Facing the Challenges of Microbiome Data through Supervised Autoencoders for the Non-invasive Disease Diagnosis with Transfer Learning techniques

Manco Francesco

February 21, 2025

1 Introduction

Autism Spectrum Disorder(ASD) is a neuropsychiatric condition complex that affects millions of individuals worldwide. Early diagnosis is crucial for timely and effective interventions, but is often difficult due to the heterogeneous nature of symptoms. In recent years, research has shown a possible correlation between the gut microbiome and ASD, suggesting that microbial profiles may offer valuable diagnostic information. This study explores the use of representation learning to improve the analysis of the microbiome and enable noninvasive diagnosis of Autism Spectrum Disorder(ASD). Specifically, a supervised autoencoder approach is adopted (SAE) to learn meaningful latent representations from microbial data, reducing dimensionality and capturing patterns relevant to classification. The supervised autoencoder is trained jointly with a classifier to optimize the data reconstruction capability and separability between ASD and TD (Typically Developing) classes, balancing reconstruction error with predictive accuracy using a combined loss function.

The representations learned from the autoencoder bottleneck , which constitute a compact but informative version of the original data, are subsequently used as input to a Random Forest (RF) classifier. This approach allows the nonlinear learning capabilities of the SAE to be exploited to extract relevant features , while the Random Forest provides interpretability and robustness in the final classification.

The combineduse of SAE and RF addresses the challenges of high dimensionality and sparsity of microbial data, improving discrimination between ASD and TD samples compared with traditional methods of statistical analysis or unsupervised dimensionality reduction (e.g. PCA). Preliminary results indicate improved performance metrics , such as F1-score ,Recall and Precision, demonstrating the potential of the proposed approach for clinical applications non invasive and identification of relevant biomarkers for early diagnosis of ASD.

2 State of the Art

In recent years, research has shown a significant connection between the gut microbiota and autism spectrum disorders(ASD). Studies have shown that children with ASD have a distinct gut microbial composition compared with neurotypical peers, characterized by reduced biodiversity and specific alterations in the bacterial community. These differences in the microbiota have been associated not only with gastrointestinal symptoms common in individuals with ASD, but also to possible influences on behavior and brain physiology via the gut-brain axis.

In parallel, the application of machine learning techniques in the analysis of biological data has opened new perspectives for the diagnosis of ASD. One promisingapproach combines the use of supervised autoencoders for extracting relevant features from microbiome data, followed by the application of classification algorithms such as random forests (Random Forests). Supervised autoencoders are neural networks designed to learn compact representations of data, preserving information essential for classification. For example, one study used an autoencoder combined with a Multilayer Perceptro to analyze brain images, achieving 70% accuracy in the diagnosis of ASD[1].

The integration of supervised autoencoders and Random Forests offers significant advantages. Autoencoders reduce the dimensionality of the data, eliminating noise and focusing on the most informative features, while Random Forests, known for their robustness and ability to handle high-dimensionality data, use these features to make accurate classifications..

In summary, the analysis of the gut microbiome through advanced machine-learning techniques machine represents a promising frontier for the noninvasive diagnosis of ASD. The combined use of supervised autoencoders and Random Forest enables the extraction and use of effectively relevant information from complex microbiome data, paving the way for earlier and more accurate diagnostic tools.

3 Business Understanding

Autism Spectrum Disorder(ASD) requires complexdiagnoses based on behavioral assessments, often delaying early intervention and limiting the effectiveness of treatment. However, emergingevidence suggests that the gut microbiome may play a significant role in the development and progression of ASD, offering new opportunities for noninvasive diagnostic approaches . This work proposes a novel framework that exploits machine learning techniques to analyze gut microbial profiles, hypothesizing a correlation between specific microbial compositions and the presence of ASD. The clinical goal is to reduce diagnostic time and identify biomarkers potential that can support early diagnosis and targeted interventions.

The developed framework is based on a transfer learning approach to overcome the technological limitations of available data. Specifically, the system exploits two types of metagenomic data: 16S rRNA data, which offer a broad-spectrum overview of the composition bacteria with relatively low cost and greater sample availability, and metagenomic shotgun data, which provide detailed information at the species level but are more expensive and limited in quantity. The transfer of knowledge between these two technologies makes it possible to exploit the rich information of 16S rRNA data to improve the representation of shotgun data, ensuring more accurate and robust analysis for pediatric applications.

