Improving Disease Prediction using Shallow Convolutional Neural Networks on Metagenomic Data Visualizations based on Mean-Shift Clustering Algorithm

Hai Thanh Nguyen[1]
College of Information and Communication Technology
Can Tho University
Can Tho, Vietnam

Toan Bao Tran[2]
Center of Software Engineering, Duy Tan University,
Da Nang, 550000 Vietnam
Institute of Research and Development, Duy Tan University,
Da Nang, 550000 Vietnam

Huong Hoang Luong[3]
Department of Information Technology
FPT University
Can Tho, Vietnam

Trung Phuoc Le[4]
Department of Information Technology
FPT University
Can Tho, Vietnam

Nghi C. Tran[5]
National Central University
Taoyuan, Taiwan

***Abstract*—Metagenomic data is a novel and valuable source for personalized medicine approaches to improve human health. Data Visualization is a crucial technique in data analysis to explore and find patterns in data. Especially, data resources from metagenomic often have very high dimension so humans face big challenges to understand them. In this study, we introduce a visualization method based on Mean-shift algorithm which enables us to observe high-dimensional data via images exhibiting clustered features by the clustering method. Then, these generated synthetic images are fetched into a convolutional neural network to do disease prediction tasks. The proposed method shows promising results when we evaluate the approach on four metagenomic bacterial species abundance datasets related to four diseases including Liver Cirrhosis, Colorectal Cancer, Obesity, and Type 2 Diabetes.**

Keywords*—Clustering algorithm; metagenomic; visualization; disease prediction; mean-shift; personalized medicine; species abundance; bacterial

I. Introduction

Human healthcare has been moving towards step by step to personalized medicine which using genetic insights and technologies. In 2020, the outbreaks of SARS-CoV-2 raises questions about the advantages of personalized medicine in general and metagenomics in particular. Personalized medicine also commonly referred to as precision medicine is the most promising approach for effective medical treatment of the individual patients based on their genetic information and medical symptoms. By combining with the traditional approaches which are based upon a policy of "one size fits all" applying the same treatments to whom with the same diseases, personalized medicine may be used to analyze and treat the disease by personalizing medicines to make them more specific, effective, and thereby improving treatment outcomes. Summarily, personalized medicine is a new approach in disease management, focusing on four essential premises: prediction, prevention, personalization, and participation [1]. Following the premises of personalized medicine, the appearance of SARS-CoV-2 may be explored based on acting on risk factors, cultures, and social determinants $(prediction)$, constrain on evolution of the virus $(prevention)$, analyzing the genetic and molecular of each the patient and giving them their personalizing medicines $(personalization)$, requiring the investment for the infrastructure, human resource training, and the cooperation of the patients $(participation)$ [2]. Several studies have indicated that many diseases are originally from genotypic so that personalized medicine is an effective treatment and can reduce the disadvantages of side effects. Many of the advantages of personalized medicine within healthcare detect and diagnose diseases, prevention of disease, and reduction of trial-and-error prescriptions.

Metagenomics that is the study of the metagenome, an application of modern genomic techniques, explores directly the communities of microbial in their natural habitats [3]. The emergence of high-throughput sequencing technology such as deep metagenomic sequencing has generated an amount of data that allowed the researchers to study both taxonomic and functional effects of microbiota on hosts [4]. The uncultured microorganisms represent the huge majority of organisms in most habitats on this planet proving by the analysis of 16rRNA sequences, it is the beginning for the development of metagenomics and led to the discovery of vast new lineages of microbial life [4], [5], [6]. The importance of understanding the microbiome has been repeatedly emphasized, thousands of human microbiome projects that have focused on the bacterial cell structure of the microbiome. The metagenomic analysis revealed variations in niche-specific abundance within the microbiome. Several studies presented the advantages of metagenomics in diagnostics and evidence-based medicine. Analyzing of Big data play a specific role in determining the causality of clinical diseases by bacteria and treating by

a suitable medicine. Therefore, in the Personalized Medicine field, metagenomics is an efficient tool to deal with numerous pressing issues and the relatives [7], [8].

II. Related Work

Metagenomic analysis has become an exciting subject for the scientific community, the primary effort on the analysis of the microbiome is the identification of microbial communities for disease or host phenotype prediction [9], [10].

Diagnostic metagenomics can be used to identify pathogens on clinical samples, outbreaks of disease or novel variant viruses. Recently, the first genome sequence of SARS-CoV-2 was conducted with metagenomic RNA sequencing, an unbiased and high-throughput method of sequencing multiple genomes [8]. As an indicator of the benefits and problems of broad screens in clinical microbiology, the well-developed blood culture contamination literature has numerous researches to conduct clinical utility studies of diagnostic metagenomics, and demonstrate associations with increased hospital costs, hospitalizations, antibiotics, surgeries, and laboratory tests [11], [12], [13], [14], [15], [16].

The study in [17] proposed a method that can detect the overlapping clusters on metagenomic sequencing data by the Bayesian multinomial matrix factorization model. The authors stated under the Bayesian framework, the number of clusters is determined by the algorithm and improving the interpretability of their detection from the available information gained from a rank tree of microbes. The cluster structures are built hierarchically based on Dirichlet-multinomial mixtures with the purpose to indicate the relative abundance of taxa through a set of latent variables. By given the binary matrix, the priors are assorted hierarchically to characterize the heterogeneity via latent features. Summarily, this approach can handle the natural microbiome data and describes the generating process of data by the Bayesian model.

