Phylogeny-Aware Deep 1-Dimensional Convolutional Neural Network for the Classification of Metagenomes

Timmy Manning*, Jyotsna Talreja Wassan†, Cintia Palu*, Haiying Wang†,
Fiona Browne†, Huiru Zheng†, Brian Kelly* and Paul Walsh*
*NSilico Life Science Ltd.
Cork, Ireland
Email: paul.walsh@nsilico.com
†School of Computing
Ulster University
Northern Ireland, United Kingdom

***Abstract*—This paper evaluates a novel approach to the integration of biological domain knowledge relating to the natural evolutionary structure of microbial community data to classifying 16S rDNA sequence samples. Specifically, we evaluate the use of phylogenetic trees in addition to amplicon sequence variant abundance in samples for the classification of a processed cattle metagenomics data set using machine learning. Further to this, we employ a class activation map of the network when applied to specific exemplars to determine, firstly, the relevance of higher level taxonomic data, and secondly, the most relevant taxa in determining the classification, according to the classifier.**

***Index Terms*—amplicon sequence variants, convolutional neural network, machine learning, metagenomics, phylogeny, Grad-CAM, visualisation**

I. Introduction

Whereas genomics is concerned with the genes or entire genome of a specific organism, metagenomics deals with evaluating the genomes of microorganisms collectively (characterising the microbial community) [1]. Such communities can be sampled from geographical locations or hosts. The differences observed in the genetic profiles of the microbiota can then be used to, for example, detect traits in a host [2]–[4] or perform functional analysis [5]–[7].

Advances in Next-Generation Sequencing (NGS) have made the application of metagenomics to large scale evaluations both practical and cost effective. This has led to research experiments which generate terabytes of data relating to the genomes and abundances of thousands of archaea and bacteria across potentially thousands of hosts. The resulting explosion in the volume of such complex big data this has produced, facilitates the need for more powerful, efficient and robust processing tools.

Phylogeny refers to the branching of taxa through evolution [8]. Phylogenetic information is typically represented on a phylogenetic tree, where the nodes represent the common ancestry [9]. All the taxa from the same kingdom identified across samples are inherently phylogenetically related at some level, with some potentially being very closely related. For example, although a species may only be in one sample, highly phylogenetically related species (possibly functionally equivalent) may be present in other samples. The fusion of phylogenetic information in a classifier therefore potentially provides a wealth of relevant information on how to interpret the presence of related taxa, as equivalent taxa can take different forms in different samples. However, the relatedness of taxons is not an aspect that is often considered when applying machine learning to metagenomic data.

Considering this, we propose a novel convolutional neural network (convnet) [10] for the functional classification of metagenomic profiles. Convnets work by learning small filters that detect *"local patterns"* in the data [11]. These filters are then applied across the data in a sliding window approach. The patterns discovered therefore become *"translation invariant"*; the filters can detect the patterns anywhere in the input, and the multiple occurrences of the pattern can be detected. A convolutional layer is usually composed of several of these filters. A convnet can contain multiple sequential convolutional layers, allowing it to detect *"patterns of patterns"*. Although originally designed with image classification in mind, they have also proven highly successful on text and natural language classification tasks [12]. Following this, we applying a Grad-CAM [13] approach to visualising the relevance of elements in the input, and evaluate its applicability to our data.

II. Related Work

Given the large amount of data that can be produced in metagenomics experiments, focus was placed on examining existing deep learning approaches. In particular, deep convnet approaches were targeted, as these algorithms are able to handle large numbers of input features, in a manner that can preclude the need for feature engineering (feature extraction) [14] and feature selection on many big data problems,

This work was supported in part by the MetaPlat project, (www.metaplat.eu) funded by H2020-MSCARISE- 2015.

allowing the network to evaluate all data, and make its own interpretations of data relevance [15]. It has also been noted that the state-of-the-art feature selection methods may have issues scaling to the larger data sets being created [16].

