Journal of Natural Sciences Research www.iiste.org
ISSN 2224-3186 (Paper) ISSN 2225-0921 (Online)
Vol.13, No.6, 2022

Using WEKA to Classify Treponemes Based on Carbohydrate Utilization and Enzymatic Activity

Andres Botero (Corresponding author)
Natural Science Department. Mercy College, 555 Broadway Dobbs Ferry, NY 10522
E-mail: abotero@mercy.edu

Nyah Tomala
Natural Science Department. Mercy College, 555 Broadway Dobbs Ferry, NY 10522
E-mail: ntomala@mercy.edu

Abstract
Prediction and classification of microbial species is an important skill for any clinical, industrial or environmental laboratory, there are many aspects including carbohydrate utilization, enzymatic activity that can be used for predicting with high accuracy bacterial species. The unique features presented by machine learning includes the possibility of classify and predict bacterial species based on their biochemical or enzymatic activities though decision tree classification, data visualization, clustering and neural network among others tools. **In this research, decision trees are applied to classify Treponema species based on enzymatic activity, visualization tools showed comparisons among multiple biochemical aspects and neural networks created patterns for carbohydrate utilization.** Treponema species are invasive pathogens causing a range of significant clinical pathologies in many cases ending in neurological complications such as in the case of syphilis produced by *Treponema pallidum*. Previous research papers explored the used of PyBact to generate a matrix which it is then evaluated through machine learning resulting in a high percentage of correct classification. Our findings indicate that decision trees are one of the most effective tools to classify bacterial species contributing significantly to any medical or diagnostic laboratory. There are several ML applications that remain to be explored not just for a particular genus but with questions involving the human microbiome, biofilm formations, and the current COVID 19 pandemic.
Keywords: Machine learning, PyBact, neural networks, decision trees, visualization, Treponema genus.
DOI: 10.7176/JNSR/13-6-04
Publication date: March 31st 2022

1.Introduction

In recent years machine learning has been used in many areas of microbiology research due to the enormous quantity of information generated that need to be analyzed **(Qu et al., 2019).** Machine learning applied multiple strategies to predict, classify, clustered, visualize results using an array of mathematical and statistical concepts such as linear regression, neural networks, decision trees (Qu et al., 2019). Prediction of microbial species is an important skill for any clinical, industrial or environmental laboratory, there are many aspects including carbohydrate utilization, enzymatic activity, G + C content, diameter and length that can be used for predicting with high accuracy bacterial species (Nantasenamat C et al., 2011 Logan, 1994). Treponemes are corkscrew invasive pathogens causing significant clinical conditions among humans and animals such is the case of venereal syphilis which encompassed three stages; an initial genital lesion is follow by a body dissemination and a third stage categorized by neurological complications (Radolf JD., 1996). The four most important treponematoses are *Treponema subsp. pallidum* producing venereal syphilis, a worldwide infection affecting adults and producing congenital infections, *Treponema subsp. pertenue* producing yaws, a tropical infection prevalent in hot and humid areas and common in children and adolescents, congenital infection is not common for this subspecies. *Treponema subsp. carateum* producing pinta, a Central and South America infection, common in children, congenital infection is not common as well in this subspecies and *Treponema subsp. endemicum* producing endemic syphilis, an infection prevalent in desertic areas, common in children and adults and rarely involve in congenital infections Treponema are categorize as chemo-organotrophs able to use an extensive number of carbohydrates such as glucose as source of energy, other species required long or short fatty acids present in serum, while other species relied on amino acids fermentation and enzymatic activity (Radolf JD., 1996). Machine learning is a subfield of artificial intelligence applied to numerous fields ranging from retail stores to scientific pursuits; microbiology and its branches such as bacteriology, virology, parasitology, mycology and immunology are not an exception to its prowess; machine learning is used for example in the prediction of vaccines, outbreaks, proteins interactions, diagnosis of infectious diseases, prevalence of conditions among local or global locations and classification of microorganisms based in biochemical or other physiological characteristics. Weka (Waikato Environment for Knowledge Analysis) is a free software composed of diverse algorithms for visualization, data analysis, regression, neural network, clustering and predicting models. (Witten and Frank, 2005). In this study, decision trees are applied

to classify Treponema species based on enzymatic activity, visualization tools showed comparisons among multiples morphological aspects and neural networks created patterns for carbohydrate utilization.

2.Materials and Methods

2.1Data compilation

The data used in this study is constituted by 17 species from Treponema genus carbohydrate utilization and 14 species from Treponema genus enzymatic activity obtained from Bergey' s Manual of Systematic Bacteriology (Krieg Noel R et al., 2010)

2.2Weka for bacterial identification

Treponema genus carbohydrate utilization and enzymatic activity was used as input data; the Weka algorithms used for identification were decision trees, multilayer Perceptron (Neural networks) and visualization. A concept map of the data mining methods used in this study is provided in Figure 1.

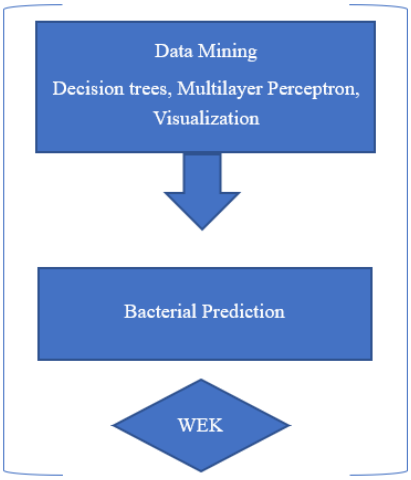

Figure 1.