DeepMicrobes is described as a state-of-the-art metagenomics tool and the first deep learning architecture that incorporates self-attention mechanisms for DNA sequence analysis. *DeepMicrobes* facilitates taxonomic classification for cohorts of interest using newly discovered species in large-scale metagenomic assembly studies. The DNA sequence was encoded into numeric matrices, these are one-hot encoding as and k-mer embedding. The convolution model, hybrid convolutional, and recurrent model take DNA sequence one-hot encoding as an input layer whereas the other as the first layer of deep neural networks. For one-hot encoding, DNA was converted into $4 \times L$ matrix. For k-mer embedding, a DNA sequence of length L was split into a list of substrings of length K with a stride of S. The authors used a stride of none for their final model, ending up with $L-K+1$ substrings. The length of K was chosen to reach a balance between the model's fitting capacity and computational resources [18].

The different approach, phylogenetic tree embedded is an interesting approach for metagenomics data analysis. Essentially, the phylogenetic tree is a $2D$ matrix populated with the relative abundance of microbial taxa in a metagenomic sample, then, to be used as an input for the CNN [19]. With this method, the constructed matrices provide better spatial and quantitative information in the metagenomic data. Besides, the authors also proposed the convolutional neural networks, namely the PopPhy-CNN and Cytoscape-a visualization method used to facilitate the examination and interpretation of the retrieved taxa on the phylogenetic tree. The authors demonstrated the feasibility of extracting features can improve the performance of SVMs compared to the other models. They also indicated the conventional vector input $1D-CNN$ does not take advantage of the biological knowledge in the phylogenetic tree. The phylogenetic information was also utilized in sparse linear discriminant models with the simultaneous use of intermediate nodes and leaves on a phylogenetic tree [20].

PhyloPhlAn 3.0 is a framework for large-scale microbial genome characterization and phylogenetic analysis on a large number of features, it scales to large phylogenies comprising $> 17,000$ microbial species and assign genomes from isolate sequencing or MAGs to species-level genome bis built from $> 230,000$ publicly available sequences. Generally, this framework is to use available references genomes, retrieve the phylogenetic markers, perform taxonomic assignment and refinement, adopt specific choices for very large scale phylogenies, and provide additional information obtained from the resulting phylogenies [21].

The data is the most limitation in machine learning, many learning algorithms require large amounts of data for the training section. However, with the data augmentation method, the performance and generalization can be improved. The authors in [22] proposed an approach for generating microbiome data by using a conditional generative adversarial network (CGAN). Additionally, synthetic datasets generated using GAN models have shown to be able to boost the performance of prediction based tasks through data augmentation [23]. CGANs are an extension of the GAN and allow the generation of samples that have certain conditions or attributes. The authors in [22] have shown this approach can improve the performance of logistic regression and MultiLayer Perceptron in predicting host phenotype. They also stated the selecting CGAN model is the limitation of this approach, it is a subjective and may miss the optimal model.

The identification based on statistical analysis to detect the different abundant taxa between disease. The authors in [24] presented a new deep learning approach, namely PopPhy-CNN, a novel convolutional neural networks (CNN) learning architecture that effectively exploits phylogenetic structure in microbial taxa. The microbial taxonomic abundance profiles have been transformed into a structured data by using a phylogenetic tree, their approach is using "Operational Taxonomic Units" (OTUs) then converting OTU vector into an input matrix for their model. OTUs are generated by clustering sequences according to a computed distance between two similar sequences and a threshold. OTUs clustering can produce high quality groups precisely due to amplicon sequences are by definition taxa-specific and different between species [25]. The clustering performance depends on the choice of threshold due to sequencing errors. Furthermore, the analysis and biologically meaningful can be problematic [26].

In this study, we present a metagenomic data visualization-based Mean-Shift algorithm to cluster features in images prepared for prediction tasks, the contributions include:

- We present a features clustering approach with Mean-Shift algorithm and compare to the other visualization methods including Fill-up with phylogenetic ordering [27] and t-Distributed Stochastic Neighbor Embedding (t-SNE) [27], [28].
- The efficient of the proposed visualization methods is evaluated on four diseases including Liver Cirrhosis (CIR), Colorectal Cancer (COL), Obesity (OBE), Type 2 diabetes (WT2) [9], [27]. The performance on the datasets with the considered diseases obtains better results in prediction tasks comparing to the-state-of-the-art such as MetAML [9], Fill-up with phylogenetic ordering and t-SNE using transparent rates with $alpha = 0.5$ and $alpha = 1$.
- We also test visualizations with a vast of colormaps including jet, rainbow, gray and customized colormaps. These color spaces exhibit various results. Color images perform the best on Liver Cirrhosis dataset and samples of Colorectal Cancer while gray scale reveals good results for Obesity and Type 2 Diabetes samples.

The remaining of this study, we introduce the visualization approaches for metagenomic data including Fill-up approach and Fill-up with Mean-Shift clustering algorithm for arranging features in Section III. In Section IV, metagenomic bacterial species abundance datasets used in the experiments are described in detail. Moreover, we present the Convolutional Neural architecture for the proposed visualization method and settings for the learning. The performance of our approach is compared to the state-of-the-art in this section. We discuss and summarize the experimental results in Section V.