To achieve these goals, a supervised autoencoder-based deep learning approach(SAE) was adopted, designed to learn compact and meaningful representations of microbial data. The autoencoder is trained on 16S rRNA data, with a structure that compresses the input information into a bottleneck of small size, optimizing the combination of reconstruction and classification through a combined loss function, balanced by a parameter alpha that adjusts the weight between targets. The resulting model is then transferred to the shotgun data, adapting the network by fine-tuning to ensure a consistent representation of the two types of data, despite the difference in dimensionality between the datasets.

The autoencoder bottleneck is subsequently used as input to a Random Forest model, which exploits the learned features to predict the presence of ASD. The choice of the Random Forest is motivated by its ability to handle high-dimensional data and its interpretability, facilitating the analysis of the most relevant features for diagnosis. In addition, the use of stratified cross-validation with k=5 ensures rigorous model evaluation, reducing the risk of overfitting and providing more reliable estimates of performance.

The experimental results obtained show that the proposed approach achieves high performance in terms of precision,recall and F1-Score. In particular, the integration of the supervised autoencoder with the Random Forest significantly improved the classification metrics compared with models based solely on raw data, suggesting that learning latent representations may improve discrimination between ASD and neurotypical subjects.

4 Data Understanding

The study uses two separate microbiome datasets for the analysis of ASD, both publicly available in the NCBI repository . The cohort consists of children with a clinical diagnosis of ASD and typical developmental (TD) subjects, with ages ranging from 2 to 13 years. Samples were obtained from annual physical examinations and include sequencing data from both 16S rRNA and metagenomic shotgun .

- **16S rRNA**: The dataset includes 255 samples (144 ASD, 111 TD) with 1322 final OTUs characterized by high dimensionality and sparsity. The mean age of the participants is 5.189 ± 0.170 years, with a distribution of 126 males and 16 females.
  - **Cost and resolution**: 16S sequencing is less expensive because it focuses on hypervariable regions of the bacterial 16S gene , requiring a lower sequencing depth. However, provides limited taxonomic resolution(typically at the genus or family) and does not allow direct analysis of the functional potential of the microbiome.
  - **Advantage of reuse**: The sample size offers greater statistical power for identifying differences in microbial composition between ASD and TD, despite the lower resolution.
- **Shotgun metagenomico**: The dataset consists of 60 samples (30 ASD, 30 TD) with 5619 OTUs obtained by species resolution metagenomic sequencing.
  - **Cost and complexity**: Shotgun sequencing is significantly more expensive because of the need to sequence the entire microbial DNA (not just the 16S gene), requiring high read depth, intensive computational resources for genome assembly, and large-volume (terabyte-scale) data storage.
  - **Analytical advantages**: Provides taxonomic resolution at the species/strain level and allows identification of metabolic pathways , antibiotic resistance genes and other functional elements, enriching the etiological analysis of ASD.

Synergy between the two approaches: The integrated reuse of 16S and shotgun data exploits the complementary advantages of the two methods. The 16S dataset , with its large cohort, allowed a robust preliminary analysis of structural disparities in the microbiome. The shotgun data, although on a small cohort, have deepened these observations, linking taxonomic differences to specific microbial functions. This hybrid strategy optimizes costs and outcomes, maximizing the impact of existing resources.

5 Data Preparation

The data were preprocessed following a structured pipeline to ensure quality and consistency among different datasets.

Preprocessing:

- Non-informative columns (e.g., indexes and identifiers) have been removed.
- A binary target variable was created to distinguish between ASD and TD samples, based on the original labels, resulting equally distributed as shown in Figure 1 and Figure 2.
- Missing values were handled by imputation based on data distribution.

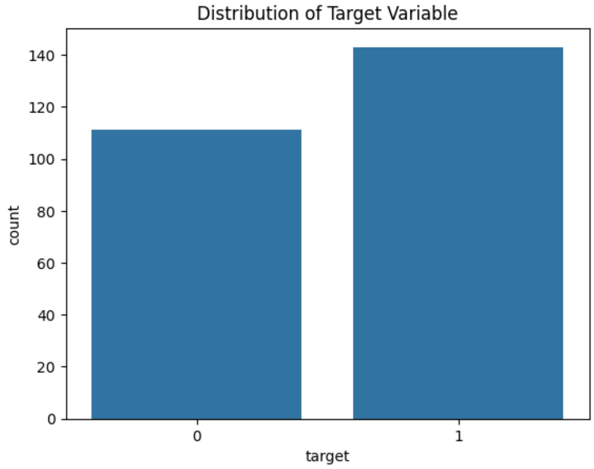

Figure 1: Distribution of the target variable for 16s.