2.3 Decision Trees

Decision trees are used to classify and predict items based on its values; root nodes corresponded to attribute conditions, leaf nodes used labels for each class, and branches represented values assume by nodes. There are several algorithms employed to develop decision trees such as C4.5 which is an extension of the ID3 (Kotsiantis. S. B. 2007). Decision trees are easy to use and understand, can be used with categorical and numerical data, additional features can easily be incorporated to create complex models, all these aspects account as advantages for using tress, however, among some disadvantages are overfitting and biased in some cases in which certain attributes dominate. Random tree is a supervised algorithm used for classification and regression that work by averaging the input data therefore reducing the overfitting common in other decision trees algorithms. Among the advantages of random trees are their high accuracy and capacity to process large amount of information making it an optimal tool in the identification and classification of bacterial species (Mishra A. K. 2016).

2.4 Neural Networks

The main goal of this powerful mathematical and computational algorithms is to mimic neurological networks by using a connectionist strategy. Artificial neurons are connected to information to generate patterns taking into consideration three important objectives, the network structure or architecture, input and output functions and the weight inherent in each connection (Kotsiantis. S. B. 2007). Neural networks are applicable on the prediction of protein structures, as well as graphic and auditory patterns recognition (Gour S and Gour M. 2014).

2.5 Clustering

Clustering has the capability of grouping data to determine patterns, specific number of groups can be created depending on the research needs. Advantages over other machine learning methods is that entire data sets can be turned into groups for expediate conclusions, as well as any data input can be use in the clustering analysis (Abernethy M.2010)

2.6 Visualization

The WEKA explorer has a variety of data visualization options to view complex information such as histograms, bar graphs, maps, images, 3D representations, tables and others. It is feasible to get the visual representation of classifications, regressions and clusters models for the easy detection of outliers. Fundamental components of data

visualization are data sampling, analysis, governance, mining and transformation (Soni R. 2013).

3.Results

This study used a 10-fold cross-validation for all data and the following classifiers: random tree, neural network, visualization and clustering. As results shown in figures 2 and 3; 14 Treponema species were correctly classify based on 19 enzymatic activities with a 78.5 % accuracy (figure 4), using WEKA random tree algorithm.

