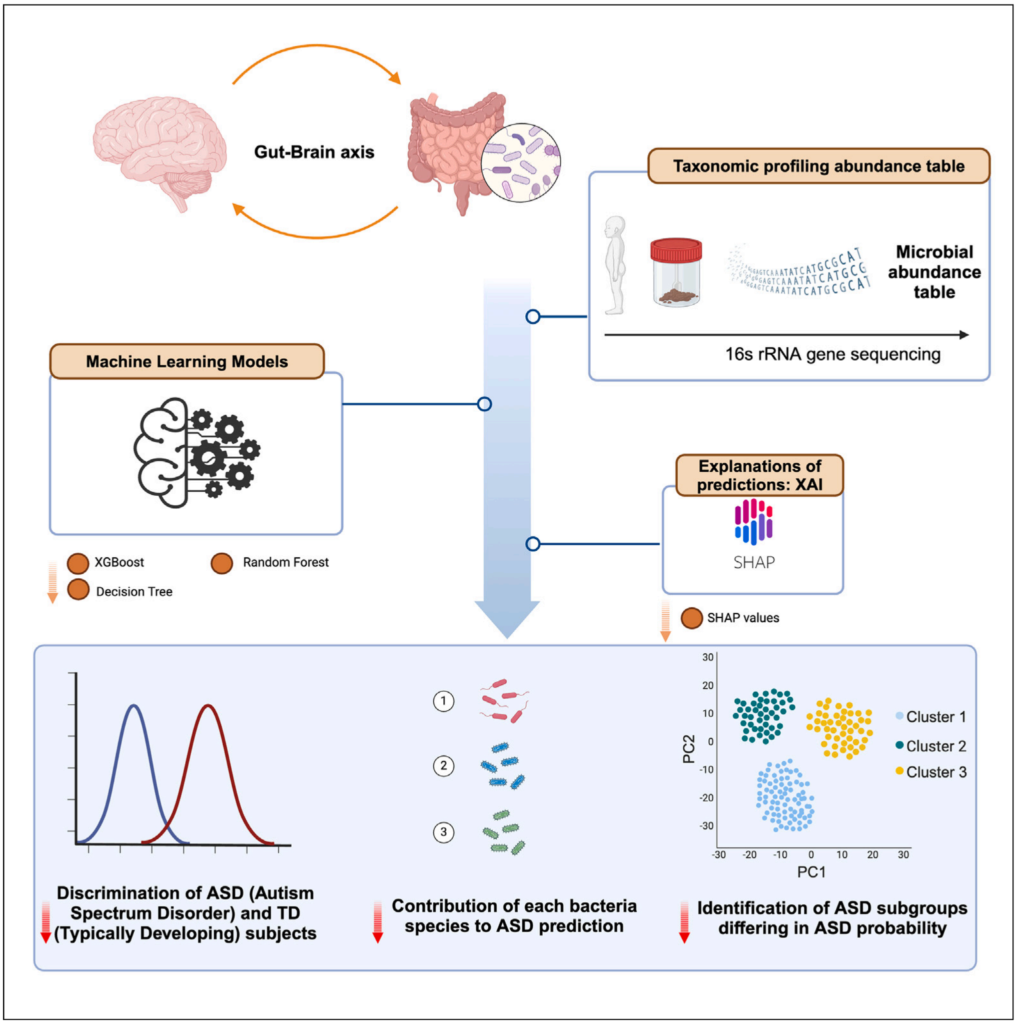
iScience

CellPress
OPEN ACCESS

Article

Personalized identification of autism-related bacteria in the gut microbiome using explainable artificial intelligence


Gut-Brain axis
Taxonomic profiling abundance table
Microbial abundance table
16s rRNA gene sequencing
Machine Learning Models
XGBoost
Random Forest
Decision Tree
Explanations of predictions: XAI
SHAP
SHAP values
1
2
3
30
20
10
0
-10
-20
-30
PC2
-30
-20
-10
0
10
20
30
PC1
Cluster 1
Cluster 2
Cluster 3
Discrimination of ASD (Autism Spectrum Disorder) and TD (Typically Developing) subjects
Contribution of each bacteria species to ASD prediction
Identification of ASD subgroups differing in ASD probability

Pierfrancesco Novielli, Donato Romano, Michele Magarelli, ..., Maria De Angelis, Roberto Bellotti, Sabina Tangaro

sabina.tangaro@uniba.it

Highlights

ML discriminates ASD and TD subjects using microbiome data

XAI identifies personalized microbiome biomarkers linked to ASD

Clustering based on SHAP embeddings reveals ASD subgroups with different ASD probabilities

Novielli et al., iScience *27*, 110709
September 20, 2024 © 2024 Published by Elsevier Inc.
https://doi.org/10.1016/j.isci.2024.110709

iScience

CellPress
OPEN ACCESS

Article

Personalized identification of autism-related bacteria in the gut microbiome using explainable artificial intelligence

Pierfrancesco Novielli,[1,2,5] Donato Romano,[1,2,5] Michele Magarelli,[1] Domenico Diacono,[2] Alfonso Monaco,[2,3] Nicola Amoroso,[2,4] Mirco Vacca,[1] Maria De Angelis,[1] Roberto Bellotti,[2,3,6] and Sabina Tangaro[1,2,6,7,*]

SUMMARY

Autism spectrum disorder (ASD) affects social interaction and communication. Emerging evidence links ASD to gut microbiome alterations, suggesting that microbial composition may play a role in the disorder. This study employs explainable artificial intelligence (XAI) to examine the contributions of individual microbial species to ASD. By using local explanation embeddings and unsupervised clustering, the research identifies distinct ASD subgroups, underscoring the disorder's heterogeneity. Specific microbial biomarkers associated with ASD are revealed, and the best classifiers achieved an AU-ROC of 0.965 ± 0.005 and an AU-PRC of 0.967 ± 0.008. The findings support the notion that gut microbiome composition varies significantly among individuals with ASD. This work's broader significance lies in its potential to inform personalized interventions, enhancing precision in ASD management and classification. These insights highlight the importance of individualized microbiome profiles for developing tailored therapeutic strategies for ASD.

INTRODUCTION

The relationship between the gastrointestinal (GI) tract and the brain, known as the gut-brain axis, is a complex and dynamic system involving neural pathways, hormones, and various molecules. This bidirectional communication was first described by Banks W.A., who highlighted the modulation of cholecystokinin secretion by bombesin.[1] Recent studies have expanded on this concept, demonstrating a significant link between brain health and the functioning of the intestinal microbiota.[2,3] The gut microbiota, a diverse community of microorganisms including bacteria, fungi, viruses, and protozoans,[4] is predominantly located in the GI tract, which hosts over 70% of these microorganisms.[5] Research has shown that the gut microbiota plays a crucial role in modulating the gut-brain axis, influencing various brain disorders, including autism spectrum disorder (ASD). ASD is characterized by challenges in language and communication, social interaction difficulties, and repetitive behaviors. It affects males more frequently than females, with a ratio of 4:1. Symptoms of ASD typically emerge between 12 and 24 months of age, though they can appear earlier if developmental delays are severe or later if symptoms are milder. Despite ongoing research, the exact causes of ASD remain unclear, with neurobiological, genetic, and environmental factors all contributing to its development. Notably, individuals with ASD often experience GI issues such as diarrhea, constipation, bloating, and gastroesophageal reflux.[6] Several studies have compared the gut microbiota of individuals with ASD to that of typically developing (TD) individuals, revealing differences in both species diversity and abundance.[7,8] To further explore these differences, we employed a machine learning (ML) approach using explainable AI (XAI) to identify specific microbiota fingerprints that can distinguish patients with ASD from TD individuals. The application of ML in microbiome analysis has gained significant attention due to advancements in high-throughput sequencing technologies, which have led to an explosion in microbiome data. ML algorithms offer powerful tools for deciphering the complex relationships within microbial communities and understanding their functional impacts. These techniques have been increasingly used to analyze microbiome composition, dynamics, and interactions, providing novel insights into their roles in human health and disease.[9–11] In this study, we utilized the XGBoost classifier to differentiate between children with ASD and TD individuals based on their gut microbiota profiles. XGBoost, a robust ML algorithm, was used to evaluate the global importance of various features in classification. The integration of explainable AI (XAI) methods allowed us to provide local explanations and assess the contribution of each feature for individual subjects,[12–14] enhancing the interpretability of our model.[15] The

[1] Dipartimento di Scienze del Suolo, della Pianta e degli Alimenti, Università degli Studi di Bari Aldo Moro, 70126 Bari, Italy
[2] Istituto Nazionale di Fisica Nucleare, Sezione di Bari, 70125 Bari, Italy
[3] Dipartimento Interateneo di Fisica "M. Merlin", Università degli Studi di Bari Aldo Moro, 70125 Bari, Italy
[4] Dipartimento di Farmacia - Scienze del Farmaco, Università degli Studi di Bari Aldo Moro, 70125 Bari, Italy
[5] These authors contributed equally
[6] These authors contributed equally
[7] Lead contact
*Correspondence: sabina.tangaro@uniba.it

https://doi.org/10.1016/j.isci.2024.110709

Table 1. Comparison between evaluation metrics of XGBoost (XGB), random forest (RF), and decision tree (DT) classifiers

	Accuracy	Sensitivity	Specificity	Precision	AU-ROC	AU-PRC
XGB	0.914 ± 0.008	0.946 ± 0.012	0.872 ± 0.012	0.905 ± 0.008	0.965 ± 0.005	0.967 ± 0.008
RF	0.921 ± 0.006	0.978 ± 0.010	0.847 ± 0.011	0.892 ± 0.007	0.964 ± 0.003	0.962 ± 0.005
DT	0.826 ± 0.016	0.857 ± 0.027	0.785 ± 0.036	0.838 ± 0.021	0.839 ± 0.024	0.840 ± 0.028

