Binning Approach based on Classical Clustering for Type 2 Diabetes Diagnosis

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

Nhi Yen Kim Phan[2]
College of Information and Communication Technology
Can Tho University
Can Tho, Vietnam

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

Nga Hong Cao[4]
Department of Computer Science and Information Engineering
National Central University
Taiwan

Hiep Xuan Huynh[5,*]
College of Information and Communication Technology
Can Tho University
Can Tho, Vietnam
*Corresponding Author

***Abstract*—In recent years, numerous studies have been focusing on metagenomic data to improve the ability of human disease prediction. Although we face the complexity of disease, some proposed frameworks reveal promising performances in using metagenomic data to predict disease. Type 2 diabetes (T2D) diagnosis by metagenomic data is one of the challenging tasks compared to other diseases. The prediction performances for T2D usually reveal poor results which are around 65% in accuracy in state-of-the-art. In this study, we propose a method combining K-means clustering algorithm and unsupervised binning approaches to improve the performance in metagenome-based disease prediction. We illustrate by experiments on metagenomic datasets related to Type 2 Diabetes that the proposed method embedded clusters generated by K-means allows to increase the performance in prediction accuracy reaching approximately or more than 70%.**

***Keywords*—Unsupervised binning; K-means clustering algorithm; metagenomics; metagenome-based disease prediction; Type 2 diabetes diagnosis**

I. Introduction

Metagenomics (Environmental Genomics, Ecogenomics or Community Genomics) is the study of genetic material recovered directly from environmental samples. Metagenomics is directly the study of communities of microbial organisms in their natural environments by applying modern genomic techniques that pass the need for isolation and lab cultivation of individual species [1], [2], [3], [4], [5], [6]. Reassembly of multiple genomes has provided insight into energy and nutrient cycling within the community, genome structure, gene function, population genetics and microheterogeneity, and lateral gene transfer among members of an uncultured community. The application of metagenomic sequence information will facilitate the design of better culturing strategies to link genomic analysis with pure culture studies. Why do we study metagenomics? As in [2] mentioned that Metagenomics has brought us discovery of novel natural products, new antibiotica, new molecules with new functions, new enzymes and bioactive molecules, what is a genome or species, diversity of life, interplay between human and microbes, how do microbial communities work and how stable are they, holistic view on biology. Metagenomics cloned specific gene sequences (usually 16S rRNA genes) to conduct data on the biodiversity of environmental samples. With traditional genetic and microbiological studies of genomes sequencing of microorganisms based on cultured lineage samples, it was found that it would be impossible to biodiversity of microorganisms. Therefore, metagenomics plays an important role in helping humans discover microbial diversity. In medicine, the microbial community plays a very important role in protecting human health. Therefore, the purpose of metagenomics is to understand the composition and activity of complex microbial groups in environmental samples through analysis of their DNA sequences. On the other hand, there are numerous data on multiple genomes that we can carry out a series of gene isolation projects depending on the purpose of the research.

Metagenomic is an improved method compared to traditional microbiology, the research of metagenomes obtained from genetic material from first samples, without the need for laboratory cultures. This method is commonly used on the human intestine because it is the place where the digestive process, metabolism and has 10 times the total number of cells of the body. Based on metagenomics, we can develop algorithms to predict disease, determine a patient's sensitivity and then offer reasonable treatments. However, the disease is complicated in diagnosis and prognosis and we only have a limited amount of data to observe.

Type 2 diabetes (T2D) is a heterogeneous metabolic disorder that damages many organs of the body. The disease tends to increase due to the influence of modern life, bad living habits. Nowadays, the prediction is not highly accurate and the treatment is commonly applied to patients diagnosed with some similar manifestations. With that treatment, we find that genetic diversity has not been effectively applied, leading to an improvement in the health of some patients. The performances on models for predicting T2D usually yield poor results.

II. Related Work

As mentioned above, metagenomics is an approach that utilizes extraction of genomic information directly from the environmental sample. So that, genetic information samples are more representative for a given environment and supplies a better insight into microbial environmental and metabolic diversity. By using next-generation sequencing in metagenomics project to determine genetic potential in microbial communities from a wealth of environmental niches, including those linked with human body and relative with human healthcare. Human microbiome in health and disease plays a significant role that has recently been given considerable observation [7], and distinct diseases have been associated with gut microbiota [7], [8], [9], [10], [11], [12], [13], [14], [15]. With respect to, experience 's Maja and et al [8] that a bias in codon usage present throughout the entire microbial community by applying definitions of translational optimization through codon usage adaptation on completely metagenomic datasets. They can be used as a powerful analytical tool for predicting community lifestyle-specific metabolism. Moreover, Maja and et al demonstrate this approach combined with machine learning, to classify microbiome samples in human gut according to the pathological condition diagnosed in the human host. In addition, predicting disease-relevant features in microbial gut metagenomes by using the principle of utilizing the prokaryotic translational optimization effect combined with the machine learning based classification and enriched gene datasets that explore a supportive method to analyzing metagenomic datasets. Authors in [8], [16] proposed methods using machine learning and deep learning to do disease prediction tasks and obtained promising results.