Reiman, Metwally and Dai [17] approached this problem by first generating a phylogenetic tree, where the leaf nodes are the abundances of operational taxonomic units (OTUs; groups of sequence read variants that differ that by less than a set threshold [18]). The abundance of a node in the tree is set as the sum of the abundance of its children. The abundance values from the tree are embedded into a two-dimensional matrix. Each row in the matrix represents a level in the phylogenetic tree, where the root node occupies the top left element in the matrix, as shown in Fig. 1. This encodes the abundance information while preserving spatial representation of the tree, where different taxonomic ranks are present in different rows, and elements derived from the same direct ancestor are placed adjacently. Empty spaces in the matrix are padded.

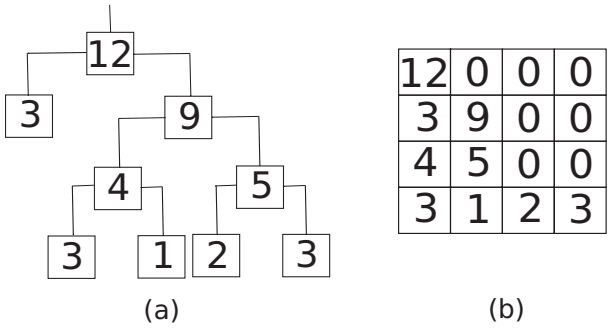

Fig. 1. (a) Simplified phylogenetic tree, with the leaves and nodes labeled with the corresponding abundances, and (b) how the abundances are mapped to the synthetic image.

The encoded matrices are used to train a two-dimensional convnet with three convolutional layers. The data comprised 4300 OTUs each in at least one of 1967 samples, taken from one of three environments, skin, gut or oral cavity, corresponding to the three possible classifications of the profile. This data is taken from the *"Moving Pictures of the Human Microbiome"* study [19]. Feature selection was carried out to reduce the number of OTUs to 1706. This approach achieved 99.47% accuracy, although the selected sites would be expected to have diverse profiles.

The Met2Img algorithm is another two-dimensional convnet approach for classification based on metagenomic profiles [20]. The images are created by first colour-coding the abundance values of the OTUs through binning based on a logarithmic scale. Under the Met2Img *"phylogenetic-sorting"* operation mode, a blank synthetic image with a number of pixels greater than the total number of OTUs is created for each sample. The complete set of OTUs are then sorted alphabetically based on their full taxonomic rank (phylum, class, order, family, genus and species). Each pixel in the image, indexed by column then row (right to left), corresponds to the abundance of the ordered OTUs, e.g., the first element in the first row corresponds to the first OTU alphabetically. Samples are then represented as individual images by colouring the pixels corresponding to the OTUs in the sample using the corresponding colour of the abundance bin. If an OTU is not detected in a sample, the corresponding pixel remains white.

This results in a more compact representation than that of Reiman, Metwally and Dai, while still generally encoding related OTUs in relative proximity. However, highly related OTUs may *"overflow"* a row, and be placed on opposite sides of the image.

The images are used to train a convent classifier with a single convolutional layer (comprised of 64 3x3 filters), one pooling layer, and an output layer with one fully connected neuron. This approach performed well and was able to outperform fully connected non-convolutional approaches and MetAML [2] on a number of evaluations on small data sets. The authors also evaluated presenting the data as a vector to a one-dimensional convnet, and noted a general reduction in performance.

However, both these encoding approaches lose a lot of information, i.e., this encoding obscures the phylogenetic relationship between elements at different taxonomic ranks. The phylogenetic tree cannot be reconstructed from the encoding. Secondly, although both approaches attempt to place elements derived from the same direct ancestor next to each other within rows, heavily divergent elements can potentially be adjacent. Given the loss in information, the classifier will be unable to differentiate the boundaries in the clusters from the encoding, which undermines the need to cluster the data based on phylogenetic relatedness.

