Gut Microbial Signatures for Early Screening of Autism Spectrum Disorder: An Interpretable Machine Learning Approach

1st Glykeria Theodorou
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
gtheodorou@biosim.ntua.gr

2nd Aris Markogiannakis
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
el20085@mail.ntua.gr

3d Maria Athanasiou
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
mathanasiou@biosim.ntua.gr

4th Konstantinos Mitsis
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
kmhtshs@biosim.ntua.gr

5th Konstantina Nikita
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
knikita@ece.ntua.gr

***Abstract*—Autism Spectrum Disorder (ASD) is a complex neurodevelopmental disorder characterized by significant phenotypic heterogeneity. Emerging evidence is associating ASD with disruptions in the gut microbiome pointing towards the gut-brain axis as a major contributor to ASD pathophysiology and offering microbial biomarkers with potential as early predictors. In this study supervised machine learning (ML) models trained and evaluated in a nested cross-validation (NCV) framework are leveraged to classify ASD based on gut microbial profiles from nine different cohorts (n=929). To address the challenges posed by high dimensionality and biological variability, compositional data augmentation techniques — Aitchison Mixup, Compositional CutMix, and Feature Dropout — were integrated into the modeling pipeline. Among the evaluated ML models, XGBoost with Feature Dropout achieved the best performance (Accuracy: 73.3%, AUC: 82.3%, F1-score: 73.3%). To interpret the ML model's predictions, SHAP values and information gain were employed, highlighting key microbial species such as *Enterobacter kobei*, *Dialister hominis*, *Leyella stercorea*, and *Comamonas kerstersii*. These results reinforce the potential of the gut microbiome as a promising source of ASD screening biomarkers and highlight the utility of interpretable ML models to extract biologically meaningful patterns in complex microbiome datasets.**

Keywords—ASD, Gut Microbiome, gut-brain axis, ML models, data augmentation

I. Introduction

Autism Spectrum Disorder (ASD) is a complex neuro-developmental disorder that affects approximately 1-2% of the global population and exhibits considerable phenotypic variability. Core features of ASD include persistent deficits in social communication and interaction, alongside restricted and repetitive behavioral patterns [1]. Although early signs of ASD typically emerge within the first three years of life [2], diagnosis is often delayed due to its high clinical heterogeneity and the disparities in access to healthcare. Another factor contributing to this is the sex disparity, with males diagnosed approximately four times more frequently than females [3]. Evidence suggests that this observed difference may not be realistic, as many females on the spectrum tend to exhibit more subtle symptom profiles and compensatory social behaviors that mask core features of ASD, resulting in underdiagnosis.

As research advances towards more timely and reliable diagnostic methods, genetic and neurobiological mechanisms are being extensively studied [4][5]. As a result, attention is shifting towards more complex and integrative underlying biological mechanisms. Among these, the gut microbiome stands out as a key contributor to ASD pathophysiology, affecting its onset and progression through the gut-brain axis [6][7]. Taking into consideration that gastrointestinal (GI) tract harbours approximately 10^{13} microbiomes [8], datasets derived from the gut microbiome studies typically present extremely high dimensionality. Machine Learning (ML) models, known for their capability of integrating and processing multimodal data, are emerging as powerful tools in unraveling the latent patterns linked to ASD phenotypes [9].

However, despite their strength, ML models face significant challenges, mainly due to the limited datasets, the high degree of variability, noise, and the frequent presence of outliers within microbiome datasets. All these inconsistencies stem from biological heterogeneity, technical differences during sample collection and preprocessing and sporadic measurement errors [10]. These factors together can distort the true biological signal and seriously constrain the performance of ML models, further complicating the identification of meaningful microbial patterns associated with ASD [11].

To address these challenges, data augmentation techniques are employed with growing frequency to generate synthetic data in order to expand microbiome datasets. In this way, the diversity of the training dataset is enhanced, reducing the risk of overfitting and improving the stability of the ML model. Equally important in this context is the interpretability of the ML model's predictions, which provides biological insight. Gut microbiome signatures uncovered through interpretable ML models, could serve as biomarkers for ASD screening, diagnosis and patient stratification. However, a limited number of interpretable approaches have been proposed so far, deploying interpretability to quantify the contribution of individual microbial features to the ML model's output [10].

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

The present study aims to identify robust gut microbial signatures associated with ASD by employing supervised ML models. To address the inherent noise and the vast biological variability of the data used, data augmentation techniques are incorporated. Additionally, by leveraging both SHAP [12] and Information Gain [13] to assess the contribution of specific microbial features to classification, emphasis is given to the ML model's interpretability. A systematic evaluation of multiple classification algorithms, each integrated with distinct augmentation strategies, is conducted, with the goal of elucidating gut microbial signatures as candidate biomarkers for ASD screening and stratification.