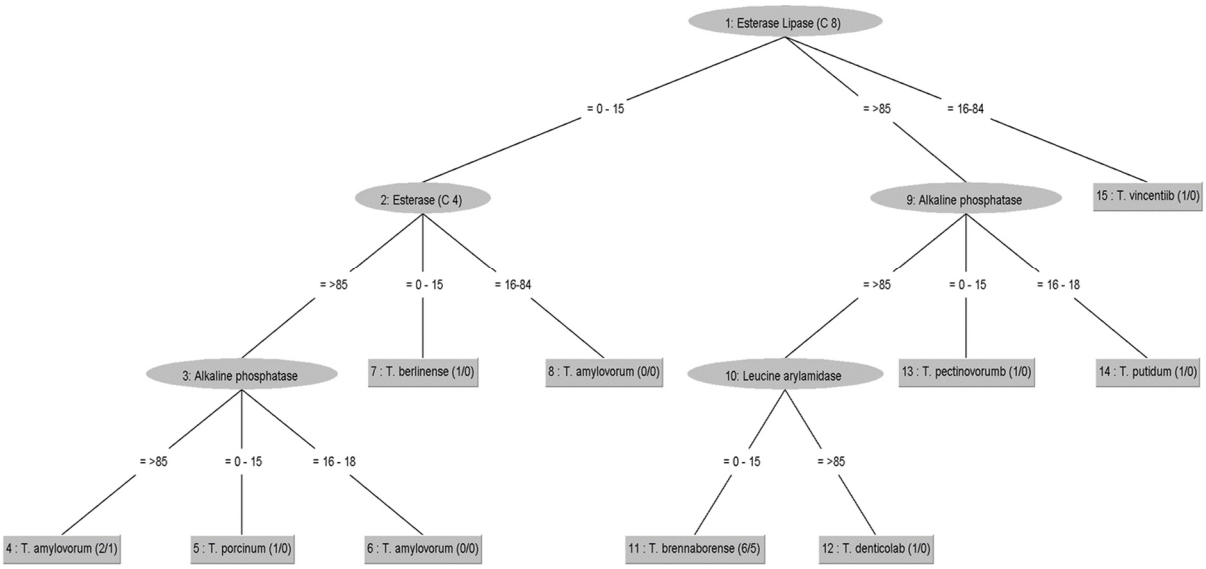

Figure 2

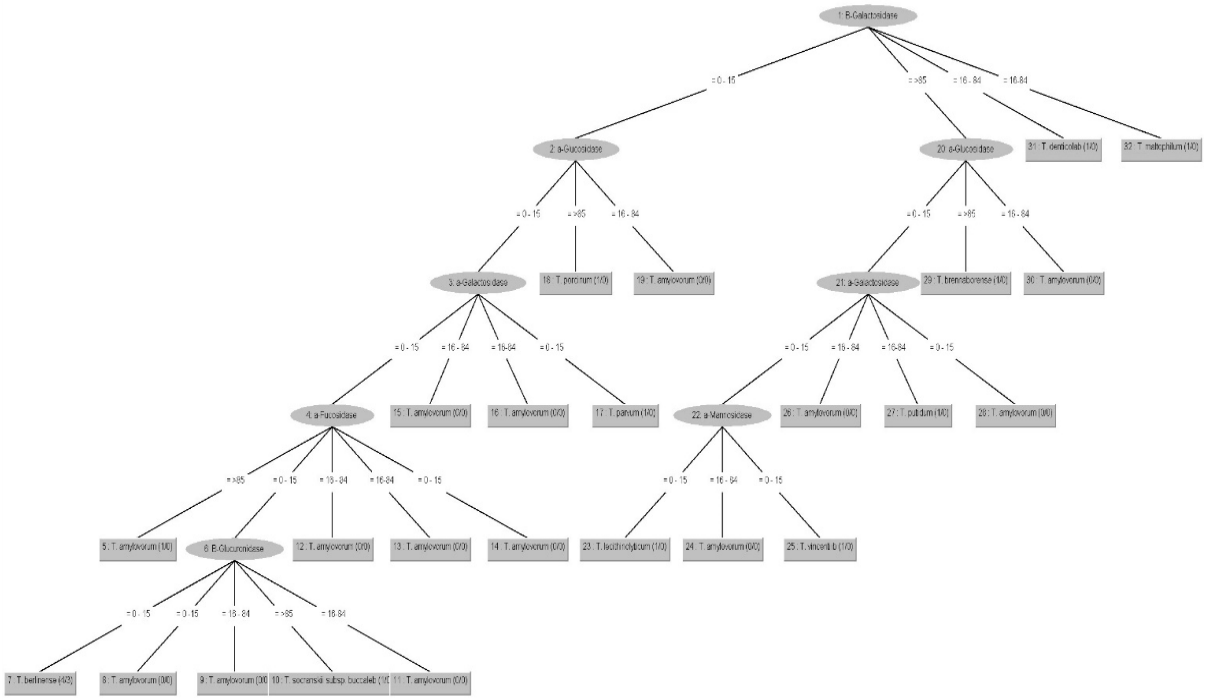

Figure 3

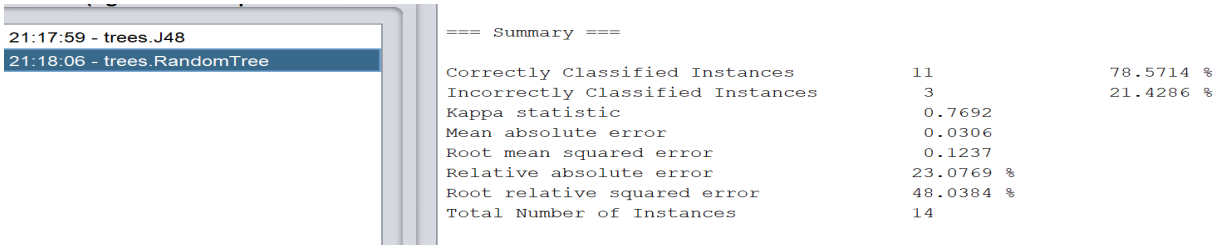

Figure 4

Results showed that random tree algorithm could be used to correctly classify Treponema species based on enzymatic activity from 14 number of instances, 11 were correctly classified which equates to 78.5714%. In another application, 17 species of Treponema species carbohydrates utilization profiles were used to generate a matrix using Pybact data generation algorythm, consequently the data set total is 1,700 instances, after using decision tree J48 the following results were obtained, 93.8% of instances were correctly classified (1,596 instances), and 6.1% of instances were incorrectly classified (104 instances), figure 5.