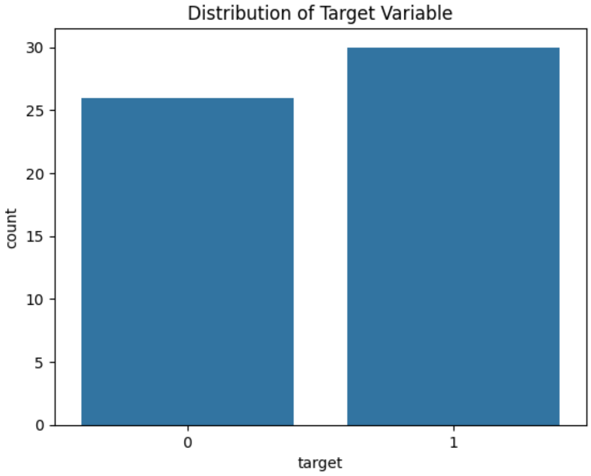

Figure 2: Distribution of the target variable for Shotgun.

Normalization: The data were normalized with **Centered Log-Ratio(CLR)** to mitigate compositionality and make the distributions more suitable for statistical analysis and model training.

Partitioning:

- The datasets were split into train/test with an 80/20 ratio using *stratified sampling*, preserving the balance of ASD/TD classes.
- A *cross-stratified validation* to improve the generalization of the model and reduce the variance of performance.

6 Modeling

The proposed solution was divided into four separate experiments, aimed at evaluating the effectiveness of different learning strategies for the noninvasive diagnosis of ASD based on the gut microbiome.

- **Baseline experiment:** Application of a model *Random Forest* directly on the concatenation of the two datasets(*16S* and *Shotgun*) in order to obtain a baseline.
- **Experiment 1:** Application of a *Random Forest* model directly on both datasets(*16S* and *Shotgun*) in order to obtain a benchmark on performance without dimensionality reduction techniques.
- **Experiment 2:** Introduction of *Principal Component Analysis(PCA)* for dimensionality reduction, followed by training a Random Forest model on the reduced representation, with the aim of evaluating the performance improvement due to the selection of the most relevant features .
- **Experiment 3:** Using a *Supervised Autoencoder(SAE)* to learn a meaningful latent feature representation , later used as input to the Random Forest model.
- **Experiment 4:** Using *Transfer Learning* techniques to transfer SAE knowledge learned on the 16S dataset to the Shotgun dataset.
- **Experiment 5:** Using techninche of **Padding** to be able to adapt feature dimensionality and use transfer learning tecninca without adaptation layer .

In all experiments , a **Stratified5-Fold Cross-Validation** to ensure robust evaluation and reduce the risk of overfitting, given the limited data size and complexity of the task.

The proposed architectures for the two SAEs are outlined below:

