Interpretable Machine Learning Decodes Soil Microbiome's Response to Drought Stress

Michelle Hagen[1], Rupashree Dass[1], Cathy Westhues[1], Jochen Blom[2], Sebastian J Schultheiss[1], Sascha Patz[1*]

[1*]Computomics GmbH, Eisenbahnstraße 1, Tübingen, 72072, Baden-Württemberg, Germany.
[2]Bioinformatics & Systems Biology, Justus Liebig University Gießen, Heinrich-Buff-Ring 58, Gießen, 35390, Hesse, Germany.

*Corresponding author(s). E-mail(s): sascha.patz@computomics.com; Contributing authors: michelle.hagen@computomics.com; rupashree.dass@computomics.com; cathy.westhues@gmail.com; jochen.blom@computational.bio.uni-giessen.de; sebastian.schultheiss@computomics.com;

Abstract

Background Extreme weather events induced by climate change, particularly droughts, have detrimental consequences for crop yields and food security. Concurrently, these conditions provoke substantial changes in the soil metagenome and affect plant health. Early recognition of soil affected by drought enables farmers to implement appropriate agricultural management practices. In this context, interpretable Machine Learning holds immense potential for drought stress classification in the soil metagenome based on marker taxa.

Results This study demonstrates that the metagenomic approach of Differential Abundance Analysis methods and Machine Learning-based Shapley Additive Explanation values provide similar information. They exhibit their potential as complementary approaches for identifying marker taxa and investigating their enrichment or depletion under drought stress in grass lineages. Additionally, the Random Forest Classifier trained on a diverse range of relative abundance data from the soil metagenome of various plant species achieves a high accuracy of 92.3 % at the genus rank for drought stress prediction. It demonstrates its generalization capacity for the lineages tested.

Conclusions In the detection of drought stress in the soil metagenome, this study emphasizes the potential of an optimized and generalized location-based ML classifier. By identifying marker taxa, this approach holds promising implications for microbe-assisted plant breeding programs and contributes to the development of sustainable agriculture practices. These findings are crucial for preserving global food security in the face of climate change.

Keywords: Metagenomics, Machine Learning, SHAP values, Differential Abundance Analysis, Soil microbiome, Drought stress

1 Background

Global food security is significantly threatened by climate change, especially in regions with limited access to food resources [1–3]. Anticipated occurrences of extreme weather events, such as droughts, are likely to increase in frequency and intensity, causing significant crop damage and threatening food availability [4, 5]. These repercussions are attributed to shortened growing seasons and substantial reductions in crop yields due to various biotic and abiotic stresses, prominently drought stress [6–8].

External perturbations, like drought stress, significantly impact the dynamics of the soil microbial community, leading to compositional shifts that are facilitated by the recruitment of beneficial microbes from the surrounding soil to the roots [9, 10]. This interaction between plants and soil microorganisms is a vital aspect of ecosystem health and stability [11–14]. Hence, this results in the opportunity to identify and interpret specific metagenomic patterns, as they have the potential to provide valuable insights into the state of both soil and plant health.

The use of Machine Learning (ML) algorithms enables the analysis of complex microbiome data by fully capturing the depth of data and identifying patterns that can discriminate between different states or conditions [15]. This can help to identify specific marker taxa that are key to understanding the intricate relationships between environmental stressors, soil health, and plant viability. Thereby, they aid in the development of more effective soil management strategies, including irrigation practices and the selection of drought-tolerant crops. Marker taxa hold significant potential for application in Synthetic Communities (SynComs), providing an innovative approach for early intervention under challenging environmental conditions, like droughts, to enhance plant resilience and growth [10, 16].

Still, the path from ML predictions to actionable insights can be challenging. ML models often resemble black boxes, with their internal decision-making obscured from users. The interpretation of the reasons for certain predictions is essential, especially for complex biological data [17–19]. This is where interpretable ML methods such as SHapley Additive ExPlanation (SHAP) values are applied [20].

The concept behind SHAP values is to distribute the credit for the model's prediction among the feature inputs based on their individual contribution, using game theory. Notably, SHAP values are model-specific yet globally constant, comprehensively taking into account interactions between features [21].

While SHAP values have been applied in clinical studies [22, 23], their use in metagenomic data is limited. For the identification of significant taxa between comparison groups, Differential Abundance Analysis (DAA) is the commonly used method in metagenomic analyses [24]. Typically used DAA tools are DESeq2 [25], ALDEx2 [26], edgeR [27], ANCOM-BC2 [28], and the non-parametric Wilcoxon rank-sum test.

DAA tools and SHAP values offer distinct approaches for detecting marker taxa. DAA tools rely on statistical tests and assume specific data distributions [29], while SHAP values are model-agnostic and applicable to any ML model [21]. Both methods aim to identify crucial features or taxa, providing insights into underlying biological mechanisms, but serve different purposes. DAA methods focus on identifying differentially abundant taxa, whereas interpretable ML using SHAP values offers importance measures based on model performance. The main objective of SHAP values is to interpret complex ML models by quantifying the contribution of each feature and explaining predictions. Their application in this study demonstrates the potential to identify key taxa in soil microbiomes as well as their role in the microbial response to drought stress.

The selection of an appropriate soil dataset was essential for this study. ML analyses thrive on datasets with many samples and informative metadata [30]. A dataset

from the work of *Naylor et al.* [31] was selected because it met these criteria exceptionally well. This dataset includes 623 samples from three soil isolation sources and investigates the effect of drought stress on 19 different crop species, including C3 and C4 plants.

Despite the remarkable achievements in applying ML to human microbiome research [32–36], its application in the context of soil metagenomics is not yet as advanced. However, the agricultural industry is increasingly recognizing the potential of ML to improve soil health and promote sustainable farming practices [37, 38]. This includes predicting plant phenotypes based on the plant and surrounding soil microbiome to detect taxa associated with plant diseases and environmental stresses [39].

Prior research has highlighted the potential of ML in agriculture, with studies identifying marker taxa for crop productivity [40] and beneficial root microbes [41]. Interestingly, the potential of ML for drought stress identification in soil microbiomes remains largely unstudied, representing a promising area for investigation.

This research aims to determine the efficacy of ML in predicting drought stress within microbial data of drought-stressed soils. The study comprises three key objectives: a) investigating the predictive capability of ML for drought stress, b) comparing the performance of interpretable ML with conventional 16S rRNA-based metagenomic analyses, and c) assessing the generalization capabilities of the trained classifier. By identifying marker taxa and deciphering microbial patterns associated with drought stress, this research addresses sustainable agriculture, improved crop productivity, and increased food security.

2 Methods

2.1 Datasets

A dataset originally curated by *Naylor et al.* [31] for their study on the impact of drought stress on the grass root microbiome was analyzed. This dataset, referred to as the 'Grass-Drought' dataset, comprises 623 samples from three isolation sources, including 'Soil,' 'Root,' and 'Rhizosphere', as well as two watering regimes, including 'Drought' and 'Control'. Samples from 18 distinct grass species within the Poaceae clade are included in the dataset. Tomato was used as an outgroup. The experimental site was located in Albany, California, characterized by silty loam soil with a pH of 5.2. Both watering regimes, 'Drought' and 'Control', were balanced, with 320 samples in the 'Control' group, receiving regular watering, and 303 samples in the 'Drought' group, experiencing conditions without water supply. All samples were sequenced using 16S rRNA amplicon sequencing of the V3-V4 region and are available under the BioProjectID PRJNA369551.

