IJHSR
ijhighschoolresearch.org
■ RESEARCH ARTICLE

Machine Learning Analysis of Microbiome Data to Predict Cocaine Addiction

Anthony B Song

Edwin O Smith High School, Mansfield-Storrs, CT 06269, USA; anthony.b.song@gmail.com
Mentors: Dr. George Weinstock, Dr. Dong-Binh Tran

ABSTRACT: Cocaine use disorder is a significant public health problem in the US and worldwide. The microbiome may influence behavioral response to cocaine via gut-brain interactions. Microbial and behavioral variables are measured on mice to test whether differences in the microbiome account for the inter-subject variation in cocaine use. However, complex data requires sophisticated data science methods to analyze effectively. This study aimed to take a novel machine learning approach to first reduce data dimension, then cluster mice according to their cocaine use patterns, and finally classify mice into the resultant clusters based on microbial features. Both linear and nonlinear dimensionality reduction methods were employed and compared to generate new coordinates of cocaine use behavior. Based on the best coordinates, K-means identified three clusters of mice: mice that became addicted quicker and administered more cocaine; mice that used fewer doses in response to different dosages; mice that required much more infusion sessions to acquire addiction eventually. An artificial neural network was used to differentiate these behavioral groups based on the abundance of 57 bacterial genera produced micro- and macro-averaged AUC of 0.73 and 0.67, respectively. Model interpretation of this network helped identify risk and protective factors. An increased abundance of *Escherichia shigella* and *Enterococcus* was associated with high-risk mice, whereas an increased abundance of *Akkemensia* and *Erysipelotrichaceae incertae sedis* was shown to be protective factors.

KEYWORDS: Microbiology; Applied Microbiology, Gut Bacteria, Machine Learning, Cocaine Addiction.

■ Introduction

The microbiome plays a key role in human health and is a predictor of various diseases.[1] Substance use and addiction are major public health problems in the United States and worldwide.[2] Particularly, cocaine abuse is associated with substantial morbidity and mortality, and cocaine overdose deaths are increasing, and in certain populations, they outnumber deaths due to heroin and opiate overdose.[3] To date, studies of the neural basis underlying substance use disorders have focused on the neurobiology of reward processing, cognitive control, and emotion regulation in the central nervous system (the brain and spinal cord).[4, 5] An emerging research direction starts to investigate the microbiome's participation in drug reward via the gut-brain axis.[6, 7] The gut-brain axis (GBA) is the bidirectional communication between the central nervous system and the virus and bacteria within a host's gut and the enteric nervous system.[8] There has been a critical gap in data science research to integrate gut microbiome data with substance use behavior to understand which microbes are associated with the development and severity of substance addiction.[9] Understanding the relationship between gut microbiota and cocaine use behavior will help explain the biological mechanism of cocaine addiction.

Although the gut and brain are separate organs, they communicate via trillions of intestinal bacteria that collectively make up the gut microbiome. Findings from humans and animals support the theory that gut microbes are critical in regulating brain function and mood.[10] Cocaine addiction reflects dysregulation of motivational, reward, and stress circuits.[11] The gut microbiome influences the same neural circuitry, suggesting gut-brain interactions in cocaine use disorders. For instance, in human studies, significantly different gut bacterial microbiomes between cocaine users and non-cocaine users were found.[12,13] Alterations of the gut microbiome also affect behavioral responses to cocaine in mice.[14] Changes in the gut microbiome and its metabolites may not only be a consequence of cocaine use but may, in fact, participate in mediating behavioral responses to cocaine.[6] Cocaine use disorder is characterized by highly heterogeneous behaviors and symptoms.[15] Microbiome opens a new pathway to examine the individual variation in cocaine use. Mice are an attractive animal model for this study because many variables can be controlled in their experiments, such as environment, diet, and lifestyle. These variables are confounders in humans.

Limited analytic tools have been available to jointly analyze the multivariate data at the microbiomic and behavior levels. Correlation and association analysis is widely used to examine concurrent patterns between gut dysbiosis and substance use.[16] However, it can be challenging to co-analyze the different data types, such as binary behavioral indicators versus real-valued microbial parameters that measure diversity and stability. Directly merging the different types of data may not maintain data integrity. Alpha and beta diversity measure, respectively, the diversity of bacteria within one sample (a mouse) and the dissimilarity between two samples (two mice), but these diversity parameters reflect the overall inter-host difference.[17] Thus, analyzing diversity parameters cannot capture the effects of individual bacteria on a host's behavior, such as cocaine use.