The metrics evaluated include accuracy, sensitivity, specificity, precision, AUC-ROC, and AUC-PRC.

motivation behind using this methodology lies in its potential to uncover novel bacterial biomarkers[16] associated with ASD, offering deeper insights into the disease's pathophysiology. By leveraging ML and XAI, we aim to improve the accuracy of ASD diagnosis and contribute to the development of targeted interventions based on gut microbiota profiles. The contributions of this article are 3-fold. First, it introduces an innovative application of XAI to microbiome analysis, emphasizing the importance of transparency and interpretability in AI-driven biomedical research. Second, it demonstrates the potential of personalized medicine to uncover unique microbiome profiles related to ASD, paving the way for customized therapeutic strategies. Third, leveraging local explanation embeddings and an unsupervised clustering method successfully clusters ASD subjects into subgroups. The chosen methodology is particularly well-suited to address the complexities and variabilities inherent in ASD. By leveraging explainable AI, we ensure that our findings are not only robust but also accessible and actionable, facilitating their translation into clinical practice. This approach represents a significant advancement over traditional methods, providing deeper insights into the microbiome-ASD relationship and highlighting the potential for personalized medicine in this field.[12–14]

RESULTS

The goal of this study is to explore the alterations of the gut microbiota in subjects with ASD to respect the TD subjects.

To identify the microbiome alterations in ASD compared to TD, a classification model based on machine learning has been trained, and the feature contributions have been discussed.

In this study, the performances of three supervised machine learning algorithms, XGBoost, Random Forest, and a simpler one such as decision tree, were compared. The best model was chosen as the one having the highest AU-PRC (Area Under Precision-Recall curve), which was XGBoost, as it outperformed the other two models in terms of AU-PRC, Specificity, and Precision. As summarized in Table 1, the most efficient model was XGBoost.

Figure 1 presents the boxplots representing the distributions of model performances across 50 runs, including metrics such as Accuracy, Sensitivity, Specificity, Precision, AU-ROC, and AU-PRC.

The feature importance has been explored in order to extract biological features with high discriminating power in the classification.

In particular, XGBoost incorporates an embedded feature importance mechanism that enables the effective analysis and optimization of selected features during the classification process.[15] The results of feature importance obtained through the XGBoost algorithm are summarized in Figure 2 where the boxplot of the importance coefficients of repetited cross-validation is represented. In XGBoost, the feature importance is calculated based on the "total gain" metric. This importance type can be defined as the cumulative gain achieved by a feature across all splits it is used in during the boosting process. During the training of the XGBoost model, each feature is evaluated for its ability to split the data and improve the model's performance. The gain is calculated for each split by considering the improvement in the model's loss function achieved through the use of that feature. The total gain for a feature is the sum of the gains across all the splits in which the feature is involved. By considering the total gain of a feature, XGBoost identifies the features that contribute the most to reducing the loss function and making accurate predictions. Higher values of total gain indicate greater importance, indicating that the feature has a more significant impact on the model's performance. The feature importance analysis for the Random Forest model is provided in Figure S1. This figure highlights the most influential features in the classification task, ranked by their importance scores. The feature importance is calculated based on the average decrease in impurity (Gini importance) across all trees in the Random Forest. The top features identified align closely with those determined by the XGBoost model, indicating consistency between the models. Specifically, the Dice index for the top 20 important features is 0.5, while for the top 5 features, the Dice index is 0.8, demonstrating a strong overlap in the most significant features.

Explainable AI solutions has been implemented, providing results that are understandable and verifiable for each subjects. SHAP methods is a local model-agnostic as explains the predictions at individual level regardless the selected models. Basically, this method learns an interpretable linear model around each test instance and estimate feature importance at local level. Figure 3 shows the most important features for classification according to the SHAP algorithm for the XGB model. The SHAP analysis for the Random Forest model is provided in the Figure S2. This plot provides a comprehensive view of the impact of each feature on the model's output, with SHAP values indicating the direction and magnitude of feature effects. The features are ranked by their average absolute SHAP values, showing the most impactful features at the top. The Shapley values are calculated by averaging across all iterations of the algorithm for each subject, taking into account the 50 repetitions. The plot offers insights into how different feature values influence the model's predictions, facilitating an understanding of the model's decision-making process. The comparison of SHAP values between Random Forest and XGBoost shows a good overlap, particularly for the top features, with a Dice index of 0.6 for the top 20 features and 0.8 for the top 5 features.

Figure 3 shows that there are OTUs, such as OTU625, for which a high relative abundance (red areas of the violin plots) is present on the positive side of the x axis, and a low relative abundance (blue areas of the violin plots) is more prevalent on the negative side of the x axis. This

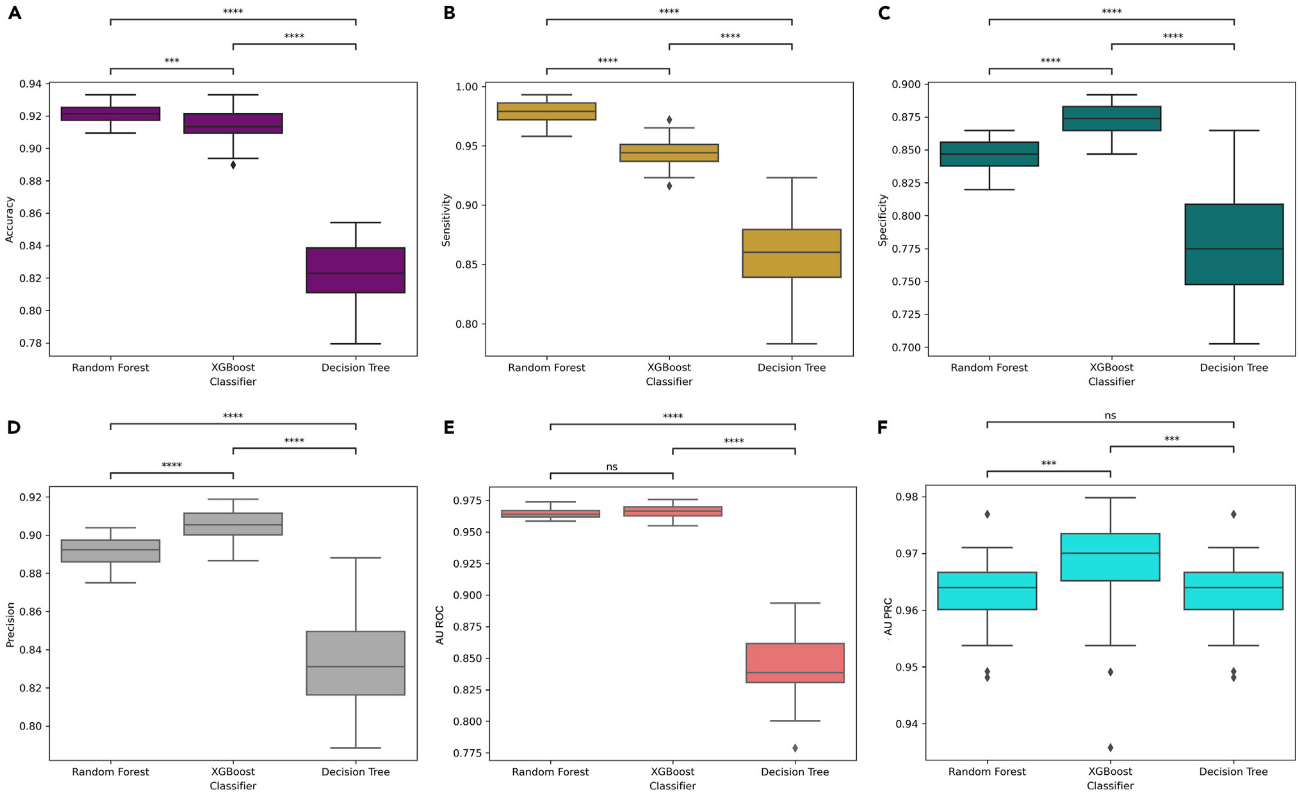


Figure 1. Comparative performance metrics of ML models across 50 runs with statistical significance annotations

(These boxplots represent the distribution of metrics for three different models, providing insights into the model performance across 50 runs of the algorithms. Panel (a) shows Accuracy, (b) shows Sensitivity, (c) shows Specificity, (d) shows Precision, (e) shows AU-ROC, and (f) shows AU-PRC. The asterisks (*) on the plots indicate the *p*-value from the statistical comparison between different distributions, obtained using the Mann-Whitney U test. The significance levels are as follows: * for *p*-values between 0.05 and 0.01, ** for *p*-values between 0.01 and 0.001, *** for *p*-values between 0.001 and 0.0001, and **** for *p*-values less than 0.0001. The Decision Tree consistently performs worse than the other two models. The Random Forest outperforms the others in terms of Accuracy and Sensitivity, while XGB outperforms the others in terms of Specificity, Precision, and AU-PRC).

means that a high relative abundance of these bacteria is generally associated with a higher probability value to ASD, whereas a low relative abundance is associated with a lower probability value t ASD. On the contrary, there are OTUs, such as OTU1301, that exhibit an opposite pattern, suggesting that a high abundance of these species is linked to a lower probability of ASD. These insights regarding the direction of effects cannot be deduced using global explanation methods such as XGBoost's built-in feature importance.