To evaluate the ML model's generalizability, its performance was assessed on a separate test dataset from *Xu et al.* [42] studying pre- and post-flowering drought stress effects on the *Sorghum bicolor* root microbiome (BioProjectID PRJNA435634), therefore referred to as the 'Sorghum-Drought' dataset. The sampling site was located in Kearney, California. To ensure the comparability of drought conditions between the Sorghum-Drought dataset and the original Grass-Drought dataset, two subsets were created: The 'Progressive Drought' subset comprised samples from the 'Control' group, along with specific time points (weeks 2 to 7 and weeks 10 to 17) from the 'Pre-Flowering Drought' and 'Post-Flowering Drought' groups, respectively. This subset comprised 278 'Control' and 210 'Drought' samples. The 'Late Drought' subset included samples from weeks 6, 7, 16, and 17 of the 'Control' group, weeks 6 and 7 of the 'Pre-Flowering Drought' group, and weeks 16 and 17 of the 'Post-Flowering Drought'

group, totaling 72 'Control' and 69 'Drought' samples. A detailed representation of the subsetting scheme can be found in the Additional file 1 in **Fig. S1**.

2.2 Data Processing

The DADA2 workflow [43] for Illumina sequenced paired-end fastq files was employed for sequence data processing, implemented in R version 4.2.3. Taxonomy was assigned using the SILVA database [44] and the Ribosomal Database Project (RDP) classifier [45] from phylum to genus rank. To enhance data quality, prevalence filtering was conducted, retaining Amplicon Sequence Variants (ASVs) present in at least 95 % of all samples, reducing the total number of ASVs from 25,415 to 3,276. Samples with low read counts were excluded, yielding a dataset of 560 samples. Rarefaction was performed, normalizing sequencing depth to the dataset's 10 % decile of 17,291 reads. Feature tables for ML for each taxonomic rank were constructed with relative abundance values per taxon across all samples and a 'Control' or 'Drought' target variable.

2.3 16S rRNA-based Metagenomic Analysis

A diversity analysis was conducted between the two watering regimes 'Control' and 'Drought'. Alpha diversity was assessed using the Shannon index with the `estimate_richness` function from the phyloseq package (version 1.44.0) [46]. Beta diversity was explored via Principal Coordinate Analysis (PCoA) based on Bray-Curtis dissimilarities with the `ordinate` and `plot_ordination` functions from phyloseq.

To identify taxonomic differences between the 'Control' and 'Drought' groups, a DAA was employed with several tools using the microbiomeMarker R package (version 1.4.0) [47], including DESeq2, ALDEx2, and edgeR, as well as ANCOM-BC2 (version 2.0.2) [28], and the non-parametric Wilcoxon rank-sum test on the ASV level. The UpSetR package (version 1.4.0) [48] was used to compare tool outcomes, and the three most suitable methods were applied to all taxonomic ranks to compare enrichment groups and Benjamini-Hochberg (BH) adjusted p-values [49].

2.4 Machine Learning

A Random Forest Classifier (RFC) was applied using the ScikitLearn Python package (version 1.1.3) [50] to all ranks to predict the samples' watering treatment using relative taxon abundances. Hyperparameter optimization was carried out through five-fold nested Cross-Validation (CV), splitting the dataset into five equally sized parts. During each fold, four parts of the dataset were used for training, while the remaining part acted as a dataset for testing the best model of each fold. The mean model performance was evaluated in terms of accuracy, F1 score, precision, recall, and Area Under the Curve (AUC) between all folds.

In order to interpret the RFC predictions, SHAP values were utilized using the SHAP Python package (version 0.41.0) [51] with the `shap.TreeExplainer` function [20]. During each fold of the nested cross-validation, feature contributions related to detecting drought stress from SHAP values were extracted. A consensus was sought across four or five of the folds, requiring alignment in the majority, to consider the enrichment information suitable for subsequent analysis. The feature contributions towards drought stress from the SHAP values were compared with taxon enrichment patterns from differential abundance testing. This was followed by a comparison of significant taxa identified by DAA methods and important taxa identified by ML.

The model performance and generalizability were tested on two independent subsets of the Sorghum-Drought dataset of *Xu et al.* [42] that has been described in

detail in the 'Datasets' section. This test dataset was processed similarly to the Grass-Drought dataset. To create feature tables, the tables were pruned to only include taxa present in the Grass-Drought dataset. Taxa that were not present in the Sorghum-Drought dataset but in the Grass-Drought dataset were added with zero counts as demonstrated in the Additional file 1 in **Tab. S1**. Model performance was evaluated, including mean accuracy, F1 score, precision, and recall, which were computed across all taxonomic ranks for both subsets.

3 Results

Alpha and Beta Diversity

Alpha diversity analysis, utilizing the Shannon index as displayed in **Fig. 1 A**, found no significant differences between the 'Control' and 'Drought' groups. This demonstrates that microbial diversity within individual samples was not significantly impacted by watering regimes. On all taxonomic ranks, no highly abundant taxon was found to be uniquely abundant to drought stress, as only differences between relative abundances between 'Control' and 'Drought' groups could be observed as shown in Additional file 1 in **Fig. S2**. Beta diversity, assessed via PCoA based on Bray-Curtis dissimilarities as shown in **Fig. 1 B**, yielded insights into the variation between the samples. The watering regime accounted for 6.8 % of the variance and could be clustered into the corresponding irrigation groups. In order to train the ML model to detect drought stress from a variety of soil samples deriving from different isolation sources and crops, the whole dataset was used without subsetting it to specific sample types. For further 16S rRNA-based metagenomic analyses with this dataset and its metadata, *Naylor et al.'s* paper itself [31] is referred to, offering interesting insights into the influence of the soil isolation source and the impact of the different crops on the root microbiome.

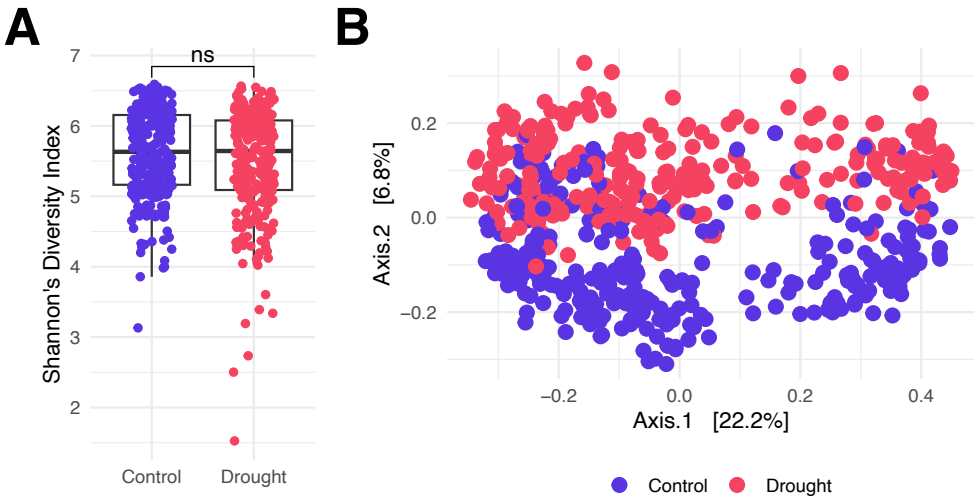

Fig. 1 Diversity Plots for the Grass-Drought Dataset. Alpha and beta diversity plots comparing the 'Control' (blue) and 'Drought' (red) watering regimes. (A) Boxplots of Shannon's Diversity Index for all samples comparing watering regimes. Significance was determined using a non-parametric Wilcoxon rank sum test (* p <0.05, ** p <0.01, *** p <0.001, **** p <0.0001). (B) Principal Coordinate plot using Bray-Curtis dissimilarities colored by the watering regimes.