DOI: 10.36838/v5i5.4

Further, microbial and behavioral data can be complex and highly dimensional, requiring dimension reduction to reveal essential data structures and relationships. Deep learning has become a favored method to model complex data.[18,19] In the microbiome study field, it has been used to mainly learn representations of microbial features, such as deepMicro,[20] a set of autoencoders used to represent microbiome data by lower-dimensional vectors. Deep learning models can also be trained to predict disease as a function of microbiome variables.[21] Still, the first question is often to determine which disease indicators and symptoms are the prediction target when the disease is multi-faceted. It can be important to design a deep learning-based framework that reduces microbiome data dimension and simultaneously predicts heterogeneous animal or human behavior.

At the Jackson Laboratory for Genomic Medicine (JAX), diversity outbred (DO) mice were assessed with their cocaine use behavior where 16S rRNA gene sequencing was performed on their feces to gather microbial features.[22] The resultant data were analyzed in this study to examine whether bacteria present in feces are associated with different cocaine use behavior and the risk of developing an addiction. DO mice can best mimic the human populations to generate a random assortment of genetics. The dataset contains counts of fecal bacteria copies in specific genera which correspond to microbial features, and nine cocaine self-administration (using catheters) behavioral parameters. These microbial features and behavioral parameters characterize each mouse. In this work, a pipeline of analytic steps is designed to first reduce the dimension of behavioral data, which helps reveal clusters of mice with more homogeneous behavioral patterns. Then microbial features are used in a classifier to predict the different clusters. Linear dimension reduction and nonlinear dimension reduction methods are used to reveal the essential dimension of the cocaine use behavior, followed by the K-means clustering method to identify clusters of mice that correspond to a cocaine use index. Then an artificial neural network (ANN)[23] is designed to take the microbial features of a mouse as inputs and calculate through layers of connected ANN nodes to predict the identified cluster assignment of the mouse.

To capture the heterogeneity of cocaine use behavior, three groups of mice were identified, showing different cocaine self-administration patterns. The ANN classifier we constructed based on gut bacteria to separate mice into their corresponding behavior clusters produced an AUC score greater than 0.65. A model interpretation strategy was used to interpret the ANN and found that *Escherichia shigella*, *Enterococcus*, *Clostridium XVIII*, *Akkemensia*, *Erysipelotrichaceae incertae sedis*, and *Fusicatenibacter* were significant in distinguishing cocaine use behavior, and might indicate risk or resilient factors for cocaine addiction.

Methods

Overview of Analytical Pipeline:

Figure 1 illustrates the proposed analytic pipeline. A dimension reduction method first projects the behavioral data into new coordinates. Using the projected data, the K-means clustering method identifies subgroups of mice that exhibit different patterns in cocaine administration. A cluster validity index is employed to evaluate the consistency within the subgroups, which helps choose the best number of clusters. An ANN classifier is then constructed to assign mice to specific subgroups based on their microbial features. The study sample is first split into training and test sets. The ANN is trained on the training set and tested for classification performance on the test set. Although deep neural networks are generally considered black boxes, researchers have explored ways to interpret the networks. A visualization method, class activation map (CAM),[24] is employed to identify the bacterial genera that play critical roles in classifying behavioral groups. This analytic pipeline allows us to better capture the heterogeneity of cocaine addiction and identify the gut bacteria that differentiate cocaine responses. All experiments in the analytical pipeline were implemented using Python version 3.9.7 based on the Scikit-Learn v0.24.2, PyTorch v1.12.0, UMAP v0.5.3, and SciPy v1.7.1 packages.