The importance ranking of features obtained from both XGB and SHAP values exhibit a substantial degree of overlap (*Jaccard Index* = 0.54), underscoring the robustness and stability of the model. This convergence confirms the consistency and reliability of the model's assessment of feature relevance. As example of local effects of feature in Figures 4A and 4B two example of subjects classified respectively as ASD and TD. For each local explanation, it can be observed the contributions of each feature (bacteria) to the prediction, arranging the bacteria in order of importance. These figures also reveal which features are confounding for the model's outcome. For instance, in Figure 4B, illustrating the SHAP values of a TD subject, there are features such as OTU910 or OTU679, according to which the subject would be classified as ASD. However, when considering the contribution of all other features, the subject is correctly classified as TD.

Autism subgroups

In this study, a comparison was made between the outcomes of PCA conducted on SHAP values and those resulting from PCA applied to microbial abundance data.[17] To accomplish this, after the computation of SHAP values, PCA was performed on these values. The first two principal components, PC1 and PC2, were subsequently calculated, followed by the determination of PC loadings, and the visualization of the results (Figure 6). For the purpose of comparison, Figure 5 also includes the results of PCA obtained from relative abundance values.

Figure 6 presents the projection of SHAP values onto the two principal components. Upon examining the findings, a consistent pattern emerges, revealing a distinct segregation between TD and ASD subjects along PC1. In contrast, the PCA results based on relative abundance data (Figure 6) do not exhibit a clear differentiation between TD and ASD samples. Additionally, ASD-associated bacteria, such as OTU625, OTU976, OTU1301, OTU390, and OTU1225, exhibit high PCA loadings in the SHAP value-based PCA (Figure 6). SHAP value embeddings,

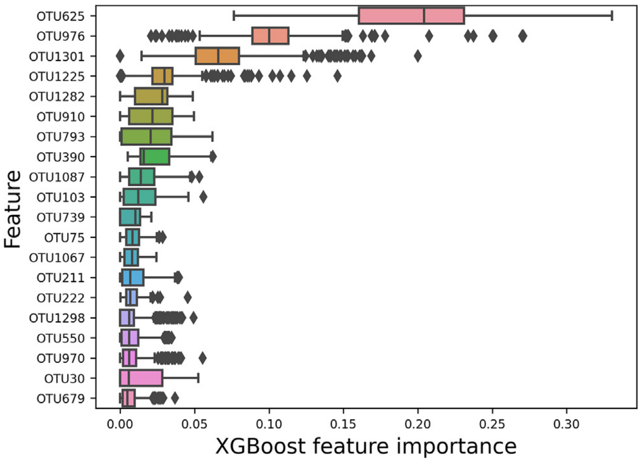


Figure 2. XGB features important plot

(Feature importance plot obtained using XGBoost. The image displays the top 20 features ranked by their importance, based on the analysis performed with cross-validation repeated 50 times. The distributions of the feature importance values are shown as boxplots for each fold of the cross-validation process, providing insights into the relative significance of these features in the predictive model).

also known as local explanation embeddings, outperform in capturing insights, making it easier to discern variations of interest (TD vs. ASD) compared to the original abundance data.

Furthermore, in Figure 7, the PCA plots of SHAP values are color-coded based on the ASD probability values obtained from the classifier. A consistent pattern emerges where individuals situated on the right side of the PC1 axis tend to exhibit lower ASD probabilities, and vice versa. This highlights an additional advantage of employing PCA on SHAP values, as it provides more interpretable results in terms of the trend in ASD probability across the PC space.

Another widely used technique for data visualization and dimensionality reduction, t-SNE, was also explored. t-SNE is a popular algorithm that originates from probability theory. It projects high-dimensional data points into 2D by aiming to give the projected data a similar distribution to the original data points, achieved through the minimization of the Kullback-Leibler divergence.[18] Although t-SNE yielded similar visualization results, PCA was preferred because, with PC loadings (Figure 7), it was possible to observe features highly correlated with the two principal components. It was observed that the features most correlated with the principal components correspond to the most important features listed in the summary plot in Figure 3.

Given the clear separation observed between ASD and TD (Figure 6) and the consistent ASD probability trend (Figure 7), it becomes plausible to cluster the ASD data points into distinct subgroups, presuming that these subgroups would manifest similar ASD probabilities. To assess this hypothesis, K-means clustering was conducted on the PC1 and PC2 values derived from the SHAP values associated with the ASD subjects. The determination of the optimal number of clusters involved the application of the Elbow method and Silhouette score. Additionally, the Silhouette score was utilized to compare the clustering outcomes derived from the first two principal components of PCA with those from the two t-SNE components and all SHAP values. The comparison revealed that clustering on all SHAP values exhibited a significantly lower Silhouette coefficient compared to clustering on the first two PCA components and t-SNE components (with the latter showing comparable results). This finding underscores the efficacy of dimensionality reduction on SHAP values, serving not only for visualization but also to enhance clustering performance.

The resulting K-means clusters of ASD subjects are shown in Figure 8. Three distinct clusters can be identified from this figure. As illustrated in Figure 9, both Cluster 1 and Cluster 2 showed higher ASD probabilities compared to the other clusters. Particularly, individuals grouped in Cluster 3 displayed the lowest median ASD probability. From this unsupervised analysis conducted on ASD subjects, has emerged a cluster (Cluster 3) in which are present all the ASD subjects misclassified as TD by the ML model, thus representing the false negatives.

DISCUSSION

The healthy adult gut microbiota (GM) is composed of four main phyla which together represent almost the totality of the bacterial population: Bacteroidetes (Gram-negative, such as *Bacteroides* and *Prevotella* genera), Firmicutes (Gram-positive aerobic and anaerobic bacteria such as *Lactobacillus*, *Clostridium*, and *Ruminococcus*), Proteobacteria (e.g., *Enterobacteriaceae*) and Actinobacteria (e.g., *Bifidobacterium*), followed by the minor phyla Fusobacteria and Verrucomicrobia. By means of data from the biggest studies on GM,[19–21] it was shown how Bacteroidetes and Firmicutes comprise around the 90% of the total bacteriome relative abundance. On average, the phylum Bacteroidetes is responsible for the proteolytic metabolism and secondary digestion of polysaccharides exerting a broad range of activities in hosts, from beneficial to pathogenic, based on specific conditions and according to the strain belonging.[22] The phylum Firmicutes, instead, predominantly contributes to the metabolism of food fatty acids and influences the nutrient intestinal absorption through the primary breakdown of complex substrates undigested by hosts.[23] In health, GM supports the immune system, aids in dietary nutrient metabolism and absorption,

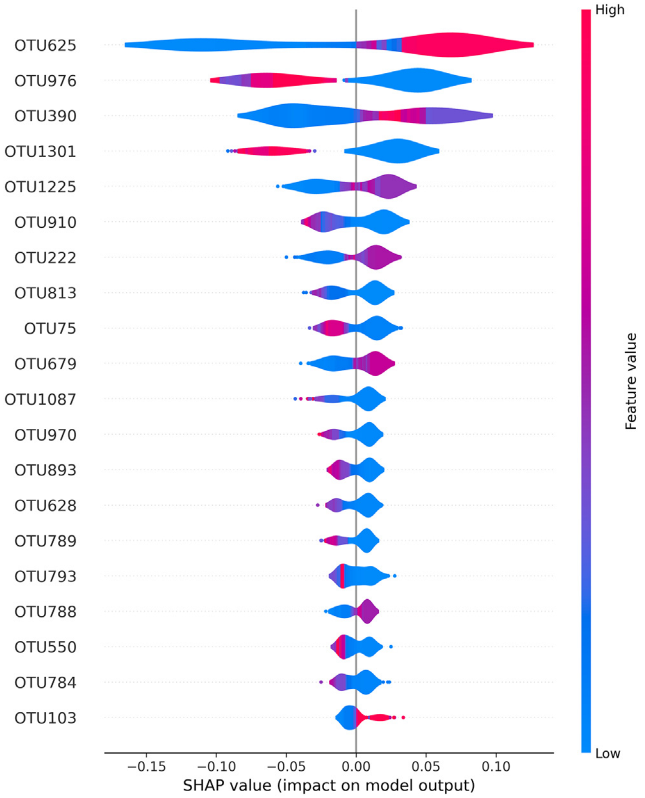


Figure 3. SHAP summary plot with violin plots

(SHAP summary plot illustrating the violin plots of the SHAP values for each feature. Each point on the plot represents a Shapley value of a subject, with the y axis indicating the corresponding feature and the x axis representing the Shapley value itself. The color gradient reflects the feature value, ranging from low to high. The features are ordered based on their mean importance, with more important features positioned toward the top. Violin plots utilize "violin-shaped" figures to illustrate the distribution and density of SHAP values for each feature, offering insights into range, variability, skewness, symmetry, and multimodality of the SHAP value distribution).

and produces beneficial molecules (e.g., specific vitamins and bacteriocins) to promote the host's well-being. However, when gut homeostasis is disrupted, whatever decrease in GM diversity, loss in keystone taxa and bloom of pathobionts leads to the status widely identified as intestinal dysbiosis.[24],[25] Dysbiosis entails the onset of adverse events, e.g., increased intestinal permeability, exogenous food-borne peptides, or neurotoxic bacterial-derived peptides and lipopolysaccharide (LPS) leading to an increased expression of inflammatory cytokines in hosts. Similarly, a broad spectrum of diseases contributes to the dysbiosis onset as the result of alterations in the host physiology, including ASD.[26]

In this study, a framework was developed to use XAI techniques in identifying the most relevant features for predicting ASD in children based on species-level microbiome abundance data. The strength of this analysis lies in the utilization of a multivariate approach to construct a machine learning framework that achieves the accurate classification of ASD versus TD children. The addition of SHAP values provides insightful information regarding the contribution of each feature. Furthermore, SHAP analysis enables the exploration of both global and local effects, offering a comprehensive understanding of the interplay between features and the classification algorithm. This capability has promising implications for personalized medicine, as it allows for an individualized assessment of the role played by each feature in the predictive model for each subject.[27]

Through a Shap analysis, the most important OTUs were identified out of 1322, deriving from the analysis of the microbiota of 254 subjects aged between 2 and 13 years, 143 of these with typical development and 111 with ASD.[8] Each OTU corresponds to a microbe, classified at domain, phylum, class, order, family, genus, and species level. In particular, the first 20 most important OTUs reported in Table 2 has been examined.

