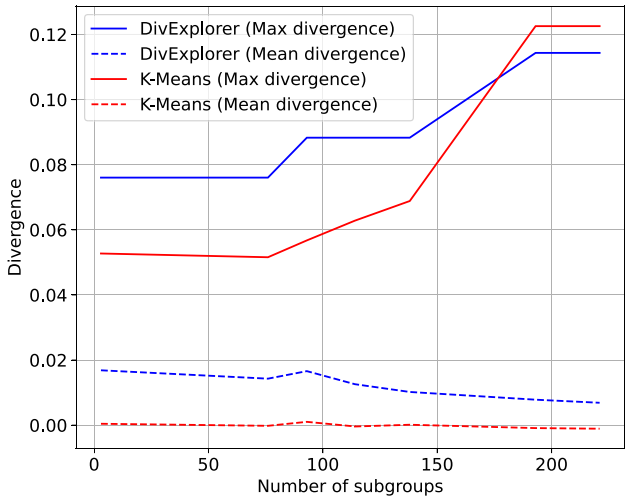


FIGURE 3. Maximum divergence (continuous lines) and mean divergence (dashed lines) for the subgroups extracted with DivExplorer (blue lines) and K-Means (red lines), as the number of extracted subgroups increases. Larger values mean that more problematic groups are discovered. A smaller number of subgroups is desired, since it can be more easily examined by domain experts.

desired since it provides a more succinct overview of the problem without overloading the domain expert with excessive information.

As such, we find DivExplorer to be the generally better alternative in terms of the usefulness of the extracted subgroups. Additionally, we emphasize that DivExplorer subgroups are generally more interpretable than those extracted with K-Means since they provide only information on specific slices of interest of the metadata under analysis instead of providing a (generally hard-to-interpret) ''centroid''. Thus, without sacrificing performance, we achieve granular and human-understandable subgroups. Due to these reasons, we obtain subgroups via DivExplorer in the following analyses.

B. INDIVIDUAL MODEL ANALYSIS: ENSEMBLE

As an initial demonstration of our framework's capabilities, we consider the ensemble model and detect and characterize all subgroups that either underperform or outperform relative to the average behavior on the entire dataset. Note that the number of samples considered in the explored subgroups is sufficient in both the train and test sets, as confirmed by Welch's t-test. The same conclusion also holds regarding the largest gains and drops in performance (Sections V-D and V-E). The subgroups showing the most significant changes in performance have been thoroughly analyzed, and Welch's t-test further supports that the sample sizes are sufficient for both training and testing.

1) NEGATIVE DIVERGENCE

We start by investigating the origins of errors made by the ensemble model. In Table 2, we present details on the top 3 itemsets exhibiting the most statistically significant negative divergence from the average performance, as measured with Welch's t-test score.

Interestingly, we find that the planet's surface gravity highly impacts the results. It appears in all three most negatively divergent subgroups, with values falling within the ranges of either 13.10 to 18.38 ms^{-2} or 4.36 to 5.58 ms^{-2}.

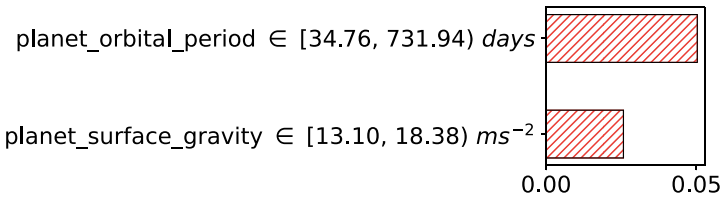
The largest drop in performance of 0.076 is observed when considering planets with orbital periods from $34.76 - 731.94$ *days* and surface gravities from 13.10 to 18.38 ms^{-2}. This subgroup analysis thus points to surface gravity, particularly its intermediate values, as prominently related to degraded prediction quality for specific planetary configurations.

When examining underperforming subgroups, it is also important to understand the relative influence of each metadata attribute in contributing to or reducing the divergence. As defined in Section IV, we quantify this significance using Shapley values. Figure 4 plots the Shapley values within the three most negatively divergent subgroups listed in Table 2. The analysis reveals that within the first subgroup (Figure 4(a)), a high planet orbital period influences the performance more than the surface gravity. In the second and third subgroups (Figures 4(b) and 4(c)), the planet's semi-major axis has a greater influence compared to surface gravity. This Shapley value attribution provides deeper context on the subgroups, indicating planet's orbital period or semi-major axis may play a more pivotal role than its surface gravity alone in certain configurations.

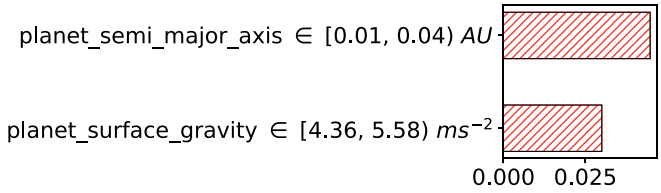
2) POSITIVE DIVERGENCE

In addition to retrieving sources of degraded performance, our framework also enables the identification of subgroups where the model (i.e., the ensemble in our analysis) notably outperforms its average predictions. Table 3 lists the three subgroups exhibiting the most improved scores. As can be seen, these subgroups contain attributes descriptive of both the host star and planet. The largest boost in performance quantified as $|0.055|$ is seen in the subgroup defined by stellar radius ranging from 1.21 to 1.35 Solar Radii $R_{\odot}$ and planetary semi-major axis spanning 0.04 to 0.05 AU. This demonstrates our approach is effective in identifying not only error-prone patterns but also cases where certain traits lead to significantly enhanced prediction quality.

