

# Comparative Deep Learning Architectures for Exploring High-Dimensional, Small-Sample Data

Marek Sztuka<sup>1\*</sup>, Joanna Szyda<sup>1,2</sup>

<sup>1</sup> Biostatistics Group, Department of Genetics, Wrocław University of Environmental and Life Science, Wrocław, Poland; <sup>2</sup> Wrocław University of Science and Technology, Department of Biomedical Engineering, Wrocław, Poland; \*Presenting and corresponding author: [marek.sztuka@upwr.edu.pl](mailto:marek.sztuka@upwr.edu.pl)

## Introduction

Deep learning (DL) has been successfully applied to pattern recognition tasks for several decades and is now a dominant analytical framework across many domains. Nevertheless, reliable pattern detection remains challenging in biological data due to noise, sparsity, and complex dependencies among features. Many biological studies are characterized by a small number of samples ( $N$ ) described by a large number of features ( $P$ ), resulting in a  $P \gg N$  problem, as exemplified by genome-wide association studies. In this study, we focus on gut microbiome abundance data as a representative case of high-dimensional, small-sample biological data. Although microbiome datasets are typically lower-dimensional than genomic data, the imbalance between sample size and feature number often persists (Eloe-Fadrosch et al., 2015). Moreover, microbiome data are sparse, zero-inflated, and compositional, such that only relative abundances are meaningful (Gloor et al., 2017). Traditional differential abundance analyses, which often test taxa independently, fail to capture higher-order co-occurrence patterns and may lead to inflated false discovery rates. To address these challenges, dimensionality reduction and embedding-based representations are commonly applied, including linear methods such as principal component analysis, nonlinear techniques such as uniform manifold approximation and t-distributed stochastic neighbor embedding, and algebraic approaches such as Singular Value Decomposition (SVD) and Non-negative Matrix Factorization (NMF). Building on these traditional methods, DL enables the construction of data-driven embeddings through architectures such as autoencoders (AE) and one-dimensional convolutional layers (Conv), which project high-dimensional microbiome profiles into lower-dimensional latent spaces to facilitate classification. However, such representations may reduce biological interpretability. Here, we compare traditional and DL-based dimensionality reduction techniques combined with shallow (SNN) and deep (DNN) neural network classifiers, alongside baseline methods. The primary objective of this study is to evaluate how different combinations of DRTs and neural network architectures perform in classifying sparse, high-dimensional, small-sample microbiome data.

## Materials & Methods

**Dataset.** The dataset originated from an experiment by Jakimowicz et al., (2025) designed to investigate whether probiotic supplementation of water and feed alters the composition of microbial communities in water, sediment, and the intestinal tract of fish. The common carp (*Cyprinus carpio*) was selected as the model organism. The experimental design consisted of five groups differing in probiotic supplementation schemes, including environmental (W1, W2) and feed-based (F1, F2) additives (Table 1). The **control group** received no supplementation. **Group 1** received the W1 additive; **Group 2** received W1 in combination with the F1 supplement; **Group 3** received W2 alone; and **Group 4** received W2 combined with F2. Each group was maintained in five independent ponds. At the end of the experiment, intestinal samples were collected from five fish per pond, resulting in 124 samples total. Microbial community profiling was performed by sequencing 16s rRNA gene. Raw reads were preprocessed and taxonomically annotated using QIIME2 (Bolyen et al., 2019),

resulting in a dataset comprising relative abundances of 126 bacterial families across 124 samples. This resulted in a data frame with a feature-to-sample ratio near one-to-one, representing a mild but non-trivial version of  $p \gg n$  structure. To account for such factors as compositional nature of the data and variable sequencing depth, Centered Log-Ratio (CLR) (Aitchison, 1982) transformation was performed.

**Table 1.** Effective microorganism supplementation scheme.

| Group         | Environmental supplementation | Feed Supplementation |
|---------------|-------------------------------|----------------------|
| Control Group | None                          | None                 |
| Group 1       | W1                            | None                 |
| Group 2       | W1                            | F1                   |
| Group 3       | W2                            | None                 |
| Group 4       | W2                            | F2                   |

**Classification baselines.** Prior to the application of DL based classification approaches, we established the following classification quality baselines. First, the classification accuracy assuming the null hypothesis of no effect of probiotic supplementation, expressed by a random assignment of ponds to classes. The expected accuracies under this scenario was 0.20 for the 5-class split. Second, a classical machine learning benchmark was implemented using the gradient-boosted decision tree classifier (XGBoost) (Chen and Guestrin, 2016) was applied. The model was configured with a maximum tree depth of 3 and a learning rate (*eta*) of 0.1. It was trained for 100 boosting iterations on the validation set and evaluated on the test set to obtain accuracy scores.