The results of the Shap analysis highlight that two Phyla, Firmicutes and Bacteroidetes, were the most representative of differences featuring the GM in ASD. The two Bacteroidetes taxa (OTU 625 and 390) belong to the class *Bacteroidia* and to the order *Bacteroidales*, differing at the family level in *Prevotellaceae* and *Porphyromonadaceae*, respectively. According to the study by Arumugam et al., *Prevotella* was identified as a keystone taxon of the enterotype 2.[28] In the GM, the *Prevotella* genus is a dietary fiber fermenter and shows potential as a biomarker due to its metabolite signature and high genetic diversity. Contributing to polysaccharide breakdown, *Prevotella* abundances were positively associated with beneficial propionate metabolism.[29] However, due to its strain variability in hosts,[30] *Prevotella* can also be

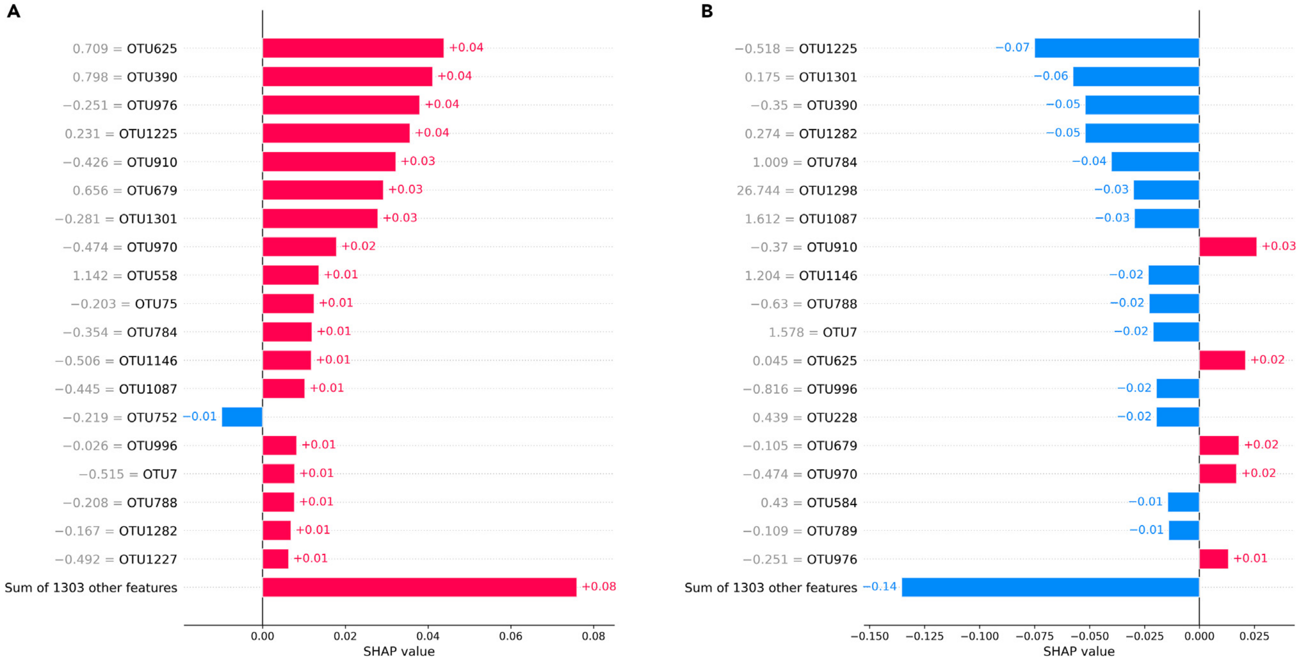


Figure 4. SHAP values bar plots

(A) SHAP values bar plot for a child with autism spectrum disorder (ASD) (A) and a child with TD (B). The plot identifies the most influential feature(s) and their impact on the ASD or TD classification. The SHAP values for these subjects are associated with a single iteration of the pipeline workflow, but the average SHAP values per subject show a similarity to these results).

considered an intestinal pathobiont, as it has been correlated with HIV[31,32] and colitis.[33] *Porphyromonadaceae* (i.e., *Barnesiella*), identified as significantly impacting the ASD fingerprint, has demonstrated the capability to produce butyrate and iso-butyrate from glucose.[34] Sixteen taxa, instead, belong to the phylum Firmicutes, fourteen of these belong to the class Clostridia, order *Clostridiales*. Among *Clostridiales*, 12 out of 14, were *Ruminococcaceae* (OTU 976, 910, 222, 75, 970, 628, 789, and 788) or *Lachnospiraceae* (OTU 1225, 679, 893, 793, 784). Species from both families are recognized as the main colonizers of GM since birth and are considered part of the core GM in humans.[35] Moreover, their genomes have been widely investigated for their fibrinolytic specialization[36] and their involvement in the metabolism of short-chain fatty acids (SCFAs).[37] Similar to *Prevotella*, *Lachnospiraceae* has shown controversial results in terms of activity in hosts, being found to increase in a wide number of patients.[36] However, considering the involvement of *Prevotella* and *Barnesiella* and the high number of OTUs

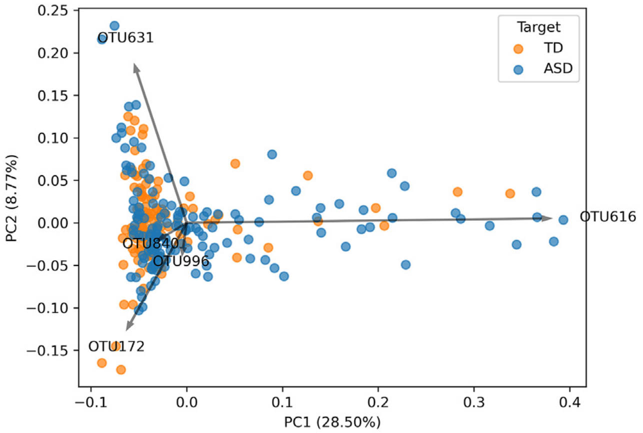


Figure 5. PCA biplot of relative abundance data for TD and ASD subjects

(PCA biplot displaying relative abundance data, where blue dots correspond to TD subjects and orange dots represent individuals with ASD. The OTUs with the largest PC loadings can be seen in the figure, showing a strong correlation with the principal components. The x and y axes represent the first two principal components (PC1 and PC2), with their explained variance ratio).

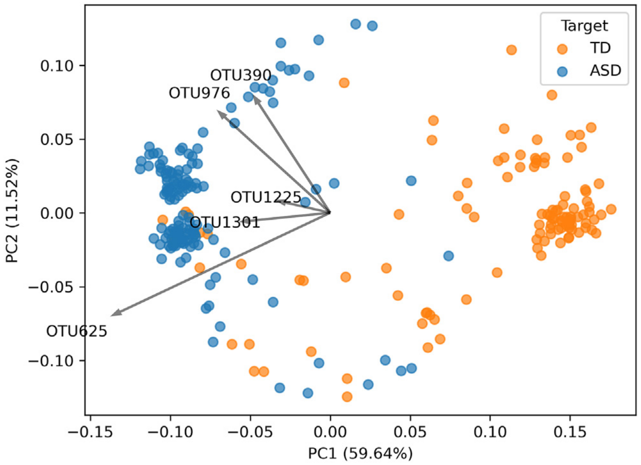


Figure 6. PCA biplot of SHAP values in TD and ASD subjects

(PCA biplot of SHAP values displaying relative abundance data, where blue dots correspond to TD subjects and orange dots represent individuals with ASD. Bacteria linked to autism spectrum disorder (ASD), including OTU625, OTU976, OTU1301, OTU390, and OTU1225, exhibit high PC loadings, indicating robust associations with the principal components. The x and y axes represent the first two principal components (PC1 and PC2), with their explained variance ratio).

from *Ruminococcaceae* and *Lachnospiraceae*, it is possible to speculate that impaired SCFA metabolism could be recognized in ASD, as suggested by previous investigations.[38] The other OTUs of Firmicutes were identified as *Megasphaera* (OTU 1301), *Christensenellaceae R-7 group* (OTU 813), and *Holdemanella* (OTU 103). The *Megasphaera* metagenome has been studied, allowing us to define its bile resistance, presence of various sensory and regulatory systems, stress response systems, membrane transporters, and antibiotic resistance.[39] Few studies have focused on the *Christensenellaceae R-7 group* and *Holdemanella*, mainly reporting correlation analyses in various experimental settings, such as obesity,[40,41] major depressive disorder,[42] and HIV.[43] The last two OTUs belonged to the Proteobacteria phylum. Sutterella (OTU 1087) has been noticed as a potential biomarker of ASD,[44] whereas OTU 550, lacking identification at the genus level, was identified as a microbe belonging to the *Enterobacteriaceae* family, a usual colonizer of the GM.