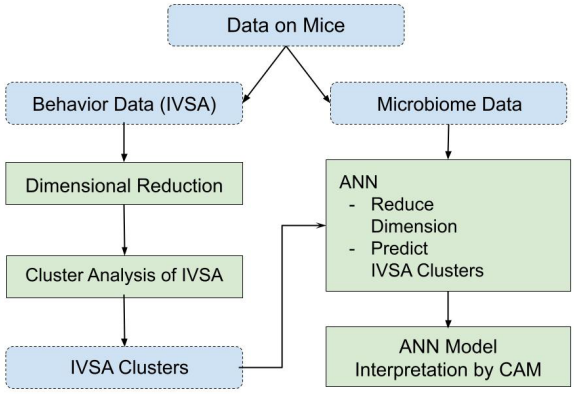

Figure 1: The pipeline of the proposed analytic steps (IVSA: intravenous self-administration of cocaine; ANN: artificial neural network; CAM: class activation map).

Dataset:

DO mice from JAX were derived from eight mouse founders. They were inter-crossed in several generations to create a population of genetically heterogeneous mice, similar to natural selection, such as in human communities.[25] A total of 175 DO mice were used in this study. The 16S ribosomal RNA of the microbes in their feces were sequenced when they were between 8 to 12 weeks. The 16S rRNA gene consists of nine hyper-variable regions (V1 – V9), each exhibiting appreciable sequence diversity among varying bacteria.[26, 27] As no region fully encapsulates all possible bacteria; the mice underwent high-density 16S rRNA sequencing on the V1 – V3 regions of fecal pellets.[28] The sequence reads were already processed at JAX using Shorelinebiome's 16s v1-v3 amplicon-kit and following rigorous protocols.[29] Ninety-six different genera were identified in this mouse population, and the gene read count for each genus was reported. Microbiome data were zero-inflated. The microbial features that exhibited constant zeros for over 99% of the mice were removed since they were too sparse to create stable classifiers. As a result, 57 microbial features - counts of bacteria copies in 57 genera - remained. The Bray-Curtis beta diversity,[30] comparing the dissimilarity

between the microbial composition of one mouse with another was calculated. The heatmap in Figure 2 displays the pairwise beta diversity values (ranging from 0 to 1). A value of 0 means that the two mice, in comparison, shared all genera, while a value of 1 means the two mice did not share any genera in common. Most of the dissimilarity indexes (70%) were in a range from 0.35 to 0.65.

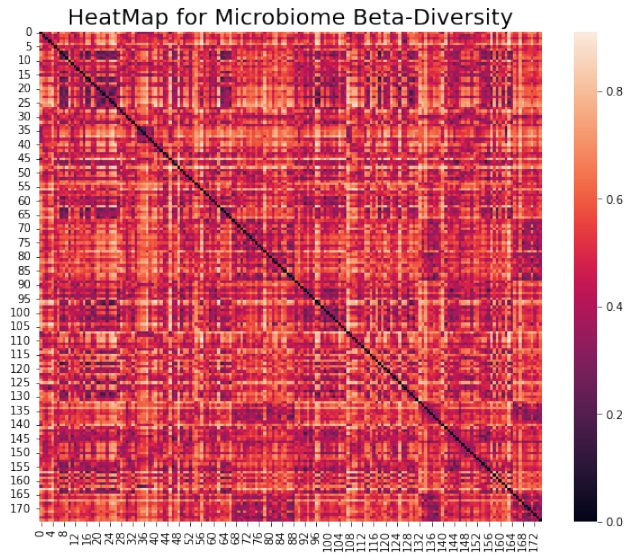

Figure 2: Heatmap of the pair-wise beta diversity of the 175 DO mice.

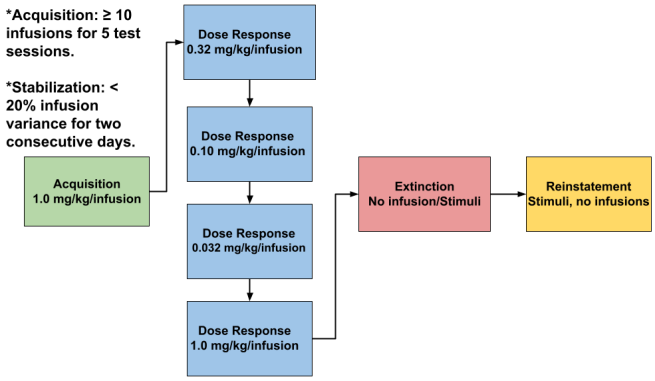

Figure 3: The cocaine IVSA testing steps conducted on JAX DO mice.