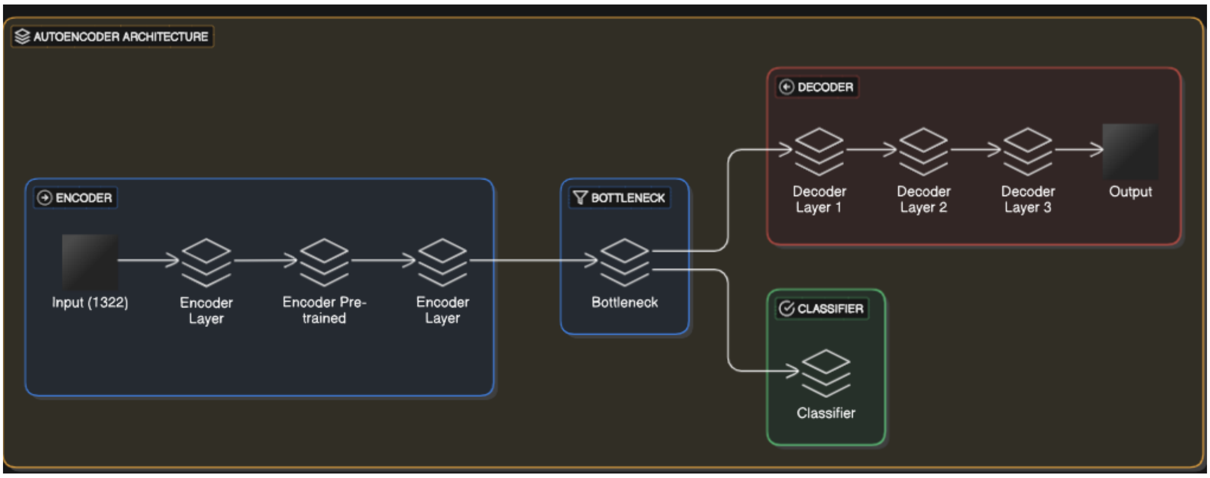

Figure 3: Architecture pre-trained model

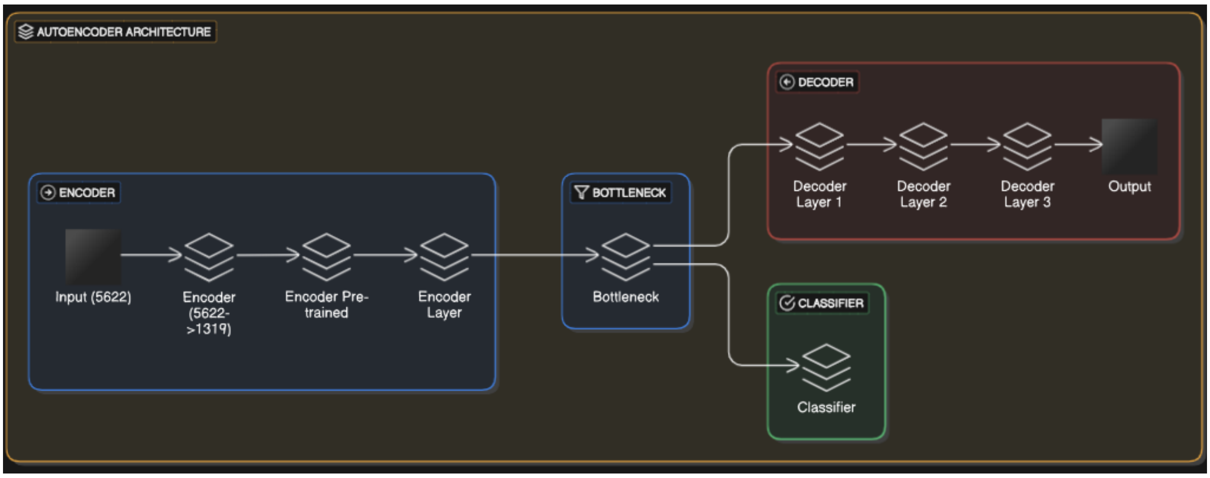

Figure 4: Architecture for Transfer Learning dataset with Adaptation Layer

6.1 Transfer Learning con SAE

The main objective of the study was to improve classification performance on the dataset *Shotgun*, which had significantly fewer instances (60 samples compared to 255 in the *16S* dataset). In this scenario, a **transfer learning** approach was chosen to take advantage of the knowledge gained from a pre-trained on the dataset *16S* and transfer it to the dataset *Shotgun*. This approach improved the effectiveness of the model in predicting targets.

One of the main difficulties in applying transfer learning was the significant difference in the number of features between the two datasets: *16S* has 1322 features , while *Shotgun* has significantly more features (5619). This gap in dimensionality would have made direct use of the pre-trained autoencoder weights impossible, since the input data in the model would have had incompatible dimensions.

To address this difficulty, a **dimensionality adaptation layer** was introduced in the initial part of the autoencoder, as can be seen in the figure 4. This projection layer was designed to reduce the dimensionality of the data from the *Shotgun* dataset (5619 features) to the feature number of the *16S* dataset (1322 features), creating compatibility between the two datasets. The idea behind this solution was to use a **dimensionality adaptation layer** that would act as a"bridge" between the two datasets, reducing the number of features in the dataset *Shotgun* and allowing the model to take advantage of the weights pre-trained on the intermediate layers , without having to re-train the entire network from scratch.

The adoption of this strategy allowed the transfer of knowledge gained from the pre-trained model on the *16S* dataset to the *Shotgun* dataset, greatly improving classification performance, thus highlighting the effectiveness of transfer learning in the context of large differences in data dimensionality.

6.2 Combined Loss Function

A key aspect of the SAE approach was the introduction of a **combined loss** , achieved by adding a classification layer within the autoencoder. The loss function adopted (MSE as reconstruction loss and BCE as classification loss) was defined as follows:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{reconstruction} + (1 - \alpha) \cdot \mathcal{L}_{classification} \tag{1}$$

The parameter α (with values between 0 and 1) controls the relative contribution of the two loss components. A value of $\alpha = 0$ emphasizes the supervised component, while a value of $\alpha = 1$ makes the model equivalent to a pure autoencoder, removing supervision. Experimental results showed that the introduction of supervision significantly improves model performance compared with the purely unsupervised approach.