II. Materials and Methods

A. Data Collection and Preprocessing

The data used in this study are obtained from publicly available datasets [14]-[23] provided by *Morton et al*. (2023) [24], which include taxonomic profiles from multiple ASD case–control cohorts spanning diverse geographical and clinical backgrounds. Species-level relative abundance profiles are generated through whole-metagenome shotgun sequencing. Only samples clearly labelled as either ASD (cases) or neurotypical (controls) are retained for downstream analysis.

After filtering and consolidation, the final dataset includes 929 samples (participants) from nine distinct cohorts, (Fig. 1) comprising 489 microbial features. Among them, 472 samples correspond to ASD cases, while 457 are neurotypical controls.

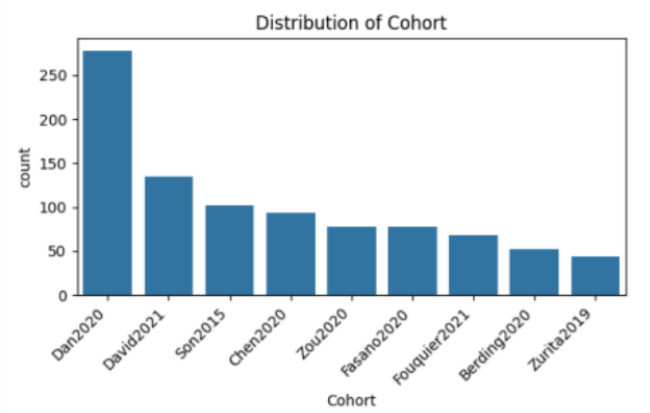

Fig.1. Distribution of data across cohorts

The dataset is predominantly composed of male participants (666 male, 177 female and 86 with unspecified gender). All participants are children, with an average age of 5.64 ± 2.62 years (Fig. 2).

The analysis aims to develop a binary classification model to distinguish children with ASD from neurotypical peers. To mitigate the risk of overfitting in subsequent modeling steps while preserving biologically relevant variation, dimensionality reduction is performed by selecting only the statistically significant microbial features based on the Kruskal–Wallis H test, a non-parametric alternative to ANOVA [25]. The Kruskal–Wallis H test does not assume normality of the underlying data distribution, making it particularly suitable for microbiome data. It ranks all observations and assesses whether the distributions of ranks differ significantly across groups. In this way, microbial features whose abundance profiles differ meaningfully between ASD and neurotypical samples can be identified. Microbial features with p-values below 0.05 are considered statistically significant and comprise the reduced feature space used for downstream ML analysis. This reduction mitigates the risk of overfitting in subsequent modeling steps while preserving biologically relevant variation.

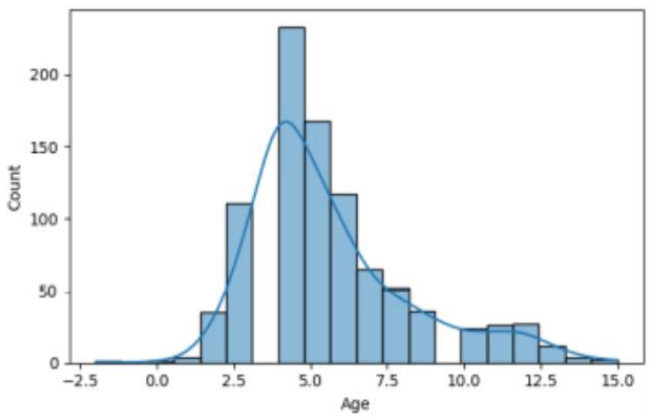

Fig.2. Distribution of data across age

After feature selection, the data are normalized by dividing each feature value by the total abundance within the corresponding sample. This transformation ensures that the data reflect relative abundances, which is appropriate for compositional datasets. It also ensures that all features are on a comparable scale, preventing those with inherently large values from dominating the model training.

The target labels are encoded into binary form using scikit-learn's LabelEncoder. Specifically, ASD cases are assigned the label 1 and neurotypical controls the label 0.

Following feature selection, the initial set of 489 microbial features are reduced to 121. These features are then used as input for all downstream modeling steps. For each ML model, multiple combinations of hyperparameters are evaluated and the three data augmentation strategies are applied.

B. Models

To classify ASD and neurotypical controls based on gut microbial abundance profiles, four supervised ML models are developed and comparatively evaluated: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and XGBoost. These classifiers are selected due to their demonstrated effectiveness in handling high-dimensional data and their capacity to capture complex, non-linear relationships among features.

A grid search approach is adopted for the identification of the optimal set of hyperparameters for each classifier. The hyperparameters' combinations presented in TABLE I are investigated for each model and the configuration with the highest average validation performance is selected.