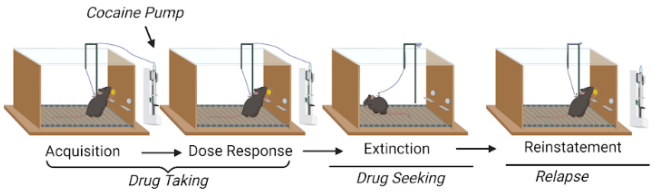

Figure 4: Illustration of the behavioral experiments. This image is directly cited from Tran *et al.* 2022.[29]

These mice were also assessed with their behavior of intravenous self-administration (IVSA) of cocaine after they were 12 weeks or older.[29] Figures 3 and 4 provide a flowchart and illustration of the behavioral experiments, respectively. In the first IVSA experimental sessions, mice pressed an active lever to receive an intravenous infusion of cocaine and an illumination stimulus of lights for five seconds. After an initial acquisition period at a 1.0 mg/kg dosage of cocaine, ≥ 10 infusions for 5 test sessions, mice fulfilled the criterion for stabilization if the number of received infusions through the active lever did not vary by more than 20% for two consecutive sessions. Following the acquisition stage, the mice progressed through the following dosages: 0.32, 0.10, 0.032, and 1.0 mg/kg/infusion. To proceed to the subsequent dosage in the series, the mice had to meet the same stabilization criteria or proceed at the same dosage until five days had passed. This period of varying dosages is known as the Dosage Response stage, and a dosage response curve (DRC) was created to record the number of active lever presses of each mouse at the different dosage response stages. The area under the curve (AUC) for these DRC was also computed to characterize the overall dosage response trend for a mouse. At the conclusion of this phase, an extinction period was executed in which active lever presses did not provide any infusion of cocaine nor activate the stimulus lights. Finally, a reinstatement period concluded the cocaine IVSA assessment, in which active lever presses were recorded, which would result in drug-paired stimuli while still lacking an infusion of cocaine. All these experiments resulted in 9 IVSA features, as listed in Table 1.

Table 1: Description of the nine variables of cocaine intravenous self-administration (IVSA).

Behavioral variable	Description of the variable
B1	Number of sessions taken to meet acquisition criteria during the acquisition stage.
B2	Number of sessions to stabilization for 1 mg/kg dose of cocaine.
B3	AUC calculated based on the numbers of infusions from the B4, B5, B6, and B7.
B4	Number of infusions at 1.0 mg/kg during the dose response stage.
B5	Number of infusions at 0.32 mg/kg during the dose response stage.
B6	Number of infusions at 0.1 mg/kg during the dose response stage.
B7	Number of infusions at 0.032 mg/kg during the dose response stage.
B8	Number of active lever presses during sessions of extinction.
B9	Number of inactive lever presses vs the number of saline active lever presses during reinstatement.

Dimension reduction:

The methods of linear dimension reduction - principal component analysis (PCA)[31] - and nonlinear dimension reduction - uniform manifold approximation and project (UMAP)[32, 33] - were used and compared. PCA maps high-dimensional data into a new coordinate system to display sample patterns. It constructs these uncorrelated coordinates by finding principal components, the linear combinations of original features. The principal components successively maximize the explained data variance. UMAP, on the other hand, first constructs a high-dimensional graph representation that connects each data point to its nearest neighbors within a variable radius. Then UMAP runs an iterative process to identify new coordinates, allowing the same topology and proximity to be preserved in the new coordinate system. In

DOI: 10.36838/v5i5.4

other words, if sample A is closer to sample B than to sample C in the original high dimensional space then the relationship should still hold in the lower dimensional space.

This study employed both methods to project the 9 IVSA features into 2 or 3 coordinates. Then the coordinates from each method were used in a cluster analysis to cluster mice into more homogeneous groups. The appropriate dimension reduction method and the best number of coordinates will be determined according to the quality of the clusters.