6.3 Training Configuration

The SAE model was trained for a total of 100 epochs, with a **early stopping** criterion set at a patience of 10 epochs to avoid overfitting. Optimization was performed using the **Adam** algorithm with an initial learning rate of 10^{-4}.

The same training settings were replicated on both datasets to ensure consistency and reproducibility of the results.

7 Evaluation

Model performance was evaluated using the following metrics: *macro average precision*, *macro average recall*, *macro average F1-score*, *weighted average precision*, *weighted average recall* and *weighted average F1-score*. These metrics provide an overall view of the predictive capabilities of the models, taking into account both the balance of classes and their individual performance.

7.1 Baseline Analysis

The Baseline experiment was carried out to make sure that simply concatenating the two datasets,with respect to users in common, would not lead to any results, since the two datasets represent two views at different granularities, so a method of fitting the two ganularities is needed to use them. Such a method we introduced is the use of a projection layer between the higher dimensionality to the lower dimensionality in the' SAE. So the Baseline involves an initial experiment on the concatenated dataset containing a total of 254 instances and 6946 features. On this dataset, a Random Forest was trained, as this model also allows NaN values to be used, obtaining the expected results shown in the following table.

	Precision	Recall	F1-Score
Macro Avg	0.329 (± 0.192)	0.503 (± 0.013)	0.388 (± 0.112)
Weighted Avg	0.358 (± 0.171)	0.559 (± 0.018)	0.427 (± 0.095)

Table 1: Precision, Recall, and F1-Score Results

7.2 Analysis of Results on Dataset 16S

The analysis of the results obtained on the dataset *16S* showed an improved significant performance due to the use of the latent representation generated by the Supervised Autoencoder (SAE) compared to the direct application of the model *Random Forest* on the entire dataset or with dimensionality reduction using *Principal Component Analysis(PCA)*.

As shown in 5 , adoption of the latent representation resulted in an average increase of **+0.03** on all metrics, demonstrating the effectiveness of the approach.

A thorough analysis of the impact of the parameter α in the combined loss showed that the optimal value turns out to be $\alpha = 0.7$, which suggests that the inclusion of the supervised component contributed significantly to the effectiveness of the model, giving greater importance to the classification component than to the simple reconstruction of the data.