The Ph-CNN approach does not attempt to represent the metagenomic data as an image [21]. Instead, it considers the fact that, in image processing, as the filters are convoluted, pixels are considered in the context of the values of their closest neighbours (the surrounding pixels). Ph-CNN replicates this concept by dynamically identifying the neighbours of an OTU based on the distances encoded in the phylogenetic tree, specifically the patristic distance. Patristic distance is defined as the sum of the lengths of the branches connecting two OTUs to their most recent common ancestor. Therefore, when a filter is convoluted onto an OTU in Ph-CNN, the abundance of the OTU is considered in the context of the abundances of these neighbours, and there is not necessarily an overlap between inputs to the filters. The input to a custom convolutional layer in Ph-CNN is the OTU abundances and the distances between the OTUs.

Although this is a very interesting approach, it is claimed that the Met2Img approach outperformed this across a number of experiments on genus level data [20]. Ph-CNN is however noted as being generally applicable to any data set for which a concept of closeness of features can be defined.

III. Methodology

A. Processing the data

The metagenomics samples we are dealing with were previously described [22]. In summary, they comprise eighty 16S paired end reads generated from rumen samples of different steer by Scotland's Rural College, Edinburgh, forty of which had nitrate as an additive to their diets at a rate of 18g of nitrate per kg dry matter. Nitrate additives are one of the most promising approaches to reducing enteric methane production by ruminal microorganisms [23]. Feed efficiency measurements are known for these samples, but they are not investigated in this research.

In contrast to the approaches outlined in the *"Related Work"* section, the processing here employs Amplicon Sequence Variants (ASVs) as opposed to OTUs. Although these are similar concepts, ASVs are considered to provide a number of advantages [18], [24]. The samples were sequenced using an Illumina high-throughput sequencing platform. The reads were imported into Qiime2 [25] which was used to demultiplex the reads from each sample. The DADA2 pipeline (Divisive Amplicon Denoising Algorithm) [26] is applied to denoise and dereplicate the reads, and filter chimeras. Based on a visual inspection, thirteen bases were trimmed from the start of both the forward and reverse reads to remove data with high error rates. Using Qiime2, the abundance of each ASV is calculated for each sample, and the phylogenetic relationship of the ASVs is generated. The data sets are divided into two groups: labelled with a *"1"* output (positive) if the sample corresponds to steer that received the nitrate additive, or *"0"* otherwise (negative).

B. Baseline Functional Classification

To set a baseline against which we can benchmark the performance of our novel convnet based approach, we evaluate the cattle data set using state-of-the-art and well regarded standard methods [27], [28]. The approaches evaluated are Random Forest (RF), regularized Logistic Regression (regularized LR), Support Vector Machines (SVMs) with Poly Kernel, Naïve Bayes (NB), Linear Discriminant Analysis (LDA), Multi-Layer Perceptron (MLP), and extreme gradient boosting of trees (XgBoost), using the caret package [29] in R. We also apply RF and LR over phylogenetically transformed data using isometric log transform (PhILR+RF/LR) as proposed by Silverman *et al.* [30].

C. The Novel Convolutional Classifier

The idea behind our classifier approach is to transform each leaf in the phylogenetic tree into a biologically pertinent sentence describing the path through the phylogenetic tree from the root to that element. Where the root represents the kingdom, and the leaf represents a species, the sentence will take the form: [kingdom] [phylum] [class] [order] [family] [genus] [species]. For example, the bacterial species ruminicola (which is abundant in many of the samples), would be represented by the sentence: *"Bacteria Bacteroidetes Bacteroidia Bacteroidales Prevotellaceae Prevotella ruminicola"* detailing the different taxonomic ranks. A separate text file is generated to hold the sentences generated from each sample. If an element is not in a sample, no corresponding sentence will be added to that samples text file.