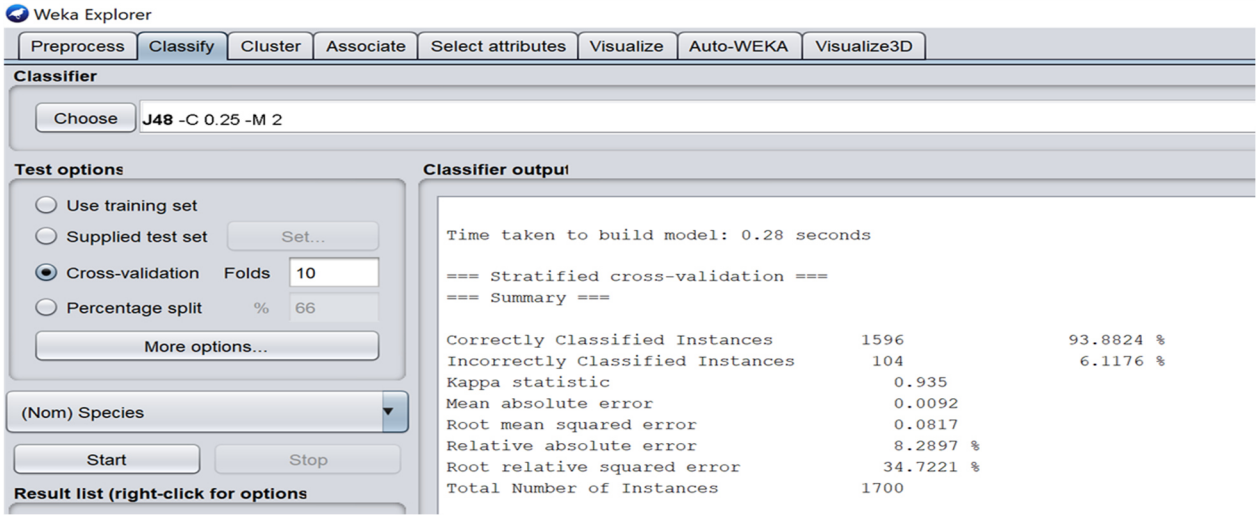

Figure 5

The multilayer perceptron, figures 6 and 7, creates models using input occurrences connected to its corresponding output, other aspects of this algorithm are the hidden layers consisting of nodes interconnected to adjacent layers based on specific weights. Figure 6 shows non-cultivable Treponema habitats neural network using a batch size (100), learning rate (0.3), hidden layers (2) and training time (500).

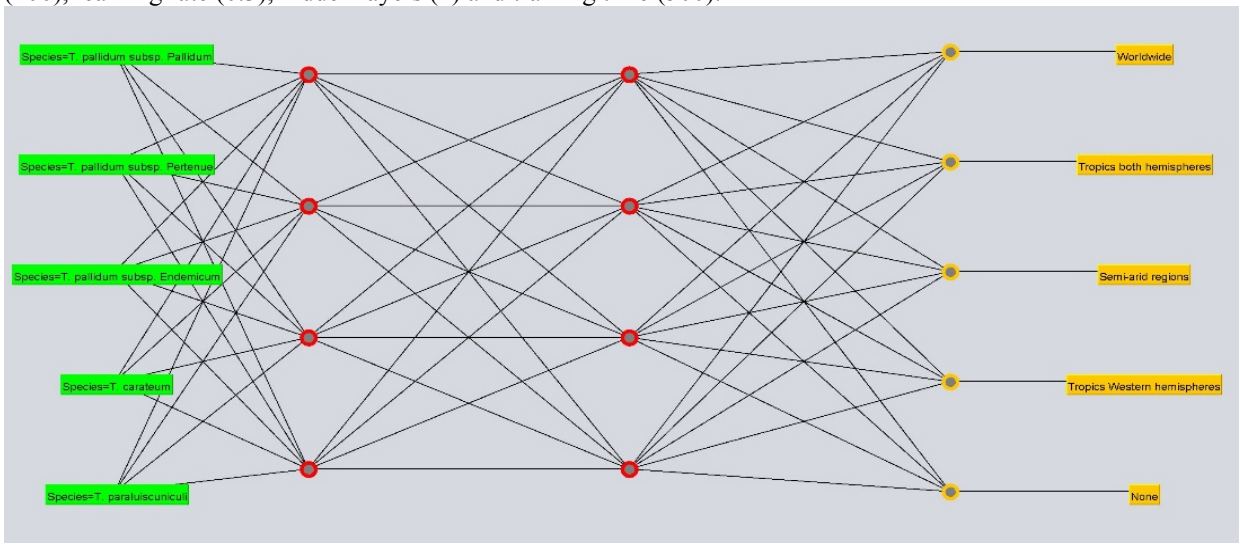

Figure 6

Figure 7 shows carbohydrate utilization by Treponema species using a batch size (100), hidden layers (4), learning rate (0.3), training time (500) and numbers of epochs (500).