Cluster analysis:

Dimension reduction often helps reveal sample grouping structures. The K-means clustering method was used to create K (=2, 3, or 4) clusters of mice. Due to the limited sample size, smaller values of K were considered. Silhouette score was used for each clustering solution to measure how well the clusters were separated.[34] The score equals $(b-a)/\max(a,b)$ where a is the average of the distances between each pair of points within a cluster, and b is the average distance between each pair of clusters. Silhouette score ranges from -1 to 1, where a higher score indicates better quality of a clustering solution. Silhouette score was employed as a metric to compare and select the number of projected coordinates and the number of clusters. Clustering solutions were also visualized by a scatter plot for comparison.

After a clustering solution was selected, a violin plot was generated to visualize each behavioral feature's cluster means, quartiles, and density. The violin plot helped detect plausible patterns in the behavioral clusters.

Classification:

An ANN[35] was created to classify the mice into the identified clusters as a function of 57 microbial features. This ANN first learned a lower dimensional representation for microbial features using one or two hidden layers and then performed classification in the last layer. The ANN used fully connected layers, each consisting of a linear mapping followed by a ReLu activation function, except the output layer. The ReLu activation takes in the output of the linear mapping and calculates a piecewise linear function called the rectified linear activation function. The output layer replaced the ReLu function with the softmax function, which computes the probability of each mouse belonging to a cluster. A mouse was assigned to the cluster for which the ANN reported the highest probability.

A hold-out set was first created to contain 20% of the data sampled with stratification (i.e., the numbers of mice in each cluster in this set were proportional to those in the full sample). This hold-out set was used to evaluate the classification performance of the ANN model. A three-fold cross-validation (CV) process[36] within the remaining 80% (training) data was used to tune the model parameters, such as the number of hidden layers and the number of hidden nodes in each hidden layer, which eventually determined the number of reduced dimensions. During the ANN training, tuning parameters such as the mini-batch size and learning rate were also determined in the process. The training data was split into three even subsets, and each time a model was trained using two subsets and tested on the remaining subset. We used the average validation classification performance to select the best model parameters. The Area Under the receiver operating characteristic Curve (AUC)[37] was used to measure the classification performance. AUC ranges from 0 to 1, and higher values indicate better classification performance. Using the selected model parameters, a final ANN was trained on all training data, and its performance on the hold-out set was reported. Important microbial features were identified by model interpretation of this final ANN.

Model interpretation:

Although ANN is generally considered as a black box, recent research has attempted to interpret its decision-making process. In particular, the class activation map[24, 38] is a powerful technique used in image classification to explain how an image is classified into a specific category. In this study, CAM visualized the cluster label predicted by the ANN for each mouse and the specific gut bacteria critical for the classification. It calculated the gradient of the model prediction by backpropagating the network output back to the input and then multiplied the gradient with the input to produce an importance weight for each microbial input feature. Using each classifier (corresponding to a cluster), an importance weight vector was computed for a mouse. Then for each cluster, the Wilcoxon rank-sum test[39] was used to test whether a microbial feature significantly differentiated a behavioral cluster from the remaining clusters by comparing the importance weights of all mice. Bonferroni Correction was used on all p-values obtained by the Wilcoxon rank-sum test to minimize the chance of a false-positive result. Specifically, we multiplied each p-value by the number of features (57) used in the classification in our experimental results. The top four features identified by the rank sum tests were shown in a bar plot for comparison.

Results and Discussions

Dimension reduction and cluster analysis:

The nine behavioral variables were first normalized to have a mean of 0 and a standard deviation of 1. PCA mapped the nine behavioral variables into 2 or 3 principal components, explaining respectively over 65% or 75% of the total variance in the behavior data. The K-means method was applied to the two sets of reduced data to obtain 2, 3, or 4 clusters, for which Silhouette scores were calculated respectively. These scores were all lower than the counterparts obtained by the nonlinear dimension reduction method UMAP. The K-means method was also used to cluster mice directly according to the nine raw features. Table 2 summarizes and compares the clustering validity scores for the raw data, and the different numbers of projected coordinates and clusters. It shows that the 3-cluster solution based on the three reduced dimensions from UMAP produced the best Silhouette score. Figure 5 shows the scatter plots of the resultant clusters in the respective coordinates calculated by PCA and UMAP for all solutions with 3 clusters, which visually confirms the same best clusters.

Table 2: The silhouette scores were calculated for the clustering solutions obtained by K-means using the reduced data from UMAP, PCA, and Raw Data.