The abundance of each element is also added to the end of each sentence. As the convolutional network is not designed to work with numerical values, the abundances are converted to a text description. Given the limited training data, a low granularity description was used. Four words were used to describe the relative abundance of an element in the sample: *trace* (value falls below the first quartile), *low* (value between first and second quartile), *high* (value between second and third quartile), and *dominant* (value above the third quartile). Abundance was recorded relative to other elements in the samples, and the abundance of that element compared to other samples where that element was detected. The two relative abundance types were prefixed with c or r, corresponding the cross samples abundance and within sample abundance respectively. This approach essentially makes the granularites different words in the different contexts to convey the different meaning, but given the highly structured nature of the sentences, this may not have been necessary. For our data set, this produced a set of 104,417 words, with 869 unique words, across 80 files. The largest file contained 1941 words, while the smallest contained 905. This approach:

- Encodes phylogenetic information
- Makes data at multiple taxonomic levels available to the classifier
- Uses highly structured sentences
- Does not require feature selection

The convolutional network is implemented using the Keras (*http://keras.io*) wrapper for a Tensorflow [31] backend. The network comprises an embedding layer, two convolutional layers (with ReLU activation), two pooling layers and a fully connected layer with a single output neuron. Binary cross entropy is used as the loss function, and ADADELTA [32] is used as the optimizer. No other special considerations were made for the network parameters. The embedding layer encodes the 869 unique words into a dense vector of 32 floating point values, and accepts inputs of up to 2000 words. The embedding layer is trained with the rest of the network to optimize the encoding. The convolutional layers contain 16 and 32 filters respectively, each of size 5. This resulted in a network with 32,689 trainable parameters. The structure of the network is given in Table I.

TABLE I
Structure and dimensionality of the Convnet.

Layer Type	Output Dimensionality	Trainable parameters
Embedding	2000*32	27488
1D Convolutional	1996*16	2576
Max pooling	399*16	0
1D Convolutional	395*32	2592
Global Max Pooling	32*1	0
Dense	1*1	33

D. Biological Relevance

Here we consider, in an unbiased data set, that elements of the input vector that the classifier deemed *"important"* in generating its output should have biological relevance. To interrogate the classifier to extract this information, we apply the Gradient-weighted Class Activation Mapping (Grad-CAM) algorithm [13]. This is an approach to generating heatmaps on the input based on network activation, so it allows you to visualise what areas of the input signal contributed to the classification generated. The heatmap is generated by plotting the activation of each filter in the final convolutional layer across its input and scaling it by its contribution to the output classification. Although this approach is intended for application to two-dimensional image processing, it should be applicable to our one-dimensional sequential data.

For this approach, the network is adapted to a two-output binary classification problem, with a *softmax* activation on the output layer, so that each class would have a corresponding high activation. All other layers and parameters were maintained as for the previous experiment. This approach generates a heatmap of 395x1 pixels, corresponding to the dimensionality of the feature map produced by each of the 32 filters in the final convolutional layer. This heatmap is then stretched to 2000x1 pixels using OpenCV (Open Source Computer Vision Library [33]) to match the dimensionality of the data fed into the network, and it is aligned against the original text.

IV. Results

A. Baseline Functional Classification

Given the limited amount of data, the evaluation was carried out using Leave-One-Out Cross Validation (LOO-CV). Under LOO-CV, for a training set with n exemplars, the algorithm is run n times, where in each run a different exemplar is removed from the set to be used for evaluating the classifier, while the remaining $n-1$ exemplars are used to train the classifier. As a minimal amount of data is reserved to evaluate the classifiers, the amount of data which can be used to actually train the classifier is increased. The results of the applied methods are listed in TABLE II, where accuracy is defined as the fraction of correctly classified samples across all n runs. Given that the data is evenly distributed between two classes, it is expected that a random classifier would achieve a value of approximately 0.5.