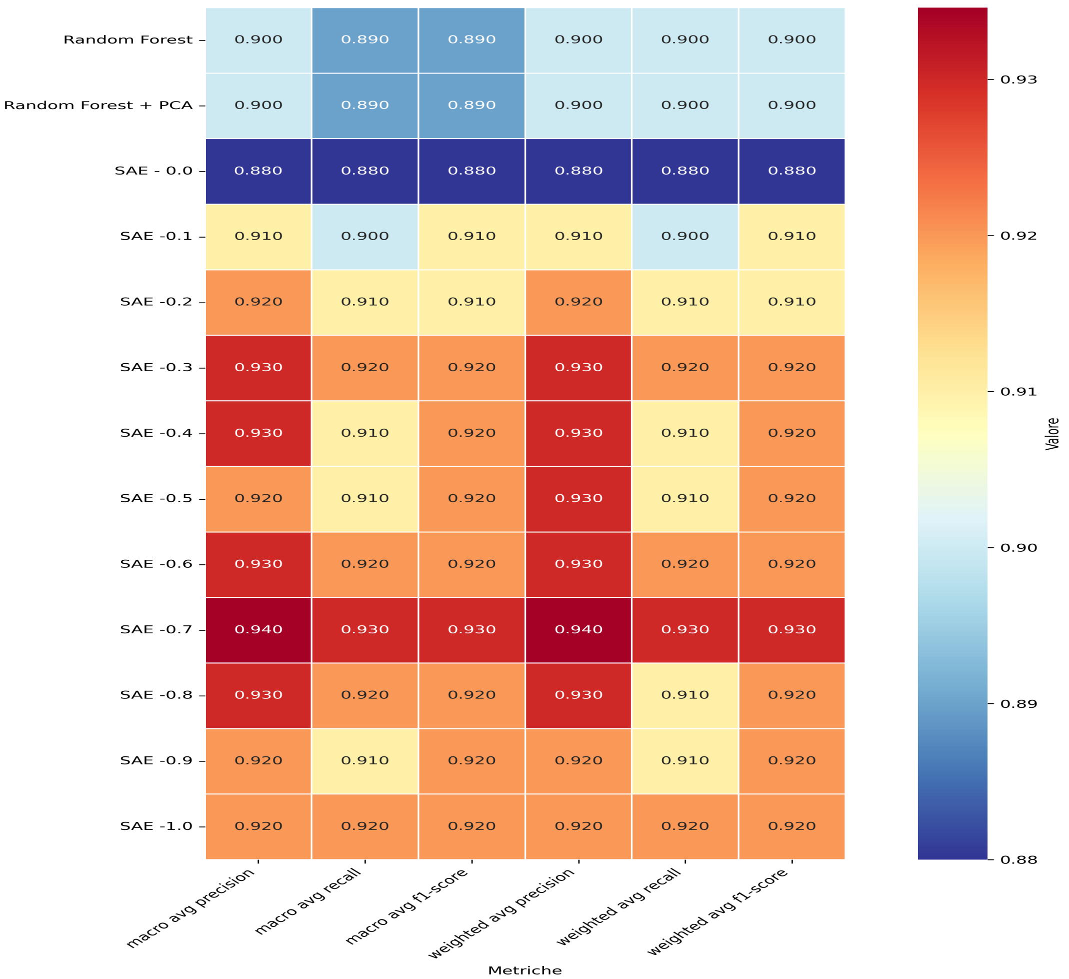

Figure 5: Heatmap of metrics for the 16S dataset.

7.3 Analysis of Results on the Shotgun Dataset

The main objective of the study was to improve performance on the dataset *Shotgun*, characterized by a small number of samples. The results were analyzed by comparing two model configurations: with and without *fine-tuning*.

As highlighted in the figure 6, the use of **transfer learning** proved crucial to exploit the pre-acquired from the dataset *16S*, allowing significant performance improvement over direct training on the reduced dataset. The analysis showed substantial improvement in metrics due to fine-tuning, confirming the effectiveness of the knowledge transfer strategy between the two datasets.

Also for the dataset *Shotgun*, analysis of the α parameter confirmed the usefulness of the supervised approach, with an optimal value of $\alpha = 0.6$. This result suggests that, despite the smaller amount of data, the inclusion of the classification component achieved performance comparable to that of the dataset *16S*, demonstrating the validity of the proposed approach.