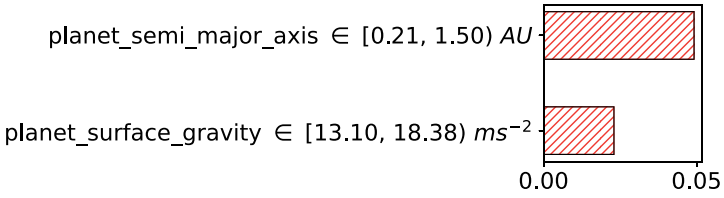
Interestingly, the item *"planet semi major axis* $\in [0.04, 0.05)$ *AU"* defines one of the best-performing subgroups as well as one of the worst, as shown previously in Table 2. Analyzing the Shapley values for each variable's contribution to subgroup divergence (Figure 5) reveals the relative contribution of each item within these positively divergent subgroups. In the most positively divergent subgroup (Figure 5(a)), the host star radius has a greater impact on performance than the planet's semi-major axis. Similarly, for the second-best subgroup (Fig. 5(b)), stellar properties are more influential drivers of the enhanced prediction quality. In contrast, planetary attributes outweigh stellar factors for the third-best subgroup (Fig. 5(c)). Once again, this Shapley analysis provides nuanced insight, showing how star



(a) 1st, $\Delta_{score} = 0.076$



(b) 2nd, $\Delta_{score} = 0.075$



(c) 3rd $\Delta_{score} = 0\,072$

FIGURE 4. ENSEMBLE. **Item contribution to the final score for (a) the subgroup with the highest negative divergence (the lower the score, the better), (b) the second, and (c) the third.**

TABLE 3. **Top-3 highest positively divergent subgroups on performance for the *ensemble*.**

Subgroup	*score*	Δ_{score}	t
{*star_radius* ∈ *[1.21, 1.35)* $R_\odot$, *planet_semi_major_axis* ∈ *[0.04, 0.05) AU*}	0.282	-0.055	3.769
{*planet_orbital_period* ∈ *[2.78, 3.66) days*, *star_distance* ∈ *[244.97, 348.43) pc*}	0.284	-0.053	2.651
{*star_mass* ∈ *[1.01, 1.07)* $M_\odot$, *planet_orbital_period* ∈ *[4.64, 6.37) days*}	0.287	-0.051	2.447

and planet attributes differently shape model performance depending on the precise configuration.

C. INDIVIDUAL MODEL ANALYSIS: WEAK LEARNER

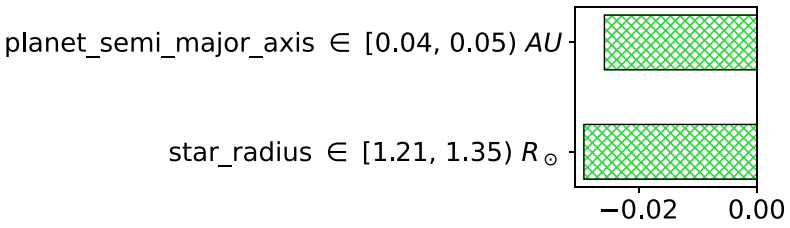
Here, we replicate the analysis for debugging an individual weak learner model, rather than an ensemble model. Since our approach is entirely model agnostic, we can examine the behavior of any model, requiring only its set of final predictions. This analysis provides additional context on how challenging subgroups for single models may be mitigated through our proposed framework by systematically evaluating their performance at the subgroup level.

1) NEGATIVE DIVERGENCE

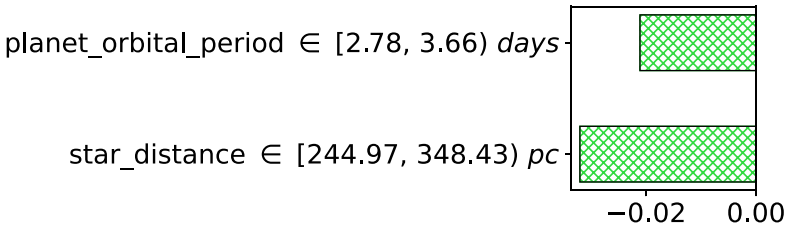
We consider the origins of errors made by an individual weak learner. Table 4 details the top 3 subgroups exhibiting the worst statistically significant performance compared to the average behavior.

Notably, we observe that the top two identified subgroups are consistent with those previously observed in the ensemble model, as indicated in Table 2. This suggests that the ensemble model encounters challenges with the same subgroups where an individual weak learner struggles, even though the ensemble consistently outperforms the individual model. Interestingly, the degree of divergence (represented by Δ_{score} in the table) between the two models is not significantly different; for the first subgroup, it's 0.076 for the ensemble (see Table 2) and 0.077 for the individual model. Similarly, for the second subgroup, we obtained a 0.075 Δ_{score} for the ensemble and 0.076 for the individual model.

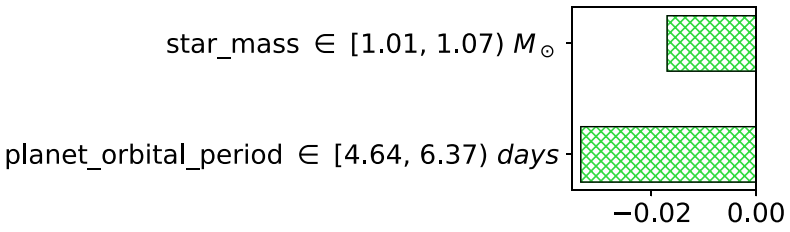
The most substantial decline in performance, a drop of 0.077, thus occurs when examining planets with orbital



(a) 1st, $\Delta_{score} = -0.055$



(b) 2nd, $\Delta_{score} = -0.053$



(c) 3rd, $\Delta_{score} = -0.051$

FIGURE 5. ENSEMBLE. Item contribution to the final score for (a) the subgroup with the highest positive divergence (the lower the score, the better), (b) the second, and (c) the third.