Model training and evaluation are conducted using a 5x3 nested cross-validation (NCV) framework to ensure unbiased performance estimation. To assess the predictive performance of the tested ML models, several evaluation metrics are utilized, including accuracy, the Area Under the Receiver Operating Characteristic Curve (AUC), and the F_1-score.

TABLE I. ML MODELS' HYPERPARAMETERS

ML Model	*Hyperparameters*	
RF	n_estimators	[100, 200, 300, 500]
	max_depth	[10, 20, 30]
	min_samples_split	[2,5]
XGBoost	n_estimators	[50, 100, 200]
	max_depth	[3, 6, 10]
	learning_rate	[0.05, 0.1, 0.2]
LR	C	[0.1, 1]
	penalty	['l1', 'l2']
	solver	['liblinear']
SVM	C	[0.1, 1]
	kernel	['rbf']
	gamma	['scale', 0.1]

C. Data Augmentation

To further enhance model generalization and address the compositional nature of gut microbiome data, three specialized data augmentation techniques are employed: Aitchison Mixup, Compositional Feature Dropout and Compositional CutMix [26]. These techniques are specifically designed for compositional data and are applied only to training sets to ensure unbiased evaluation.

1) Aitchison Mixup

Aitchison Mixup generates new training samples by blending existing ones in a transformed space that respects the compositional nature of microbiome data. Each new datapoint is sampled as follows:

- Draw a class c from the class prior and draw $\lambda \sim U(0,1)$.
- Draw two training points i_1, i_2 such that:

$$y_{i_1} = y_{i_2} = c \tag{1}$$

  uniformly at random.
- Set:

$$x^{aug} = (\lambda \odot x_{i_1}) \oplus ((1-\lambda) \odot x_{i_2}) \tag{2}$$

$$y^{aug} = c \tag{3}$$

2) Compositional Feature Dropout

Compositional Feature Dropout generates augmented data by randomly removing parts of a sample's microbial profile, i.e. selecting and zeroing out a subset of taxa. This approach functions as a form of feature-level dropout. The modified sample is then re-normalized to remain within the compositional space. Each new datapoint is generated as follows:

- Draw $\lambda \sim U(0,1)$. Draw training point i uniformly at random and set $\tilde{x} = x_i$.
- For each $j \in \{1, \dots, D\}$, draw $I_j \sim Bernoulli(\lambda)$ and set $\tilde{x}_j = 0$ if $I_j = 0$.
- Set:

$$x^{aug} = \frac{\tilde{x}}{\sum_{i=1}^{D} \tilde{x}_i} \tag{4}$$

$$y^{aug} = y_i \tag{5}$$

3) Compositional CutMix

Compositional CutMix is a hybrid augmentation strategy that combines ideas from both Aitchison Mixup and Compositional Feature Dropout. Similar to Aitchison Mixup, it generates new training samples by combining pairs of datapoints from the same class. However, instead of blending the samples linearly, it takes complementary subsets from each sample and then renormalizes the resulting composition to ensure it remains valid. Each new datapoint is generated as follows:

- Draw a class c from the class prior and draw $\lambda \sim U(0,1)$.
- Draw two training points i_1, i_2 such that $y_{i_1} = y_{i_2} = c$, uniformly at random.
- For each $j \in \{1, \dots, D\}$, draw $I_j \sim Bernoulli(\lambda)$ and set:

$$\tilde{x_j} = x_{i1j}, \text{ if } I_j = 0 \tag{6}$$

$$\tilde{x_j} = x_{i2j}, \text{ if } I_j = 1 \tag{7}$$

- Set:

$$x^{aug} = \frac{\tilde{x}}{\sum_{i=1}^{D} \tilde{x}_i} \tag{8}$$

$$y^{aug} = c \tag{9}$$

The impact of each method on classification performance is quantitatively assessed using the same evaluation metrics to ensure fair comparison. This setup enables a systematic analysis of how compositional data augmentation influence ML model generalization, particularly under challenging conditions such as noise and sparsity.

D. Interpretability

To interpret the generated predictions and identify the most influential microbial features, the SHAP (SHapley Additive exPlanations) framework, a widely used method for interpreting ML models' predictions, is utilized. SHAP works by assigning each feature a contribution score that reflects how much it influenced a particular prediction. These scores are based on Shapley values, a concept from cooperative game theory originally designed to fairly divide rewards among players depending on their individual contributions. In the context of ML, SHAP uses the same idea to fairly attribute the ML model's output to the input features, helping us understand why a model made a certain decision.