Moreover, in this study was explored the potential use of SHAP values for enhancing the interpretability of PCA outcomes in the context of microbiome-disease data. Microbiome datasets typically exhibit high dimensionality, characterized by a considerably larger number of features (taxa) compared to the total number of samples. The reduction of data dimensionality and its projection onto a lower-dimensional space offer valuable insights for data exploration and visualization purposes. Specifically, PCA frequently serves as a means to visualize sample similarities within a 2D or 3D space, representing a valuable preliminary step prior to sample clustering or classification.[45]

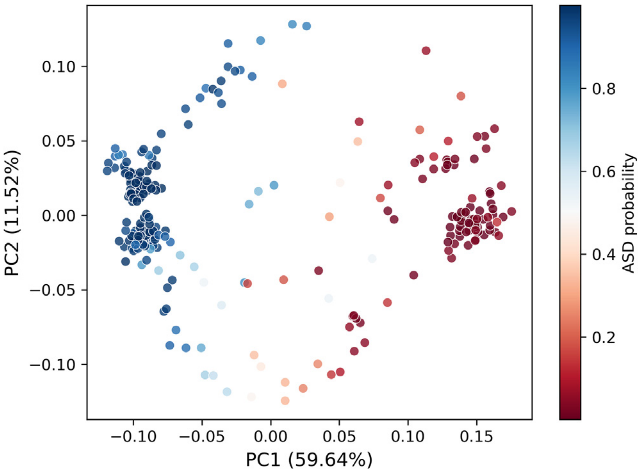


Figure 7. PCA plots of SHAP values with ASD probability overlay

(PCA plots illustrating SHAP values overlaid with ASD probability, revealing a distinct pattern of escalating ASD likelihood from the right to the left side of PC1. The x and y axes represent the first two principal components (PC1 and PC2), with their explained variance ratio).

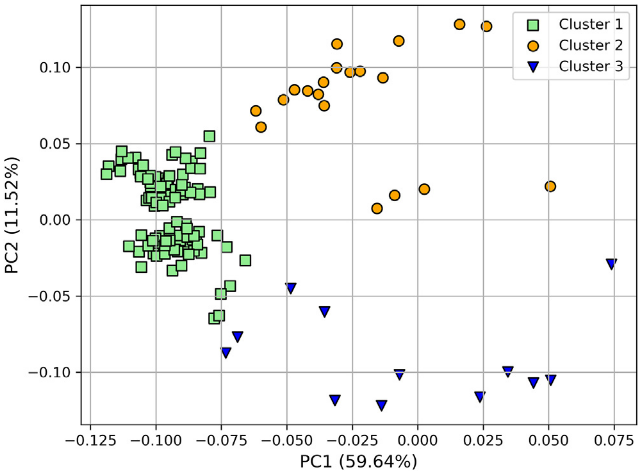


Figure 8. K-means clustering outcome on local explanation embeddings for ASD individuals
(Outcome of K-means clustering applied to the local explanation embeddings of individuals with ASD. The x and y axes represent the first two principal components (PC1 and PC2), with their relative explained variance).

In the PCA plot of microbial relative abundance data, a clear separation between TD and ASD samples is not evident. On the contrary, when PCA is applied to SHAP values, a more pronounced separation becomes observable. Additionally, a trend of increasing ASD probability along PC1 was also observed, as subjects located on the left side of PC1 exhibit a higher probability of ASD, and vice versa.

Furthermore, by leveraging local explanation embeddings and an unsupervised clustering method, it was possible to cluster ASD subjects into subgroups. This analysis revealed a cluster with a lower ASD probability, which allowed the identification of false negatives. The recognition of false negatives holds substantial clinical implications, necessitating a deeper exploration of its ramifications for patient care. This observation prompts a critical examination of potential factors contributing to false negatives, thereby offering valuable insights for refining the accuracy of ASD classification. The identification of a cluster with lower ASD probability not only contributes to the understanding of ASD heterogeneity but also presents opportunities for tailoring diagnostic strategies and interventions based on individualized profiles within this subgroup.

Conclusion

In summary, this study has explored the application of explainable artificial intelligence (XAI) in the context of gut microbiome-based autism spectrum disorder (ASD) classification. The utilization of the SHapley Additive exPlanations (SHAP) algorithm has demonstrated its capacity to derive personalized feature importance, enabling the identification of potential bacterial biomarkers associated with ASD. The proposed

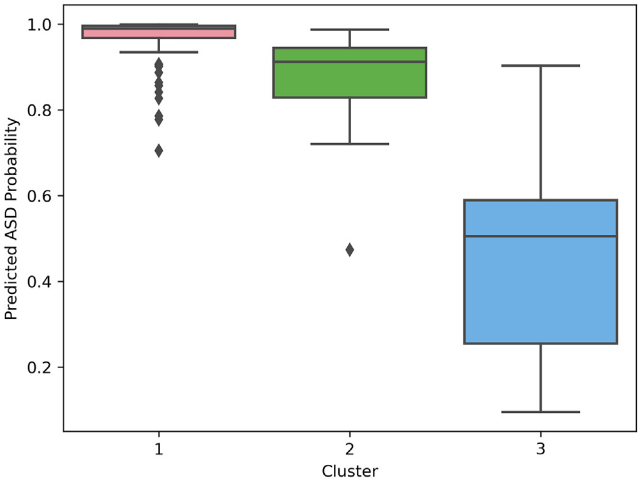


Figure 9. Boxplots of ASD probability distributions across different clusters
(Boxplots of the distributions of ASD probabilities among different clusters).

Table 2. Classification of the first 20 OTUs deriving from the Shap analysis

OTU	Phylum	Class	Order	Family	Genus	Species
625	Bacteroidetes	*Bacteroidia*	*Bacteroidales*	*Prevotellaceae*	*Prevotella 2*	*Uncultured bacterium*
976	Firmicutes	*Clostridia*	*Clostridiales*	*Ruminococcaceae*	*Ruminococcaceae UCG-014*	*Unclassified bacterium*
390	Bacteroidetes	*Bacteroidia*	*Bacteroidales*	*Porphyromonadaceae*	*Barnesiella*	*Uncultured bacterium*
1301	Firmicutes	*Negativicutes*	*Selenomonadales*	*Veillonellaceae*	*Megasphaera*	*Uncultured bacterium*
1225	Firmicutes	*Clostridia*	*Clostridiales*	*Lachnospiraceae*	*Lachnoclostridium*	*Uncultured bacterium*
910	Firmicutes	*Clostridia*	*Clostridiales*	*Ruminococcaceae*	*Ruminococcaceae UCG-014*	*Unclassified bacterium*
222	Firmicutes	*Clostridia*	*Clostridiales*	*Ruminococcaceae*	*Intestinimonas*	*Intestinimonas butyriciproducens*
813	Firmicutes	*Clostridia*	*Clostridiales*	*Christensenellaceae*	*Christensenellaceae R-7 group*	*Uncultured bacterium*
75	Firmicutes	*Clostridia*	*Clostridiales*	*Ruminococcaceae*	*Subdoligranulum*	*Uncultured bacterium*
679	Firmicutes	*Clostridia*	*Clostridiales*	*Lachnospiraceae*	*Lachnospira*	*Uncultured bacterium*
1087	Proteobacteria	*Betaproteobacteria*	*Burkholderiales*	*Alcaligenaceae*	*Sutterella*	*Gutmetagenome Sutterella*
970	Firmicutes	*Clostridia*	*Clostridiales*	*Ruminococcaceae*	*Ruminiclostridium 6*	*Unclassified bacterium*
893	Firmicutes	*Clostridia*	*Clostridiales*	*Lachnospiraceae*	*Anaerostipes*	*Anaerostipes hadrus*
628	Firmicutes	*Clostridia*	*Clostridiales*	*Ruminococcaceae*	*Ruminiclostridium 6*	*Eubacterium siraeum DSM*$_1$5702
789	Firmicutes	*Clostridia*	*Clostridiales*	*Ruminococcaceae*	*Ruminococcus 1*	*Uncultured bacterium*
793	Firmicutes	*Clostridia*	*Clostridiales*	*Lachnospiraceae*	*Lachnospiraceae NK4A136 group*	*Uncultured bacterium*
788	Firmicutes	*Clostridia*	*Clostridiales*	*Ruminococcaceae*	*Ruminococcaceae UCG-004*	*Uncultured bacterium*
550	Proteobacteria	*Gammaproteobacteria*	*Enterobacteriales*	*Enterobacteriaceae*	*Unclassified bacteria*	*Unclassified bacteria*
784	Firmicutes	*Clostridia*	*Clostridiales*	*Lachnospiraceae*	*Eubacterium xylanophilum*	*Unclassified bacterium*
103	Firmicutes	*Erysipelotrichia*	*Erysipelotrichales*	*Erysipelotrichaceae*	*Holdemanella*	*Uncultured bacterium*

The OTU column indicates the top 20 most impacting OTUs in the predictive model. The Phylum, Class, Order, Family, Genus and Species columns indicate the classification of bacteria corresponding to each OTU.

methodology extends its utility to facilitate data exploration in the domain of microbiome-disease association, particularly through the generation of interpretable Principal Component Analysis (PCA) results. Notably, this approach effectively unveils distinct ASD subgroups with varying probabilities. This study contributes valuable insights and tools to advance our understanding of ASD and microbiome interactions.

Limitations of the study

As a pilot study, our research is subject to several limitations. First, it is an observational study wherein we examine the microbiota of certain children without any intervention procedures. Given the nature of this study, establishing a cause-and-effect relationship between microbiota dysbiosis and the onset of autism is challenging. Consequently, this study is confined to exploring the association between these two phenomena, and a clinical intervention study will be necessary to investigate causal relationships. Another limitation of this study is the lack of

Table 3. Demographic characteristics of the study participants

	ASD (111)	TD (143)	*p*-value
Age	5.09 ± 1.99	4.94 ± 1.85	0.59
Gender	99 M/12 F	130 M/13 F	0.65

The Mann-Whitney test was performed for age, while the chi-squared test for gender.

external validation using gut microbiota data from other geographical locations. The current study's dataset was sourced entirely from Dan et al. 2020,[8] and no external datasets were available for testing the model's performance. Future research should aim to include external data to evaluate the generalizability and robustness of the model across different populations. Access to diverse datasets will help in understanding the influence of geographical and environmental factors on the gut microbiome composition in ASD and TD children.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - ○ Metataxonomic (16S rRNA gene) analysis
  - ○ Machine learning based classification
  - ○ Evaluation metrics
  - ○ SHapley Additive exPlanations (SHAP) analysis
  - ○ Clustering SHAP values based
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.110709.