Comparative Analysis of DAA Tools

This study's comprehensive approach to DAA encompassed five distinct methods: DESeq2, ANCOM-BC2, ALDEx2, edgeR, and the non-parametric Wilcoxon rank-sum test (**Fig. 2**). All methods used False Discovery Rate (FDR)-corrected p-values with BH correction and an alpha threshold <0.05. A total of 2,356 ASVs were identified as significantly differentially abundant. Strikingly, 441 ASVs were identified by all five methods, highlighting a core set of differentially abundant taxa. EdgeR and the non-parametric Wilcoxon rank-sum test identified 485 and 318 unique ASVs, respectively. On the other hand, ANCOM-BC2, ALDEx2, and DESeq2 showed consistent results with no or only a small number of uniquely identified ASVs and were therefore used for DAA on all taxonomic ranks. An increased level of consistency between the three tools was visible as displayed in Additional file 1, **Fig. S3**.

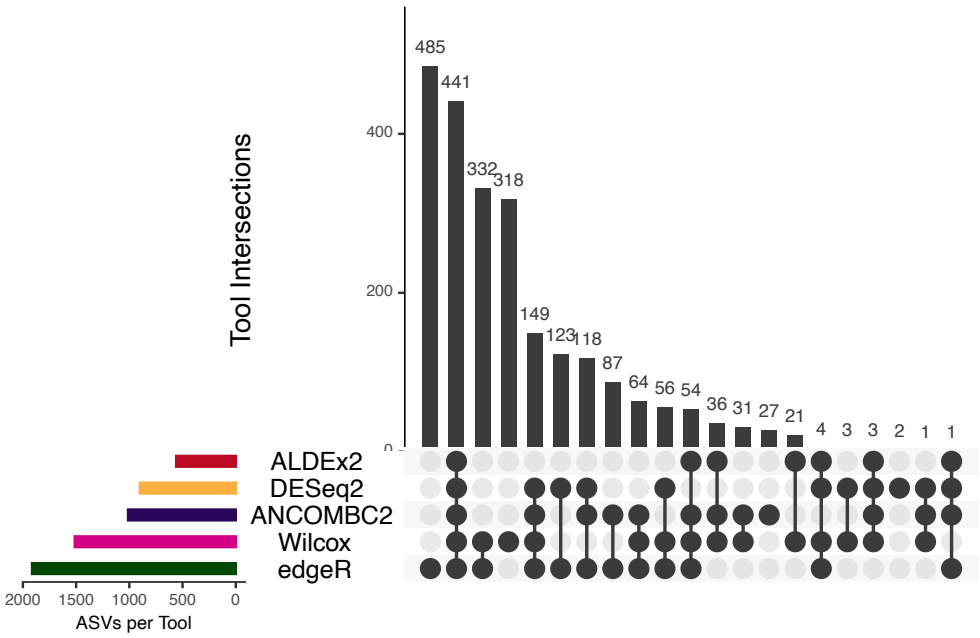

Fig. 2 ASV Intersections between different DAA Tools on ASV level of the Grass-Drought Dataset. Upset plots displaying the overlap and uniqueness of significant taxa on ASV level identified by DAA methods (ALDEx2, DESeq2, ANCOM-BC2, non-parametric Wilcoxon rank-sum test, and edgeR). The horizontal bars show the total number of taxa for each tool, while the vertical bars show the number of shared taxa between corresponding sets, sorted by the total number of shared taxa. All tools use an alpha threshold of 0.05 for significance.

The RFC shows remarkable classifying performance across all ranks

Machine learning using the trained RFC demonstrated remarkable performance scores in predicting drought stress in the soil metagenome. **Tab. 1** shows, that across all taxonomic ranks, the RFC consistently delivered exceptional results, with a mean accuracy surpassing 90 %.

The genus level proved to be the most effective input, achieving an accuracy of 0.923 ± 0.029, an F1 score of 0.921 ± 0.030, and a recall of 0.954 ± 0.029. Family-level analysis excelled in precision, with a score of 0.902 ± 0.038. Furthermore, the AUC underscored the robust performance of the RFC, with a mean AUC of 0.980 ±

0.010 at the genus level. The corresponding Reciever Operating Characteristic (ROC) curves can be found in Additional file 1 in **Fig. S4**.

Tab. 1 Random Forest Classifier Performance of the Grass-Drought Dataset. Table displaying the mean accuracy, F1 score, precision, recall, and AUC of the classifier on different taxonomic ranks of the Grass-Drought dataset, with the best-performing rank for each metric marked in bold.

Metric	Phylum	Class	Order	Family	Genus
Accuracy	0.900 ± 0.025	0.911 ± 0.020	0.914 ± 0.023	0.921 ± 0.017	**0.923 ± 0.029**
F1 score	0.895 ± 0.029	0.906 ± 0.023	0.912 ± 0.024	0.919 ± 0.017	**0.921 ± 0.030**
Precision	0.891 ± 0.041	0.902 ± 0.032	0.890 ± 0.038	**0.902 ± 0.038**	0.892 ± 0.041
Recall	0.899 ± 0.024	0.911 ± 0.022	0.936 ± 0.022	0.939 ± 0.031	**0.954 ± 0.029**
AUC	0.960 ± 0.020	0.960 ± 0.010	0.970 ± 0.010	0.970 ± 0.010	**0.980 ± 0.010**

Interpretable ML and DAA as complementary approaches in marker taxa identification

DAA tools (ANCOM-BC2, ALDEx2, DESeq2), and SHAP values emerged as powerful methodologies for investigating taxon enrichment and feature importance, respectively. Both approaches, as displayed in **Fig. 3**, consistently identified taxa responsible for driving differences between the 'Control' and 'Drought' groups, which can be considered marker taxa for drought stress. Overall, the proportion of matches in the enrichment assignments of DAA tools and SHAP values for all identified taxa ranged from 79.59 % on the order level to 82.65 % on the genus level.

The significance and importance of microbial taxa in the dataset were explored using both DAA significances and SHAP value feature importances of the RFC. Comparing the sorted top 25 genera with adjusted significance levels from ANCOM-BC2 and their corresponding mean absolute SHAP values, it was found that the most significant and also important taxon with more than a two-fold difference to the next taxon was the genus *Kribbella* (**Fig. 3**).