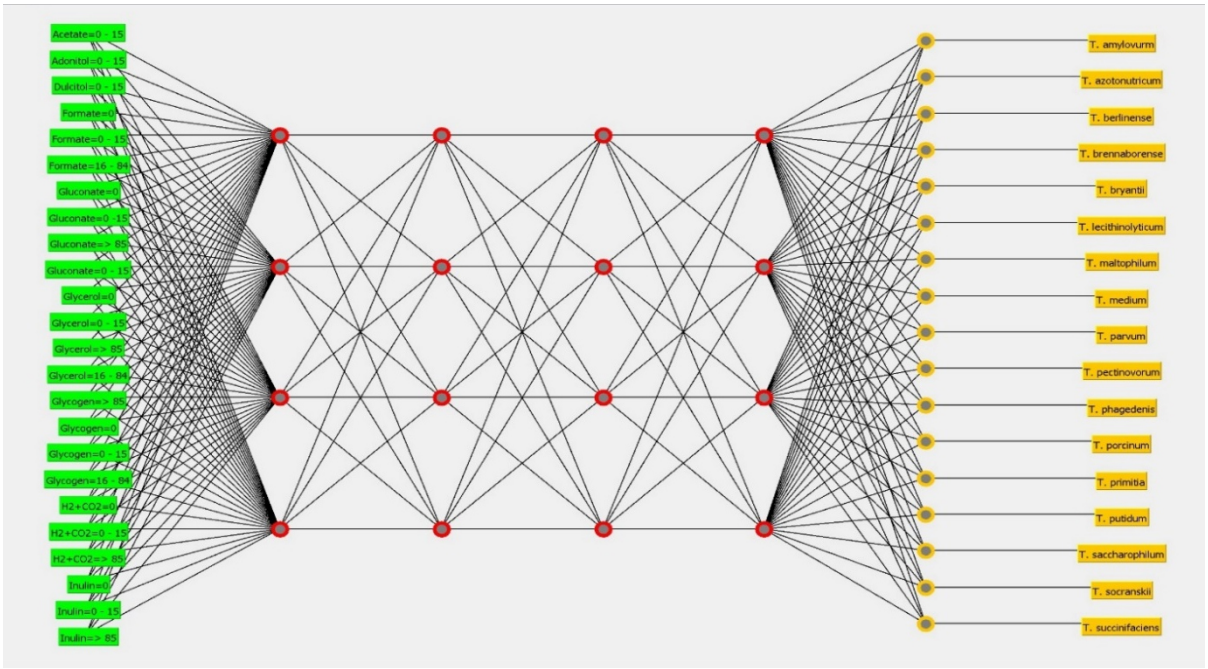

Figure 7

1,700 instances generated through PyBact algorithm for the Treponema species utilization of carbohydrates was used for Simple K means clustering for the Acetate, it shows two clusters, one for the positive utilization of acetate (cluster 1; 613 instances for a 36%) for *T. bryantii*, *T. saccharophilum*, and *T. succinifaciens*; for the remaining species, acetate utilization is negative (cluster 0; 1,087 instances for a 64%), time taken to build model was 0.08 seconds, figure 8.

Other clustering for maltose shows positive utilization (cluster 1; 942 instances for a 55%), and for negative utilization (cluster 0; 758 instances for a 45%), time taken to build model was 0.01 seconds, figure 9.