III. Features Arrangement based on Mean-shift Clustering in Fill-up method

Data visualization is a strong method to interpret data. Each visualization method will be used to represent the abundance or presence of data. In this study, we propose a visualization method based on Mean-shift clustering algorithm to arrange features in images, thereby making it easier for observing the distribution of the features. Therefore, we expect to improve the performance of disease prediction task with the proposed visualizations.

A. Metagenomic Visualization by Fill up Approach

Fill-up [27] is an effective solution for visualizing metagenomic data into images. The main idea of this method is to arrange features into a square matrix which has minimum size to fit all features and contains arranged abundance or presence values in a right-to-left order by row top-to-bottom. The order to arrange species can follow the phylogenetic-sorting or another type of ordering.

t-SNE is also a technique for visualizing metagenomic data. t-SNE not only captures the local structure of the higher dimension but also preserves the global structures of the data like clusters.

In order to convert continuous values into discrete values (for coloring features on images), we use a binning technique. Binning is a data pre-processing method, the key goal is to reduce the effects of minor observation errors, it has a smoothing effect on the input data and may also reduce the chances of overfitting in case of small datasets. In this study, Species Bin (SPB) [27] is implemented and investigated to pre-process values before visualizing them onto an image. With the binning technique, the features were visualized into images by Fill-up with phylogenetic-sorting or t-Distributed Stochastic Neighbor Embedding (t-SNE) in [27].

B. Mean-shift Clustering in Fill-up Method

Algorithm 1 Algorithm for features clustering based on Mean-shift algorithms

Input:

- D: a data matrix where each row is a sample and each column represents a feature

Output:

- B: an array containing a list of strings combining between the labels of generated clusters and order of features sorted by phylogenetic ordering.

Begin

Step 1: Sort D so that features along with their data follow phylogenetic ordering. Save the list containing the order of features according to phylogenetic ordering to P.

Step 2: Transpose D: $D_1 = t(D)$. Because we want to group features into clusters, we transpose D so that each feature at this time is considered as a "data point" for clustering.

Step 3: Run Mean-shift clustering algorithm on D_1 to indicate clusters for features. Each feature is assigned to a cluster. A cluster can contain one or more features.

. The labels of clusters contained features are saved to L. L includes information on clusters which each feature belongs to. For example, the 1st feature belongs to cluster 5, the 2nd feature is labeled to cluster 1, and so on.

Step 4:
. We concatenate labels of clusters for features L and their phylogenetic ordering:

$$B[i] = string(L[i]) +' \_' + string(P(i))$$
$$With : i = 0..\#features$$

Return B

End

Mean-shift [29] is an unsupervised learning algorithm. In principle, the algorithm iteratively assigns each data point towards the closest cluster centroid and direction to the closest cluster centroid is determined by where most of the points nearby are at. So each iteration each data point will move closer to where the most points are at, which is or will lead to the cluster center. When the algorithm stops, each point is assigned to a cluster. Assume we have:

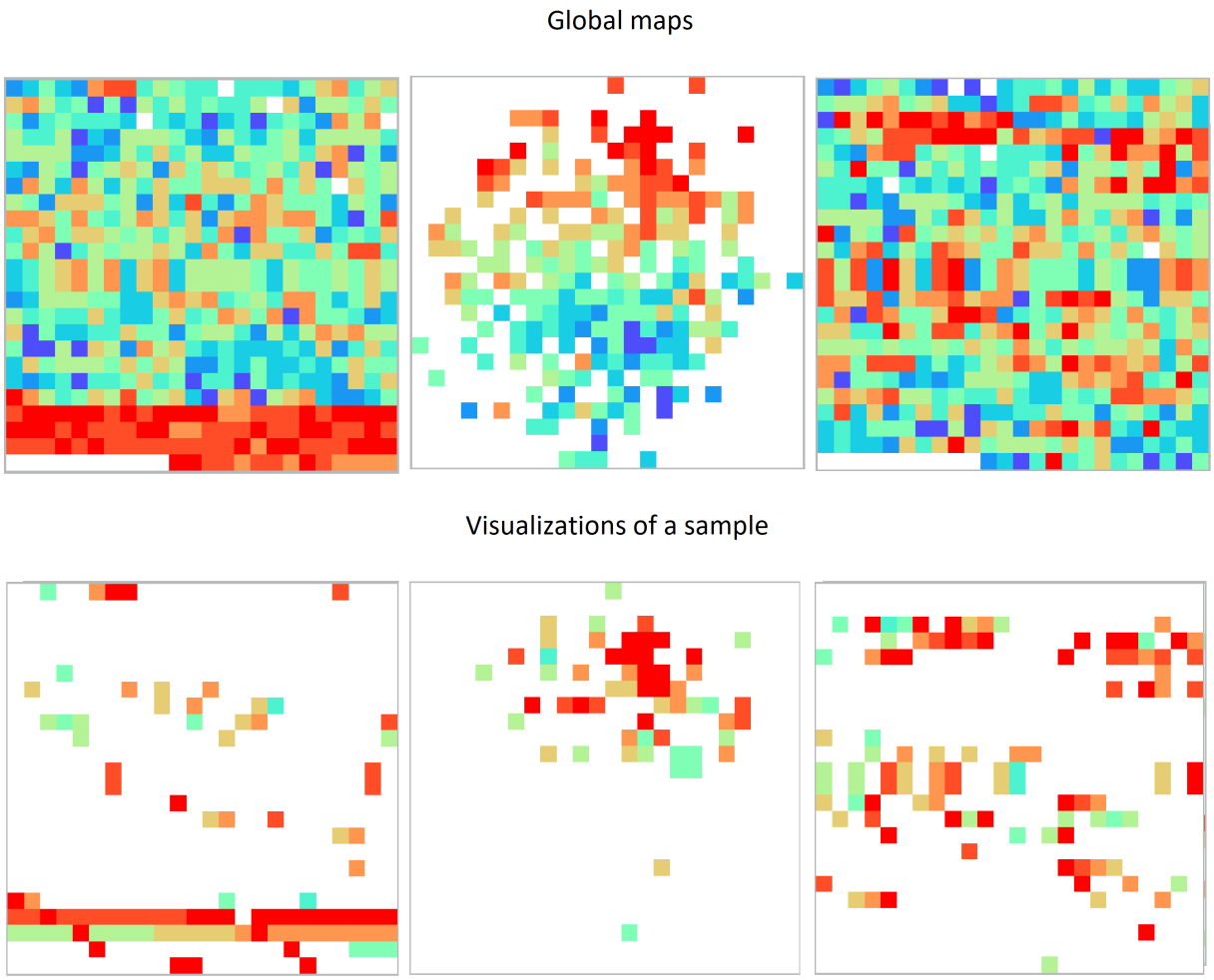

Fig. 1. Liver Cirrhosis samples (CIR dataset, details in Table I) Visualization comparison using various features arrangements including (left-to-right) Mean-shift, t-SNE, and features ordered based on phylogenetic information. The first row: global images. The second row: visualizations of a sample.

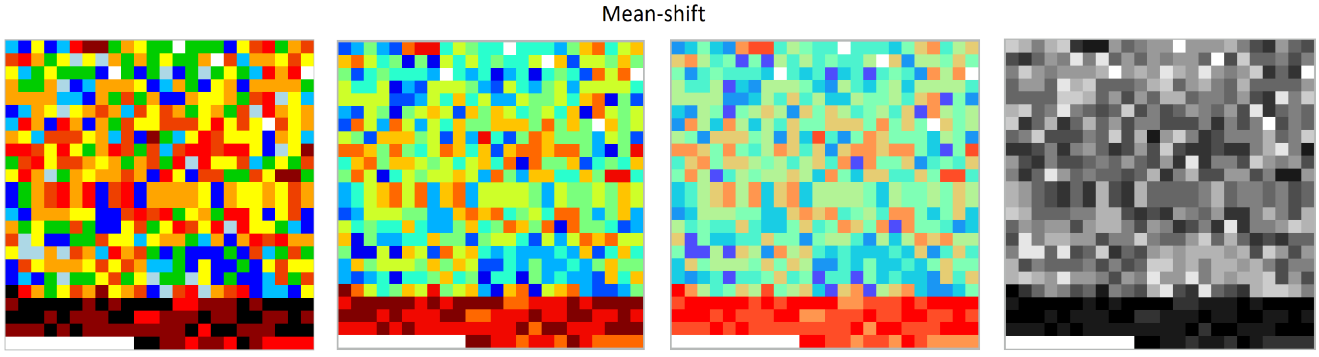

Fig. 2. Visualization of the global maps from Liver Cirrhosis samples (CIR dataset) using various color spaces including custom, jet, rainbow and gray scale with Mean-shift clustering.

- Initial estimate x.
- Gaussian kernel function:

$$K(x_i - x) = e^{-c\|\mathbf{x_i} - \mathbf{x}^2\|}$$

This function determines the weight of nearby points for re-estimation of the mean.

The weighted mean of the density in the window determined by K is:

$$m(x) = \frac{\sum_{x_i \in N_{(x)}} K(x_i - x) x_i}{\sum_{x_i \in N_{(x)}} K(x_i - x)}$$

Where:

- $N_{(x)}$ is the neighborhood of x.

$m(x) - x$ is called mean shift [29] and $x \leftarrow m(x)$, and repeats the estimation until $m(x)$ converges.

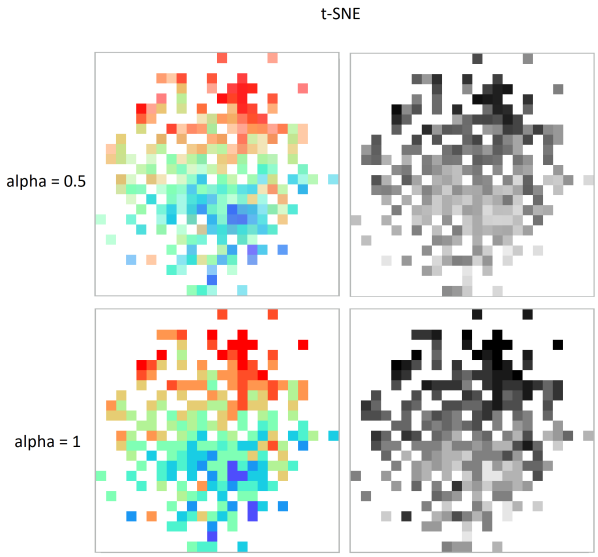

Fig. 3. Visualization comparison between rainbow colormap and gray scale images using t-SNE on Liver Cirrhosis samples. Top: t-SNE with $alpha = 0.5$. Bottom: t-SNE with $alpha = 1$.