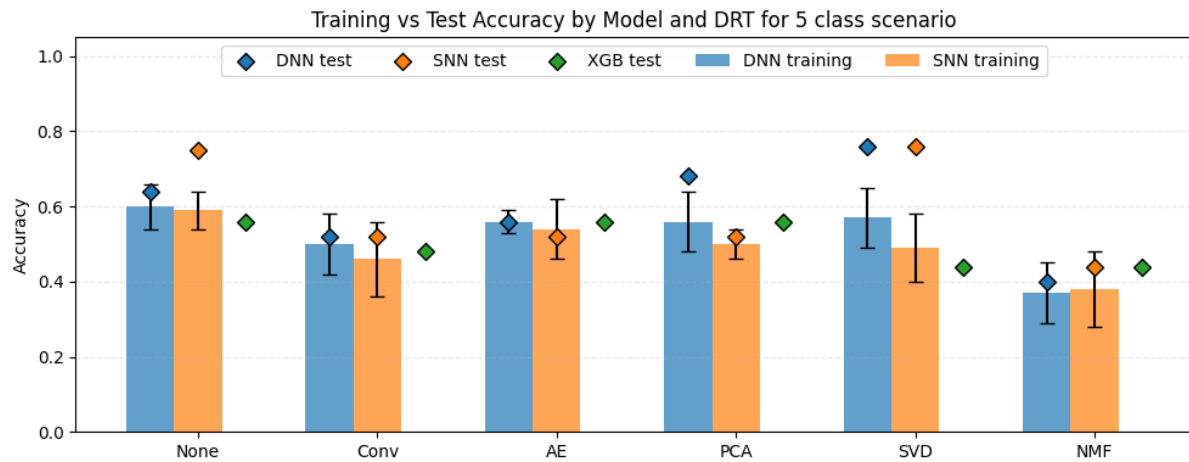
**Classification approaches.** In this study, five different DRT's were applied. To ensure consistency in the modeling pipeline all of them were first fitted on the validation set and then applied to both validation and test set. First was the Principal Component Analysis (PCA) (Pearson, 1901). In order to find the best number of principal components (PC's) the method was first run iteratively with an increasing number of components, with a cumulative explained variance threshold of 90% as the initial target. Once the number of PCs required to meet this threshold was identified, a range of nearby component counts was also evaluated. These transformed datasets were then used as input for neural network classifiers to assess classification performance. Based on these results, the final number of principal components was selected to balance dimensionality reduction with classification accuracy. The second dimensionality reduction technique explored in this study was SVD. In the similar to PCA fashion, the number of components was determined based on variance threshold, and then nearby component counts were explored with regard to validation accuracy. The third DRT applied in this study was NMF. Since NMF requires strictly non-negative input values, specifically for this approach the dataset was shifted after CLR transformation. NMF was implemented using the scikit-learn package (Pedregosa et al., 2012), with the maximum number of components set to 99. After factorization, Gini coefficients (Hurley and Rickard, 2009) were calculated for each component to assess sparsity. Components with Gini values below 0.9 were discarded, and the remaining components were used as the reduced representation of the dataset. The fourth dimensionality reduction was in form of 1-D Conv layers implemented using TensorFlow (Abadi et al., 2016) with keras API (Chollet, 2015). To enable pattern recognition, the input data was formatted such that each individual animal's bacterial abundance represented a one-dimensional "image". First Conv layer had two kernels of size 25 followed by another Conv layer with one kernel of size 4. Both of those layers utilised Rectified Linear Unit (ReLU) as an activation function as well as kernel

regularization in a form of L2. Next was a 1-D MaxPooling layer with pool size of 4. Conv section was trained together with an DL architecture. Last DRT employed in this study was DL based embedding using AE architecture. Those are the class of unsupervised artificial networks designed to learn compressed efficient representation of the data. Such architecture consists of two parts: encoder which is tasked to embed input data into latent space vector of smaller dimensionality, and decoder which reconstructs original input from this compressed representation. In this study, the AE architecture was symmetrical, with both encoder and decoder composed of three fully connected layers using linear activation functions. The encoder comprised layers with 64, 32, and 19 nodes, resulting in a 19-dimensional embedding vector for each sample. Mean squared error (MSE) was used as the loss function, and the ADAM optimizer was applied during training, which continued until the loss function plateaued.