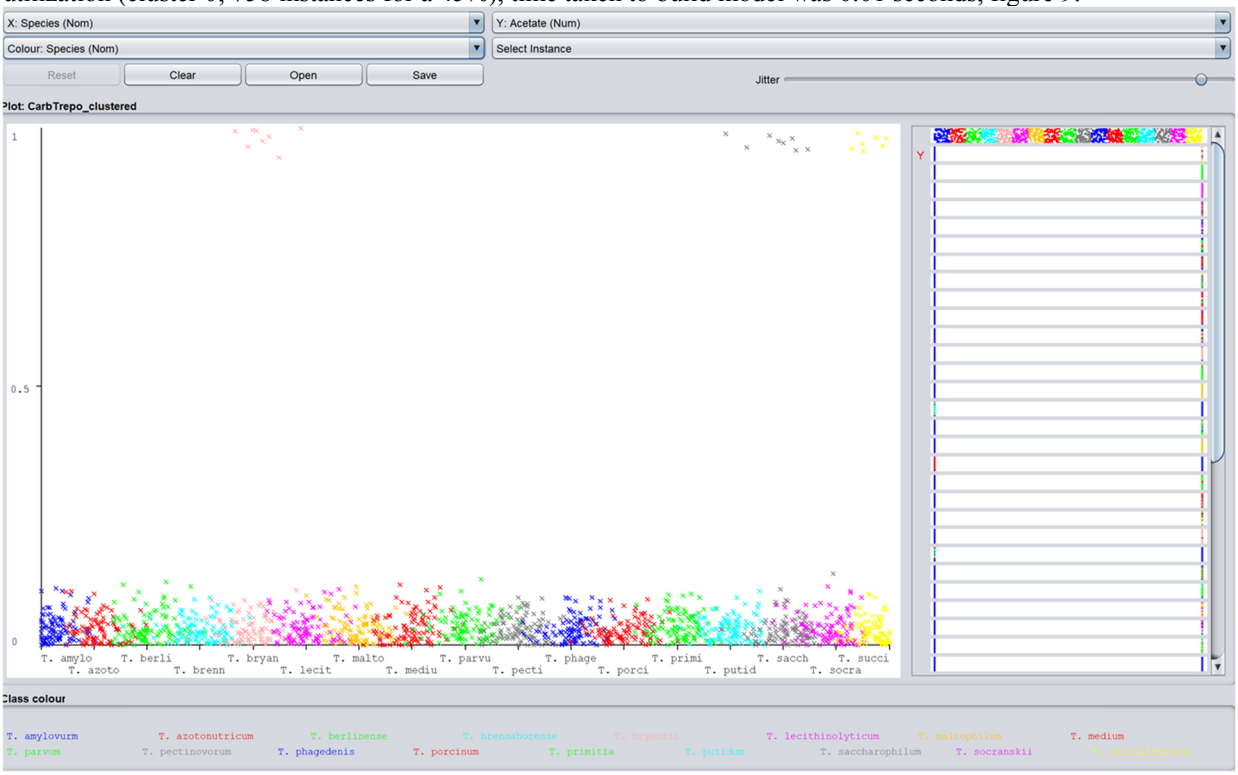

Figure 8

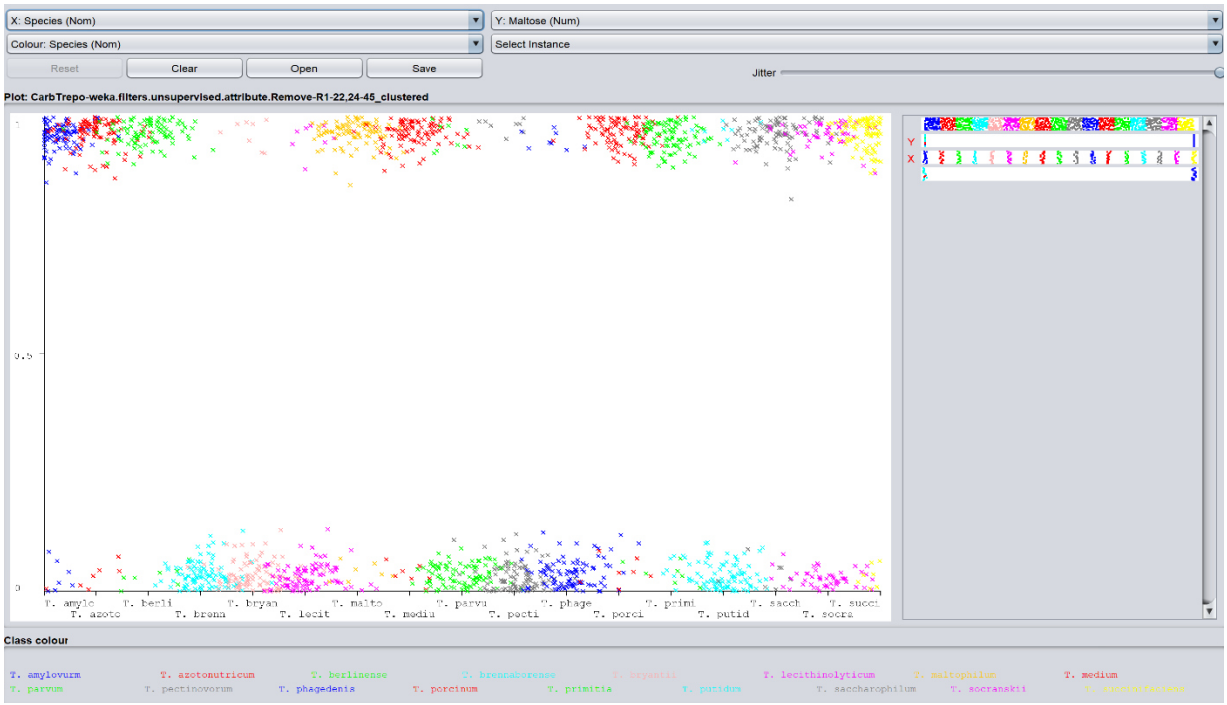

Figure 9

The WEKA explorer has a variety of adjustable (plot size, point size, jitter, subsample %) visualization options to view information such as the use of plot matrix for Treponema species utilization of acetate and maltose, the subsample percentage used in this representation was 15%, figure 10.

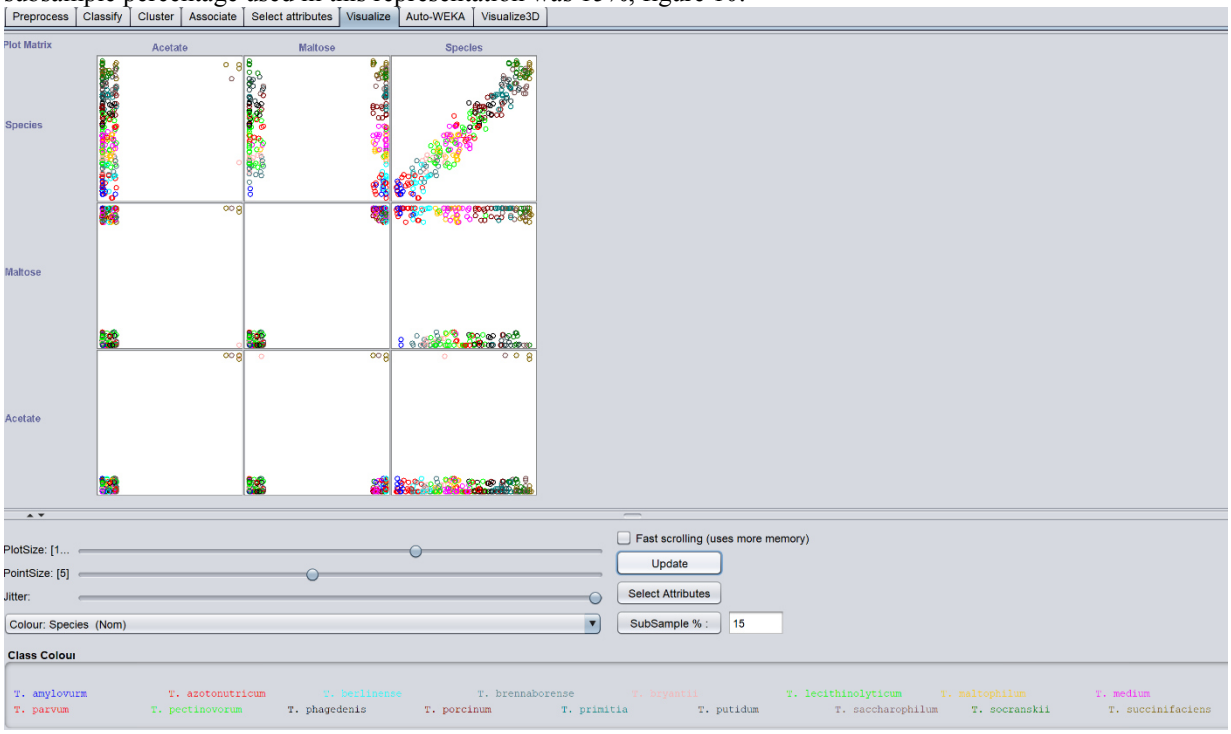

Figure 10

4.Discussion

Machine learning is gaining a highly importance in research, specially in the area of Microbiology due to the multitude of microbial species and their multiple phenotypes and niches. Even more significant is the possibility of applying all this knowledge to solve public health issues such the current crisis with Covid 19 pandemic. Machine learning tools and algorithms had been used for the prediction of new drugs, identification of species present in polymicrobial infections and as diagnosis among other applications. In this paper we explored the use of machine learning to classify Treponema genus based on carbohydrate utilization and enzyme activity applying