The synthetic metagenomic images are generated by Fill-up and t-SNE method in [27]. In this study, use of Mean-shift algorithm, we expect to improve the performance by finding regions containing a high density of data and group them into a cluster with smallest non-overlapping boundaries. This approach is performed as shown in Algorithm 1. After clustering, we obtain an array B which contains the arranged features order by the labeled clusters along with information on phylogenetic ordering. Information on phylogenetic embedded in synthetic metagenomic images is based on the alphabetical order as described in [27]. In our method, if features are in the same cluster, we consider the alphabetical order of features to place them close together. By combining between order of features sorted by cluster labels and phylogenetic ordering, we expect to improve the quality of visualizations as well as the prediction performance of deep learning algorithm on the proposed visualizations.

In order to visualize features, we use 10 colors in gray-scale, rainbow, jet, and custom colormap [27]. In Fig. 1 displays the comparison between clustered features on global and sample images from Liver Cirrhosis samples (CIR dataset, see details in Table I) based on mentioned visualization methods in rainbow colormap. The global map which is an image visualizing average value of each feature of all samples in training set. From left-to-right and top-to-bottom, the first two images in Fig. 1 shows the global and sample image visualized by Fill-up combining the clustering method, in the middle contains images represent the global map and a sample visualization of t-SNE embedding. The last ones are visualized by Fill-up with phylogenetic ordering. We only use samples from training set to cluster features and build coordinates for all features. These coordinates are carried out to generate all images for samples of both training set and test sets.

Fig. 2 illustrates the representation of clustered features in different colors. The images in Fig. 2 are global images from CIR dataset which mentioned above with Fill-up and clustering method, from left-to-right custom, jet, rainbow, and gray colormaps. More specific, the custom colormap is built based on jet combined to black with distinctive colors. Furthermore, we also visualize the global images with t-SNE exhibited in Fig. 3, the images on the top are generated by t-SNE with $alpha = 0.5$ while the second row shows the images with $alpha = 1$. The first column, we use rainbow colormap while gray scale is applied for images in the other. The difference between t-SNE with $alpha = 0.5$ and $alpha = 1$ is the problem of the overlapped points. t-SNE suffers overlapped issues where the visualization exists numerous points which are hidden by other points. In order to reduce this negative affect, the alpha value is deployed in the RGBA color space to indicate the transparency of a colour. The alpha value ranges from 0 to 1 where 0 is completely transparent while alpha value of 1 is not transparent at all. By choosing $alpha = 0.5$, the futures are mixed-up if they are overlapped. Otherwise, with $alpha = 1$, some features can be hidden by other features.

IV. Experimental Results

A. Benchmark Datasets

We evaluated our approach on four bacterial species abundance datasets [9], [27] which are related to four diseases including Liver Cirrhosis (CIR), Colorectal Cancer (COL), Obesity (OBE), and Type 2 diabetes samples from western women (WT2). Details are in Table I. For each sample, species abundance (feature) is represented as a real number and the total abundance of all species in a sample sums to 1:

$$\sum_{i=1}^{k} f_i = 1$$

With:

- k is the number of features for a sample.
- f_i is the value of the i-th feature.

Table I presents the details of all considered datasets including the numbers of features, samples, and some extra

TABLE I. FOUR CONSIDERED BACTERIAL SPECIES ABUNDANCE DATASETS DESCRIPTION

Diseases	**Liver Cirrhosis**	**Colorectal Cancer**	**Obesity**	**Type 2 diabetes**
Datasets name	**CIR**	**COL**	**OBE**	**WT2**
#Features	542	503	465	381
#Samples	232	121	253	96
#Patients	118	48	164	53
#Controls (healthy)	114	73	89	43
Ratio of patients	0.51	0.40	0.65	0.55
Ratio of Controls (healthy)	0.49	0.60	0.35	0.45
Minimum size of images	24×24	23×23	22×22	20×20

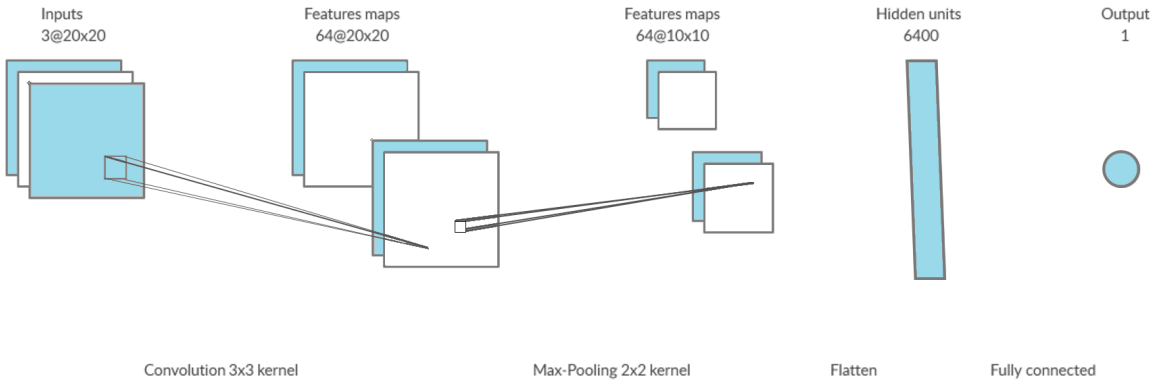

Fig. 4. A shallow convolutional Neural Network Architecture for metagenomic images on color images of WT2 samples.