K-means clustering is an unsupervised learning algorithm. From the input data without the label to be clustered and the number of clusters to be divided, we will use the algorithm to divide the data into clusters of similar properties. Applications of clustering algorithms have been used commonly to resolve data clustering. Based on clustering methods, we can obtain a meaningful intuition of the structure of the data. Moreover, we can use "Cluster-then-predict". That means, we observe generated clusters, then different models will be built for various subgroups if there exists a wide variation in the behaviors of a variety of subgroups. Numerous studies in biological computation tasks have been applying k-mean to do specific analyses. Authors in [17] used k-mean to process Microarray data for bioinformatics tasks. [18] also implemented k-mean to cluster biological sequences by first converting them into an intermediate binary format where Hamming distance is used as the metric of comparison. The research in [19] presented enhanced k-mean to do Bioinformatics Data Clustering. In 2019, a study [20] introduce a modified sparse K-means clustering method to detect risk genes involved with Type II Diabetes Mellitus. From some previous results, we can see potential benefits to leverage k-mean in bioinformatics tasks.

In recent years, the application of machine learning algorithms to study metagenomic has become popular and the accuracy of diagnosis has been improved over time. In this article, we propose the application of the K-means clustering algorithm in the binning approach to improve the accurate results in predicting T2D. We leverage k-mean clustering as a tool to support binning data. By identifying clusters which can exist in the data, we hope to improve the performance via using a binning approach. Our study's contribution is multi-fold:

- We present results of various binning approaches on Type Diabetes disease using metagenomic data which appear as a very big challenge for diagnosis.
- The work aims to illustrate a potential advantage of using clustering algorithms to identify breaks for binning approaches to obtain a better result in T2D prediction compared to other binning methods.
- The results reveal high performances of state-of-the-art in deep learning algorithms, the Convolutional neural network, compared to traditional neural networks such as Multi-Layer Perceptron. Convolutional Neural networks can work efficiently even on one-dimensional data.
- Most cases, machine learning outperforms deep learning algorithms. For numeric data formed in 1D, classical machine learning reveals a robust prediction ability.
- Previous studies have not investigated the efficiency of classic machine learning with binning approaches. Our study proves by using Random Forest that it is possible be the best choice to select machine learning combining approaches to improve prediction performance on numeric species abundance datasets.

The remaining of this study, we present a short description of two considered T2D datasets in Section III. Furthermore, methods which we choose will be introduced in Section IV. Experimental Results of our proposed methods in this paper are illustrated in Section V. Finally, Section VI and Section VII discuss the results and summarize important remarks for this research.

III. Data Benchmarks for Metagenomic Analysis

We run the experiments on metagenomic abundance data that indicates how present (or absent) is an OTU (Operational taxonomic unit) in human gut. The abundance datasets are obtained using default parameters of MetaPhlAn2 described as detailed in [14].

A little more detail of the process of generating abundance shown in Fig. 1, the stool sample collected from human is fetched into machines to extract total Deoxyribo Nucleic acid (DNA). DNA then is sequenced to create millions of reads. The new generation sequencing techniques can process millions of sequencing reads in parallel. These reads are mapped to a catalog of references including all known gut microbial genes and known bacterial at levels of species, genus and so on. The techniques also indicate the presence and abundance of each gene and each species in any samples. As revealed in numerous studies, species abundance and genes abundance can distinguish patients and healthy controls. Moreover, genes and species can be leveraged to develop robust tools for diagnosis and prognosis.

We evaluated our approach on the disease of Type 2 Diabetes with two datasets. The first one (T2D1) includes 344 Chinese individuals [22], and 96 western women are in other