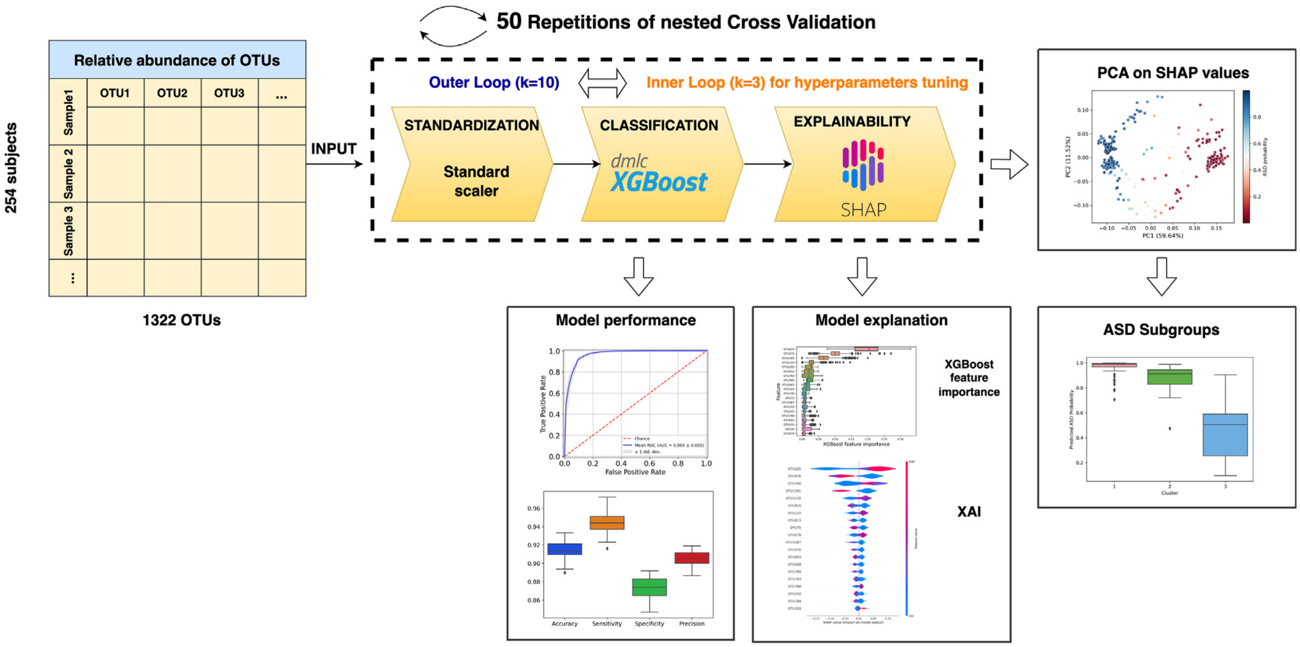


Figure 10. Schematic workflow of the performed analyses

ACKNOWLEDGMENTS

The authors thank ReCaS-Bari for providing the computing facilities. This research was funded by: the National Institute for Nuclear Physics (INFN), next_AIM (Artificial Intelligence in Medicine: next steps) research project (INFN-CSN5), https://www.pi.infn.it/aim (accessed on 30 October 2023); The University of Bari, XAI4Microbiome, Project code S30 – CUP H99J21017720005, Horizon Europe Seeds; National Recovery and Resilience Pan (NRRP), Mission 4 Component 2 Investment 1.4—Call for tender No. 3138 of 16 December 2021 of Italian Ministry of University and Research funded by the European Union—NextGenerationEU. Award Number: Project code: CN00000013, Concession Decree No. 1031 of 17 February 2022 adopted by the Italian Ministry of University and Research, CUP H93C22000450007, Project title: "National Centre for HPC, Big Data and Quantum Computing." The research was also funded by project called "Explained artificial intelligence for the identification of new biomarkers for microbiome-driven atrial fibrillation – Acronym: XAI_FA_Microbiomarker. Concession Decree No. 1319 of 25/07/2024 adopted by the Italian Ministry of University and Research, Unique Project Code (CUP) B83C22003920001.

AUTHOR CONTRIBUTIONS

Conceptualization: P.N., D.R, and S.T.; methodology: P.N., D.R, and S.T.; software: P.N. and D.R.; writing—original draft preparation, P.N., D.R., M.M., and S.T.; writing—review and editing, P.N., D.R., M.M., D.D., A.M., N.A., M.V., M.D.A., R.B., and S.T.; visualization, P.N., D.D., and S.T.; supervision: R.B and S.T.; funding acquisition, S.T. All authors have read and agreed to the published version of the article.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 2, 2024
Revised: July 5, 2024
Accepted: August 7, 2024
Published: August 13, 2024

REFERENCES

1. Banks, W.A. (1980). Evidence for a cholecystokinin gut-brain axis with modulation by bombesin. Peptides *1*, 347–351. https://doi.org/10.1016/0196-9781(80)90013-3.
2. Bercik, P., Collins, S.M., and Verdu, E.F. (2012). Microbes and the gut-brain axis. Neuro Gastroenterol. Motil. *24*, 405–413. https://doi.org/10.1111/j.1365-2982.2012.01906.x.
3. Shahin, K., Soleimani-Delfan, A., He, Z., Sansonetti, P., and Collard, J.M. (2023). Metagenomics revealed a correlation of gut phageome with autism spectrum disorder. Gut Pathog. *15*, 39. https://doi.org/10.1186/s13099-023-00561-0.
4. Sekirov, I., Russell, S.L., Antunes, L.C.M., and Finlay, B.B. (2010). Gut microbiota in health and disease. Physiol. Rev. *90*, 859–904. https://doi.org/10.1152/physrev.00045.2009.
5. Ley, R.E., Peterson, D.A., and Gordon, J.I. (2006). Ecological and evolutionary forces shaping microbial diversity in the human intestine. Cell *124*, 837–848. https://doi.org/10.1016/j.cell.2006.02.017.
6. Kang, V., Wagner, G.C., and Ming, X. (2014). Gastrointestinal dysfunction in children with autism spectrum disorders. Autism Res. *7*, 501–506. https://doi.org/10.1002/aur.1386.
7. De Angelis, M., Piccolo, M., Vannini, L., Siragusa, S., De Giacomo, A., Serrazzanetti, D.I., Cristofori, F., Guerzoni, M.E., Gobbetti, M., and Francavilla, R. (2013). Fecal microbiota and metabolome of children with autism and pervasive developmental disorder not otherwise specified. PLoS One *8*, e76993. https://doi.org/10.1371/journal.pone.0076993.
8. Dan, Z., Mao, X., Liu, Q., Guo, M., Zhuang, Y., Liu, Z., Chen, K., Chen, J., Xu, R., Tang, J., et al. (2020). Altered gut microbial profile is associated with abnormal metabolism activity of autism spectrum disorder. Gut Microb. *11*, 1246–1267. https://doi.org/10.1080/19490976.2020.1747329.
9. Golob, J.L., Oskotsky, T.T., Tang, A.S., Roldan, A., Chung, V., Ha, C.W., Wong, R.J., Flynn, K.J., Parraga-Leo, A., Wibrand, C., et al. (2023). Microbiome preterm birth dream challenge: Crowdsourcing machine learning approaches to advance preterm birth research. Preprint at medRxiv. https://doi.org/10.1101/2023.03.07.23286920.
10. Bellando-Randone, S., Russo, E., Venerito, V., Matucci-Cerinic, M., Iannone, F., Tangaro, S., and Amedei, A. (2021). Exploring the oral microbiome in rheumatic diseases, state of art and future prospective in personalized medicine with an ai approach. J. Personalized Med. *11*, 625. https://doi.org/10.3390/jpm11070625.
11. Papoutsoglou, G., Tarazona, S., Lopes, M.B., Klammsteiner, T., Ibrahimi, E., Eckenberger, J., Novielli, P., Tonda, A., Simeon, A., Shigdel, R., et al. (2023). Machine learning approaches in microbiome research: Challenges and best practices. Front. Microbiol. *14*, 1261889. https://doi.org/10.3389/fmicb.2023.1261889.
12. Lombardi, A., Diacono, D., Amoroso, N., Biecek, P., Monaco, A., Bellantuono, L., Pantaleo, E., Logroscino, G., De Blasi, R., Tangaro, S., and Bellotti, R. (2022). A robust framework to investigate the reliability and stability of explainable artificial intelligence markers of mild cognitive impairment and alzheimer's disease. Brain Inform. *9*, 17. https://doi.org/10.1186/s40708-022-00165-5.
13. Lombardi, A., Diacono, D., Amoroso, N., Monaco, A., Tavares, J.M.R.S., Bellotti, R., and Tangaro, S. (2021). Explainable deep learning for personalized age prediction with brain morphology. Front. Neurosci. *15*, 674055. https://doi.org/10.3389/fnins.2021.674055.
14. Bellantuono, L., Monaco, A., Amoroso, N., Lacalamita, A., Pantaleo, E., Tangaro, S., and Bellotti, R. (2022). Worldwide impact of lifestyle predictors of dementia prevalence: An explainable artificial intelligence analysis. Frontiers in big Data *5*, 1027783. https://doi.org/10.3389/fdata.2022.1027783.
15. Chen, T., and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794. https://doi.org/10.1145/2939672.2939785.
16. Yagin, F.H., Cicek, İ.B., Alkhateeb, A., Yagin, B., Colak, C., Azzeh, M., and Akbulut, S. (2023). Explainable artificial intelligence model for identifying covid-19 gene biomarkers. Comput. Biol. Med. *154*, 106619. https://doi.org/10.1016/j.compbiomed.2023.106619.
17. Rynazal, R., Fujisawa, K., Shiroma, H., Salim, F., Mizutani, S., Shiba, S., Yachida, S., and Yamada, T. (2023). Leveraging explainable ai for gut microbiome-based colorectal cancer classification. Genome Biol. *24*, 21. https://doi.org/10.1186/s13059-023-02858-4.
18. Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-sne. J. Mach. Learn. Res. 9, 2579–2605.
19. (2012). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207–214. https://doi.org/10.1038/nature11234.
20. Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., Arumugam, M., Kultima, J.R., Prifti, E., Nielsen, T., et al. (2014). An integrated catalog of reference genes in the human gut microbiome. Nat. Biotechnol. *32*, 834–841. https://doi.org/10.1038/nbt.2942.