TABLE 4. Top-3 highest negatively divergent subgroups on performance for a *weak learner*.

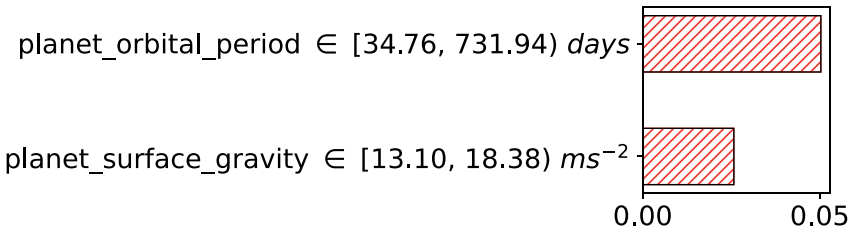
Subgroup	*score*	Δ_{score}	t
{*planet_orbital_period* ∈ *[34.76, 731.94) days, planet_surface_gravity* ∈ *[13.10, 18.38)* ms^{-2}}	0.424	0.077	2.929
{*planet_semi_major_axis* ∈ *[0.01, 0.04) AU, planet_surface_gravity* ∈ *[4.36, 5.58)* ms^{-2}}	0.423	0.076	2.770
{*planet_semi_major_axis* ∈ *[0.21, 1.50) AU, planet_mass* ∈ *[0.01, 0.02)* M_J}	0.420	0.073	2.990

TABLE 5. Top-3 highest positively divergent subgroups on performance for a *weak learner*.

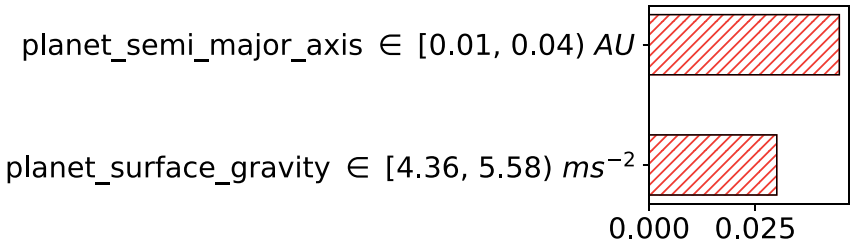
Subgroup	*score*	Δ_{score}	t
{*planet_orbital_period* ∈ *[2.78, 3.66) days, star_distance* ∈ *[244.97, 348.43) pc*}	0.290	-0.057	3.105
{*planet_semi_major_axis* ∈ *[0.04, 0.05) AU, star_distance* ∈ *[244.97, 348.43) pc*}	0.293	-0.054	3.263
{*planet_orbital_period* ∈ *[4.64, 6.37) days, planet_mass* ∈ *[0.29, 0.35)* M_J}	0.295	-0.052	3.001

periods ranging from 34.76 to 731.94 *days* and surface gravities from 13.10 to 18.38 ms^{-2}.

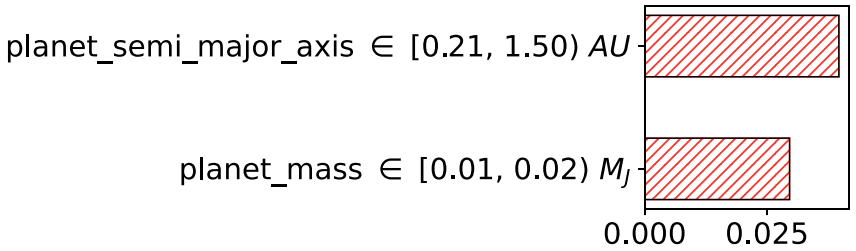
As we have previously highlighted, when assessing underperforming subgroups, it is also crucial to understand the



(a) 1st, $\Delta_{score} = 0.077$



(b) 2nd, $\Delta_{score} = 0.076$



(c) 3rd, $\Delta_{score} = 0.073$

FIGURE 6. WEAK LEANER. Item contribution to the final score for (a) the subgroup with the highest negative divergence (the lower the score, the better), (b) the second, and (c) the third.

relative importance of each metadata attribute in contributing to or reducing the performance divergence. We measure this importance using Shapley values. Figure 6 depicts the Shapley values within the three most negatively divergent subgroups as listed in Table 4.

This analysis reveals that in the first subgroup (Figure 6(a)), a high planet orbital period carries more explanatory weight than surface gravity. In the second and third subgroups (Figures 6(b) and 6(c)), the planet's semi-major axis yields a higher influence compared to surface gravity and mass, respectively.

2) POSITIVE DIVERGENCE

As we have already anticipated, our framework does not just identify sources of decreased performance and allows us to detect subgroups where the model performs notably better than its average predictions. This demonstrates *Ex(o)plain*'s effectiveness in providing interpretable insights into where the model excels at a granular data level.

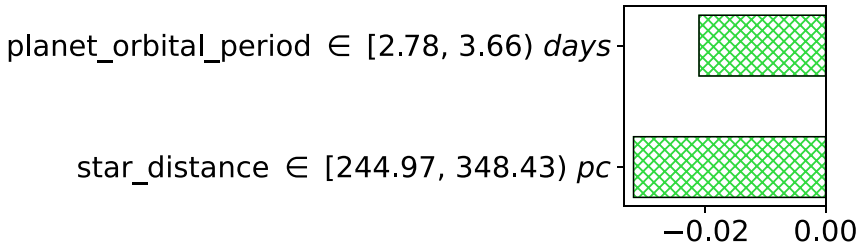
Table 5 presents the top three configurations that exhibit the most significant improvement scores for the weak learner under analysis. These subgroups include attributes of both the host star and the planet. However, in contrast to the negatively divergent subgroups, the ones identified here are very different from those found for the ensemble model, as shown in Table 3. This implies that the subgroups where an ensemble exhibits improved performance are not necessarily the same as those where its individual weak learners perform at their best.