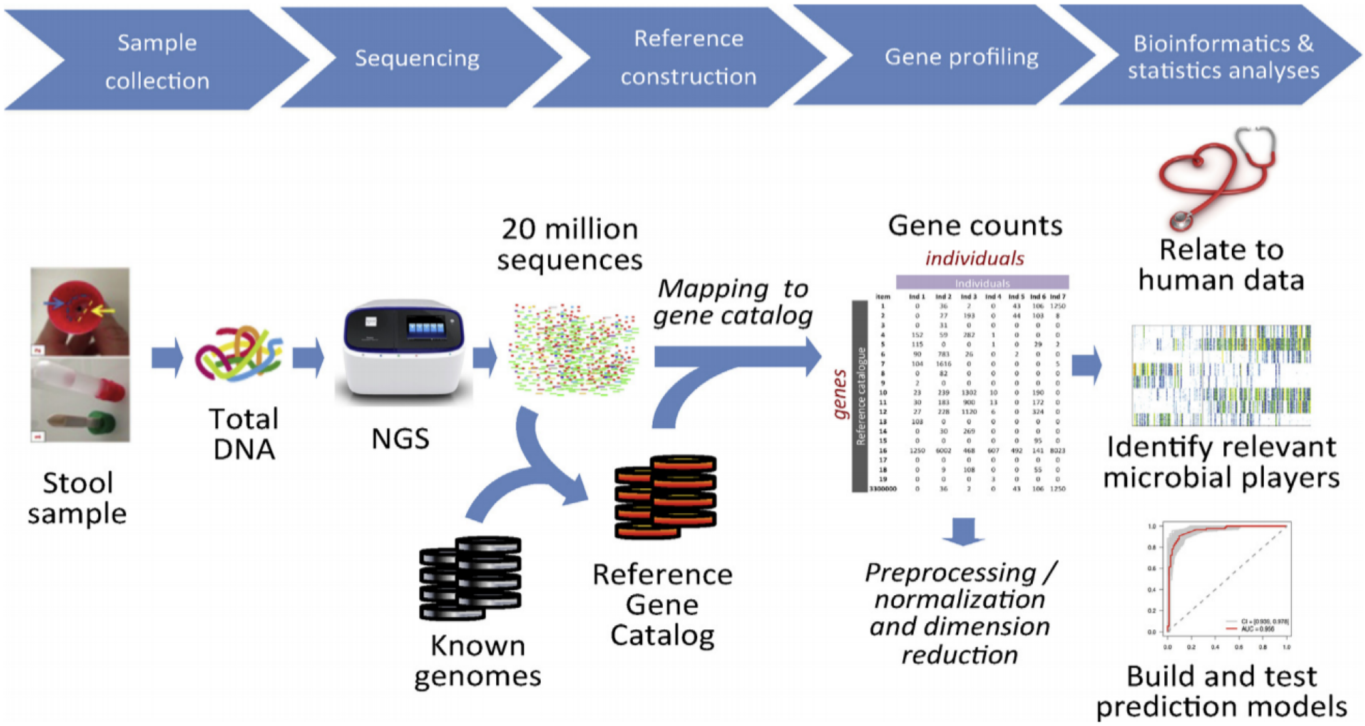

Fig. 1. Quantitative metagenomic data to explore human gut microbiome [21]

TABLE I. BINNING APPROACHES PERFORMANCE COMPARISON IN AVERAGE OF ACCURACY (VAL_ACC) AND MATTHEWS CORRELATION COEFFICIENT (VAL_MCC) ON TEST SETS USING MULTI-LAYER PERCEPTRON

Datasets	T2D1	T2D2
#Samples	344	96
#Features	572	381
#patients who affected by T2D	170	174
#controls/healthy individuals	53	43

dataset (T2D2) [23]. The datasets are characterized by bacterial species abundance. For each sample in each dataset, species abundance is a relative proportion and formed as a real number. The total abundance of all features in each sample is equal to 1. More details are shown in Table I. We consider to investigate on T2D because it is considered as one of the most changeling disease prediction tasks.

Let D be the set of considered datasets, D = $\{d_1, d_2\}$, with d_1 = T2D1, d_2 = T2D2, d = 1..2

S_i = $\{s_1, s_2, ..., s_n\}$ includes n samples corresponding to d_i

F_i = $\{f_1, f_2, ..., f_m\}$ includes m features corresponding to d_i

P_i = $\{p_1, p_2, ..., p_k\}$ includes k patients who affected by T2D corresponds to d_i

C_i = $\{c_1, c_2, ..., c_k\}$ includes x controls / healthy individuals that correspond to d_i

$$Matrix(C) = \begin{pmatrix} d_1 & S_1 & F_1 & P_1 & C_1 \\ d_2 & S_2 & F_2 & P_2 & C_2 \end{pmatrix}$$

$$= \begin{pmatrix} T2D1 & 344 & 572 & 170 & 53 \\ T2D2 & 96 & 381 & 174 & 43 \end{pmatrix}$$

Total abundance of all features in one sample is sum up 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.

IV. BINNING APPROACHES

A. Binning Approaches for Metagenomic Data

Some binning approaches were introduced in [24] including Species bins (SPB) based on species abundance distribution on 6 datasets, binning based on equal width and the method based on equal frequency.

- Species Bins (**SPB**) are conducted from data distribution of six metagenomic bacterial species abundance datasets related to various diseases. Authors in [25] observed that original species abundance almost follows the zero-inflated distribution. When they convert data with a scaler using log-transformed (with logarithm base 4), the scaled data is more normally-distributed (see a example of the raw species abundance and log-transformed (with logarithm base 4) of two considered datasets of T2D shown in Fig. 2).