21. Manor, O., Dai, C.L., Kornilov, S.A., Smith, B., Price, N.D., Lovejoy, J.C., Gibbons, S.M., and Magis, A.T. (2020). Health and disease markers correlate with gut microbiome composition across thousands of people. Nat. Commun. *11*, 5206. https://doi.org/10.1038/s41467-020-18871-1.
22. Zafar, H., and Saier, M.H., Jr. (2021). Gut bacteroides species in health and disease. Gut Microb. *13*, 1–20. https://doi.org/10.1080/19490976.2020.1848158.
23. Sun, Y., Zhang, S., Nie, Q., He, H., Tan, H., Geng, F., Ji, H., Hu, J., and Nie, S. (2023). Gut firmicutes: Relationship with dietary fiber and role in host homeostasis. Crit. Rev. Food Sci. Nutr. *63*, 12073–12088. https://doi.org/10.1080/10408398.2022.2098249.
24. Vangay, P., Ward, T., Gerber, J.S., and Knights, D. (2015). Antibiotics, pediatric dysbiosis, and disease. Cell Host Microbe *17*, 553–564. https://doi.org/10.1016/j.chom.2015.04.006.
25. Levy, M., Kolodziejczyk, A.A., Thaiss, C.A., and Elinav, E. (2017). Dysbiosis and the immune system. Nat. Rev. Immunol. *17*, 219–232. https://doi.org/10.1038/nri.2017.7.
26. Ho, L.K.H., Tong, V.J.W., Syn, N., Nagarajan, N., Tham, E.H., Tay, S.K., Shorey, S., Tambyah, P.A., and Law, E.C.N. (2020). Gut microbiota changes in children with autism spectrum disorder: a systematic review. Gut Pathog. *12*, 6–18. https://doi.org/10.1186/s13099-020-0346-1.
27. Behrouzi, A., Nafari, A.H., and Siadat, S.D. (2019). The significance of microbiome in personalized medicine. Clin. Transl. Med. *8*, 1–9. https://doi.org/10.1186/s40169-019-0232-y.
28. Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D.R., Fernandes, G.R., Tap, J., Bruls, T., Batto, J.-M., et al. (2011). Enterotypes of the human gut microbiome. nature *473*, 174–180. https://doi.org/10.1038/nature09944.
29. Nakayama, J., Yamamoto, A., Palermo-Conde, L.A., Higashi, K., Sonomoto, K., Tan, J., and Lee, Y.-K. (2017). Impact of westernized diet on gut microbiota in children on leyte island. Front. Microbiol. *8*, 197. https://doi.org/10.3389/fmicb.2017.00197.
30. Ley, R.E. (2016). Prevotella in the gut: choose carefully. Nat. Rev. Gastroenterol. Hepatol. *13*, 69–70. https://doi.org/10.1038/nrgastro.2016.4.
31. Dillon, S.M., Lee, E.J., Kotter, C.V., Austin, G.L., Gianella, S., Siewe, B., Smith, D.M., Landay, A.L., McManus, M.C., Robertson, C.E., et al. (2016). Gut dendritic cell activation links an altered colonic microbiome to mucosal and systemic t-cell activation in untreated hiv-1 infection. Mucosal Immunol. *9*, 24–37. https://doi.org/10.1038/mi.2015.33.
32. Lozupone, C.A., Rhodes, M.E., Neff, C.P., Fontenot, A.P., Campbell, T.B., and Palmer, B.E. (2014). Hiv-induced alteration in gut microbiota: driving factors, consequences, and effects of antiretroviral therapy. Gut Microb. *5*, 562–570. https://doi.org/10.4161/gmic.32132.
33. Scher, J.U., Sczesnak, A., Longman, R.S., Segata, N., Ubeda, C., Bielski, C., Rostron, T., Cerundolo, V., Pamer, E.G., Abramson, S.B., et al. (2013). Expansion of intestinal prevotella copri correlates with enhanced susceptibility to arthritis. Elife *2*, e01202. https://doi.org/10.7554/eLife.01202.
34. Sakamoto, M., Takagaki, A., Matsumoto, K., Kato, Y., Goto, K., and Benno, Y. (2009). Butyricimonas synergistica gen. nov., sp. nov. and butyricimonas virosa sp. nov., butyric acid-producing bacteria in the family 'porphyromonadaceae'isolated from rat faeces. Int. J. Syst. Evol. Microbiol. *59*, 1748–1753. https://doi.org/10.1099/ijs.0.007674-0.
35. Tap, J., Mondot, S., Levenez, F., Pelletier, E., Caron, C., Furet, J.-P., Ugarte, E., Muñoz-Tamayo, R., Paslier, D.L.E., Nalin, R., et al. (2009). Towards the human intestinal microbiota phylogenetic core. Environ. Microbiol. *11*, 2574–2584. https://doi.org/10.1111/j.1462-2920.2009.01982.x.
36. Biddle, A., Stewart, L., Blanchard, J., and Leschine, S. (2013). Untangling the genetic basis of fibrolytic specialization by lachnospiraceae and ruminococcaceae in diverse gut communities. Diversity *5*, 627–640. https://doi.org/10.3390/d5030627.
37. Vacca, M., Celano, G., Calabrese, F.M., Portincasa, P., Gobbetti, M., and De Angelis, M. (2020). The controversial role of human gut lachnospiraceae. Microorganisms *8*, 573. https://doi.org/10.3390/microorganisms8040573.
38. De Angelis, M., Francavilla, R., Piccolo, M., De Giacomo, A., and Gobbetti, M. (2015). Autism spectrum disorders and intestinal microbiota. Gut Microb. *6*, 207–213. https://doi.org/10.1080/19490976.2015.1035855.
39. Shetty, S.A., Marathe, N.P., Lanjekar, V., Ranade, D., and Shouche, Y.S. (2013). Comparative genome analysis of megasphaera sp. reveals niche specialization and its potential role in the human gut. PLoS One *8*, e79353. https://doi.org/10.1371/journal.pone.0079353.
40. Alcazar, M., Escribano, J., Ferré, N., Closa-Monasterolo, R., Selma-Royo, M., Feliu, A., Castillejo, G., Luque, V.; Obemat20 Study Group, and Muñoz-Hernando, J., et al. (2022). Gut microbiota is associated with metabolic health in children with obesity. Clin. Nutr. *41*, 1680–1688. https://doi.org/10.1016/j.clnu.2022.06.007.
41. Romaní-Pérez, M., López-Almela, I., Bullich-Vilarrubias, C., Rueda-Ruzafa, L., Gomez Del Pulgar, E.M., Benítez-Páez, A., Liebisch, G., Lamas, J.A., and Sanz, Y. (2021). Holdemanella biformis improves glucose tolerance and regulates glp-1 signaling in obese mice. Faseb. J. *35*, e21734. https://doi.org/10.1096/fj.202100126R.
42. Dong, Z., Shen, X., Hao, Y., Li, J., Xu, H., Yin, L., and Kuang, W. (2022). Gut microbiome: A potential indicator for predicting treatment outcomes in major depressive disorder. Front. Neurosci. *16*, 813075. https://doi.org/10.3389/fnins.2022.813075.
43. Yamada, E., Martin, C.G., Moreno-Huizar, N., Fouquier, J., Neff, C.P., Coleman, S.L., Schneider, J.M., Huber, J., Nusbacher, N.M., McCarter, M., et al. (2021). Intestinal microbial communities and holdemanella isolated from hiv+/- men who have sex with men increase frequencies of lamina propria ccr5+ cd4+ t cells. Gut Microb. *13*, 1997292. https://doi.org/10.1080/19490976.2021.1997292.
44. Wang, L., Christophersen, C.T., Sorich, M.J., Gerber, J.P., Angley, M.T., and Conlon, M.A. (2013). Increased abundance of sutterella spp. and ruminococcus torques in feces of children with autism spectrum disorder. Mol. Autism. *4*, 42–44. https://doi.org/10.1186/2040-2392-4-42.
45. Ringnér, M. (2008). What is principal component analysis? Nat. Biotechnol. *26*, 303–304. https://doi.org/10.1038/nbt0308-303.
46. Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Ann. Stat. *28*, 337–407.
47. Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. Ann. Stat. *29*, 1189–1232.
48. Lim, S., and Chi, S. (2019). Xgboost application on bridge management systems for proactive damage estimation. Adv. Eng. Inf. *41*, 100922. https://doi.org/10.1016/j.aei.2019.100922.
49. Schaffer, C. (1993). Selecting a classification method by cross-validation. Mach. Learn. *13*, 135–143.
50. Bergstra, J., and Bengio, Y. (2012). Random search for hyper-parameter optimization. J. Mach. Learn. Res. *13*, 281–305.
51. Varma, S., and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. BMC Bioinf. *7*, 1–8. https://doi.org/10.1186/1471-2105-7-91.
52. Saito, T., and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. PLoS One *10*, e0118432. https://doi.org/10.1371/journal.pone.0118432.
53. Ozenne, B., Subtil, F., and Maucort-Boulch, D. (2015). The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. J. Clin. Epidemiol. *68*, 855–859. https://doi.org/10.1016/j.jclinepi.2015.02.010.
54. Lundberg, S.M., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. *30*, 4768–4777. https://doi.org/10.48550/arXiv.1705.07874.
55. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. Nat. Mach. Intell. *2*, 56–67. https://doi.org/10.1038/s42256-019-0138-9.
56. Mrukwa, G., and Polanska, J. (2022). Divik: divisive intelligent k-means for hands-free unsupervised clustering in big biological data. BMC Bioinf. *23*, 538. https://doi.org/10.1186/s12859-022-05093-z.
57. Kodinariya, T.M., and Makwana, P.R. (2013). Review on determining number of cluster in k-means clustering. Int. J. *1*, 90–95.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
The OTU abundance data have been downloaded from Gene Expression Omnibus (GEO) - NCBI (GEO Database: accession number GSE113690)	https://doi.org/10.1080/19490976.2020.1747329	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE113690
All analyzed data and code	This study	https://doi.org/10.5281/zenodo.12826282
Software and algorithms		
Python software version 3.11.5	Python Software Foundation	https://www.python.org/
Python package: numpy version 1.24.4	Python Software Foundation	https://numpy.org/
Python package: pandas version 1.5.3	Python Software Foundation	https://pandas.pydata.org/
Python package: matplotlib version 3.7.2	Python Software Foundation	https://matplotlib.org/
Python package: shap version 0.43.0	Python Software Foundation	https://shap.readthedocs.io/en/latest/
Python package: sklearn version 1.2.2	Python Software Foundation	https://scikit-learn.org/stable/
Python package: xgboost version 2.0.2	Python Software Foundation	https://xgboost.readthedocs.io/en/stable/
Python package: seaborn version 0.12.2	Python Software Foundation	https://seaborn.pydata.org/
Python package: scipy version 1.10.1	Python Software Foundation	https://scipy.org/
Python package: statannotations version 0.2.3	Python Software Foundation	https://statannotations.readthedocs.io/en/latest/statannotations.html