The most substantial performance boost, quantified at 0.057, is found in the subgroup defined by a stellar distance ranging from 244.97 to 348.43 *pc* and a planetary orbital period spanning from 2.78 to 3.66 *days*.

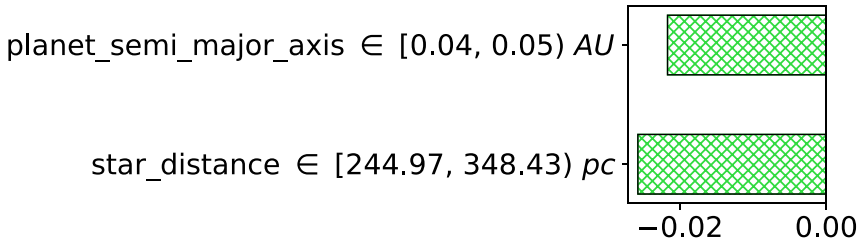
By looking again at the Shapley values, in the top subgroup (Figure 7(a)), the host star's distance has a more significant impact on performance than the planet's orbital period. Similarly, for the second-best subgroup (Figure 7(b)), stellar properties play a more substantial role w.r.t. planetary ones in driving the improved prediction quality.

D. MODELS COMPARISON ANALYSIS: ENSEMBLE VS. WEAK LEARNER

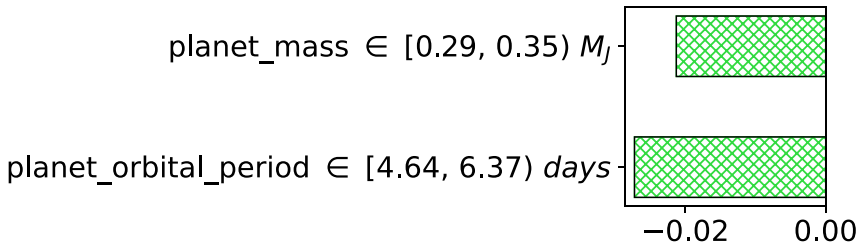
Overall, the ensemble's score of 0.336 improves upon the single model's 0.347, indicating better predictive quality.



(a) 1st, $\Delta_{score} = -0.057$



(b) 2nd, $\Delta_{score} = -0.054$



(c) 3rd, $\Delta_{score} = -0.052$

FIGURE 7. WEAK LEARNER. Item contribution to the final score for (a) the subgroup with the highest positive divergence (the higher the score, the worse), (b) the second, and (c) the third.

TABLE 6. Performance gain when transitioning from *individual* to *ensemble* models on itemsets where performance increases the most (↑), decreases (↓), or remains equal (=).

	Subgroups	$gain_{score}$	$individual_{score}$	$ensemble_{score}$
↑	{*star_radius* ∈ *[1.62, 6.30)* $R_\odot$, *planet_surface_gravity* ∈ *[18.38, 244.88)* ms^{-2}}	-0.041	0.381	0.340
=	{*star_radius* ∈ *[0.30, 0.72)* $R_\odot$, *planet_mass* ∈ *[0.01, 0.02)* M_J}	0.00	0.323	0.323
↓	{*star_radius* ∈ *[0.72, 0.81)* $R_\odot$, *planet_semi_major_axis* ∈ *[0.04, 0.05)* *AU*}	0.017	0.327	0.344

We recall that lower values for the K-S test indicate better performance. This increase in score further suggests that the ensemble indeed surpasses the individual learners' performance.

To analyze subgroup dynamics, we calculate the gain as the difference in performance between single and ensemble models within each data cohort. Our analysis found performance increased in 91.83% of the retrieved subgroups and decreased in the 8.16% of them. This indicates that while the overall K-S score of the ensemble approach is better than the one of the single model, as shown in Figure 2, there are some subgroups for which performance decreases. One possible explanation for this behavior is that the characteristics of such subgroups may not be well-captured by the ensemble, leading to a lack of consensus among the models, which can reduce the ensemble's ability to generalize. Figure 8

TABLE 7. Performance gain when changing one *individual weak learner* with another on itemsets where performance increases the most (↑), decreases (↓), or remains equal (=).

	Subgroups	$gain_{score}$	$individual1_{score}$	$individual2_{score}$
↑	{*star_radius* ∈ *[1.21, 1.35)* $R_\odot$, *star_temperature* ∈ *[6234.80, 10170.00)* K}	-0.027	0.358	0.331
=	{*star_radius* ∈ *[0.30, 0.72)* $R_\odot$, *planet_mass* ∈ *[0.01, 0.02)* M_J}	0.00	0.323	0.323
↓	{*star_radius* ∈ *[1.62, 6.30)* $R_\odot$, *star_temperature* ∈ *[4830.40, 5187.00)* K}	0.048	0.370	0.418