	# of Coordinates	# of Clusters	Silhouette Score
UMAP	2	2	0.62
		3	0.66
		4	0.36
	3	2	0.54
		3	**0.83**
		4	0.66
PCA	2	2	0.39
		3	0.41
		4	0.42
	3	2	0.33
		3	0.33
		4	0.34
Raw	9	2	0.49
		3	0.35
		4	0.23

The three clusters from the selected solution consisted of 89, 53, and 33 mice, respectively. Figure 6 shows the three violin plots each for a cluster. Each violin plot illustrates the cluster mean, quartile, and probability density curve of each of the nine behavioral features. Because all features B1-B9 have been normalized to have zero mean when a behavioral feature has a smaller mean (<0) for a cluster, it indicates that mice in the cluster endorse the specific behavior less than an average mouse does; or otherwise, mice in the cluster endorse the behavior more than an average mouse. For the first and the largest cluster, because B1 and B2 have negative mean values, mice in this cluster used fewer numbers of infusion sessions than the sample average to meet the acquisition criteria and to stabilize on the initial dosage of 1mg/kg, indicating that these mice were quicker and easier to get addicted to cocaine. These mice had relatively larger numbers of infusions during the dosage response (DR) stages. Significantly, the high tails of the violin bars show that this cluster has some mice reaching very high numbers of DR infusions and taking more lever presses during extinction sessions with easier reinstatement. Thus, this cluster corresponds to a "*high-risk*" group of mice. In the second cluster, mice used the average number of sessions to reach acquisition and stabilize. Still, they took less cocaine (i.e., the number of infusions) in response to the different dosages and especially much less for smaller dosages. Hence, this cluster is named the "*low-use*" group. The third cluster shows an opposite trend to the first cluster on variables B1 and B2. Mice in the third cluster needed substantially more sessions to meet acquisition and stabilize on the initial dosage. They acquired average DR infusions but had a smaller number of infusions for a lower dosage of 0.032 mg/kg. Thus, this cluster is overall a "*low-risk*" group.

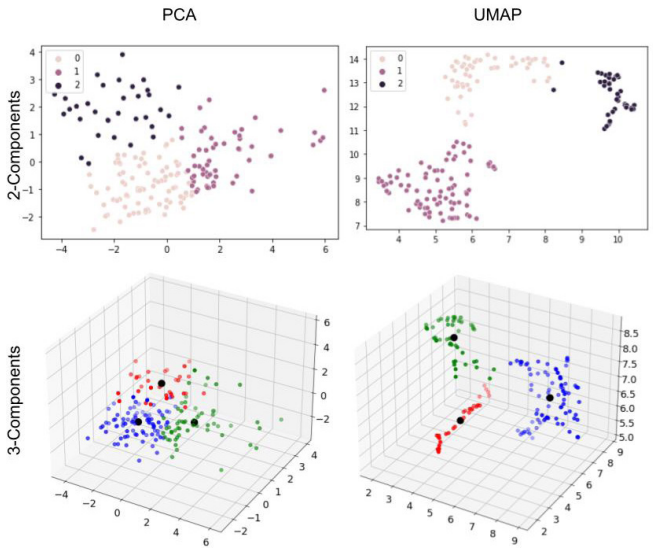

Figure 5: Comparison of clusters (n=3) obtained from the 2 or 3 projected coordinates of PCA and UMAP in scatter plots.

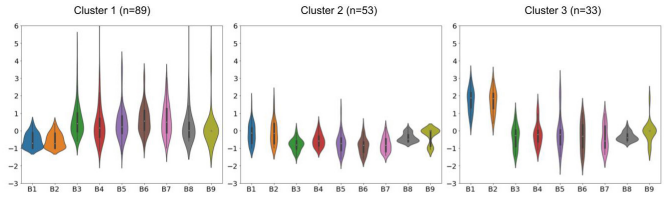

Figure 6: Violin plots of the three clusters show the cluster mean, quartile, and the distribution density of each behavioral feature for the clusters.