B. The Novel Convolutional Classifier

The novel approach is evaluated using LOO-CV, as per the baseline benchmark. The networks are trained for 30 epochs on each fold. Of the data used to train the network, 8 exemplars are randomly selected set aside to detect overtraining (validation set). After each epoch, the loss is calculated on the 8 reserved exemplars. After training, the weights of the network are reverted to the points at which the loss on the reserved 8 exemplars was at its lowest, and the network is evaluated on the final exemplar. A box-plot of the distribution of accuracies for each of 20 runs of the LOO-CV algorithm are presented in Fig. 2.

TABLE II
Baseline performance of a number of approaches in LOO-CV on the cattle data set.

Approach	Model (LOO-CV)	Accuracy
Based on Raw Species Abundances	RF	**0.975**
	Regularized LR	**0.975**
	SVMs with Poly Kernel	0.913
	NB	0.963
	LDA	0.938
	MLP	0.650
	XgBoost	**0.987**
Based on Integrating Phylogeny	PhILR+RF	0.700
	PhILR +LR	0.775

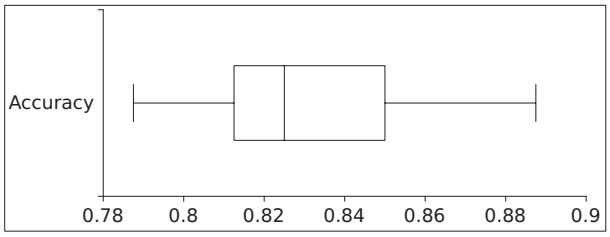

Fig. 2. Boxplot of the accuracies achieved by the convolutional neural network over 20 runs of LOO-CV on the cattle metagenomic data.

C. Biological Relevance

For each sample, the heatmap is generated from a network trained on the remainder of the data, with eight samples reserved for detecting overtraining, as detailed previously. The output is cast to a *"jet"* colour scale, where low activation is shown in blue, and high activation is shown in red. The three most highly activated sections of the heatmap generated from the largest positive sample are given in Fig 3.

V. Discussion

A. Baseline Functional Classification

The results indicate that RF, regularised LR, SVMs, NB and LDA, allow for more accurate predictions. XgBoost with the parameters of *nrounds = 50, max_depth = 2, eta = 0.4, gamma = 0, colsample_bytree = 0.6, min_child_weight = 1* and *subsample = 1*, attained the best model over raw abundance counts in this use case. It should be noted that the baseline approaches that inherently apply feature selection (XgBoost, Regularized LR, and RF) out-performed the other approaches evaluated on this data set, with XgBoost providing the highest level of accuracy. This is in-line with observations in the literature [28].

B. The Novel Convolutional Classifier

The novel convnet approach was out-performed by a number of the benchmark approaches, but we observed that it did perform better than machine learning methods of Regularized LR or RF over PhILR (itself a phylogeny-based approach) [30]. Hence, including phylogenetic context in functional analysis