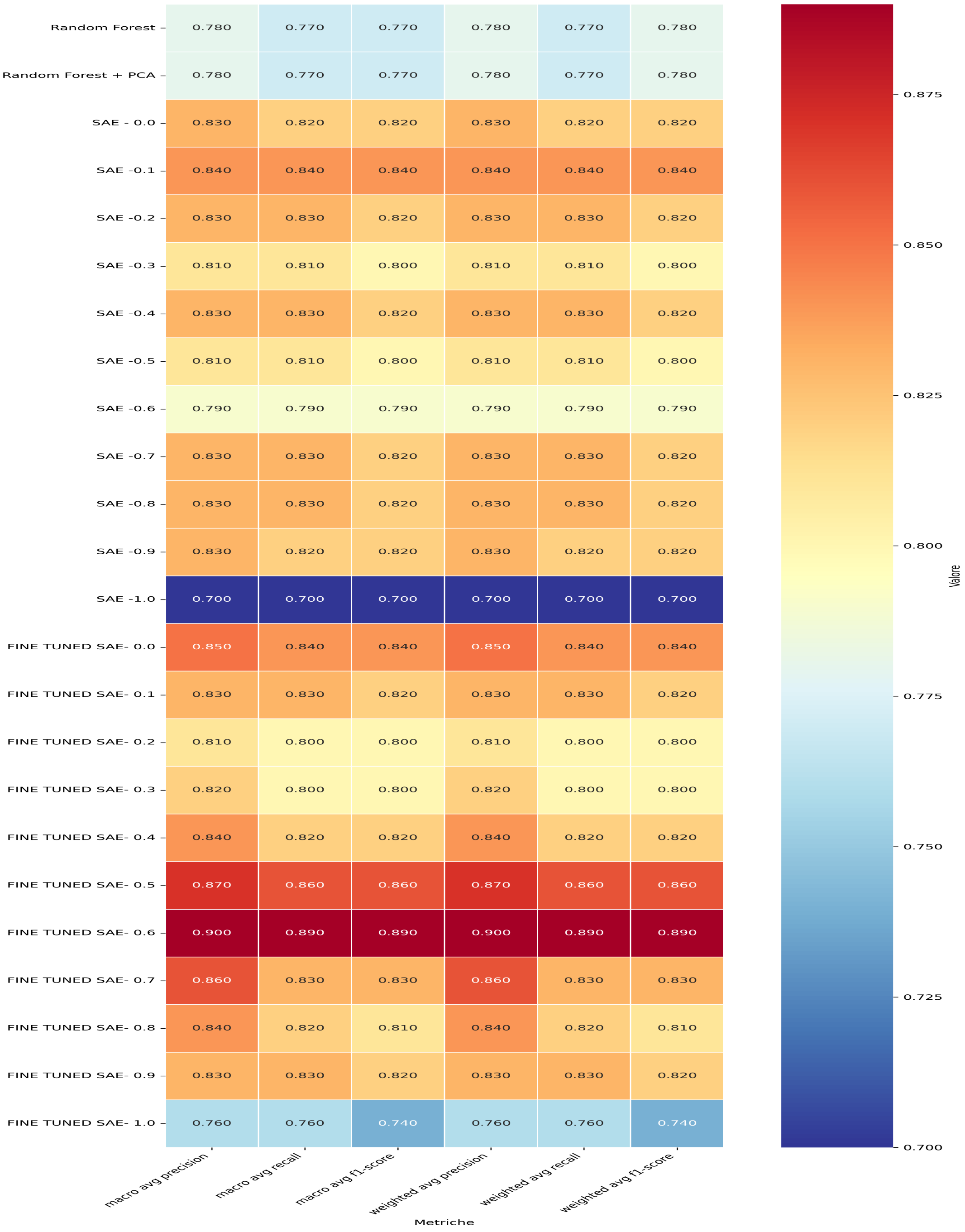

Figure 6: Heatmap of metrics for the Shotgun dataset.

7.4 Analysis Experiment with Padding

The application of padding, implemented in order to avoid the inclusion of an adaptation layer , produced unsatisfactory results, as shown in the figure 7 . The introduction of a significant amount of padding in the initial dataset had a negative impact on the performance of the model, limiting its ability to capture meaningful representations of the data. In particular, padding increased the dimensionality of the data without adding useful information, involving the inclusion of neutral values that did not contribute to the propagation of information signals through the neural network .

This choice had a deleterious effect on the learning phase, as the model struggled to distinguish between relevant values and the padding itself, reducing the quality of learned features . Furthermore, the approach of applying excessive padding did not improve the generalization of the model, as it did not preserve discriminative information for the classification task , especially in the context of the ”Shotgun” dataset, which contains fewer instances than the first dataset. In fact, the high number of padding introduced a bias in the data, preventing the model from focusing on meaningful patterns . This phenomenon made it difficult for the model to learn effectively, leading to a reduction in overall performance.