Classification and model interpretation:

When training the ANN to classify the mice into the respective high-risk, low-use, and low-risk clusters based on the microbial features, several choices for the number of hidden layers (L=1 or 2) and the number of hidden nodes were compared to determine the ANN architecture. Because dimension reduction was also a study target when training the ANN, the number of hidden nodes was set to be smaller than the number of microbial features. When a single hidden layer was employed, 3, 5, and 8 hidden nodes were tested. When two hidden layers were used, a gradual dimension reduction was desirable, so the first hidden layer had more hidden nodes than the second hidden layer. Table 3 compares the different choices in terms of the AUC values, each of which was calculated by averaging over all three clusters and all three cross-validation splits. From this table, the choice of 2 hidden layers with 20 and 5 hidden nodes in the first and second layers produced the best classifier with an average AUC of 0.65, which was then selected. During the tuning of the model parameters, mini-batch sizes of 30, 50, and 60 and learning rates of 0.01 and 0.001 were also tested. For the selected ANN model, mini-batch size = 60 and learning rate = 0.01 were the best choices.

DOI: 10.36838/v5i5.4

Table 3: Comparison of average AUC values for the different choices of hidden layers and hidden nodes (where L denotes the number of hidden layers, and H1 and H2 represent the number of hidden nodes in the first and second hidden layers, respectively).

# of hidden layers	# of hidden nodes		Average AUC
L = 1	H1 = 3		0.50
	H1 = 5		0.56
	H1 = 8		0.53
L = 2	H1 = 10	H2 = 3	0.51
		H2 = 5	0.58
		H2 = 8	0.57
	H1 = 20	H2 = 3	0.51
		H2 = 5	**0.65**
		H2 = 8	0.60

The full training set was then utilized for training an ANN model with the selected parameter setting. The resultant model was tested on the hold-out set, producing AUC values of 0.66 (in classifying the high-risk mice from the rest), 0.72 (low-use mice versus the rest), and 0.59 (low-risk mice versus the rest), and the micro- and macro-averaged AUC over the three clusters of 0.73 and 0.67, respectively. The classification performance for separating the low-risk mice from the rest was the worst, partly due to the small cluster size (n=33). Figure 7 shows the test ROC curves for the three class outputs of the ANN model and the related micro- and macro-average ROC curves. These results provide evidence that variation in the abundance of gut bacteria can explain to some level the difference in the behavioral response to cocaine because the microbial features show predictive power to the different behavioral clusters.

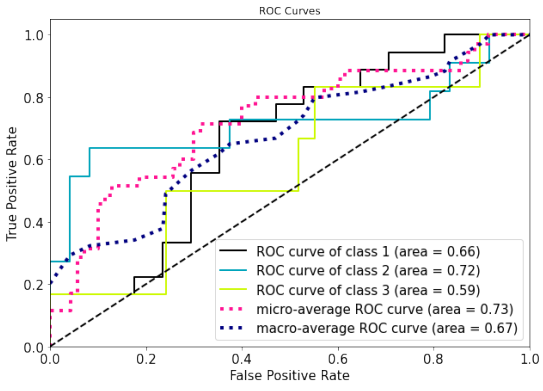

Figure 7: Receiver operating characteristics (ROC) curves for each behavioral cluster and the related macro- and micro-ROC curves.

For every mouse that this ANN correctly classified, a CAM was computed using the backpropagation of the ANN output with respect to the input features so one could evaluate which features were responsible for the correct class assignment. Then these CAMs were gathered for each cluster to apply the Wilcoxon rank sum test for each microbe genus. This process identified quite a few microbial features that showed significant statistical power in distinguishing one cluster from the rest in a nonlinear interplay with other features specified by the ANN. Figure 8 lists the top four microbial features identified for each cluster. For the high-risk cluster, the strongest features identified had p-values reaching 10^{-8} (Bonferroni corrected) For the low-risk cluster, the significant features had p-values reaching 10^{-6} (Bonferroni corrected). A conjecture was that the low-use cluster might be positioned in between the *high-risk* and *low-risk* clusters, so identifying features that differentiate the *low-use* mice from both the *high-risk* and *low-risk* mice did not produce significant Bonferroni corrected p-values (the minimal p-value was ~0.8).