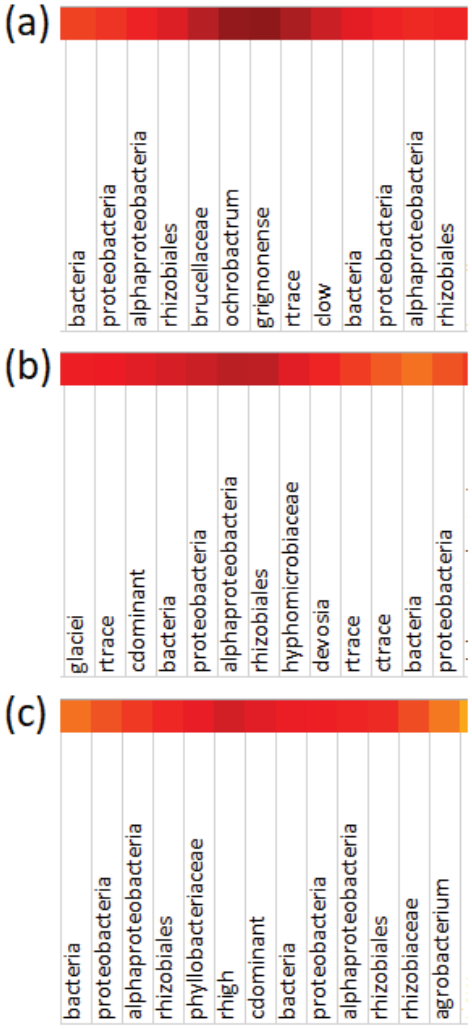

Fig. 3. The three most highly activated regions of the heatmap generated using Grad-CAM from the largest sample in the cattle data set.

with the proposed method based on CNNs, has potential to improve the classification over microbiomes.

For what we wanted to achieve here, the data set is too small, with over-training acknowledged as an issue given the large number of trainable parameters and the small amount of training data available. Indeed, training data volume is acknowledged as one of the key factors in determining the quality of a learned system [34]. Nevertheless, the highly structured nature of the sentences and the limited vocabulary (< 1000 words) appear to have allowed the application of convolutionary networks to successfully classify many of the samples. Therefore, this evaluation can be taken as a *proof-of-concept*, and it is still considered that this approach may indeed give superior performance when applied to larger data sets.

The high level of variation in performance may be partially attributable to the small samples size, which may lead to sampling effects in the division of the data. As an aside, reducing the number of words used to the 700 most common (assuming others had low statistical power), appears to show a slight decrease in performance, although this was not thoroughly investigated.

C. Biological Relevance

The three heatmap sections in Fig. 3 correspond to sentences describing the species grignonense, the genus devosia, and the family phyllobacteriaceae respectively. We therefore infer that the presence of these elements in the sample have some biological relevance to the classification. Ochrobactrum grignonense, for example, identified as the most active point on the heatmap, is noted as being capable of denitrification [35]. This suggests that the approach is indeed credible.

There are several other interesting points to note from these heatmap sections. Firstly, it is noted that the activation in Fig 3 (b) is centred on class level, even though the sentence relates to genus level information. This demonstrates that the classifier is able to make use of higher rank taxonomic data encoded in the sentences. Secondly, of the three sections, only phyllobacteriaceae places the focus on the abundance (within sample abundance here), while abundance is relatively lowly activated in genus devosia. This suggests that only the presence of taxa, and not the actual abundance may be relevant, in some situations.

An interesting aside here is that this approach shows the activations that suggest that a sample could belong to the other class, by using the gradient of the corresponding activation. This may prove to be a valuable source of information in understanding the problem, the functioning of the classifier and manually validating the results.

VI. Conclusions

Given the tentative nature with which we can evaluate the performance of the classifier on such a small data set, future work will prioritise identifying rich and large scale biological data sets. This should allow us to, firstly, gauge the volume of data that is required for a convolutional approach such as this to be meaningfully applied, and secondly, to properly evaluate the true potential performance of the algorithm on larger data sets relative to other approaches. Although this algorithm uses ASV data, it should also work with OTU data, greatly increasing its applicability given the vast amounts of older data publicly available. It is acknowledged however that the reduced accuracy of OTUs may limit performance and the meaningfulness of any biological relevance extracted [18]. The *"Related Work"* section has highlighted a number of data sets and sources that are promising.

A larger data set may also facilitate the use of higher granularity encoding of the numeric values and allow further fine tuning of the algorithm. Additionally, the phylogenetic tree encodes numeric values representing the evolutionary distance between linked elements. Encoding this data may provide additional relevant information to the classifier, but would increase the number of trainable parameters and the size of the generated text descriptions.