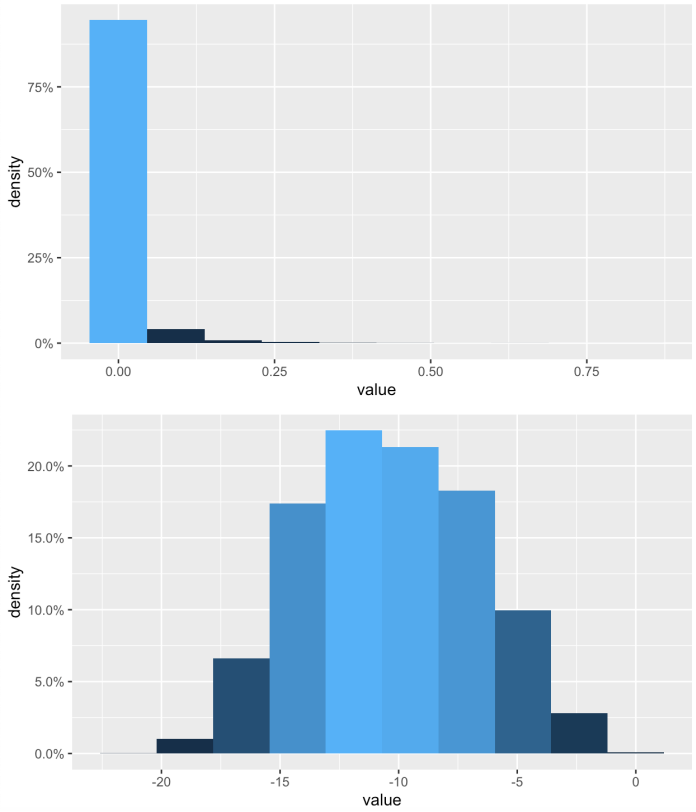

Fig. 2. Species abundance distribution of two considered T2D datasets. The top chart show original species abundance data distribution illustrates zero-inflated distribution. The other reveals a normally-distributed when we do log-transformed (with logarithm base 4) on this data.

From that, authors proposed breaks for binning where each break is the one that in the logarithm base 4 is equivalent to a fold increase from the previous bin. A little more detail, the first breaks will start at 0 and 10^{-7} (the minimum values of six considered datasets), the next break will be $4 * 10^{-7}$ and so on. This bins seem to be efficient for the prediction.

- A commonly-wided way is equal width binning (**EQW**). This technique is rather simple. The breaks are identified based on the width of the considered range of values. Let's say, we want to discretize 5 bins for a range of [Min,Max] with Min=0 and Max=0.5. The width of each bin is equal and computed by $\frac{Max-Min}{5} = 0.1$. Breaks in this example will be $0, 0.1, 0.2, 0.3, 0.4$.
- Binning based on frequency of values is also an effective method. The method is equal frequency binning (**EQF**) where each bin can contain approximately the number of elements. Therefore, the interval width can be very different. The breaks can be $0.1, 0.11, 0.2, 0.5$ and so on, for example, depending on the value distribution.
- The last binning described in this section is binary bins. This method only considers whether the value of that feature is greater 0 or not. Since it determines the Presence of feature in the samples, we also call it "**PR**".

B. Binning based on K-means Algorithm

With different distributions of data, the clustering algorithm is a crucial tool to identify groups in data. Determining groups for binning, we hope to improve the performance by identifying various areas which have high data density. K-means clustering is a common method in cluster analysis and data mining. The purpose of this method is to partition n elements into clusters such that each element of the cluster has the closest mean value, acting as the cluster's prototype. This method is performed based on the smallest Euclidean distance between the elements and the central element of the group. Assume each object has m attributes. Each object's properties are like coordinates of an m-dimensional space; each object is a point on that space. Euclidean distance is calculated by the formula:

$$\partial_{ji} = \sqrt{\sum_{s=1}^{m}(x_{is} - x_{js})^2}$$

With

- ai = (xi1, xi2, ... xim) i = 1..n - the ith object to be classified
- cj = (xj1, xj2, ... xjm) j = 1..k - central element group j

The central element is determined by the average of the elements in the group. Initially, these elements will be randomly selected and after each addition of objects to groups, the central elements will be recalculated. To calculate cij - the j coordinate of the group i central element, we have the formula:

$$c_{ij} = \frac{\sum_{s=1}^{t} x_{sj}}{t}$$

With:

- $j = 1..m$ (m is the number of properties)
- xsj - jth attribute of element s (s = 1..t)

Binning with K-means clustering, we will get better results than the methods mentioned earlier. Suppose we need to binning with n = 10 (the numbers of bins). This method is performed as follows:

Algorithm 1 Algorithm for identifying the list of binning breaks based on clustering algorithm, K-Means

Input: n - number of clusters, matrix C to find bin breaks
Output: B - array containing list of n bin breaks found
Begin

Step 1: Initialize data
- Convert matrix C to 1-dimensional array.
- Remove 0 or uncountable values in array.
- Sort the array in ascending values.