Additionally, for tree-based models such as RF and XGBoost, global feature importance is assessed using model-specific internal metrics. In RF, feature importances derive from the "feature importances attribute", which reflects the mean decrease in impurity (e.g., Gini index) across all trees, averaged and normalized per feature. XGBoost computes feature importance based on the average gain, defined as the mean improvement in the loss function brought by all splits involving a given feature, aggregated across the entire ensemble. These metrics provide complementary insights into which microbial features most strongly influenced the ML model's predictions. An overview of the proposed pipeline is presented in Fig. 3.

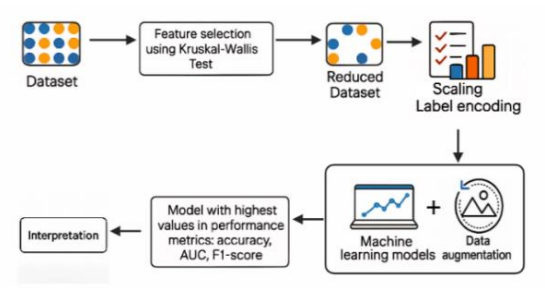

Fig. 3. Overview of the ML pipeline used in this study.

III. RESULTS

A. Model Performance

A comparative summary of the results obtained across all ML models is provided in TABLE II.

Based on these results, XGBoost with Compositional Feature Dropout applied to the training set emerges as the best-performing ML model among those evaluated, with an Accuracy of 73.3%, an AUC of 82.3% and a F_1-score of 73.3%.

In Fig. 4, a confusion matrix is also generated for the XGBoost to provide deeper insight into its classification performance. Out of 472 ASD cases, 347 are correctly identified as ASD (TP), while 125 are misclassified as neurotypical controls (FN). Conversely, among the 457 neurotypical controls, 334 are correctly classified (TN), and 123 are incorrectly predicted as ASD (FP).

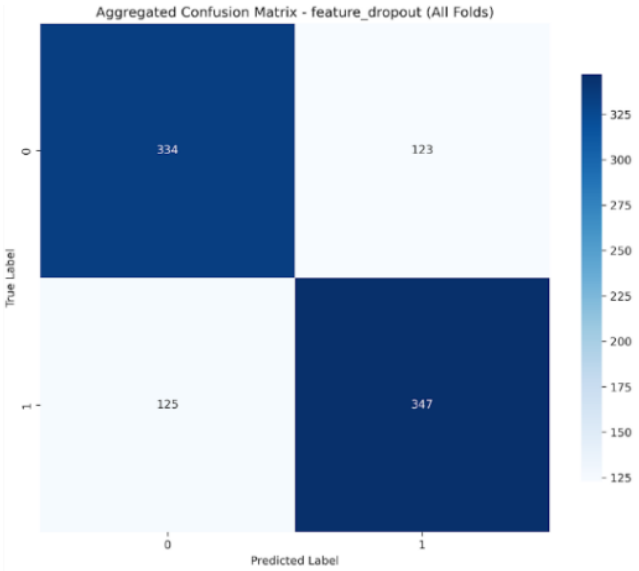

Fig. 4 Aggregated Confusion Matrix of XGBoost

These results correspond to a sensitivity of 73.5%, reflecting the model's ability to correctly detect ASD cases, and a specificity of 73.1%, indicating its effectiveness in identifying neurotypical individuals. The near-equal values of these two metrics demonstrate that the model maintains balanced performance across both classes, which is a critical requirement in clinical or screening scenarios where both FP and FN carry significant implications.

TABLE II. PERFORMANCE METRICS OF THE EVALUATED ML MODELS (VALUES REPRESENT THE MEAN ± STD COMPUTED ACROSS THE FIVE OUTER FOLDS OF THE NCV FRAMEWORK.)

Model	Augmentation Technique	Accuracy	AUC	F1-score
LR	-	0.617 (±0.052)	0.653 (±0.056)	0.616 (±0.052)
	Aitchishon Mixup	0.625 (±0.04)	0.67 (±0.046)	0.624 (±0.04)
	Compositional Feature Dropout	0.619 (±0.044)	0.663 (±0.048)	0.618 (±0.043)
	Compositional Cutmix	**0.635 (±0.043)**	**0.664 (±0.049)**	**0.634 (±0.043)**
SVM	-	0.635 (±0.033)	0.668 (±0.035)	0.631 (±0.035)
	Aitchishon Mixup	0.63 (±0.029)	0.689 (±0.04)	0.625(±0.029)
	Compositional Feature Dropout	**0.637 (±0.016)**	**0.677 (±0.03)**	**0.634 (±0.017)**
	Compositional Cutmix	0.63 (±0.019)	0.676 (±0.024)	0.628 (±0.02)
RF	-	0.717 (±0.014)	0.811 (±0.016)	0.717 (±0.014)
	Aitchishon Mixup	**0.729 (±0.021)**	**0.819 (±0.023)**	**0.729 (±0.021)**
	Compositional Feature Dropout	0.720 (±0.02)	0.81 (±0.037)	0.719 (±0.02)
	Compositional Cutmix	0.721 (±0.025)	0.812 (±0.025)	0.721 (±0.025)
XGBoost	-	0.708 (±0.011)	0.806 (±0.016)	0.708 (±0.012)
	Aitchishon Mixup	0.723 (±0.016)	0.822 (±0.02)	0.723 (±0.016)
	Compositional Feature Dropout	**0.733 (±0.026)**	**0.823 (±0.018)**	**0.733 (±0.026)**
	Compositional Cutmix	0.73 (±0.029)	0.826 (±0.026)	0.73 (±0.029)