Regarding the approach to determining the biologically relevant elements in the input vector, although it does appear to function as intended, it does produce quite inexact results. This is due to a number of reasons. Firstly, the approach was originally designed for visualisation on images. Secondly, it is a noisy approach due to overlaps in the filter processing, the requirement to stretch the activation map to match the input, and inconsistent activation level across samples. Interpretation of this data is currently approached manually, focusing on one sample at a time. Future work may optimise this approach to one-dimensional data, and allow us to automatically generate information across all samples and classes.

REFERENCES

[1] C. S. Riesenfeld, P. D. Schloss, and J. Handelsman, "Metagenomics: genomic analysis of microbial communities," *Annu. Rev. Genet.*, vol. 38, pp. 525–552, 2004.

[2] E. Pasolli, D. T. Truong, F. Malik, L. Waldron, and N. Segata, "Machine learning meta-analysis of large metagenomic datasets: tools and biological insights," *PLoS computational biology*, vol. 12, no. 7, p. e1004977, 2016.

[3] G. S. Ginsburg and H. F. Willard, "Genomic and personalized medicine: foundations and applications," *Translational research*, vol. 154, no. 6, pp. 277–287, 2009.

[4] H. Wang, H. Zheng, F. Browne, R. Roehe, R. J. Dewhurst, F. Engel, M. Hemmje, and P. Walsh, "Analysis of rumen microbial community in cattle through the integration of metagenomic and network-based approaches," in *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*. IEEE, 2016, pp. 198–203.

[5] P. Walsh, C. Palu, B. Kelly, B. Lawor, J. T. Wassan, H. Zheng, and H. Wang, "A metagenomics analysis of rumen microbiome," in *Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2077–2082.

[6] J. T. Wassan, H. Wang, F. Browne, P. Wash, B. Kelly, C. Palu, N. Konstantinidou, R. Roehe, R. Dewhurst, and H. Zheng, "An integrative approach for the functional analysis of metagenomic studies," in *International Conference on Intelligent Computing*. Springer, 2017, pp. 421–427.

[7] J. T. Wassan, H. Wang, F. Browne, and H. Zheng, "Microbial abundance analysis and phylogenetic adoption in functional metagenomics," in *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2017 IEEE Conference on*. IEEE, 2017, pp. 1–8.

[8] S. Whelan, P. Liò, and N. Goldman, "Molecular phylogenetics: state-of-the-art methods for looking into the past," *TRENDS in Genetics*, vol. 17, no. 5, pp. 262–272, 2001.

[9] D. R. Maddison, K.-S. Schulz, and W. P. Maddison, "The tree of life web project," *Zootaxa*, vol. 1668, no. 1, pp. 19–40, 2007.

[10] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, L. Wang, G. Wang *et al.*, "Recent advances in convolutional neural networks," *arXiv preprint arXiv:1512.07108*, 2015.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[12] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[13] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization." in *ICCV*, 2017, pp. 618–626.

[14] O. Faust, Y. Hagiwara, T. J. Hong, O. S. Lih, and U. R. Acharya, "Deep learning for healthcare applications based on physiological signals: a review," *Computer methods and programs in biomedicine*, 2018.

[15] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," *Briefings in bioinformatics*, vol. 18, no. 5, pp. 851–869, 2017.

[16] V. Bolón-Canedo, D. Rego-Fernández, D. Peteiro-Barral, A. Alonso-Betanzos, B. Guijarro-Berdiñas, and N. Sánchez-Maroño, "On the scalability of feature selection methods on high-dimensional data," *Knowledge and Information Systems*, pp. 1–48, 2018.

[17] D. Reiman, A. Metwally, and Y. Dai, "Using convolutional neural networks to explore the microbiome," in *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*. IEEE, 2017, pp. 4269–4272.

[18] B. J. Callahan, P. J. McMurdie, and S. P. Holmes, "Exact sequence variants should replace operational taxonomic units in marker-gene data analysis," *The ISME journal*, vol. 11, no. 12, p. 2639, 2017.