Step 2: Using the K-means algorithm with a total number of clusters n - 1. We have array A containing the grouped elements.

Step 3: Construct array B containing n bin breaks
- Find n - 2 bin breaks by calculating the average of two boundaries in two adjacent groups.

$$B[i] = \frac{(max(A[i-1]) + min(A[i]))}{2}$$

With: $i = 1..n-1$
- Add 0 and 1 to array B.
- Sort the array in ascending values

End

For easier comparisons, all binning approaches in this study are implemented with the same number of bin (10 bins) for all classifiers. We underline that the breaks for binning are conducted using the training sets to avoid overfitting issues.

V. Experiments

For comparing the efficiency binning approaches in improving T2D prediction performance on various learning algorithms, each learning architecture is presented in each separated table. Table II gives results using MLP while Table III illustrates the performance of CNN1d. The last table (Table IV), we present the best results with Random Forest and also compare to state-of-the-art in MetAML [14]. The datasets used was described in Section III. The details of models used in the experiments and results are presented as following.

A. Learning Models for Comparison

In order to evaluate and compare the efficiency on a wide range of learning models, we propose to use 3 different learning algorithms. A state-of-the-art in machine learning is Random Forest that is implemented to run the experiments on the datasets. Moreover, as a traditional neural network, Multi-Layer Perceptron (MLP) is also leveraged for the comparison. We also evaluate one-dimensionality convolutional neural network (CNN1D) on considered datasets.

- Previous studies, most successful methods applied to numeric omics datasets are known mainly Random Forest (RF). Authors in [14] introduced MetAML using Random Forest and obtained the best results among considered algorithms. Applying the same parameters proposed in [14], we use 500 trees for this algorithm for the learning.
- The MLP is used in this study with parameters proposed in [16] including one hidden layer and 128 neural.
- CNN1D consists of one one-dimensional convolutional layer of 128 filters followed by a max pooling of 2 and ending by a fully connected layer. MLP and CNN1D use Adam optimizer function with a batch size of 16. Other parameters are also the same with a default learning rate of 0.001 and epoch patience of 5 for early stopping technique (for reducing overfitting issues).

B. Metrics for Comparison

The performances are assessed by 10-fold cross validation. We compute Average Accuracy and Average Matthews Correlation Coefficient (MCC) as performance measurement for evaluating the generalization of the classifiers. Training and test sets are exactly the same for each classifier, or we can say that the same folds are used for all classifiers. With this technique, the changes when comparing performance of any two classifiers could be computed directly as the difference in metrics within each test fold.

Accuracy is a common measurement for models's performance while MCC is considered as a good performance evaluation score for biology datasets and helps to evaluate whether the model is going well or not. As in [28], the authors said that "among the common performance evaluation scores, MCC is the only one which correctly takes into account the ratio of the confusion matrix size". Matthews correlation coefficient score is computed as following formula:

With:

- TP stands for True Positive
- TN is True Negative
- FP: False Positive
- FN: False Negative

Matthews Correlation Coefficient score is computed by:

$$MCC = \frac{TP.TN - FP.FN}{\sqrt{(TP+FP).(TP+FN).(TN+FP).(TN+FN)}}$$

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

The model reaches the best when $mcc = 1$ while the worst value is $mcc = -1$. Authors in [28] recommended using this metric for evaluating the algorithm performance.

C. Experimental Results

1) Evaluation binning approaches with MLP: We are considering two diseases T2D1 and T2D2 with results using MLP in Table II. As a result, the binning approach with K-means in both diseases achieved val_acc and val_mcc values higher than all other approaches EQW, PR, SPB. Considering dataset dataset T2D1, K-means is significantly higher than SPB. Specifically, val_acc is higher than val_acc of SPB is 0.034 and of val_mcc is 0.044. For approaches like EQW, PR or EQF, the K-Means approach returns values with relatively good disparities. Considering dataset dataset T2D2, val_acc of K-means is more than 0.069, val_mcc is 1.46 times higher than

TABLE II. BINNING APPROACHES PERFORMANCE COMPARISON IN AVERAGE OF ACCURACY (VAL_ACC) AND MATTHEWS CORRELATION COEFFICIENT (VAL_MCC) ON TEST SETS USING MULTI-LAYER PERCEPTRON

val_acc	val_mcc	Dataset	Approach
0.686	**0.379**	**T2D1**	**k-means**
0.681	0.371	T2D1	EQW
0.663	0.353	T2D1	PR
0.658	0.34	T2D1	EQF
0.652	0.335	T2D1	SPB
0.727	**0.459**	**T2D2**	**k-means**
0.714	0.437	T2D2	EQW
0.667	0.339	T2D2	PR
0.705	0.414	T2D2	SPB
0.652	0.314	T2D2	EQF