B. *Interpretation on ML model's Decisions*

To gain further insight into the XGBoost model's decision-making process, we leverage SHAP values and information gain to extract significant features and enhance interpretability.

Fig. 5 depicts the obtained SHAP summary plot, where SHAP values quantify the influence of each microbial feature on the model's ASD prediction, with color gradients representing relative abundance (red for high, blue for low), and the violin shapes showing the distribution of impact across all test samples.

Fig. 6 presents the obtained feature ranking based on XGBoost's information gain.

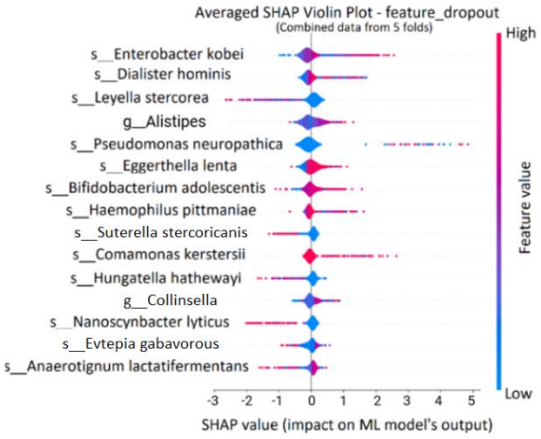

Fig.5. Aggregated SHAP violin plots for the top predictive microbial features.

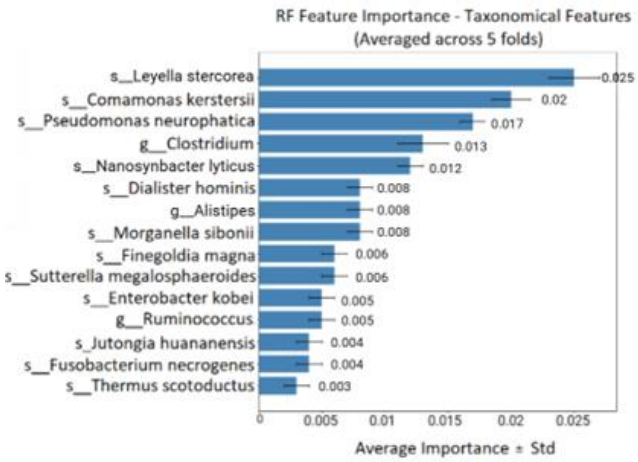

Fig.6. Feature importance based on information gain

These results reveal a clear consensus between the two methods regarding several microbial features that stand out as key predictors. Specifically, the microbial species *Enterobacter kobei*, *Dialister hominis*, *Leyella stercorea*, *Pseudomonas neuropathica*, *Nanosynbacter lyticus* and *Comamonas kerstersii* and some not yet identified species of the genus *Alistipes,* consistently demonstrate strong and stable influences in the SHAP value distributions across the dataset, while also exhibit high importance scores according to the information gain metric. This agreement underscores the robustness of these features in driving the ML model's predictions.

Further analysis of the SHAP summary plot indicates that taxa such as *Enterobacter kobei, Eggerthella lenta, Haemophilus pittmaniae, Bifidobacterium adolescentis* and *Comamonas kerstersii* display high SHAP values in samples with elevated abundance, indicating a consistent and positive contribution to ASD classification. In contrast, species as *Nanosynbacter lyticus* and *Anaerotignum lactatifermentans* show the opposite trend: their higher abundance is associated with negative SHAP values, suggesting they are more prevalent in neurotypical profiles. Other species, including *Leyella stercorea*, *Sutterella stercoricanis, Hungatella hathewayi, Evtepia gabavorous* and *Anaerotignum lactatifermentans* and some not yet identified species of the genus *Alistipes* exhibit broader and asymmetric SHAP distributions, indicating that their influence varies depending on the sample and could follow non-linear or context-sensitive patterns. These observations reveal that while some microbes serve as stable predictive markers, others have nuanced, individual-specific effects, thus highlighting the heterogeneity of the ASD microbiome and supporting the need for personalized interpretation strategies that account for both shared and patient-specific microbial signals.