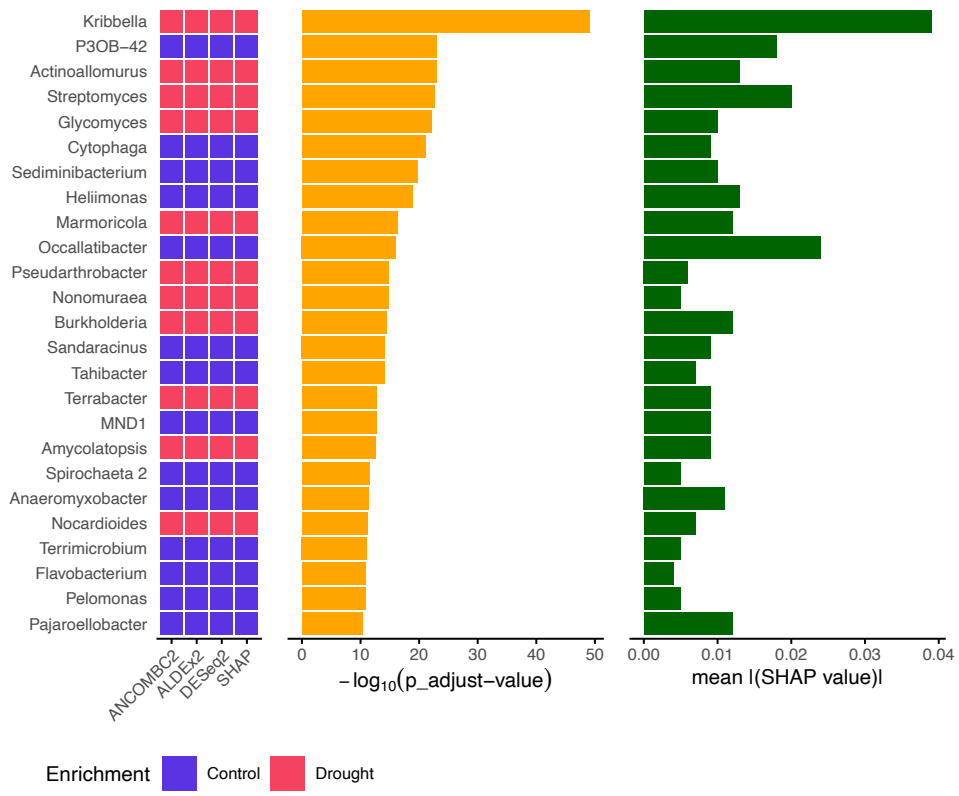

Fig. 3 Genera Enrichment, Significance and Importance by DAA Tools and SHAP Values of the Grass-Drought Dataset. Binary heatmap showing the enrichment of the top 25 significant genera from ANCOM-BC2 between 'Control' (blue) and 'Drought' (red) groups for the three methods used for DAA (DESeq2, ANCOM-BC2, ALDEx2) with an alpha <0.05, and SHAP values obtained from the RFC. Corresponding bar plots comparing -$\log_{10}$(p_adjust) values (orange) and mean(|SHAP value|) (green).

In a direct comparison with the feature importances, it was shown that the order of the taxa identified as important differed greatly from that of ANCOM-BC2 in some genera. This effect was also visible at higher taxonomic ranks (Additional file 1 **Fig. S5** and **Fig. S6**). Especially the genera *Streptomyces* and *Occallatibacter* did not stand out from the results of ANCOM-BC2 but were prioritized considerably higher by their SHAP values.

The trained RFC generalizes to unseen samples from a different dataset

To assess the generalizability of the trained RFC model, it was applied to a test dataset from *Xu et al.* (2018), focusing on *Sorghum bicolor* root microbiomes subjected to drought stress. First, the trained RFC underwent testing using samples exhibiting advanced drought stress conditions, referred to as the 'Late Drought' subset. These samples were expected to demonstrate the most noticeable changes in the relative abundances of the taxa. Stable accuracies across all taxonomic ranks could be detected, as displayed in **Tab. 2**, with a notable improvement towards the family level. The family level achieved the highest accuracy (0.854 ± 0.017), while also excelling in F1 score (0.855 ± 0.021) and precision (0.830 ± 0.029). The order level exhibited the best recall (0.925 ± 0.031).

Tab. 2 Late Drought Classifier Performance of the Sorghum-Drought Dataset. Table displaying the mean accuracy, F1 score, precision, and recall of the classifier on different taxonomic ranks on the Late Drought subset. The best-performing rank for each metric is marked in bold.

Metric	Phylum	Class	Order	Family	Genus
Accuracy	0.797 ± 0.029	0.672 ± 0.009	0.828 ± 0.027	**0.854 ± 0.017**	0.811 ± 0.031
F1 score	0.788 ±0.033	0.541 ± 0.016	0.841 ± 0.025	**0.855 ± 0.021**	0.810 ± 0.033
Precision	0.807 ± 0.029	0.862 ± 0.027	0.771 ± 0.022	**0.830 ± 0.029**	0.806 ± 0.049
Recall	0.771 ± 0.048	0.394 ± 0.017	**0.925 ± 0.031**	0.998 ± 0.064	0.826 ± 0.096

Due to the classifier's outstanding performance with the 'Late Drought' subset, testing extended to another subset containing various drought stress levels, referred to as the 'Progressive Drought' subset. This subset contained samples of the complete course of the drought period with associated controls. Here, the order level displayed the best F1 score (0.754 ± 0.024) and recall (0.887 ± 0.018), while the family level yielded the highest accuracy (0.768 ± 0.018) and precision (0.692 ± 0.014) as shown in **Tab. 3**. For both subsets, it was noticeable that the best performance was not observed at the genus level, but at the family or order level.

Tab. 3 Progressive Drought Classifier Performance of the Sorghum-Drought Dataset. Table displaying the mean accuracy, F1 score, precision, and recall of the classifier on different taxonomic ranks on the Progressive Drought subset. The best-performing rank for each metric is marked in bold.

Metric	Phylum	Class	Order	Family	Genus
Accuracy	0.678 ± 0.029	0.613 ± 0.004	0.750 ± 0.029	**0.768 ± 0.018**	0.711 ± 0.044
F1 score	0.625 ± 0.031	0.374 ± 0.016	**0.754 ± 0.024**	0.753 ± 0.030	0.697 ± 0.029
Precision	0.628 ± 0.037	0.614 ± 0.007	0.656 ± 0.028	**0.692 ± 0.014**	0.647 ± 0.055
Recall	0.623 ± 0.038	0.270 ± 0.015	**0.887 ± 0.018**	0.831 ± 0.076	0.773 ± 0.106

4 Discussion

This study employed ML to predict the irrigation state of soil samples based on their microbial community composition and aimed to discover specific marker taxa for drought stress. In terms of alpha and beta diversity analysis, it is crucial to note that drought stress primarily influenced the relative abundance of taxa rather than causing a complete abolishment or appearance of certain taxa [52], which aligns with the study's focus on the 'Control' and 'Drought' labels. This suggests that the classifier was trained by emphasizing variations in taxon abundance rather than focusing on the presence or absence of specific taxa.

Regarding the comparison of interpretable ML with conventional DAA tools for marker taxa identification, the focus was on finding the most suitable DAA tools for the Grass-Drought dataset from a variety of popular methods, namely ALDEx2, ANCOM-BC2, DESeq2, edgeR, and the non-parametric Wilcoxon rank-sum test. This approach is generally recommended for DAA, as it is not possible to find the true number of significant taxa in real-world data sets like it is the case with mock data [53, 54]. The used DAA methods made different assumptions about the data distribution [53]. For instance, DESeq2 and edgeR assume a negative binomial distribution, while ALDEx2 and ANCOM-BC2 assume a Gaussian distribution. The non-parametric Wilcoxon rank-sum test, on the other hand, does not make any distribution assumptions.

On the ASV level, out of 3,276 total assigned ASVs, 71.9 % were identified as significant by at least one of the five DAA methods. However, only 13.46 % of these significant ASVs were detected by all five methods, suggesting a substantial proportion

of ASVs being uniquely identified by specific tools, possibly indicating false discoveries [53]. Specifically, edgeR and the Wilcoxon rank-sum test uniquely identified a high number of ASVs not detected by other tools, which can be an indicator for many false discoveries and unreliable results [55].

In contrast, ANCOM-BC2, DESeq2, and ALDEx2 exhibited more reliable results, with fewer uniquely identified ASVs. These findings are in line with previous studies that have highlighted the reliability of these three methods in controlling the FDR [53, 54, 56, 57]. Out of these three tools, ANCOM-BC2 was selected as the DAA method to compare its results directly with those of interpretable ML, as it showed the most overlap in detected ASVs with the other two selected DAA tools.