Fig. 5. Performance Comparison of different colormaps on all considered metagenomic datasets using Mean-shift for features arrangement in metagenomic visualization

information. We calculate the ceiling of Square Root of the numbers and then of features to feed into a 2D matrix. For

instant, on CIR dataset we have 542 features, so the 2D matrix shape should be 24×24 to contain 542 features because $\sqrt{542} = 23.28$ and the ceiling of 23.28 is 24.

B. Learning Model and Settings

Our classification tasks are carried out by a shallow deep learning network, a Convolutional Neural Network (CNN) as illustrated in Fig. 4. The architecture contains one Convolutional layer with 64 kernels of 3×3, followed by a Max-Pooling layer of 2×2 (stride 2) and a Fully Connected layer. CNN is implemented with Adam optimizer, the default learning rate is 0.001, and the network uses a batch of size 16. The architecture is suggested from [27]. To avoid overfitting issues, if the $loss$ is not improved after every consecutive 5 epochs, we will stop the training section by using the Early Stopping method. In the opposite case, training can run up to 500 epochs. To evaluate the performance, we compute average accuracy (ACC) on 10-fold-cross-validation. The same folds are used for all classifiers. We calculate the accuracy which is the fraction of true predictions by the following formula:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Where:

- TP: True Positive
- TN: True Negative
- FP: False Positive
- FN: False Negative

C. Disease Classification of Mean-shift Clustering with Fill up on Various Diseases

The efficiency of arranging features based on Mean-shift is evaluated in various colormaps, namely gray, jet, rainbow, and custom. The last one is combined between black and jet colormap. Fig. 5 illustrates the average accuracy of the methods on four considered datasets. Generally, each colormap gives a satisfying result each individual dataset. The the jet colormap exhibits a quite good and reaches the highest average performance on four datasets while gray scale works well on OBE and WT2 and the rainbow achieves the highest performance on COL while the custom colormap gives exceptional results on CIR with the performances of 0.926.

D. State-of-the-art Comparison

The performance comparison of Mean-shift clustering, t-SNE, and phylogenetic ordering [27] are illustrated in Fig. 6 and Fig. 7. The chart in Fig. 7 reveals the accuracy on four considered datasets using rainbow colormap while the results with gray images are shown in the other. As observed, the in Fig. 6 features arrangements based on Mean-shift clustering demonstrates its advantages on 3 out of 4 datasets using both rainbow in comparing to phylogenetic ordering.

Furthermore, we also summarize the result with results of the jet and custom colormaps, and compare to MetAML [9], a computation framework for metagenomic analysis based on classic machine learning algorithms such as Random Forests and Support Vector Machines in Table II. On CIR dataset, Mean-shift clustering method reaches the accuracy of 0.926

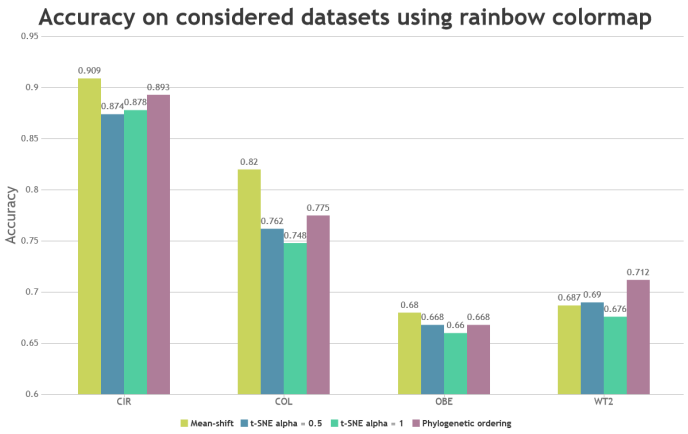

Fig. 6. Performance of different visualization approaches using rainbow colormap on four considered datasets (details in Table II).

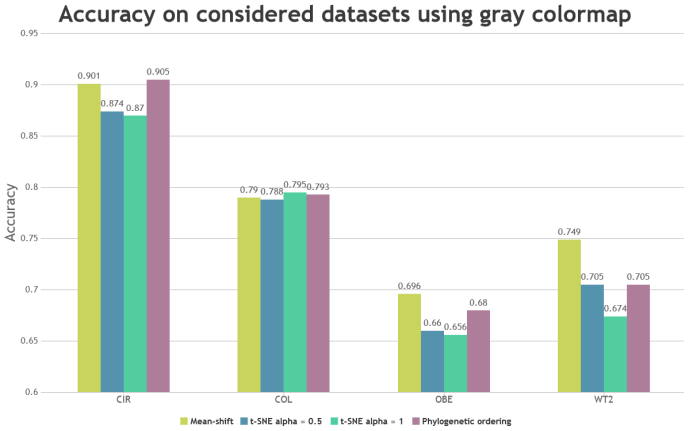

Fig. 7. Visualization methods Comparison in ACC on the considered datasets using gray scale images (details revealed in Table II).

while MetAML, t-SNE ($alpha = 1$) and Fill-up using phylogenetic ordering reveal the results of 0.877, 0.853 and 0.897 respectively. The color images which are jet, rainbow, and custom give quite better results than gray images on Liver Cirrhosis and Colorectal Cancer samples while the results with gray scale are slight better for OBE and WT2 datasets.