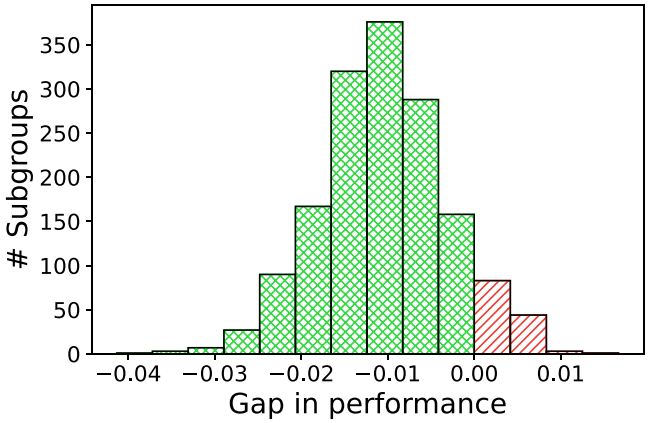


FIGURE 8. Distribution of gain ensemble-individual. Cross-hatched green denotes performance improvement when going from the individual model to the ensemble (the lower the score, the better), while red indicates performance decrease.

depicts this gain distribution when transitioning from the individual weak learner to the ensemble model, with green cross-hatching indicating improved subgroups and red worsening.

Table 6 highlights the top subgroups where transitioning to the ensemble most significantly increases (↑), negligibly changes (=), or decreases (↓) performance. The largest improvement (−0.041) involved planets with surface gravities in [18.38, 244.88) ms^{-2} orbiting stars of radii [1.62, 6.30) $R_\odot$. This comparative analysis at the subgroup level reinforces the ensemble's ability to enhance predictions across diverse planetary configurations.

As with prior analyses, these subgroups formed by combining planetary and stellar features exhibit varying behavior across different models. To gain more insights and understand which attributes contribute the most, we investigate Shapley values, that quantify each feature's influence on subgroup gain when transitioning from one model (i.e., the weak learner) to the other (i.e., the ensemble).

Figure 9 shows Shapley values within the largest positive (Figure. 9(a)) and negative (Figure 9(b)) gain subgroups. In the subgroup with the most improved performance, stellar radius has a greater effect than planetary surface gravity or semimajor axis. Interestingly, in the second subgroup, where individual and ensemble performance us unchanged, planetary mass clearly contributes more than the stellar radius (Figure 9(b)).

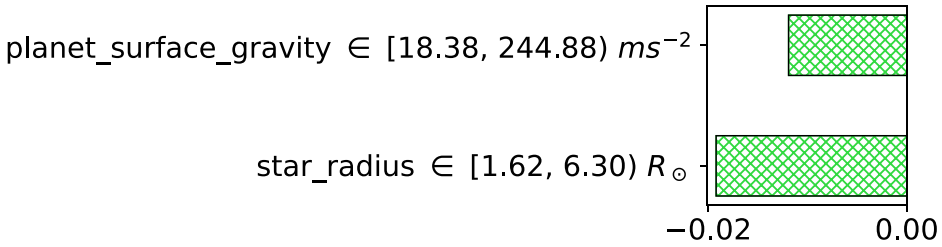
E. MODEL COMPARISON ANALYSIS BETWEEN TWO WEAK LEARNERS

We conduct here an additional comparative analysis between two distinct individual weak learners rather than against the ensemble model. We aim to highlight the advantages and limitations observed when substituting one baseline model for another. By exploring subgroup performance behavior between weak learners without ensemble averaging, our goal is to provide further context on how model selection impacts predictions across diverse exoplanet configurations. Such analyses offer complementary insight to the previous experiments benchmarking individual models against ensembles, reinforcing the effectiveness of our proposed framework for conducting detailed, interpretable debugging of model behavior at a granular level.

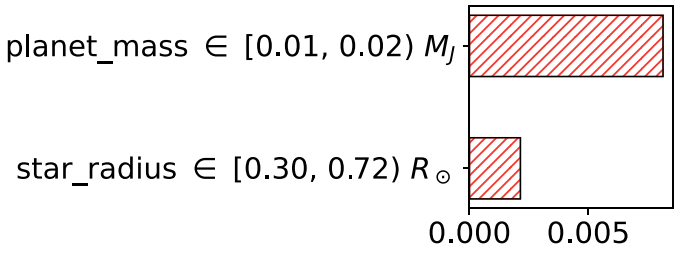
We compare a weak model with an overall score of 0.347 and another model with a K-S score of 0.352. While we could have compared any two models, we chose models with similar K-S scores for this comparison. Similar results and insights can be observed with different models as well. At the overall level, these two models are quite similar in performance. We recall that in the K-S test, lower values indicate better performance. Our analysis revealed that transitioning from the first to the second model is advantageous for 32.01% of the examined subgroups while not for 67.98% of them.

Figure 10 illustrates this gain distribution when transitioning from one individual weak learner to the other, with green cross-hatching indicating improved subgroups and red signifying a decline.

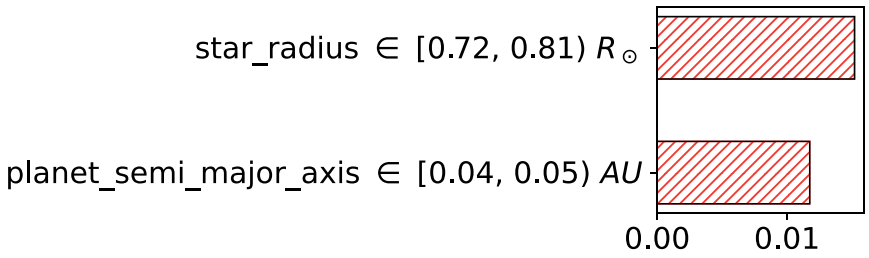
Table 7 highlights the top subgroups where the transition from the first weak learner model to the second most significantly increases (indicated by ↑), negligibly changes (indicated by =), or decreases (indicated by ↓) performance. The most notable improvement (a reduction of −0.027) is observed in the subgroup comprising planets orbiting stars with radii in the range of [1.21, 1.35) $R_\odot$ and temperatures in the range of [6234.80, 10170.00) K. Interestingly, the highest decrease in performance (equal to 0.048) is found in the subgroup including the same two metadata, but with different



(a) $\uparrow \Delta_{score} = -0.041$: Highest improvement



(b) $= \Delta_{score} = 0.00$: No improvement



(c) $\downarrow \Delta_{score} = 0.017$: Highest decrease

FIGURE 9. SUBGROUP GAIN. Item contribution to the gain when comparing the ensemble with individual models. (a) the subgroup with the highest improvement when transitioning to the ensemble, (b) the subgroup for which the models perform equally, and (c) the subgroup with the highest decrease.