For interpretable ML, a RFC was chosen as it is recognized as a top-performing classifier for handling high-dimensional and sparse data, such as metagenomic datasets with hundreds to thousands of features and non-linear relationships between features and the target variable [58–62].

The classifier, trained on a dataset containing soil samples from a variety of soil isolation sources, crops, and drought stress levels, exhibited exceptional performance across all taxonomic ranks, directly addressing objective a) of this study. As the taxonomic rank descended from higher (e.g., phylum) to lower levels (e.g., genus), the granularity and resolution of the features increased. At the genus level, which represented the lowest taxonomic rank, the classifier exhibited the highest mean accuracy of 92.3 %, as well as the highest F1 score, recall, and AUC, demonstrating its effectiveness in capturing true positive instances. While there was a slight increase in overall accuracy from the phylum to the genus level, the genus level provided more specific insights into microbial diversity and potential marker taxa for drought stress.

In the exploration of marker taxa for drought stress in the soil metagenome, interpretable ML was employed using SHAP values in the nested CV of each taxonomic rank. Enrichment and feature importance results were compared with the output of the three most suitable DAA tools for this dataset. At higher taxonomic ranks, the agreement between DAA enrichment and SHAP value contribution was less definitive. Among the DAA tools, taxa were often not classified as significant by all three tools, as seen with Bacteroidota at the phylum level. Similarly, SHAP values did not always provide clear results between the loops of the nested CV, making it challenging to assign enrichment to either 'Control' or 'Drought', as observed with Firmicutes at the phylum level. In some cases, the results from DAA and interpretable ML differed, such as with Verrucomicrobiae and Armanimonadota at the phylum level, which were classified as enriched in 'Control' by DAA tools and enriched under 'Drought' by SHAP. According to the literature, both phyla were found to be more enriched under irrigation, but the same study concluded that both taxa have the potential to assist plants under drought conditions [63]. However, at lower, more specific ranks such as family and genus levels, all enrichment information among the top 25 taxa is consistent. This consistency highlights that SHAP values can be equally useful for the discovery of specific marker taxa under stress conditions, effectively fulfilling objective b).

Furthermore, the rankings of taxa between DAA and ML approaches were compared. While the order of significant taxa differed, the genus *Kribbella* consistently emerged as most significant and important, displaying a two-fold increase compared to the next relevant genus. Although being a poorly studied genus, *Kribbella* has shown potential in promoting plant growth [64, 65], making it a promising marker taxon for drought stress.

Additionally, in the direct comparison of significances and feature importances, certain taxa were detected with greater prominence, as evidenced by a peak in their mean absolute SHAP value compared to the significance assigned by the DAA analysis, like the genera *Streptomyces* and *Occallatibacter*. *Streptomyces*, a dominant genus in soil

microbiomes, has been associated with drought stress and plant health in dry environments [66–68]. The genus *Occallatibacter* showed depletion under drought conditions and was considered an important feature for the prediction of drought stress, although further research is needed to understand its specific impact on soil metagenomes under drought, as it has only been observed under heat stress and no-stress conditions [69, 70].

SHAP values and DAA tools use different underlying approaches for the identification of important or significant taxa. In the context of this study, it is not possible to determine which approach is more suitable, but the overall results suggest that both methods provide important information for the identification of marker taxa. Therefore, these approaches should be seen as complementary rather than interchangeable, with each providing valuable insights into metagenomic data analysis.

To evaluate the generalization capabilities of this study's classifier, its performance was tested on another drought stress dataset. The classifier's performance was assessed with samples undergoing several weeks of drought stress as the most impactful differences were expected between the two watering regimes. The Late Drought subset exhibited an accuracy score of 0.854 ± 0.017 at the family level. Therefore, the classifier's effectiveness and robustness across the entire spectrum of drought stress levels of the Sorghum-Drought dataset was explored by predicting drought stress in the Progressive Drought subset. Remarkably, the results consistently demonstrated the classifier's outstanding performance in both scenarios. The Progressive Drought subset achieved an accuracy score of 0.768 ± 0.018 at the family level, indicating the model's reliability in classifying drought stress regardless of the drought stress level involved. In contrast to the Grass-Drought dataset, where the best performance was achieved at the genus level as the lowest taxonomic rank with the highest granularity, the subsets of the Sorghum-Drought test dataset did not yield the best classification results on this rank. The best performance was observed on the order and family levels. This can be attributed to the inherent sparsity on the genus level due to the addition of taxa with zero counts to create feature tables for the prediction with equal feature inputs, as displayed in the Additional file 1 in **Tab. S1**.

These results emphasize the classifier's adaptability across diverse drought stress conditions, reinforcing its utility as a valuable tool for drought stress classification, in line with the objectives outlined in part c) of this study. Even though the classifier was trained with a dataset containing 16S rRNA metagenomic data of different drought stress levels, soil isolation sources, and a variety of plants, the approach might vary based on input data from other sequencing regions or plants that the classifier was not trained on. Such differences may emerge due to differences in the estimation of microbial diversity [71, 72]. The effect of the soil isolation source on the prediction also offers scope for further investigations to improve the classifier's predictive capabilities and potentially develop a classifier tailored to specific soil isolation sources [73]. Expanding the sample size could further enhance the classifier's generalizability, as a more extensive representation of taxa would increase the likelihood of encountering taxa abundances present in unknown samples during new predictions. However, when dealing with large datasets, selecting a comprehensive range of core taxa is recommended. By selecting the most important taxa as features [74] and training the classifier accordingly, the introduction of sparsity in the feature tables when predicting new data can be prevented. Furthermore, including samples from different locations introduces another dimension, as variations in microbial composition across diverse geographic locations and climates [75] can impact the classifier's performance. With the upcoming possibilities of 16S rRNA long-read sequencing, it is recommended to create a location- and long-read sequencing-based classifier to generate an individual classifier with a reduced bias. Further evaluation with different ML models, more data, and subsequent feature selection seem interesting applications for future research.

5 Conclusions

In conclusion, with the ongoing threat of extreme weather events, notably droughts [76, 77], it is indispensable to explore innovative methods for understanding the impact of the soil microbiome on agriculture and ecosystems. The primary accomplishment of this study is the creation of a location-based classifier for drought stress in the soil metagenome. Demonstrating remarkable generalization capabilities, the classifier assesses drought stress across various drought stress levels and is applicable to different grasses.

The application of this study's generalized ML model extends beyond the classification of drought stress, facilitating precision agriculture, including the optimization of irrigation strategies. Further, in microbe-assisted plant breeding programs, the discovery of marker taxa for drought stress using interpretable ML with SHAP values provides farmers and breeders with valuable insights for the definition of microbial strains for targeted bioinoculation approaches. Here, a deeper understanding of the plant growth-promoting functions associated with drought stress-related taxa holds promise for future advancements. This knowledge could play a pivotal role in enhancing plant adaptation to drought stress, strengthening the plant immune system against yield losses, and reducing susceptibility to pathogens [78].