We also compute the average accuracy on four investigated datasets for the comparison in the last column of Table II. In general, as shown in the table, visualization methods with Fill-up based on Mean-shift clustering algorithm (including average values of 0.771, 0.777, 0.784, 0.788 with customized, rainbow, gray scale, and jet colormaps, respectively) outperform MetAML, t-SNE and Fill-up using phylogenetic ordering with the values of 0.757 of MetAML, 0.774 and 0.741 being the best results of Fill-up with phylogenetic ordering and t-SNE, respectively. Jet colormap appears to be the most efficiency while custom colormap with customized distinctive colors shows the worst among the considered color spaces. However, we noted that the best accuracy is on CIR dataset with an average accuracy of 0.926 using custom colormap.

TABLE II. COMPARISON WITH THE-STATE-OF-THE-ART. **BOLD RESULTS** ARE BETTER PERFORMANCE THAN THE METHOD OF FILL-UP WITH PHYLOGENETIC ORDERING.

Approaches	Color space	CIR	COL	OBE	WT2	AVG
MetAML [9]	-	0.877	0.805	0.644	0.703	0.757
t-SNE with $alpha = 1$ [27]	gray	0.870	0.795	0.656	0.674	0.749
Fill-up phylogenetic ordering [27]	gray	0.905	0.793	0.680	0.705	0.770
Our approach	gray	0.901	0.790	**0.696**	**0.749**	**0.784**
t-SNE with $alpha = 1$ [27]	jet	0.879	0.748	0.661	0.660	0.737
Fill-up phylogenetic ordering [27]	jet	0.903	0.798	0.681	0.713	0.774
Our approach	jet	**0.913**	**0.799**	**0.695**	**0.745**	**0.788**
t-SNE with $alpha = 1$ [27]	rainbow	0.878	0.748	0.660	0.676	0.741
Fill-up phylogenetic ordering [27]	rainbow	0.893	0.775	0.668	0.712	0.762
Our approach	rainbow	**0.909**	**0.820**	**0.690**	0.687	**0.777**
t-SNE with $alpha = 1$ [27]	custom	0.853	0.771	0.660	0.661	0.736
Fill-up phylogenetic ordering [27]	custom	0.897	0.782	0.673	0.707	0.765
Our approach	custom	**0.926**	**0.791**	0.656	**0.712**	**0.771**

V. DISCUSSION AND CONCLUSION

We presented an approach to visualize high-dimensional data using features arrangement with Mean-shift and compare the results to the state-of-the-art. The method reveals encouraging results. We obtain better results on all considered datasets compared to Fill-up with phylogenetic ordering and t-SNE images classified by deep learning algorithm and MetAML with a classic machine learning algorithm. As seen from the experiments, features which are clustered to arrange close together show benefits to improve the performance both in visualizations and in classification tasks. Although t-SNE also groups similar features, it suffers the issue of overlapped points. However, for gray images on Colorectal cancer samples, t-SNE achieves a slightly better result comparing to others. Further research can work on t-SNE to investigate approaches to enhance performance.

Various colormaps are carried out to compare different methods. The results depend on different datasets for the classification tasks. Customised colors obtain the highest average accuracy with 0.926 on CIR dataset while it shows only an average accuracy of 0.656 on OBE dataset. It is clear that visualization methods are good solutions for Liver Cirrhosis, Colorectal Cancer prediction but Predicting Obesity and Type 2 diabetes is still facing challenges with metagenomic data. However, the performance on Cirrhosis samples and Colorectal cancer samples also reveal great potentials of metagenomic in disease prediction with personalized medicine.

Our study only runs the classification tasks with shallow deep learning architectures. Advancements in deep learning techniques have been increasing their efficiency on numerous fields. In the future, further research should investigate on deeper architectures and more sophisticated techniques to improve the performance on synthetic metagenomic visualizations classification tasks.

REFERENCES

[1] Sagner M, McNeil A, Puska P, Auffray C, Price ND, Hood L, et al. The P4 health spectrum - a predictive, preventive, personalized and participatory continuum for promoting healthspan. Prog Cardiovasc Dis. 2017;59:506–521. doi: 10.1016/j.pcad.2016.08.002. 2016.

[2] Crisci, Carlos D et al. “A Precision Medicine Approach to SARS-CoV-2 Pandemic Management.” Current treatment options in allergy, 1-19. 8 May. 2020, doi:10.1007/s40521-020-00258-8. 2020.

[3] Chen K, Pachter L. Bioinformatics for whole-genome shotgun sequencing of microbial communities. PLoS Comput Biol. 2005;1(2):106-112. doi:10.1371/journal.pcbi.0010024

[4] Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev. 2004;68(4):669-685. doi:10.1128/MMBR.68.4.669-685.2004

[5] Ma, Bing & France, Michael & Ravel, Jacques. (2020). Meta-Pangenome: At the Crossroad of Pangenomics and Metagenomics. doi10.1007/978-3-030-38281-0_9.

[6] Jang SJ, Ho PT, Jun SY, Kim D, Won YJ. Dataset supporting description of the new mussel species of genus Gigantidas (Bivalvia: Mytilidae) and metagenomic data of bacterial community in the host mussel gill tissue. Data Brief. 2020;30:105651. Published 2020 Apr 29. doi:10.1016/j.dib.2020.105651. 2020

[7] Alfredo D. Guerron et al. “Performance and Improvement of the DiaRem Score in Diabetes Remission Prediction - A Study with Diverse Procedure Types”, May. 2020, doi:https://doi.org/10.1016/j.soard.2020.05.010. 2020.