[19] J. G. Caporaso, C. L. Lauber, E. K. Costello, D. Berg-Lyons, A. Gonzalez, J. Stombaugh, D. Knights, P. Gajer, J. Ravel, N. Fierer *et al.*, "Moving pictures of the human microbiome," *Genome biology*, vol. 12, no. 5, p. R50, 2011.

[20] T. H. Nguyen, E. Prifti, Y. Chevaleyre, N. Sokolovska, and J.-D. Zucker, "Disease classification in metagenomics with 2d embeddings and deep learning," *arXiv preprint arXiv:1806.09046*, 2018.

[21] D. Fioravanti, Y. Giarratano, V. Maggio, C. Agostinelli, M. Chierici, G. Jurman, and C. Furlanello, "Phylogenetic convolutional neural networks in metagenomics," *BMC bioinformatics*, vol. 19, no. 2, p. 49, 2018.

[22] C.-A. Duthie, S. Troy, J. Hyslop, D. Ross, R. Roehe, and J. Rooke, "The effect of dietary addition of nitrate or increase in lipid concentrations, alone or in combination, on performance and methane emissions of beef cattle," *animal*, vol. 12, no. 2, pp. 280–287, 2018.

[23] C. Yang, J. A. Rooke, I. Cabeza, and R. J. Wallace, "Nitrate and inhibition of ruminal methanogenesis: microbial ecology, obstacles, and opportunities for lowering methane emissions from ruminant livestock," *Frontiers in microbiology*, vol. 7, p. 132, 2016.

[24] J. T. Nearing, G. M. Douglas, A. M. Comeau, and M. G. Langille, "Denoising the denoisers: an independent evaluation of microbiome sequence error-correction approaches," *PeerJ*, vol. 6, p. e5364, 2018.

[25] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon *et al.*, "Qiime allows analysis of high-throughput community sequencing data," *Nature methods*, vol. 7, no. 5, p. 335, 2010.

[26] B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes, "Dada2: high-resolution sample inference from illumina amplicon data," *Nature methods*, vol. 13, no. 7, p. 581, 2016.

[27] A. Statnikov, M. Henaff, V. Narendra, K. Konganti, Z. Li, L. Yang, Z. Pei, M. J. Blaser, C. F. Aliferis, and A. V. Alekseyenko, "A comprehensive evaluation of multicategory classification methods for microbiomic data," *Microbiome*, vol. 1, no. 1, p. 11, 2013.

[28] J. T. Wassan, H. Wang, F. Browne, and H. Zheng, "A comprehensive study on predicting functional role of metagenomes using machine learning methods," *IEEE/ACM transactions on computational biology and bioinformatics*, 2018.

[29] M. Kuhn, "Building predictive models in r using the caret package," *Journal of Statistical Software, Articles*, vol. 28, no. 5, pp. 1–26, 2008. [Online]. Available: https://www.jstatsoft.org/v028/i05

[30] J. D. Silverman, A. D. Washburne, S. Mukherjee, and L. A. David, "A phylogenetic transform enhances analysis of compositional microbiota data," *Elife*, vol. 6, p. e21887, 2017.

[31] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: a system for large-scale machine learning." in *OSDI*, vol. 16, 2016, pp. 265–283.

[32] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[33] I. Culjak, D. Abram, T. Pribanic, H. Dzapo, and M. Cifrek, "A brief introduction to opencv," in *MIPRO, 2012 proceedings of the 35th international convention*. IEEE, 2012, pp. 1725–1730.

[34] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *null*. IEEE, 2003, p. 958.

[35] M. Lebuhn, W. Achouak, M. Schloter, O. Berge, H. Meier, M. Barakat, A. Hartmann, and T. Heulin, "Taxonomic characterization of ochrobactrum sp. isolates from soil samples and wheat roots, and description of ochrobactrum tritici sp. nov. and ochrobactrum grignonense sp. nov." *International Journal of Systematic and Evolutionary Microbiology*, vol. 50, no. 6, pp. 2207–2223, 2000.