6 List of abbreviations

ASVs Amplicon Sequence Variants
AUC Area Under the Curve
BH Benjamini-Hochberg
CV Cross-Validation
DAA Differential Abundance Analysis
FDR False Discovery Rate
ML Machine Learning
PCoA Principal Coordinate Analysis
RDP Ribosomal Database Project
RFC Random Forest Classifier
ROC Reciever Operating Characteristic
SHAP SHapley Additive ExPlanation
SynComs Synthetic Communities

Supplementary Information

Additional file 1

- File name: Additional file 1
- File format: .pdf
- Title of data: Supplementary Figures and Tables
- Description of data: **Fig. S1:** Weekly Watering Scheme of the Sorghum-Drought Test Dataset. Watering scheme for the three groups 'Control' (C), 'Pre-Flowering Drought' (PrD), and 'Post-Flowering Drought' (PoD). Dashed lines indicate watering and white cells indicate no watering. The test dataset was subset into the 'Progressive Drought' subset (red) and 'Late Drought' subset (blue). **Tab. S1:** Feature Table Pruning between the Datasets. Table displaying differences in the total number of different taxa between the Grass-Drought dataset as the dataset the RFC has been trained on and the Sorghum-Drought dataset as the test dataset. A high number of taxa added with zero counts in order to bring the feature tables of the test dataset into a suitable format for prediction also increases the level of potential sparsity. **Fig. S2:** Relative Abundances per Rank of the Grass-Drought

Dataset. Bar plots displaying the relative abundance of the top 10 taxa between the 'Control' and 'Drought' groups on (A) Phylum, (B) Class, (C) Order, (D) Family, and (E) Genus level in alphabetical order. **Fig. S3:** Significant Taxa Intersections between DAA Tools per Rank of the Grass-Drought Dataset. Upset plots displaying the overlap and uniqueness of significant taxa identified by the three DAA methods 'ALDEx2', 'DESeq2', and 'ANCOM-BC2' on (A) Phylum, (B) Class, (C) Order, (D) Family, and (E) Genus level. The horizontal bars show the total number of taxa for each tool, while the vertical bars show the number of shared taxa between corresponding sets, sorted by the total number of shared taxa. All tools use an alpha threshold of 0.05 for significance. **Fig. S4:** ROC Curves per Rank of the Grass-Drought Dataset. Receiver Operating Characteristic (ROC) curve at (A) Phylum, (B) Class, (C) Order, (D) Family, and (E) Genus level, showing the area under the curve (AUC) for each fold of the nested cross-validation. The ROC curve displays the best model for each fold and the mean AUC. **Fig. S5:** Taxon Enrichment, Significance, and Importance by DAA Tools and SHAP Values of the Grass-Drought Dataset. Binary heatmap showing the enrichment of the top significant genera from ANCOM-BC2 on (A) Phylum and (B) Class level between 'Control' (blue) and 'Drought' (red) groups for the three methods used for DAA (DESeq2, ANCOM-BC2, ALDEx2) with an alpha <0.05, and SHAP values obtained from the RFC. Empty cells display no significant enrichment. Corresponding bar plots comparing -$\log_{10}$(p_adjust) values (orange) and mean(|SHAP value|) (green). **Fig. S6:** Taxon Enrichment, Significance, and Importance by DAA Tools and SHAP Values of the Grass-Drought Dataset. Binary heatmap showing the enrichment of the top significant genera from ANCOM-BC2 on (A) Order and (B) Family level between 'Control' (blue) and 'Drought' (red) groups for the three methods used for DAA (DESeq2, ANCOM-BC2, ALDEx2) with an alpha <0.05, and SHAP values obtained from the RFC. Empty cells display no significant enrichment. Corresponding bar plots comparing -$\log_{10}$(p_adjust) values (orange) and mean(|SHAP value|) (green).

Declarations

6.1 Ethics approval and consent to participate

The study incorporated 16S rRNA metagenomic data from plants and soil that is publicly available and was conducted in accordance with the present ethical guidelines as of the date of publication.

6.2 Consent for publication

Not applicable

6.3 Availability of data and material

The datasets supporting the conclusions of this article are available in the NCBI Short Read Archive under the BioProjectID PRJNA369551 ('Grass-Drought' dataset), and PRJNA435634 ('Sorghum-Drought' dataset). All analysis scripts are available on GitHub (https://github.com/Computomics/SoilMicrobiomeDroughtML).

6.4 Competing interests

The authors MH, RD, CW, SJS, and SP, currently or formerly employed by Computomics GmbH, and JB of the Justus Liebig University Giessen declare that they have no competing interests.

6.5 Funding

RD, CW, and SJS are/were supported by funds of the Federal Ministry of Education and Research (BMBF), Germany [01|21038]. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication.

6.6 Authors' contributions

MH conceived, designed, and performed the analysis with inputs from SP, RD, CW, SJS, and JB. MH drafted the manuscript and it was reviewed and edited by all authors. All authors read and approved the final manuscript.

6.7 Acknowledgements

Not applicable

6.8 Authors' information (optional)

Not applicable

References

[1] Wheeler T, von Braun J. Climate Change Impacts on Global Food Security. Science. 2013;341(6145):508–513. https://doi.org/10.1126/science.1239402.

[2] Schmidhuber J, Tubiello FN. Global food security under climate change. Proceedings of the National Academy of Sciences. 2007;104(50):19703–19708. https://doi.org/10.1073/pnas.0701976104.

[3] Myers S, Fanzo J, Wiebe K, Huybers P, Smith M. Current guidance underestimates risk of global environmental change to food security. The BMJ. 2022;378:e071533. https://doi.org/10.1136/bmj-2022-071533.

[4] Trenberth KE, Dai A, van der Schrier G, Jones PD, Barichivich J, Briffa KR, et al. Global warming and changes in drought. Nature Climate Change. 2014;4(1):17–22. https://doi.org/10.1038/nclimate2067.

[5] Kempf M. Enhanced trends in spectral greening and climate anomalies across Europe. Environmental Monitoring and Assessment. 2023;195(2):260. https://doi.org/10.1007/s10661-022-10853-8.

[6] Raza A, Razzaq A, Mehmood SS, Zou X, Zhang X, Lv Y, et al. Impact of Climate Change on Crops Adaptation and Strategies to Tackle Its Outcome: A Review. Plants. 2019;8(2):34. https://doi.org/10.3390/plants8020034.

[7] Moriondo M, Giannakopoulos C, Bindi M. Climate change impact assessment: the role of climate extremes in crop yield simulation. Climatic Change. 2011;104(3):679–701. https://doi.org/10.1007/s10584-010-9871-0.

[8] Battisti DS, Naylor RL. Historical warnings of future food insecurity with unprecedented seasonal heat. Science (New York, NY). 2009;323(5911):240–244. https://doi.org/10.1126/science.1164363.

[9] Ossowicki A, Raaijmakers JM, Garbeva P. Disentangling soil microbiome functions by perturbation. Environmental Microbiology Reports. 2021;13(5):582–590. https://doi.org/10.1111/1758-2229.12989.

[10] Ali S, Tyagi A, Park S, Mir RA, Mushtaq M, Bhat B, et al. Deciphering the plant microbiome to improve drought tolerance: Mechanisms and perspectives. Environmental and Experimental Botany. 2022;201:104933. https://doi.org/10.1016/j.envexpbot.2022.104933.

[11] Berendsen RL, Pieterse CMJ, Bakker PAHM. The rhizosphere microbiome and plant health. Trends in Plant Science. 2012;17(8):478–486. https://doi.org/10.1016/j.tplants.2012.04.001.

[12] Xiong W, Song Y, Yang K, Gu Y, Wei Z, Kowalchuk GA, et al. Rhizosphere protists are key determinants of plant health. Microbiome. 2020;8(1):27. https://doi.org/10.1186/s40168-020-00799-9.

[13] Gao M, Xiong C, Gao C, Tsui CKM, Wang MM, Zhou X, et al. Disease-induced changes in plant microbiome assembly and functional adaptation. Microbiome. 2021;9(1):187. https://doi.org/10.1186/s40168-021-01138-2.