[8] Hongyu Chen, Sanjeev Kumar Awasthi, Tao Liu, Zengqiang Zhang. Mukesh Kumar Awasthi, “An assessment of the functional enzymes and corresponding genes in chicken manure and wheat straw composted with addition of clay via meta-genomic analysis”, Industrial Crops and Products, vol. 153, 2020, doi:https://doi.org/10.1016/j.indcrop.2020.112573

[9] Pasolli et al. “Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights”. PLoS Comput. Biol. 2016;12(7):e1004977. Published 2016 Jul 11. doi:10.1371/journal.pcbi.1004977. 2016.

[10] Syed Hamid Jalal Shaha, Aamir Humayun Malik, Bing Zhang, Yiming Bao, Javaria Qazi, “Metagenomic analysis of relative abundance and diversity of bacterial microbiota in Bemisia tabaci infesting cotton crop in Pakistan”, May 2020, doi:https://doi.org/10.1016/j.meegid.2020.104381

[11] Hasman, Henrik et al. “Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples.” Journal of clinical microbiology vol. 52,1 (2014): 139-46. doi:10.1128/JCM.02452-13. 2014.

[12] Self WH, Speroff T, Grijalva CG, et al. Reducing blood culture contamination in the emergency department: an interrupted time series quality improvement study. Acad Emerg Med. 2013;20(1):89-97. doi:10.1111/acem.12057

[13] Hall KK, Lyman JA. Updated review of blood culture contamination. Clin Microbiol Rev. 2006;19(4):788-802. doi:10.1128/CMR.00062-05

[14] Gander RM, Byrd L, DeCrescenzo M, et al. Impact of blood cultures drawn by phlebotomy on contamination rates and health care costs in

a hospital emergency department. J Clin Microbiol. 2009;47:1021–1024. 2009. DOI:10.1128/JCM.02162-08

[15] Bates DW, Goldman L, Lee TH. Contaminant blood cultures and resource utilization. The true consequences of false-positive results. JAMA. 1991;265(3):365-369.

[16] van der Heijden YF, Miller G, Wright PW, Shepherd BE, Daniels TL, Talbot TR. Clinical impact of blood cultures contaminated with coagulase-negative staphylococci at an academic medical center. Infect Control Hosp Epidemiol. 2011;32(6):623-625. doi:10.1086/660096

[17] Fangting Zhou, Kejun He, Qiwei Li, Robert S. Chapkin, Yang Ni, "Bayesian biclustering for microbial metagenomic sequencing data via multinomial matrix factorization," arXiv:2005.08361, May 2020

[18] Qiaoxing Liang, Paul W Bible, Yu Liu, Bin Zou, Lai Wei, DeepMicrobes: taxonomic classification for metagenomics with deep learning, NAR Genomics and Bioinformatics, Volume 2, Issue 1, March 2020, lqaa009, https://doi.org/10.1093/nargab/lqaa009

[19] D. Reiman, A. Metwally, J. Sun and Y. Dai, "PopPhy-CNN: A Phylogenetic Tree Embedded Architecture for Convolutional Neural Networks to Predict Host Phenotype from Metagenomic Data," in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2020.2993761. 2020.

[20] Fukuyama J, Rumker L, Sankaran K, et al. Multidomain analyses of a longitudinal human microbiome intestinal cleanout perturbation experiment. PLoS Comput Biol. 2017;13(8):e1005706. Published 2017 Aug 18. doi:10.1371/journal.pcbi.1005706. 2017.

[21] Asnicar, F., Thomas, A.M., Beghini, F. et al. Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. Nat Commun 11, 2500 (2020). https://doi.org/10.1038/s41467-020-16366-7. 2020.

[22] Reiman, Derek and Dai, Yang, "Using Conditional Generative Adversarial Networks to Boost the Performance of Machine Learning in Microbiome Datasets," bioXiv:2020.05.18.102814, https://doi.org/10.1101/2020.05.18.102814, May 2020.

[23] Che, Z., Cheng, Y., Zhai, S., Sun, Z., & Liu, Y. (2017). Boosting Deep Learning Risk Prediction with Generative Adversarial Networks for Electronic Health Records. 2017 IEEE International Conference on Data Mining (ICDM), 787-792.

[24] D. Reiman, A. Metwally, J. Sun and Y. Dai, "PopPhy-CNN: A Phylogenetic Tree Embedded Architecture for Convolutional Neural Networks to Predict Host Phenotype from Metagenomic Data," in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2020.2993761. 2020.

[25] Soueidan, Hayssam and Macha Nikolski. "Machine learning for metagenomics: methods and tools." arXiv: Genomics (2015): n. pag.

[26] (2014). Molecular Markers in Phylogenetic Studies-A Review. Journal of Phylogenetics & Evolutionary Biology. 02. 10.4172/2329-9002.1000131.

[27] Thanh Hai Nguyen, Edi Prifti, Nataliya Sokolovska, Jean-Daniel Zucker. Disease Prediction using Synthetic Image Representations of Metagenomic data and Convolutional Neural Networks. The 13th IEEE-RIVF International Conference on Computing and Communication Technologies 2019, Da Nang 20-22/03/2019; pp 231-236; 2019; ISBN 978-1-5386-9313-1. IEEE Xplore. 2019.

[28] Maaten, Laurens van der and Geoffrey E. Hinton. "Visualizing Data using t-SNE." (2008).

[29] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," in IEEE Transactions on Information Theory, vol. 21, no. 1, pp. 32-40, January 1975, doi: 10.1109/TIT.1975.1055330.