several machine learning algorithms such as decision tree classification (random tree and J48), multilayer perceptron, clustering (simple K means) and visualization (Plot matrix). PyBact free software was employed as well in this research to generated matrix based on the carbohydrate utilization of Treponema species and then use this as a model for machine learning analysis; classifier showed a great accuracy on predicting bacterial species, however there are several applications that remain to be discovered not just for a particular genus but with even bigger microbial communities such those belonging to the human microbiome or those participating in biofilm formations. **Future research directions are to apply machine learning to classify Treponemes based in other biochemical tests such volatile fatty acids or hydrogen sulfide production. Another important direction is to use machine learning algorithms to explore phylogenic relationships among Treponemes species based on genomic sequences; finding the right algorithm can be used to implement quick diagnosis in microbiology laboratories or medical settings struggling with fastidious organisms.**

Acknowledge
We would like to express our special thanks of gratitude to Mercy College Natural Science Department, as well as the Computer Science Department which help us doing the research and downloading the different programs use in this research.

References

1. Al-Shahib, Ali, Rainer Breitling, and David R. Gilbert. "Predicting Protein Function by Machine Learning on Amino Acid Sequences – a Critical Evaluation." *BMC Genomics* 8, no. 1 (March 20, 2007): 78. https://doi.org/10.1186/1471-2164-8-78.
2. IBM Developer. "Classification and Clustering," May 12, 2010. https://developer.ibm.com/technologies/analytics/articles/os-weka2/.
3. "Comparative Study Using Neural Networks for 16S Ribosomal Gene Classification | Journal of Computational Biology," August 1, 2021. https://www.liebertpub.com/doi/full/10.1089/cmb.2019.0436.
4. "Data Mining: Practical Machine Learning Tools and Techniques," July 19, 2021. https://www.cs.waikato.ac.nz/ml/weka/book.html.
5. Douglas A. Jabs, Rubens Belfort Jr, Bahram Bodaghi, Elizabeth Graham, Gary N. Holland, Susan L. Lightman, Neal Oden, Alan G. Palestine, Justine R. Smith, Jennifer E. Thorne, Brett E. Trusko. THE STANDARDIZATION OF UVEITIS NOMENCLATURE (SUN) WORKING GROUP. "Classification Criteria for Syphilitic Uveitis". *American Journal of Ophthalmology* (2021) 28, 182 -191
6. Frank, E., M. Hall, L. Trigg, G. Holmes, and I. H. Witten. "Data Mining in Bioinformatics Using Weka." *Bioinformatics* 20, no. 15 (October 12, 2004): 2479–81. https://doi.org/10.1093/bioinformatics/bth261.
7. Gewehr, Jan E., Martin Szugat, and Ralf Zimmer. "BioWeka—Extending the Weka Framework for Bioinformatics." *Bioinformatics* 23, no. 5 (March 1, 2007): 651–53. https://doi.org/10.1093/bioinformatics/btl671.
8. Goodswen, Stephen J, Joel L N Barratt, Paul J Kennedy, Alexa Kaufer, Larissa Calarco, and John T Ellis. "Machine Learning and Applications in Microbiology." *FEMS Microbiology Reviews*, no. fuab015 (March 16, 2021). https://doi.org/10.1093/femsre/fuab015.
9. Heather R. Elder, PhD, MPH,* Susan Gruber, PhD,† Sarah J. Willis, PhD, MPH,*‡ Noelle Cocoros, DSc, MPH,‡ Myfanwy Callahan, MD, MPH,§ Elaine W. Flagg, PhD, MS,¶ Michael Klompas, MD, MPH,‡‖ and Katherine K. Hsu, MD, MPH*** "Can Machine Learning Help Identify Patients at Risk for Recurrent Sexually Transmitted Infections?". *Sexually Transmitted Diseases* (2021) 48(1), 56 -62
10. Huang, Lei, and Tong Wu. "Novel Neural Network Application for Bacterial Colony Classification." *Theoretical Biology and Medical Modelling* 15, no. 1 (December 2, 2018): 22. https://doi.org/10.1186/s12976-018-0093-x.
11. Khamaru, Ananda, Sunil Karforma, and Soumendranath Chatterjee. "Intelligent Neural Network for Bacteria Classification: An Innovation in Artificial Neural Network" 8, no. 11 (2019): 10.
12. Kim, Eun-Hye, Seunghoon Kim, Hyun-Joo Kim, Hyoung-oh Jeong, Jaewoong Lee, Jinho Jang, Ji-Young Joo, et al. "Prediction of Chronic Periodontitis Severity Using Machine Learning Models Based On Salivary Bacterial Copy Number." *Frontiers in Cellular and Infection Microbiology* 10 (2020). https://doi.org/10.3389/fcimb.2020.571515.
13. Kotsiantis, S B. "Supervised Machine Learning: A Review of Classification Techniques," n.d., 20.
14. Lancashire, L., O. Schmid, H. Shah, and G. Ball. "Classification of Bacterial Species from Proteomic Data Using Combinatorial Approaches Incorporating Artificial Neural Networks, Cluster Analysis and Principal Components Analysis." *Bioinformatics (Oxford, England)* 21, no. 10 (May 15, 2005): 2191–99. https://doi.org/10.1093/bioinformatics/bti368.
15. M L, Chayadevi, and Gt Raju. "Data Mining, Classification and Clustering with Morphological Features of Microbes." *International Journal of Computer Applications* 52 (August 1, 2012): 1–5.