[14] Xie J, Dawwam GE, Sehim AE, Li X, Wu J, Chen S, et al. Drought Stress Triggers Shifts in the Root Microbial Community and Alters Functional Categories in the Microbial Gene Pool. Frontiers in Microbiology. 2021;12.

[15] Kumar R, Yadav G, Kuddus M, Ashraf GM, Singh R. Unlocking the microbial studies through computational approaches: how far have we reached? Environmental Science and Pollution Research. 2023;30(17):48929–48947. https://doi.org/10.1007/s11356-023-26220-0.

[16] Miller T, Mikiciuk G, Kisiel A, Mikiciuk M, Paliwoda D, Sas-Paszt L, et al. Machine Learning Approaches for Forecasting the Best Microbial Strains to Alleviate Drought Impact in Agriculture. Agriculture. 2023;13(8):1622. https://doi.org/10.3390/agriculture13081622.

[17] Watson DS. Interpretable machine learning for genomics. Human Genetics. 2022;141(9):1499–1513. https://doi.org/10.1007/s00439-021-02387-9.

[18] Bifarin OO. Interpretable machine learning with tree-based shapley additive explanations: Application to metabolomics datasets for binary classification. PLOS ONE. 2023;18(5):e0284315. https://doi.org/10.1371/journal.pone.0284315.

[19] Conard AM, DenAdel A, Crawford L. A spectrum of explainable and interpretable machine learning approaches for genomic studies. WIREs Computational Statistics. 2023;15(5):e1617. https://doi.org/10.1002/wics.1617.

[20] Lundberg SM, Erion GG, Lee SI. Consistent Individualized Feature Attribution for Tree Ensembles. arXiv preprint arXiv:180203888. 2019;https://doi.org/10.48550/arXiv.1802.03888.

[21] Lundberg S, Lee SI. A Unified Approach to Interpreting Model Predictions. Advances in neural information processing systems. 2017;30.

[22] Tan CCS, Acman M, van Dorp L, Balloux F. Metagenomic evidence for a polymicrobial signature of sepsis. Microbial Genomics. 2021;7(9):000642. https://doi.org/10.1099/mgen.0.000642.

[23] Rynazal R, Fujisawa K, Shiroma H, Salim F, Mizutani S, Shiba S, et al. Leveraging explainable AI for gut microbiome-based colorectal cancer classification. Genome Biology. 2023;24(1):21. https://doi.org/10.1186/s13059-023-02858-4.

[24] Cappellato M, Baruzzo G, Camillo BD. Investigating differential abundance methods in microbiome data: A benchmark study. PLOS Computational Biology. 2022;18(9):e1010467. https://doi.org/10.1371/journal.pcbi.1010467.

[25] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;15(12):550. https://doi.org/10.1186/s13059-014-0550-8.

[26] Fernandes AD, Macklaim JM, Linn TG, Reid G, Gloor GB. ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq. PLOS ONE. 2013;8(7):e67019. https://doi.org/10.1371/journal.pone.0067019.

[27] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–140. https://doi.org/10.1093/bioinformatics/btp616.

[28] Kaul A, Mandal S, Davidov O, Peddada SD. Analysis of Microbiome Data in the Presence of Excess Zeros. Frontiers in Microbiology. 2017;8.

[29] Yang L, Chen J. A comprehensive evaluation of microbial differential abundance analysis methods: current status and potential solutions. Microbiome. 2022;10(1):130. https://doi.org/10.1186/s40168-022-01320-0.

[30] Rajput D, Wang WJ, Chen CC. Evaluation of a decided sample size in machine learning applications. BMC Bioinformatics. 2023;24(1):48. https://doi.org/10.1186/s12859-023-05156-9.

[31] Naylor D, DeGraaf S, Purdom E, Coleman-Derr D. Drought and host selection influence bacterial community dynamics in the grass root microbiome. The ISME Journal. 2017;11(12):2691–2704. https://doi.org/10.1038/ismej.2017.118.

[32] Oh TG, Kim SM, Caussy C, Fu T, Guo J, Bassirian S, et al. A Universal Gut-Microbiome-Derived Signature Predicts Cirrhosis. Cell Metabolism. 2020;32(5):878–888.e6. https://doi.org/10.1016/j.cmet.2020.06.005.

[33] Zhou YH, Gallins P. A Review and Tutorial of Machine Learning Methods for Microbiome Host Trait Prediction. Frontiers in Genetics. 2019;10:579. https://doi.org/10.3389/fgene.2019.00579.

[34] Barnard E, Shi B, Kang D, Craft N, Li H. The balance of metagenomic elements shapes the skin microbiome in acne and health. Scientific Reports. 2016;6(1):39491. https://doi.org/10.1038/srep39491.

[35] Lee SJ, Rho M. Multimodal deep learning applied to classify healthy and disease states of human microbiome. Scientific Reports. 2022;12(1):824. https://doi.org/10.1038/s41598-022-04773-3.

[36] Robertson R, Church J, Edens T, Mutasa K, Geum HM, Baharmand I, et al. The fecal microbiome and rotavirus vaccine immunogenicity in rural Zimbabwean infants. Vaccine. 2021;39. https://doi.org/10.1016/j.vaccine.2021.07.076.

[37] Meshram V, Patil K, Meshram V, Hanchate D, Ramkteke SD. Machine learning in agriculture domain: A state-of-art survey. Artificial Intelligence in the Life Sciences. 2021;1:100010. https://doi.org/10.1016/j.ailsci.2021.100010.

[38] Dhaliwal DS, Williams MM. Sweet corn yield prediction using machine learning models and field-level data. Precision Agriculture. 2023;https://doi.org/10.1007/s11119-023-10057-1.

[39] Deng Z, Zhang J, Li J, Zhang X. Application of Deep Learning in Plant–Microbiota Association Analysis. Frontiers in Genetics. 2021;12:697090. https://doi.org/10.3389/fgene.2021.697090.

[40] Chang HX, Haudenshield JS, Bowen CR, Hartman GL. Metagenome-Wide Association Study and Machine Learning Prediction of Bulk Soil Microbiome and Crop Productivity. Frontiers in Microbiology. 2017;8. https://doi.org/10.3389/fmicb.2017.00519.

[41] Jin T, Wang Y, Huang Y, Xu J, Zhang P, Wang N, et al. Taxonomic structure and functional association of foxtail millet root microbiome. GigaScience. 2017;6(10):1–12. https://doi.org/10.1093/gigascience/gix089.

[42] Xu L, Naylor D, Dong Z, Simmons T, Pierroz G, Hixson KK, et al. Drought delays development of the sorghum root microbiome and enriches for monoderm bacteria. Proceedings of the National Academy of Sciences. 2018;115(18):E4284–E4293. https://doi.org/10.1073/pnas.1717308115.

[43] Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods. 2016;13(7):581–583. https://doi.org/10.1038/nmeth.3869.

[44] Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research. 2012;41(D1):D590–D596. https://doi.org/10.1093/nar/gks1219.

[45] Wang Q, Garrity GM, Tiedje JM, Cole JR. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Applied and Environmental Microbiology. 2007;73(16):5261–5267. https://doi.org/10.1128/AEM.00062-07.

[46] McMurdie PJ, Holmes S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLOS ONE. 2013;8(4):e61217. https://doi.org/10.1371/journal.pone.0061217.

[47] Cao Y, Dong Q, Wang D, Zhang P, Liu Y, Niu C. microbiomeMarker: an R/Bioconductor package for microbiome marker identification and visualization. Bioinformatics. 2022;38(16):4027–4029. https://doi.org/10.1093/bioinformatics/btac438.