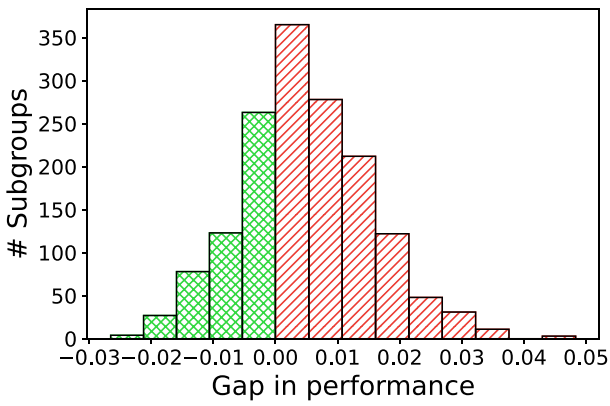


FIGURE 10. Distribution of gain individual-individual. Cross-hatched green denotes performance improvement when going from the first weak learner model to the second (the lower the score, the better), while red indicates performance decrease.

range values (stars with radii in the range of [1.62, 6.30) $R_\odot$ and temperatures in the range of [4830.40, 5187.00) K).

F. GLOBAL DIVERGENCE

We finally provide a global evaluation of each item's impact on two facets: the performance of the ensemble model, and the performance gain achieved when transitioning from individual to the ensemble model. This assessment uses global Shapley value $\mathcal{S}_g(i)$, where positive values indicate including an item i in subgroup I (when the item $i \notin I$) typically degrades predictive performance.

Figure 11 depicts the 15 items with the strongest performance influence on the ensemble based on $\mathcal{S}_g$. The greatest gains are observed for narrow ranges of small planet semi-major axis (i.e., $\in [0.06, 0.07)$ AU) and surface gravity (i.e., $[8.19, 9.51)$ ms^{-2}). Conversely, higher range values for these attributes relate to decreased K-S scores.

Additionally, Figure 12 displays the top 15 items most substantially impacting gain when transitioning from the weak individual model to the ensemble model based on such global Shapley values. Differently from before, this comparison shows that the greatest improvements and degradations are linked to host star characteristics. The peak increase is associated with high stellar temperatures

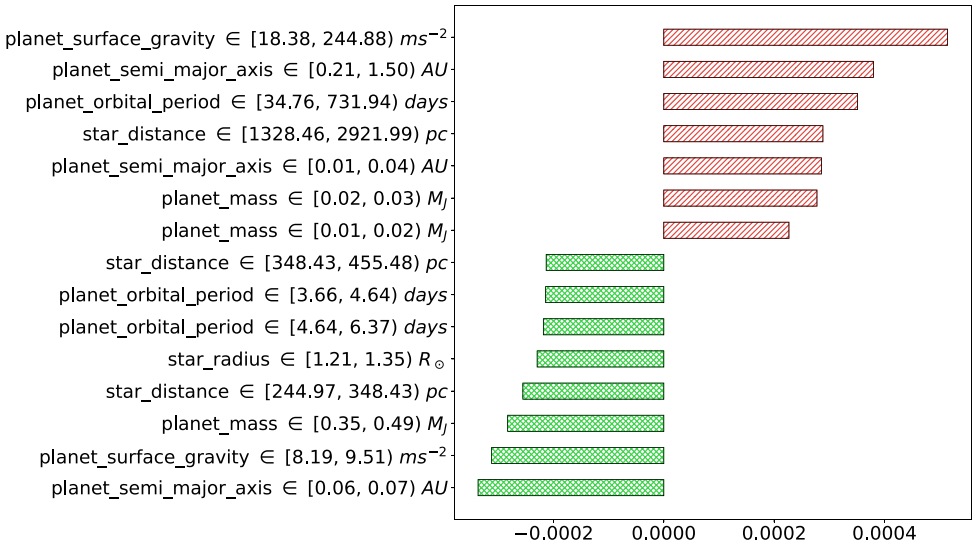


FIGURE 11. Global Shapley values of the top-15 items for the ensemble. Cross-hatched green denotes performance improvement, while red indicates performance decrease.