TABLE III. BINNING APPROACHES PERFORMANCE COMPARISON IN AVERAGE OF ACCURACY (VAL_ACC) AND MATTHEWS CORRELATION COEFFICIENT (VAL_MCC) ON TEST SETS USING CNN1D

val_acc	val_mcc	Dataset	Approach
0.692	**0.392**	**T2D1**	**k-means**
0.678	0.363	T2D1	EQW
0.677	0.367	T2D1	PR
0.652	0.323	T2D1	EQF
0.649	0.316	T2D1	SPB
0.740	**0.473**	**T2D2**	**k-means**
0.707	0.413	T2D2	EQW
0.700	0.397	T2D2	PR
0.687	0.382	T2D2	SPB
0.674	0.346	T2D2	EQF

EQF. The value of EQW in this disease is the second most in approach and is 0.022 different from when using K-Means. In summary, the results when binning with K-Means cluster using Multi-Layer Perceptron, we will get the best results compared to the remaining methods.

2) Evaluation binning approaches with Convolutional Neural Network on 1D data: Table III shows the performance using CNN1D. When using the One-Dimensional Convolutional Neural Network, the results of K-Means are 0.692 for val_acc, 0.740 for val_mcc, respectively. Both results are better than using Multi-Layer Perceptron (val_acc = 0.686, val_mcc = 0.727). In T2D1, the result of K-Means is much higher than the next EQW value, namely 0.014 difference for val_acc and 0.076 for val_mcc compared to K-Means. The value of val_acc of K-Means compared to the lowest value in this disease of SPB is 0.076 and of val_mcc is 0.043. In T2D2, the lowest valued approach for this disease is EQF. Val_acc value is more than 0.066, val_mcc of K-Means is 1.367 more than EQF. The difference between the values of EQW and K-Means is quite good, respectively 0.033 for val_acc, 0.06 for val_mcc. In summary, when using the One-Dimensional Convolutional Neural Network, the K-Means approach results in better results when using the Multi-Layer Perceptron and this result is still the best result compared to the other approach.

3) Random Forest obtains promising results with the proposed binning, compared to state-of-the-art MetAML: We also used the Random Forest for results comparison in Table III. Similar to the previous two tables, when binning with K-means we obtain very good results compared to using other approaches. A previously used framework, MetAML, K-means, gave val_acc more than 0.036 for T2D1 and 0.056 for T2D2. Considering T2D1, K-means val_acc is more than 0.04 and val_mcc is 0.07 more than SPB. The second result in the table for both diseases is the PR approach. The difference in value between K-means and PR is quite good. K-means has val_acc more than 0.014, val_mcc is more than 0.017 than PR. Considering T2D2, val_acc is 0.107 and val_mcc is 1.683 times higher than SPB results. K-means has val_acc more than 0.023, val_mcc is more than 0.032 than PR. In short, when choosing K-means as an approach, we will get better results than some common approaches such as PR, EQW, EQF or SPB, especially the approach used was MetAML.

TABLE IV. BINNING APPROACHES PERFORMANCE COMPARISON IN AVERAGE OF ACCURACY (VAL_ACC) AND MATTHEWS CORRELATION COEFFICIENT (VAL_MCC) ON TEST SETS USING RANDOM FOREST

val_acc	val_mcc	Dataset	Approach
0.700	**0.400**	**T2D1**	**k-means**
0.686	0.383	T2D1	PR
0.680	0.370	T2D1	EQF
0.674	0.357	T2D1	EQW
0.660	0.330	T2D1	SPB
0.664		T2D1	*MetAML*
0.759	**0.515**	**T2D2**	**k-means**
0.736	0.483	T2D2	PR
0.720	0.440	T2D2	EQW
0.690	0.370	T2D2	EQF
0.652	0.306	T2D2	SPB
0.703			*MetAML*

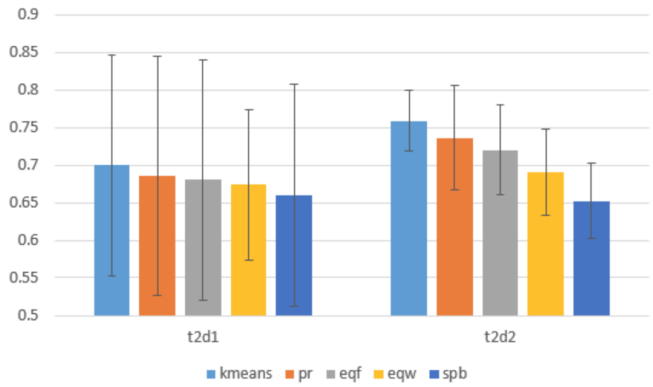

Fig. 3. Performance Comparison in Average Accuracy of different binning approaches including EQF, EQW, K-means, PR and SPB. Standard deviations are shown in error bar.