[48] Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. UpSet: Visualization of Intersecting Sets. IEEE Transactions on Visualization and Computer Graphics.

2014;20(12):1983–1992. https://doi.org/10.1109/TVCG.2014.2346248.

[49] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological). 1995;57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

[50] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;p. 2825–2830.

[51] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence. 2020;2(1):56–67. https://doi.org/10.1038/s42256-019-0138-9.

[52] Naylor D, Coleman-Derr D. Drought Stress and Root-Associated Bacterial Communities. Frontiers in Plant Science. 2018;8.

[53] Nearing JT, Douglas GM, Hayes MG, MacDonald J, Desai DK, Allward N, et al. Microbiome differential abundance methods produce different results across 38 datasets. Nature Communications. 2022;13(1):342. https://doi.org/10.1038/s41467-022-28034-z.

[54] Wallen ZD. Comparison study of differential abundance testing methods using two large Parkinson disease gut microbiome datasets derived from 16S amplicon sequencing. BMC Bioinformatics. 2021;22(1):265. https://doi.org/10.1186/s12859-021-04193-6.

[55] Thorsen J, Brejnrod A, Mortensen M, Rasmussen MA, Stokholm J, Al-Soud WA, et al. Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. Microbiome. 2016;4(1):62. https://doi.org/10.1186/s40168-016-0208-8.

[56] Quinn TP, Crowley TM, Richardson MF. Benchmarking differential expression analysis tools for RNA-Seq: normalization-based vs. log-ratio transformation-based methods. BMC Bioinformatics. 2018;19:274. https://doi.org/10.1186/s12859-018-2261-8.

[57] Lin H, Peddada SD. Analysis of compositions of microbiomes with bias correction. Nature Communications. 2020;11(1):3514. https://doi.org/10.1038/s41467-020-17041-7.

[58] Cawley GC, Talbot NLC. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. Journal of Machine Learning Research. 2010;11(70):2079–2107.

[59] Capitaine L, Genuer R, Thiébaut R. Random forests for high-dimensional longitudinal data. Statistical Methods in Medical Research. 2021 Jan;30(1):166–184. https://doi.org/10.1177/0962280220946080.

[60] Smith MR, Martinez T, Giraud-Carrier C. The potential benefits of data set filtering and learning algorithm hyperparameter optimization. In: Proceedings of the 2015 International Conference on Meta-Learning and Algorithm Selection - Volume 1455. MetaSel'15. Aachen, DEU: CEUR-WS.org; 2015. p. 3–14.

[61] Gao Y, Zhu Z, Sun F. Increasing prediction performance of colorectal cancer disease status using random forests classification based on metagenomic shotgun sequencing data. Synthetic and Systems Biotechnology. 2022;7(1):574–585. https://doi.org/10.1016/j.synbio.2022.01.005.

[62] Chen X, Ishwaran H. Random forests for genomic data analysis. Genomics. 2012;99(6):323–329. https://doi.org/10.1016/j.ygeno.2012.04.003.

[63] Jang SW, Yoou MH, Hong WJ, Kim YJ, Lee EJ, Jung KH. Re-Analysis of 16S Amplicon Sequencing Data Reveals Soil Microbial Population Shifts in Rice Fields under Drought Condition. Rice. 2020;13(1):44. https://doi.org/10.1186/s12284-020-00403-6.

[64] Mehmood MA, Fu Y, Zhao H, Cheng J, Xie J, Jiang D. Enrichment of bacteria involved in the nitrogen cycle and plant growth promotion in soil by sclerotia of rice sheath blight fungus. Stress Biology. 2022;2(1):32. https://doi.org/10.1007/s44154-022-00049-y.

[65] Siebielec S, Siebielec G, Klimkowicz-Pawlas A, Galazka A, Grzadziel J, Stuczyński T. Impact of Water Stress on Microbial Community and Activity in Sandy and Loamy Soils. Agronomy. 2020;10(9):1429. https://doi.org/10.3390/agronomy10091429.

[66] Barka EA, Vatsa P, Sanchez L, Gaveau-Vaillant N, Jacquard C, Klenk HP, et al. Taxonomy, Physiology, and Natural Products of Actinobacteria. Microbiology and Molecular Biology Reviews. 2015;80(1):1–43. https://doi.org/10.1128/MMBR.00019-15.

[67] Tóth Z, Táncsics A, Kriszt B, Kröel-Dulay G, Ónodi G, Hornung E. Extreme effects of drought on composition of the soil bacterial community and decomposition of plant tissue. European Journal of Soil Science. 2017;68(4):504–513. https://doi.org/10.1111/ejss.12429.

[68] Abbasi S, Sadeghi A, Safaie N. Streptomyces alleviate drought stress in tomato plants and modulate the expression of transcription factors ERF1 and WRKY70 genes. Scientia Horticulturae. 2020;265:109206. https://doi.org/10.1016/j.scienta.2020.109206.

[69] Liu L, Lin W, Zhang L, Tang X, Liu Y, Lan S, et al. Changes and Correlation Between Physiological Characteristics of Rhododendron simsii and Soil Microbial Communities Under Heat Stress. Frontiers in Plant Science. 2022;13:950947. https://doi.org/10.3389/fpls.2022.950947.

[70] Faist H, Trognitz F, Antonielli L, Symanczik S, White PJ, Sessitsch A. Potato root-associated microbiomes adapt to combined water and nutrient limitation and have a plant genotype-specific role for plant stress mitigation. Environmental Microbiome. 2023;18(1):18. https://doi.org/10.1186/s40793-023-00469-x.

[71] Zhou J, Wu L, Deng Y, Zhi X, Jiang YH, Tu Q, et al. Reproducibility and quantitation of amplicon sequencing-based detection. The ISME Journal. 2011;5(8):1303–1313. https://doi.org/10.1038/ismej.2011.11.

[72] Abellan-Schneyder I, Matchado MS, Reitmeier S, Sommer A, Sewald Z, Baumbach J, et al. Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing. mSphere. 2021;6(1):e01202–20. https://doi.org/10.1128/mSphere.01202-20.

[73] Fierer N. Embracing the unknown: disentangling the complexities of the soil microbiome. Nature Reviews Microbiology. 2017;15(10):579–590. https://doi.org/10.1038/nrmicro.2017.87.

[74] Chen RC, Dewi C, Huang SW, Caraka RE. Selecting critical features for data classification based on machine learning methods. Journal of Big Data. 2020;7(1):52. https://doi.org/10.1186/s40537-020-00327-4.

[75] Chen H, Ma K, Lu C, Fu Q, Qiu Y, Zhao J, et al. Functional Redundancy in Soil Microbial Community Based on Metagenomics Across the Globe. Frontiers in Microbiology. 2022;13(1). https://doi.org/10.3389/fmicb.2022.878978.

[76] Spinoni J, Vogt JV, Naumann G, Barbosa P, Dosio A. Will drought events become more frequent and severe in Europe? International Journal of Climatology. 2018;38(4):1718–1736. https://doi.org/10.1002/joc.5291.

[77] Ault TR. On the essentials of drought in a changing climate. Science. 2020;368(6488):256–260. https://doi.org/10.1126/science.aaz5492.

[78] No JH, Nishu SD, Hong JK, Lyou ES, Kim MS, Wee GN, et al. Raman-Deuterium Isotope Probing and Metagenomics Reveal the Drought Tolerance of the Soil Microbiome and Its Promotion of Plant Growth. mSystems. 2022 Feb;7(1):e01249–21. Publisher: American Society for Microbiology. https://doi.org/10.1128/msystems.01249-21.