In addition to the aforementioned shared microbial features, some distinct sets of microbial taxa are uniquely identified through the SHAP value analysis and the XGBoost information gain, respectively. Notably, the SHAP analysis highlights unique contributions from the species *Eggerthella lenta, Bifidobacterium adolescentis, Haemophilus pittmaniae*, *Sutterella stercoricanis*, *Hungatella hathewayi, Evtepia gabavorous, Anaerotignum lactatifermentans* and unidentified species from the genus *Collinsella.* Conversely, information gain identifies unique taxa including *Morganella sibonii, Finegoldia magna, Sutterella megalosphaeroides, Jutongia huaianensis, Fusobacterium necrogenes, Thermus scotoductus* as well as unidentified species from the genera *Clostridium* and *Ruminococcus.*

These discrepancies highlight the complementary advantages of the two methods: information gain quantifies the global importance of features based on their overall contribution to model splits, whereas SHAP values capture local, instance-specific effects and interactions that may not be reflected in global metrics.

IV. DISCUSSION

In this study, an interpretable ML pipeline is implemented to explore gut microbiome-derived signatures associated with ASD. Among the evaluated ML models, XGBoost with Compositional Feature Dropout achieves the highest performance and demonstrates its ability to detect ASD cases while correctly rejecting neurotypical controls.

To better understand how the ML model generates its predictions, two complementary interpretability methods are applied: SHAP values and information gain. Interestingly, both approaches consistently identified the following microbial species: the microbial species *Enterobacter kobei, Dialister hominis, Leyella stercorea, Pseudomonas neuropathica, Nanosynbacter lyticus, Comamonas kerstersii* and some not yet identified species of the genus *Alistipes* as top contributors. The convergence of results from both approaches highlights these species as promising pre-screening biomarkers, offering an interpretable and data-driven foundation for early ASD risk stratification. Such biomarkers could potentially be used ahead of traditional clinical diagnostics to guide more targeted neurodevelopmental assessments.

In addition to these shared features, the two methods reveal distinct yet informative perspectives. SHAP values offer fine-grained, instance-level insights, identifying taxa whose influence varied across individual samples. For example, *Hungatella hathewayi*, and species within the *Alistipes* genus exhibit non-linear and sometimes contradictory effects, contributing to ASD predictions in some cases, but not in others. Such patterns suggest more complex relationships, likely influenced by interactions with other microbial species or host-specific factors. These nuanced effects would be overlooked by relying solely on global metrics.

On the other hand, information gain captures species that consistently contributed to decision-making across the model as a whole, such as *Jutongia huaianensis* and *Thermus scotoductus*, which do not appear in the SHAP-based ranking. This divergence illustrates how global and local interpretability methods can work together to uncover both broad and subtle microbial influences.

A key aspect of our pipeline is its consideration of the compositional nature of microbiome data. We apply augmentation techniques tailored for this data type (Aitchison Mixup, Compositional CutMix, and Compositional Feature Dropout) which aim to reduce overfitting and enhance model stability. These approaches are especially helpful in managing the challenges posed by the high-dimensional and sparse feature spaces common in microbiome research.

Our findings are consistent with recent studies that have employed ML models on gut microbiome datasets for ASD prediction [27–29]. These studies have also identified ASD-associated microbial signatures, further supporting the potential of gut microbiome profiling as an early, non-invasive screening tool.

A significant challenge in this study is the integration of data from nine heterogeneous cohorts, which, while increasing representativeness, also introduces potential batch effects arising from variations in sample collection, sequencing protocols, and geography-specific biases that may confound biological signals. These findings highlight the need for more refined harmonization strategies, such as domain-adversarial approaches, and emphasize the importance of validating models on fully independent cohorts to ensure clinical translatability.

Future work should focus on validating these findings in independent, multi-cohort datasets to assess generalizability. Integrating host-derived data, including immune, genetic, and multi-omics layers such as metabolomics and transcriptomics may further clarify the microbiome's role in ASD. Ultimately, interpretable and personalized models could support the development of microbiome-informed screening tools or adjunctive strategies in ASD care.

V. Conclusion

In this study, an interpretable ML pipeline is implemented to explore gut microbiome-derived signatures associated with ASD. Among the evaluated ML models, XGBoost with Compositional Feature Dropout achieves the highest performance and demonstrates its ability to detect ASD cases while correctly rejecting neurotypical controls. To better understand how the ML model generates its predictions, two complementary interpretability methods were applied: SHAP values and information gain, which consistently highlighted key microbial species such as *Enterobacter kobei*, *Dialister hominis*, *Leyella stercorea*, and *Comamonas kerstersii*.