**Neural networks.** Two neural networks were employed and tasked to classify samples based on reduced by DRT or original dataset. First DNN consisted of 4 Dense layers with one dropout layer in between, followed by Softmax output layer. All of the dense layers employed ReLU activation function and L2 kernel regularization. Second architecture; SNN utilised only one dense layer with same parameters as DNN followed by output Softmax layer. Both of the networks were trained, validated and tested on corresponding datasets using ADAM optimizer. Main metric used for performance comparison was classification accuracy (acc).

## Results

For the 5-class split, the highest test accuracy (0.76) was achieved by both DNN and SNN models combined with SVD-based DRT, as well as by the SNN trained without DRT. The next-best performance was observed for PCA-DNN, which reached a test accuracy of 0.68. The lowest test performance was obtained for NMF, with accuracies of 0.40 and 0.44 for DNN and SNN, respectively. This poor performance is most likely caused by the incompatibility of NMF with necessary shift of values generated by the CLR transformation. For the remaining DRTs, test accuracies varied only moderately, ranging from 0.64 (None-DNN) to 0.52 (PCA-SNN), as shown in Fig. 1. 5-fold cross-validation (CV) results revealed that the models trained without DRT showed the most stable and consistently high performance for both SNN and DNN, achieving mean accuracies close to 0.60 with a standard deviation of approximately 0.06. This suggests that the dataset may be too small or sparse, for some of the DRT's to generalize reliably. The weakest CV performance was again observed for NMF, followed by Conv-based representations. The limited effectiveness of Conv layers likely reflects the absence of meaningful structural patterns in taxonomic abundance data. AE-based representations exhibited moderate but notably stable performance. Test accuracies of 0.56 and 0.52 were obtained for DNN and SNN models, respectively, with corresponding validation accuracies of 0.56 and 0.54 and standard deviations of 0.03 (DNN) and 0.08 (SNN). Although AE-based models did not achieve the highest peak accuracies, their consistency suggests notable robustness under sparse conditions and complex patterns. Across all DRTs, CV performance of DNN and SNN models was comparable, with accuracy differences remaining within standard deviation, indicating no consistent advantage of deeper architectures under small-sample conditions. The XGBoost baseline showed performance similar to NN models in most settings, with the notable exception of SVD-based representations, where its performance was substantially lower. All evaluated approaches significantly exceeded the theoretical worst-case accuracy of 0.20 for the 5-class task.



**Figure 1.** Validation (bar with whiskers) and test Accuracy for Shallow (orange) and Deep (blue) neural network architectures with different dimensionality reduction techniques. Baseline XGboost (green). For five class classification task.

## Conclusions

Results from this study indicate that under high-dimensional, small-sample conditions, simpler NN architectures and unreduced feature spaces can perform competitively with more complex DRT-based models; however, when appropriately matched to the data structure, certain DRTs, such as SVD, can substantially improve classification performance. Further evaluation on larger or simulated datasets is necessary to better characterize the conditions under which DRT consistently benefits DL-based microbiome classification. It is worth mentioning that effective microorganism treatment introduced some shifts in microbial community structure that could be recognised by machine learning models.

## References

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, et al. TensorFlow: A system for large-scale machine learning. doi:10.48550/ARXIV.1605.08695
- Aitchison, J., (1982). The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 44, 139–160. doi:10.1111/j.2517-6161.1982.tb01195.x
- Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, et al. (2019) . Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology* 37, 852–857. doi:10.1038/s41587-019-0209-9
- Chen, T., Guestrin, C., (2016). XGBoost: A Scalable Tree Boosting System. doi:10.48550/ARXIV.1603.02754
- Chollet, F., (2015). Keras.
- Eloe-Fadrosch, E.A., Brady, A., Crabtree, J., Drabek, E.F. et al. (2015). Functional Dynamics of the Gut Microbiome in Elderly People during Probiotic Consumption. *mBio* 6, e00231-15. doi:10.1128/mBio.00231-15
- Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V., Egozcue, J.J., (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology* 8, 2224. doi:10.3389/fmicb.2017.02224
- Hurley, N., Rickard, S., (2009). Comparing Measures of Sparsity. *IEEE Transactions on Information Theory* 55, 4723–4741. doi:10.1109/TIT.2009.2027527
- Jakimowicz, M., Sidorczuk, K., Huyben, D., Hildebrand, F., Napora-Rutkowski, et al. (2025). Supplementation with effective microorganisms in earthen ponds affects common carp growth but not overall microbial communities. doi:10.1101/2025.07.07.663436
- Pearson, K., (1901). LIII. *On lines and planes of closest fit to systems of points in space*. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2, 559–572. doi:10.1080/14786440109462720
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, et al. (2012). Scikit-learn: Machine Learning in Python. doi:10.48550/ARXIV.1201.0490