4) Random Forest obtains better results compared to neural networks: The chart in Fig. 3 shows the results being conducted from two datasets of T2D. We use five approaches for testing, namely, EQF, EQW, K-Means, PR, SPB. Considering T2D1 disease, the K-means approach has the largest Average Accuracy value, reaching 0.7. SPB has a value of Average Accuracy is 0.66, this is the smallest value and smaller than K-Means 0.34. Similarly, for T2D2 disease, the Average Accuracy of K-Means value is 0.759, the highest among the remaining approaches. This value is higher than the next PR value of 0.023. The Average Accuracy of SPB is less than 0.107 compared to K-Means.

The chart in Fig. 4 shows the results Average MCC value on 2 datasets of T2D and 5 approaches. K-Means has the highest Average MCC value on both datasets and 0.4 for T2D1 and 0.515 for T2D2. Average MCC value of K-Means greater than SPB in T2D1 is 0.07, 1,683 times that of T2D2. The disparity with the next high value of PR is also quite clear,

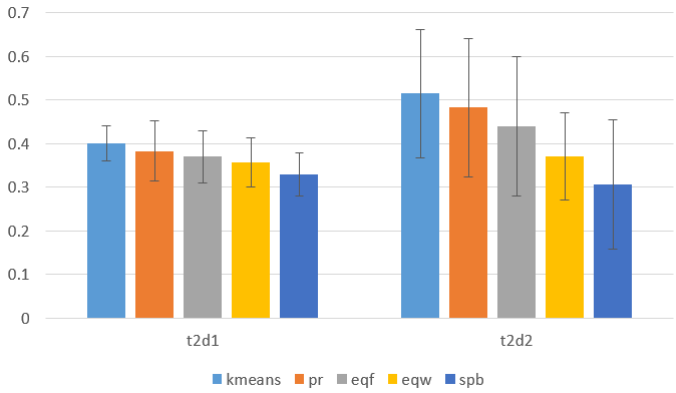

Fig. 4. Performance Comparison in Average MCC of different binning approaches including EQF, EQW, K-means, PR and SPB.

namely, 0.017 for T2D1 and 0.032 for T2D2.

VI. Discussion

From collected results, we can see that RF obtains the best among considered models. These results are similar to [25] where authors also have attempted to apply deep learning but the performance in T2D disease is still worse than RF. This reflects a fact as mentioned in [26]: "the deep learning approaches may not be suitable for metagenomic applications". As stated in [27], we are facing challenges when applying deep learning to solve biological and clinical tasks because of limited data availability, result interpretation and hyper-parameters tuning for deep learning algorithms.

Although PR only considers whether a bacterial species exists in a patient or, it revels a better performance (using RF) than several other binning methods such as SPB, EQW, EQF. From results, we can propose medical examinations for T2D only determining the existence of bacterial species in human body for the diagnosis. These examinations can be simpler than computing quantitative compositions of bacterial.

In most situations, SPB performs poor performance compared to the others because SPB was conducted from species abundance distribution from various diseases. Each disease should be considered independently because one disease can have its own complexity, characteristics as well as data density.

VII. Conclusion

We introduce a novel binning approach using a classical clustering algorithm such as K-means. As shown from the comparison results among considered existing binning approaches such as binning based on species distribution, based on width and frequency and binary bins, we can see the encouraging results in use of clustering methods for identifying breaks for binning to enhance the prediction performance.

The analysis of two architectures of one-dimensional convolutional neural network and Multi-layer Perceptron shows that convolutional neural network not only achieve a good performance on images but also obtain a promising result compared to traditional neural network such as MLP.

As some results in previous studies, classic machine learning such as Random Forest still works better more complex models such as MLP and CNN1D in T2D diagnosis by metagenomic data. Further research can investigate more deeper and sophisticated models to improve the performance.

Using classic clustering algorithm K-means with default parameters in binning gives encouraging results. This could promote studies to go deeper in use of clustering methods to generate breaks for binning. This illustrate that there are potentials in exploring density data to improve not only for T2D disease but also for other diseases.

References

[1] Kevin Chen, Lior Pachter. Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities. 2005.

[2] DeLong EF Microbial population genomics and ecology. Curr Opin Microbiol 5: 520–524. 2002.

[3] Handelsman J, Metagenomics: Application of genomics to uncultured microorganisms. Microbiol Mo lBiol Rev 68: 669-684. 2004

[4] Riesenfeld CS, Schloss P, Handelsman J, Metagenomics: Genomic analysis of microbial communities. Annu Rev Genet 38: 525–552. 2004.

[5] Rodriguez-Valera F, Environmental genomics, the big picture? FEMS Microbiol Lett 231: 153–158. 2004.

[6] Streit WR, Schmitz RA, Metagenomics—The key to the uncultured microbes. Curr Opin Microbiol 7: 492–498. 2004.