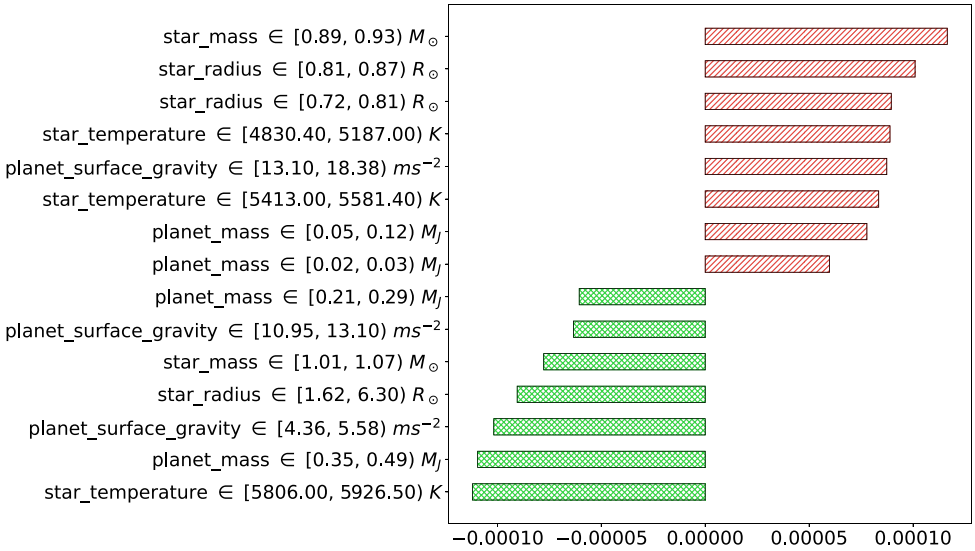


FIGURE 12. Global Shapley values (top-15 items) of the gain when transitioning from individual to ensemble. Cross-hatched green denotes performance improvement, while red indicates performance decrease.

in the range of 5806.00 to 5926.50 K. Conversely, the highest reductions involve stars with a relatively small mass ($[0.89, 0.93)\ M_\odot$) and radii showing either $[0.81, 0.87)\ R_\odot$ or $[0.72, 0.81)\ R_\odot$ ranges.

VI. CONCLUSION

In this work, we explored the adoption of interpretability techniques to offer descriptions of the situations of degraded performance that can occur when estimating atmospheric parameters of exoplanets. We do so by identifying frequent subgroups (as defined by auxiliary data regarding the star/planet system) for which a significant variation in performance can be observed.

Although the proposed results concern a synthetic dataset, we argue that the same technique can and should be used to provide better insights regarding the contexts that can produce degraded (or improved) reconstructions. This is helpful both as a way of weighting the results obtained (as well as their validity) and for identifying situations in the current experimental design that could benefit from additional work.

We additionally show that ensembles of learners, as is well known in the literature, produce overall better performance when compared to single weak learners. We further provide descriptions of the most relevant changes in subgroups' performance that justify this change.

We plan on further expanding the process of identifying degraded performance to produce more meaningful descriptions that do not rely on the discretization of the available auxiliary information. We are also interested in improving the performance of the learners adopted by, for example, leveraging loss functions that better reflect the nature of the estimated parameters.