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Sabina Tangaro (sabina.tangaro@uniba.it).

Materials availability

This study did not generate new unique materials.

Data and code availability

- The OTU abundance data have been downloaded from Gene Expression Omnibus (GEO) - NCBI (accession number GSE113690). The DOI is listed in the key resources table.
- All original code has been deposited at Zenodo and is publicly available as of the date of publication. The DOI is listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

The analyzed dataset is composed of 254 children aged 1–13 years: 111 affected by ASD and 143 TD.[8] The average age of children with Autism Spectrum Disorder (ASD) is 5.09 years, while the average age of children with Typical Development (TD) is 4.94 years. Regarding gender, there are 99 males and 12 females in the ASD group, and 130 males and 13 females in the TD group. The main informations about it are summarised in Table 3.

Obtained from the 16S rRNA gene sequence, the relative abundance of 1322 gut microbiota operational taxonomic units (OTUs) were available for each subject.

METHOD DETAILS

Metataxonomic (16S rRNA gene) analysis

As reported by authors,[8] the 16S rRNA gene V4 region-specific primer are 515F GTGCCAGCMGCCGCGGTAA and 806A GGACTACHVGG GTWTCTAAT and Illumina TruSeq DNA PCR-Free Library Preparation Kit (Illumina, USA) was used to generate sequencing libraries. Raw fastq files were merged and Operational taxonomic units (OTUs) were clustered using Uparse 7.1 (http://drive5.com/uparse) according to a sequences similarity $\geq$ 97%, then annotated using the RDP classifier algorithm (http://sourceforge.net/projects/rdp-classifier) according to the GreenGene version 13.5 database.

Machine learning based classification

The framework implemented is composed of three parts: (i) the classification using XGBoost classifier; (ii) the identification of the most important features in the classification of each single patient by using the SHAP algorithm; (iii) the clustering of ASD subject by means of Kmeans algorithm applied to first two principal components of SHAP values. In Figure 10 the schematic workflow of analysis has been shown.

EXtreme Gradient Boosting (XGBoost) is a supervised learning algorithm based on the ensemble concept: it exploits an aggregation of weak knowledge models to obtain a more accurate prediction. Each new tree is trained on the errors of the previous one in a process called boosting.

This type of ML relies on simple basic classifiers, such as decision trees, also called weak models. The peculiarity of boosting is to focus on training examples that are more difficult to classify. From the training dataset, a subset of examples is extracted without replacement to train a weak classifier. A weak second classifier is trained with another subset of examples randomly drawn without remittance from the training dataset, adding 50% of the previous misclassified examples. A weak third classifier is trained on all examples on which the first two classifiers disagree. The final classification is achieved by combining the three weak learning models via majority voting.[46]

XGBoost classifier is a learning algorithm based on gradient boosting. It generates a strong classifier by iteratively updating parameters of the former classifier to decrease the gradient of loss function.[15,47,48]

Evaluation metrics

The performance of machine learning algorithms depends on the samples used in the training phase. To resolve this dependence and obtain statistically robust results, a 10-fold cross-validation was applied to partition the available dataset. Each of the partitions was used as a test set, while the remaining nine were used as a training set.[49] In order to optimize the machine learning model, hyperparameter tuning was conducted using random search with the *RandomizedSearchCV* function from scikit-learn.[50] The following parameters were varied: "max depth" $= [3, 5, None]$, "colsample bytree" $= [0.3, 0.5, 0.7, 0.9]$, and "n estimators" ranged from 50 to 300 with steps of 50. To find the best model, this hyperparameter optimization was performed using nested cross-validation to eliminate bias in the estimation of test error.[51] Therefore, within each training phase, an inner 3-fold cross-validation was conducted. The whole sequence was repeated 50 times, by dividing the dataset with different partitions between the various iterations.

The metrics used to evaluate the performance of models were.

- Accuracy:

$$ACC = \frac{TP+TN}{TP+FP+TN+FN} \qquad \text{(Equation 1)}$$

- Sensitivity:

$$SENS = \frac{TP}{TP+FN} \qquad \text{(Equation 2)}$$

- Specificity:

$$SPEC = \frac{TP+TN}{TP+FP+TN+FN} \qquad \text{(Equation 3)}$$

- Precision:

$$PREC = \frac{TP}{TP+FP} \quad \text{(Equation 4)}$$

- AU-ROC: is the area under the Receiver Operating Characteristic curve, which plots the true positive rate against the false positive rate.
- AU-PRC: is the area under the Precision-Recall curve, which plots precision against recall (sensitivity).

The best model was chosen as the one scoring higher in more metrics, with particular attention to the highest AU-PRC, which is more informative than AU-ROC in evaluating imbalanced data classification problems.[52,53]

SHapley Additive exPlanations (SHAP) analysis

Explainable Artificial Intelligence (XAI) refers to the development of AI systems in such a way that the outputs can be understood and explained by humans. AI models are often complex and difficult to interpret, therefore XAI focuses on creating algorithms that can articulate the decision-making process, reasons and logic behind the results, offering clear and easy-to-understand explanations to users.

We used SHapley Additive exPlanations method, a XAI algorithm borrowed from game theory.[54,55] The SHAP value of a feature is evaluated by measuring whether its inclusion or exclusion from the model affects the algorithm's performance on the validation set:

$$\Phi_j(x) = \sum_{F \subseteq S - \{j\}} \frac{|F|!(|S| - |F| - 1)!}{|S|!} [f_x(F \cup j) - f_x(F)] \quad \text{(Equation 5)}$$

where x is an instance, the sum is over all the subsets S of features which include the feature j, $\frac{|F|!(|S| - |F| - 1)!}{|S|!}$ is a weight parameter that multiplies all of the permutations of S! by the potential permutations of the remaining class that doesn't belong to S, while $f_x(F \cup j)$ and $f_x(F)$ denote respectively the classification score obtained by including and non-including feature j.

Clustering SHAP values based

Principal Component Analysis (PCA) is a commonly used technique for dimensionality reduction in the context of data visualization and exploration. PCA was applied on SHAP values obtained from the classifier. The first two principal components were used to cluster ASD patients by using k-means algorithm.[56] K-means is an unsupervised learning method to group data points into clusters based on their similarity. The number of clusters has been identified through the Elbow method and the Silhoette score.[57] The Elbow method helps to determine the appropriate value of clusters by plotting the variance explained (inertia) as a function of the number of clusters and looking for the "elbow" point in the plot. The point where the variance explained begins to level off or form an "elbow" is considered a good estimate for the optimal number of clusters. The optimization method for determining the number of clusters based on the Silhouette score relies on finding the partition with the highest Silhouette Coefficient. The Silhouette Coefficient is computed by considering the mean intra-cluster distance and the mean nearest-cluster distance for each sample.

QUANTIFICATION AND STATISTICAL ANALYSIS

To ensure the robustness and reliability of our comparative analysis of machine learning models, we conducted a statistical significance test on the performance metrics. The Mann-Whitney-Wilcoxon test was employed in a two-sided manner. We set the threshold for statistical significance at $P < 0.05$.