References

[1] Ousley, O., Cermak, T. Autism Spectrum Disorder: Defining Dimensions and Subgroups. Curr Dev Disord Rep 1, 20–28 (2014). https://doi.org/10.1007/s40474-013-0003-1

[2] Pelling, N.J., & Burton, L.J. (Eds.). (2017). The Elements of Psychological Case Report Writing in Australia (1st ed.). Routledge. https://doi.org/10.4324/9781351258043

[3] Loomes R, Hull L, Mandy WPL. What Is the Male-to-Female Ratio in Autism Spectrum Disorder? A Systematic Review and Meta-Analysis. J Am Acad Child Adolesc Psychiatry. 2017 Jun;56(6):466-474. doi: 10.1016/j.jaac.2017.03.013. Epub 2017 Apr 5. PMID: 28545751.

[4] A Jeremy Willsey, Matthew W State, Autism spectrum disorders: from genes to neurobiology, Current Opinion in Neurobiology, Volume 30, 2015, Pages 92-99, ISSN 0959-4388, https://doi.org/10.1016/j.conb.2014.10.015

[5] Katherine W Eyring, Daniel H Geschwind, Three decades of ASD genetics: building a foundation for neurobiological understanding and treatment, *Human Molecular Genetics*, Volume 30, Issue 20, 15 October 2021, Pages R236R244, https://doi.org/10.1093/hmg/ddab176

[6] Morton, J.T., Jin, DM., Mills, R.H. et al. *Multi-level analysis of the gut–brain axis shows autism spectrum disorder-associated molecular and microbial profiles*. Nat Neurosci 26, 1208–1217 (2023). https://doi.org/10.1038/s41593-023-01361-0

[7] Wong GC, Montgomery JM, Taylor MW. The Gut-Microbiota-Brain Axis in Autism Spectrum Disorder. In: Grabrucker AM, editor. Autism Spectrum Disorders [Internet]. Brisbane (AU): Exon Publications; 2021 Aug 20. Chapter 8. Available from: https://www.ncbi.nlm.nih.gov/books/NBK573606/

[8] Leviatan, S., Shoer, S., Rothschild, D. *et al.* An expanded reference map of the human gut microbiome reveals hundreds of previously unknown species. *Nat Commun* 13, 3863 (2022). https://doi.org/10.1038/s41467-022-31502-1

[9] Mingbang Wang, Ceymi Doenyas, Jing Wan, Shujuan Zeng, Chunquan Cai, Jiaxiu Zhou, Yanqing Liu, Zhaoqing Yin, Wenhao Zhou, Virulence factor-related gut microbiota genes and immunoglobulin A levels as novel markers for machine learning-based classification of autism spectrum disorder, Computational and Structural Biotechnology Journal, Volume 19, 2021, Pages 545-554, ISSN 2001-0370, https://doi.org/10.1016/j.csbj.2020.12.012

[10] Kumar, B., Lorusso, E., Fosso, B., & Pesole, G. (2024). A comprehensive overview of microbiome data in the light of machine learning applications: categorization, accessibility, and future directions. *Frontiers in microbiology*, *15*, 1343572.

[11] Walsh C, Stallard-Olivera E, Fierer N. Nine (not so simple) steps: a practical guide to using machine learning in microbial ecology. mBio. 2024 Feb 14;15(2):e0205023. doi: 10.1128/mbio.02050-23. Epub 2023 Dec 21. PMID: 38126787; PMCID: PMC10865974.

[12] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

[13] Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106.

[14] Maria Fernanda Zurita, Paul A Cardenas, Marıa Elena Sandoval, Marıa Caridad Pena, Marco Fornasini, Nancy Flores, Marcia H Monaco, Kirsten Berding, Sharon M Donovan, Thomas Kuntz, Jack A Gilbert, and Manuel E Baldeon. Analysis of gut microbiome, nutrition and immune status in autism spectrum disorder: a case-control study in ecuador. Gut Microbes, 11(3):453–464, May 2020

[15] Zhou Dan, Xuhua Mao, Qisha Liu, Mengchen Guo, Yaoyao Zhuang, Zhi Liu, Kun Chen, Junyu Chen, Rui Xu, Junming Tang, Lianhong Qin, Bing Gu, Kangjian Liu, Chuan Su, Faming Zhang, Yankai Xia, Zhibin Hu, and Xingyin Liu. Altered gut microbial profile is associated with abnormal metabolism activity of autism spectrum disorder. Gut Microbes, 11(5):1246–1267, September 2020.