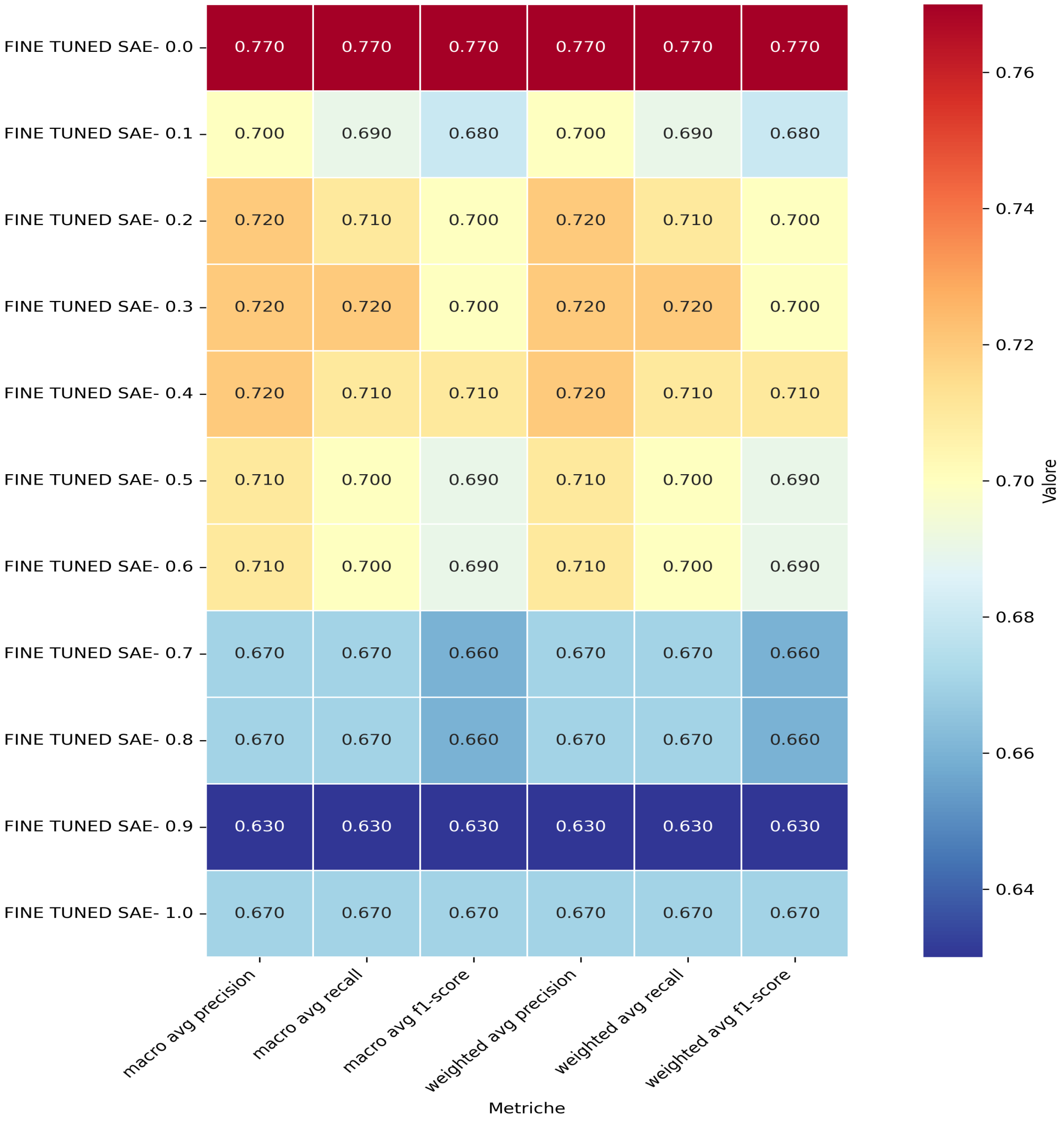

Figure 7: Metrics heatmap for Shotgun dataset, using padding

8 Conclusions

In this study , we explored different modeling strategies for classification of ASD and TD samples using two distinct microbial datasets : *16S rRNA* and *Shotgun metagenomic.* The experiments conducted showed that the approach based on the use of a Supervised Autoencoder (SAE) for the extraction of latent representations resulted in significant performance improvement over the directapplication of a Random Forest, the use of dimensionality reduction techniques such as PCA or simple concatenation of the two datasets.

A key result of our study was the effectiveness of **transfer learning** between the two datasets. In particular, the transfer of weights from the intermediate layers of the trained autoencoder on the dataset *16S* resulted in a marked improvement of the metrics on the dataset *Shotgun*, characterized by a limited number of instances. This approach demonstrated how knowledge gained from larger and richer datasets can be exploited to improve the quality of forecasts on smaller datasets.

The introduction of the **combined loss**, which balances the data reconstruction and the supervised through the parameter α, allowed for a better trade-off between unsupervised and supervised learning. The results showed that an optimal value of α between 0.6 and 0.7 allows maximizing the performance, emphasizing the importance of the classification component without compromising the generalization capability of the model.

Finally, the comparative analysis of the different strategies tested confirmed that the SAE approach with fine-tuning is the most effective solution, allowing for superior results and more stable performance through *stratified5-fold cross-validation.*

8.1 FutureWork

Future research directions could include:

- The exploration of more complex deep learning architectures, such as convolutional neural networks or transformers, for better capture of latent relationships in microbial data.
- The integration of additional multi-omics data sources to enrich the available information and improve the robustness of the model.
- The implementation of data augmentation strategies to address the sparsity of samples in the dataset *Shotgun*, improving the generalization of the model.
- The optimization of preprocessing and feature engineering processes to achieve even more informative representations.

In conclusion, our work provides a solid basis for the use of techniques of deep learning in the context of microbiome analysis for ASD, offering concrete insights for future developments in precision medicine.

References

[1] Anibal S. Heinsfeld, Alexandre R. Franco, R. Cameron Craddock, Augusto Buchweitz, and Felipe Meneguzzi. Identification of autism spectrum disorder using deep learning and the abide dataset. *NeuroImage: Clinical*, 17:16–23, 2018.