REFERENCES

[1] S. Seager and D. D. Sasselov, "Theoretical transmission spectra during extrasolar giant planet transits," *Astrophysical J.*, vol. 537, no. 2, pp. 916–921, Jul. 2000.
[2] J. P. Gardner, J. C. Mather, M. Clampin, R. Doyon, M. A. Greenhouse, H. B. Hammel, J. B. Hutchings, P. Jakobsen, S. J. Lilly, K. S. Long, and J. L. Lunine, "The James webb space telescope," *Space Sci. Rev.*, vol. 123, pp. 485–606, 2006.
[3] G. Tinetti, P. Drossart, P. Eccleston, P. Hartogh, A. Heske, J. Leconte, G. Micela, M. Ollivier, G. Pilbratt, L. Puig, and D. Turrini, "The science of ariel (atmospheric remote-sensing infrared exoplanet large-survey)," in *Space Telescopes and Instrumentation 2016: Optical, Infrared, and Millimeter Wave*, vol. 9904. Bellingham, WA, USA: SPIE, 2016, pp. 658–667.
[4] M. Aubin, C. Cuesta-Lazaro, E. Tregidga, J. Viaña, C. Garraffo, I. E. Gordon, M. López-Morales, R. J. Hargreaves, V. Y. Makhnev, J. J. Drake, D. P. Finkbeiner, and P. Cargile, "Simulation-based inference for exoplanet atmospheric retrieval: Insights from winning the ariel data challenge 2023 using normalizing flows," 2023, *arXiv:2309.09337*.
[5] N. Madhusudhan, *Atmospheric Retrieval of Exoplanets*. Cham, Switzerland: Springer, 2018, pp. 2153–2182.
[6] C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt, "Advances in variational inference," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 2008–2026, Aug. 2019.
[7] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 44, pp. 22071–22080, Oct. 2019.
[8] E. Pastor, L. de Alfaro, and E. Baralis, "Looking for trouble: Analyzing classifier behavior via pattern divergence," in *Proc. Int. Conf. Manage. Data*, New York, NY, USA, Jun. 2021, pp. 1400–1412.
[9] Q. Changeat and K. H. Yip, "ESA-ariel data challenge NeurIPS 2022: Introduction to exo-atmospheric studies and presentation of the Atmospheric Big Challenge (ABC) database," *RAS Techn. Instrum.*, vol. 2, no. 1, pp. 45–61, Jan. 2023, doi: 10.1093/rasti/rzad001.
[10] M. Morvan, A. Tsiaras, N. Nikolaou, and I. P. Waldmann, "PyLightcurve-torch: A transit modeling package for deep learning applications in PyTorch," *Publications Astronomical Soc. Pacific*, vol. 133, no. 1021, Mar. 2021, Art. no. 034505.
[11] J. E. Krick, J. Fraine, J. Ingalls, and S. Deger, "Random forests applied to high-precision photometry analysis with spitzer IRAC," *Astronomical J.*, vol. 160, no. 3, p. 99, Sep. 2020.
[12] S. M. Lawler and B. Gladman, "Debris disks in Kepler exoplanet systems," *Astrophysical J.*, vol. 752, no. 1, p. 53, Jun. 2012.
[13] A. Koudounas, F. Giobergia, and E. Baralis, "Time-of-flight cameras in space: Pose estimation with deep learning methodologies," in *Proc. IEEE 16th Int. Conf. Appl. Inf. Commun. Technol. (AICT)*, Oct. 2022, pp. 1–6.
[14] F. Ardevol Martinez, M. Min, I. Kamp, and P. I. Palmer, "Convolutional neural networks as an alternative to Bayesian retrievals," 2022, *arXiv:2203.01236*.
[15] K. H. Yip, Q. Changeat, B. Edwards, M. Morvan, K. L. Chubb, A. Tsiaras, I. P. Waldmann, and G. Tinetti, "On the compatibility of ground-based and space-based data: Wasp-96 B, an example," *Astronomical J.*, vol. 161, no. 1, p. 4, 2020.
[16] K. Hou Yip, Q. Changeat, A. Al-Refaie, and I. Waldmann, "To sample or not to sample: Retrieving exoplanetary spectra with variational inference and normalising flows," 2022, *arXiv:2205.07037*.
[17] F. Giobergia, A. Koudounas, and E. Baralis, "Reconstructing atmospheric parameters of exoplanets using deep learning," in *Proc. IEEE 17th Int. Conf. Appl. Inf. Commun. Technol. (AICT)*, Oct. 2023, pp. 1–6.
[18] Y. Chung, T. Kraska, N. Polyzotis, K. H. Tae, and S. E. Whang, "Automated data slicing for model validation: A big data–AI integration approach," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 12, pp. 2284–2296, Dec. 2020.
[19] S. Sagadeeva and M. Boehm, "SliceLine: Fast, linear-algebra-based slice finding for ML model debugging," in *Proc. Int. Conf. Manage. Data*, New York, NY, USA, Jun. 2021, pp. 2290–2299.
[20] A. Koudounas, E. Pastor, G. Attanasio, V. Mazzia, M. Giollo, T. Gueudre, L. Cagliero, L. de Alfaro, E. Baralis, and D. Amberti, "Exploring subgroup performance in end-to-end speech models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.
[21] A. Koudounas, E. Pastor, G. Attanasio, V. Mazzia, M. Giollo, T. Gueudre, E. Reale, L. Cagliero, S. Cumani, L. de Alfaro, E. Baralis, and D. Amberti, "Towards comprehensive subgroup performance analysis in speech models," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 32, pp. 1468–1480, 2024.
[22] K. H. Yip, Q. Changeat, N. Nikolaou, M. Morvan, B. Edwards, I. P. Waldmann, and G. Tinetti, "Peeking inside the black box: Interpreting deep-learning models for exoplanet atmospheric retrievals," *Astronomical J.*, vol. 162, no. 5, p. 195, Nov. 2021.
[23] Q. Changeat and K. H. Yip, "ESA-ariel data challenge NeurIPS 2022: Introduction to exo-atmospheric studies and presentation of the atmospheric big challenge (ABC) database," *RAS Techn. Instrum.*, vol. 2, no. 1, pp. 45–61, Jan. 2023.
[24] R. Potthast, "A survey on sampling and probe methods for inverse problems," *Inverse Problems*, vol. 22, no. 2, pp. R1–R47, Apr. 2006.
[25] Q. Changeat, A. F. Al-Refaie, B. Edwards, I. P. Waldmann, and G. Tinetti, "An exploration of model degeneracies with a unified phase curve retrieval analysis: The light and dark sides of WASP-43 B," *Astrophysical J.*, vol. 913, no. 1, p. 73, May 2021.
[26] A. F. Al-Refaie, Q. Changeat, I. P. Waldmann, and G. Tinetti, "TauREx 3: A fast, dynamic, and extendable framework for retrievals," *Astrophysical J.*, vol. 917, no. 1, p. 37, Aug. 2021.
[27] A. N. Kolmogorov, "Sulla determinazione empirica di una legge didistribuzione," *Giorn Dell'inst Ital Degli Att*, vol. 4, pp. 89–91, Aug. 1933.
[28] N. Smirnov, "Table for estimating the goodness of fit of empirical distributions," *Ann. Math. Statist.*, vol. 19, no. 2, pp. 279–281, Jun. 1948.

ALKIS KOUDOUNAS (Graduate Student Member, IEEE) received the B.S. degree in computer and automation engineering from the Universitã Politecnica delle Marche, in 2019, and the M.S. degree in computer engineering from the Politecnico di Torino, in 2021, where he is currently pursuing the Ph.D. degree in computer engineering, with previous internship experience in research on object detection and pose estimation with Thales Alenia Space, Turin, and in optimization algorithms with TUAT, Tokyo. His current research interests include spoken language understanding, audio-text multimodal understanding, and explainable AI.

FLAVIO GIOBERGIA (Member, IEEE) received the master's degree in computer engineering from the Politecnico di Milano and the Ph.D. degree in computer and control engineering from the Politecnico di Torino. He is currently an Assistant Professor with the Politecnico di Torino and an Adjunct Professor with the Università del Piemonte Orientale. His current research interests include machine unlearning, large language models, and machine learning applications in the field of space science.

ELENA BARALIS (Member, IEEE) received the master's degree in electrical engineering and the Ph.D. degree in computer engineering from the Politecnico di Torino. She is currently the Protector with the Politecnico di Torino, where she has been a Full Professor, since January 2005. She has published over 200 papers in international journals and conference proceedings. Her current research interests include data mining, specifically on mining algorithms for big data and stream data analysis.

Open Access funding provided by 'Politecnico di Torino' within the CRUI CARE Agreement