[7] Maja Fabijanić and Kristian Vlahoviček, Oliviero Carugo and Frank Eisenhaber (eds.), Data Mining Techniques for the Life Sciences, Methods in Molecular Biology, vol. 1415, DOI 10.1007/978-1-4939-3572-7_26, © Springer Science+Business Media New York 2016.

[8] Edwards RA, Rohwer F, Viral metagenomics. Nat Rev Microbiol 3: 504–510. 2005.

[9] NIH HMP Working Group, Peterson J, Garges S et al, The NIH Human Microbiome Project. Genome Res 19:2317– 2323. 2009. doi:10.1101/gr.096651.109. 2009.

[10] Garrett WS, Gallini CA, Yatsunenko T et al, Enterobacteriaceae act in concert with the gut microbiota to induce spontaneous and maternally transmitted colitis. Cell Host Microbe 8:292–300. doi:10.1016/j.chom.2010.08.004. 2010.

[11] Karlsson FH, Fåk F, Nookaew I et al, Symptomatic atherosclerosis is associated with an altered gut metagenome. Nat Commun 3:1245. doi:10.1038/ncomms2266. 2012.

[12] Qin N, Yang F, Li A et al, Alterations of the human gut microbiome in liver cirrhosis. Nature 513:59–64. doi:10.1038/nature13568. 2014.

[13] Turnbaugh PJ, Gordon JI, The core gut microbiome, energy balance and obesity. J Physiol 587:4153–4158. doi:10.1113/ jphysiol.2009.174136. 2009.

[14] E. Pasolli, D. T. Truong, F. Malik, L. Waldron & N. Segata; Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights; PLoS Comput. Biol. 12, p. e1004977. 2016.

[15] Steve Miller, Charles Chiu, Kyle G. Rodino, Melissa B. Miller; Point-Counterpoint: Should We Be Performing Metagenomic Next-Generation Sequencing for Infectious Disease Diagnosis in the Clinical Laboratory?. DOI: 10.1128/JCM.01739-19. Journal of Clinical Microbiology. 2020.

[16] Thanh Hai Nguyen, Jean-Daniel Zucker. Enhancing Metagenome-based Disease Prediction by Unsupervised Binning Approaches. The 2019 11th International Conference on Knowledge and Systems Engineering (KSE-IEEE), ISBN: 978-1-7281-3003-3, pp 381-385. 2019.

[17] Hanaa M. Hussain et al. FPGA implementation of K-means algorithm for bioinformatics application: An accelerated approach to clustering Microarray data. 2011 NASA/ESA Conference on Adaptive Hardware and Systems (AHS). 2011.

[18] Timothy et al. K-Means Clustering of Biological Sequences. ADCS 2017: Proceedings of the 22nd Australasian Document Computing Symposium. 2017.

[19] Jasmin T. Jose1 et al. Case Study on Enhanced K-Means Algorithm for Bioinformatics Data Clustering. International Journal of Applied Engineering Research ISSN 0973-4562. 2017.

[20] Vijayalakshmi K., Padmavathamma M. (2019) Design and Implementation of Modified Sparse K-Means Clustering Method for Gene Selection of T2DM. In: Computational Intelligence and Big Data Analytics. SpringerBriefs in Applied Sciences and Technology. Springer, Singapore. 2019.

[21] Stanislav Dusko Ehrlich. The human gut microbiome impacts health and disease. PubMed. 339(7-8):319-23. doi: 10.1016/j.crvi.2016.04.008. PMID: 27236827. 2016

[22] Karlsson FH, Tremaroli V, Nookaew I, Bergström G, Behre CJ, Fagerberg B, et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature 2013;498(7452):99–103. pmid:23719380. 2013.

[23] Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A. 2013 metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 2012;490(7418):55–60. pmid:23023125. 2012.

[24] Le Chatelier E, Nielsen T, Qin J et al Richness of human gut microbiome correlates with metabolic markers. Nature 500:541–546. doi:10.1038/nature12506. 2013.

[25] Thanh Hai Nguyen et al.; Disease Classification in Metagenomics with 2D Embeddings and Deep Learning; In Proceedings of CAp, France 2018.

[26] G. Ditzler, R. Polikar & G. Rosen; Multi-Layer and Recursive Neural Net- works for Metagenomic Classification; IEEE Trans. Nanobioscience 114, p. 608–616. 2015.

[27] Fioravanti, D., Giarratano, Y., Maggio, V. et al. Phylogenetic convolutional neural networks in metagenomics. BMC Bioinformatics 19, 49. https://doi.org/10.1186/s12859-018-2033-5. 2018.

[28] Baghban, H. and Rahmani, A.M. A heuristic on job scheduling in grid computing environment. In Grid and Cooperative Computing, 2008. GCC'08. Seventh International Conference on (pp. 141-146). IEEE. October, 2008.