https://doi.org/10.5120/8187-1547.
16. Madhulatha, T. Soni. "An Overview on Clustering Methods." *ArXiv:1205.1117 [Cs]*, May 5, 2012. http://arxiv.org/abs/1205.1117.
17. Mishra, Ajay Kumar, and Bikram Kesari Ratha. "Study of Random Tree and Random Forest Data Mining Algorithms for Microarray Data Analysis," 2016, 3.
18. Montavon, Grégoire, Wojciech Samek, and Klaus-Robert Müller. "Methods for Interpreting and Understanding Deep Neural Networks." *Digital Signal Processing* 73 (February 2018): 1–15. https://doi.org/10.1016/j.dsp.2017.10.011.
19. Nantasenamat, Chanin, Likit Preeyanon, Chartchalerm Isarankura-Na-Ayudhya, and Virapong Prachayasittikul. "PyBact: An Algorithm for Bacterial Identification." *EXCLI Journal* 10 (2011): 240–45.
20. Nurlaila, Ika, Wahyu Irawati, Kartika Purwandari, and Bens Pardamean. "K-Means Clustering Model to Discriminate Copper-Resistant Bacteria as Bioremediation Agents." *Procedia Computer Science*, 5th International Conference on Computer Science and Computational Intelligence 2020, 179 (January 1, 2021): 804–12. https://doi.org/10.1016/j.procs.2021.01.068.
21. Patil, Pritee Chunarkar, Pradnya Suresh Panchal, Shweta Madiwale, and Vidya Sunil Tale. "An Analysis of Non-Cultivable Bacteria Using WEKA." *Bioinformation* 16, no. 8 (August 31, 2020): 620–24. https://doi.org/10.6026/97320630016620.
22. Peiffer-Smadja, N., S. Dellière, C. Rodriguez, G. Birgand, F.-X. Lescure, S. Fourati, and E. Ruppé. "Machine Learning in the Clinical Microbiology Laboratory: Has the Time Come for Routine Practice?" *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 26, no. 10 (October 2020): 1300–1309. https://doi.org/10.1016/j.cmi.2020.02.006.
23. Pircher, Thomas, Bianca Pircher, Eberhard Schlücker, and Andreas Feigenspan. "The Structure Dilemma in Biological and Artificial Neural Networks." *Scientific Reports* 11, no. 1 (March 10, 2021): 5621. https://doi.org/10.1038/s41598-021-84813-6.
24. Qu, Kaiyang, Fei Guo, Xiangrong Liu, Yuan Lin, and Quan Zou. "Application of Machine Learning in Microbiology." *Frontiers in Microbiology* 10 (2019). https://doi.org/10.3389/fmicb.2019.00827.

———. "Application of Machine Learning in Microbiology." *Frontiers in Microbiology* 10 (April 18, 2019): 827. https://doi.org/10.3389/fmicb.2019.00827.

25. Sarkar, Archana, and Prashant Pandey. "River Water Quality Modelling Using Artificial Neural Network Technique." *Aquatic Procedia*, INTERNATIONAL CONFERENCE ON WATER RESOURCES, COASTAL AND OCEAN ENGINEERING (ICWRCOE'15), 4 (January 1, 2015): 1070–77. https://doi.org/10.1016/j.aqpro.2015.02.135.
26. **Schaefer, J., Lehne, M., Schepers, J. *et al.* "The use of machine learning in rare diseases: a scoping review". *Orphanet J Rare Dis* 15, 145 (2020). https://doi.org/10.1186/s13023-020-01424-6**
27. Soni, Rajesh. "Visualization of Behavioral Model Using WEKA" 3, no. 3 (2013): 4.
28. **Sriavastava, Santosh Kumar, Dr Yogesh Kumar Sharma, and Dr Sheo Kumar. "Using Of WEKA Tool In Machine Learning: A Review." *International Journal of Advanced Science and Technology* 29, no. 08 (May 15, 2020): 4456–66.**
29. **Weis, C. V., C. R. Jutzeler, and K. Borgwardt. "Machine Learning for Microbial Identification and Antimicrobial Susceptibility Testing on MALDI-TOF Mass Spectra: A Systematic Review." *Clinical Microbiology and Infection* 26, no. 10 (October 1, 2020): 1310–17. https://doi.org/10.1016/j.cmi.2020.03.014.**
30. Wilson-Welder, J. H., Elliot, M. K., Zuerner, R. L. *et al*. "Biochemical and molecular characterization of Treponemes phagedensis-like spirochete isolated from a bovine digital dermatitis lesion". *BMC Microbiol* 13, 280 (2013).
31. **Yulin Song , Vincent Hervé , Renate Radek , Fabienne Pfeiffer, Hao Zheng and Andreas Brune. "Characterization and phylogenomic analysis of Breznakiella homolactica gen. nov. sp. nov. indicate that termite gut treponemes evolved from non-acetogenic spirochetes in cockroaches". *Environmental Microbiology* (2021) 23(8), 4228–4245**