[16] Jennifer Fouquier, Nancy Moreno Huizar, Jody Donnelly, Cody Glickman, Dae-Wook Kang, Juan Maldonado, Rachel A Jones, Kimberly Johnson, James B Adams, Rosa Krajmalnik-Brown, and Catherine Lozupone. The gut microbiome in autism: Study-Site effects and longitudinal analysis of behavior change. mSystems, 6(2), April 2021

[17] Rong Zou, Fenfen Xu, Yuezhu Wang, Mengmeng Duan, Min Guo, Qiang Zhang, Hongyang Zhao, and Huajun Zheng. Changes in the gut microbiota of children with autism spectrum disorder. Autism Res., 13(9):1614–1625, September 2020.

[18] Dan Bai, Benjamin Hon Kei Yip, Gayle C Windham, Andre Sourander, Richard Francis, Rinat Yoffe, Emma Glasson, Behrang Mahjani, Auli Suominen, Helen Leonard, et al. Association of genetic and environmental factors with autism in a 5-country cohort. JAMA psychiatry, 76(10):1035–1043, 2019

[19] Dae-Wook Kang, James B Adams, Ann C Gregory, Thomas Borody, Lauren Chittick, Alessio Fasano, Alexander Khoruts, Elizabeth Geis, Juan Maldonado, Sharon McDonough-Means, Elena L Pollard, Simon Roux, Michael J Sadowsky, Karen Schwarzberg Lipson, Matthew B Sullivan, J Gregory Caporaso, and Rosa Krajmalnik-Brown. Microbiota transfer therapy alters gut ecosystem and improves gastrointestinal and autism symptoms: an open-label study. Microbiome, 5(1):10, January 2017.

[20] Dae-Wook Kang, James B Adams, Devon M Coleman, Elena L Pollard, Juan Maldonado, Sharon McDonough-Means, J Gregory Caporaso, and Rosa Krajmalnik-Brown. Long-term benefit of microbiota transfer therapy on autism symptoms and gut microbiota. Sci. Rep., 9(1):5821, April 2019.

[21] Jiang Zhu, Xueying Hua, Ting Yang, Min Guo, Qiu Li, Lu Xiao, Ling Li, Jie Chen, and Tingyu Li. Alterations in gut vitamin and amino acid metabolism are associated with symptoms and neurodevelopment in children with autism spectrum disorder. J. Autism Dev. Disord., July 2021.

[22] Son JS, Zheng LJ, Rowehl LM, Tian X, Zhang Y, Zhu W, Litcher-Kelly L, Gadow KD, Gathungu G, Robertson CE, Ir D, Frank DN, Li E. Comparison of Fecal Microbiota in Children with Autism Spectrum Disorders and Neurotypical Siblings in the Simons Simplex Collection. PLoS One. 2015

[23] Kirsten Berding and Sharon M Donovan. Dietary patterns impact temporal dynamics of fecal microbiota composition in children with autism spectrum disorder. Front Nutr, 6:193, 2019

[24] Morton, J.T., Jin, DM., Mills, R.H. et al. *Multi-level analysis of the gut–brain axis shows autism spectrum disorder-associated molecular and microbial profiles*. Nat Neurosci 26, 1208–1217 (2023). https://doi.org/10.1038/s41593-023-01361-0

[25] (2008). Kruskal-Wallis Test. In: The Concise Encyclopedia of Statistics. Springer, New York, NY. https://doi.org/10.1007/978-0-387-32833-1_216

[26] E. Gordon-Rodriguez, T. Quinn, and J. P. Cunningham, "Data Augmentation for Compositional Data: Advancing Predictive Models of the Microbiome," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., Curran Associates, Inc., 2022, pp. 20551–20565. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022file/81a28be483155f802ddef448d6fc4b57-Paper-Conference.pdf

[27] Wan Y, Wong OWH, Tun HM, Su Q, Xu Z, Tang W, Ma SL, Chan S, Chan FKL, Ng SC. Fecal microbial marker panel for aiding diagnosis of autism spectrum disorders. Gut Microbes. 2024 Jan-Dec;16(1):2418984. doi: 10.1080/19490976.2024.2418984. Epub 2024 Oct 28. PMID: 39468837; PMCID: PMC11540074.

[28] Temiz, M.; Bakir-Gungor, B.; Ersoz, N.S.; Yousef, M. Machine Learning-Based Prediction of Autism Spectrum Disorder and Discovery of Related Metagenomic Biomarkers with Explainable AI. Appl. Sci. 2025, 15, 9214. https://doi.org/10.3390/app15169214

[29] P. Novielli et al., "Personalized identification of autism-related bacteria in the gut microbiome using explainable artificial intelligence," iScience, vol. 27, no. 9, Sep. 2024, doi: 10.1016/j.isci.2024.110709