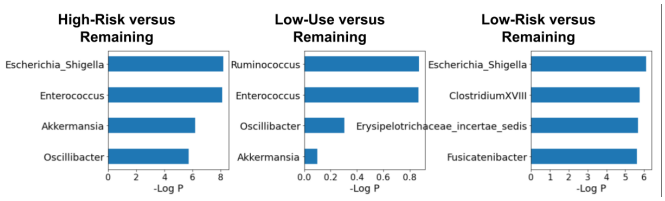

Figure 8: Top microbial features identified by applying the Wilcoxon rank sum test to CAMs are useful in differentiating mice in one cluster from those in the remaining clusters.

The identified significant features were further examined with their cluster means to study how they help separate a cluster of mice from the remaining mice. Figure 9 shows three groups of bar plots. Microbial features were also normalized to have zero mean. The bacterial genus *Escherichia shigella* was identified as the most significant feature in separating the high-risk mice from the remaining mice and the low-risk mice from the rest. This feature exhibited clearly increased abundance for the *high-risk* mice on average. Its decreased abundance was significant for the *low-risk* mice, so it was the most useful feature to discriminate between these two groups of mice. Overall, an increased abundance of *Escherichia shigella*, *Enterococcus*, and *Oscillibacter*, as well as a decreased abundance of *Akkermansia* helped differentiate *high-risk* mice from other mice. A decreased abundance of *Escherichia shigella*, *Fusicatenbacter*, and *Clostridium XVIII* but an increased abundance of *Erysipelotrichaceae incertae sedis* were significantly associated with the low-risk mice.

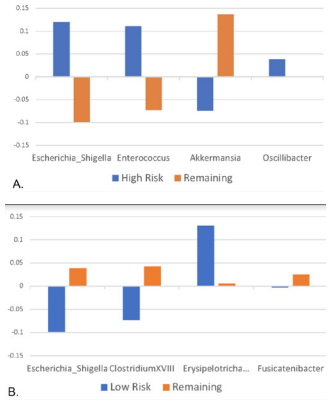

Figure 9: Grouped bar plots to compare the cluster means of the microbial features identified by the ANN. Features differentiating *high-risk* and *low-* respectively, from the remaining mice, are shown in graphs A and B. *Erysipelotricha…* corresponds to *Erysipelotrichaceae incertae sedis.*

A dataset significantly overlapping with the one used in the present work was analyzed in a prior study[29] where the dataset was clustered into two groups: mice that acquired IVSA (addiction) and those that did not acquire IVSA (non-addiction). Standard hypothesis tests, Wilcoxon and Kruskal-Wallis rank sum tests,[40] were used in that study to test whether the microbiome and which specific microbiota affect addiction-predisposing behaviors and cocaine intravenous self-administration. The study identified that an increased abundance of *Barnesiella*, *Ruminococcus*, and *Clostridium IV* and decreased abundance of Robinsoniella differentiated addiction mice from non-addiction mice. The present study identified another genus in the Clostridium family (*Clostridium XVIII*) with an increased abundance in high-risk mice, which confirms one of the findings in the earlier study. The other identified bacteria genera were not reported previously and may warrant a future examination using separate samples.

Conclusion

In this study, an analytic pipeline, consisting of three major components: dimension reduction, cluster analysis, and classification, has been proposed to model the relationship between microbiome and cocaine use behavior. Linear and non-linear dimension reduction methods were tested so that the K-means method could find clusters of mice based on their behavioral patterns. The non-linear dimension reduction method, UMAP, was shown to achieve better silhouette cluster validity scores in this example for creating homogeneous groups with scores much higher than those based on principal components. The K-means algorithm found three clusters, characterizing mice with increased risk for addiction, low cocaine use, and low addiction risk, respectively. While it is possible to create a neural network based on the microbiome to directly predict the behavioral variables, because of the high heterogeneity in cocaine use behavior, it would be difficult to interpret which microbe is related to the risk of addiction. Instead, the proposed pipeline allows a refined classification of cocaine use patterns to identify the gut bacteria that potentially relate to the classification. A neural network with fully connected layers was tuned and trained to classify mice into one of the three categories based on the abundance of 57 bacterial genera and produced high AUC values, which justifies evaluating the critical features used in this neural network. Using a model interpretation strategy, it was found that an increased abundance of *Escherichia shigella*, *Enterococcus*, *Oscillibacter*, *Fusicatenbicater*, *Clostridium XVIII*, and decreased abundance of *Akkemensia* and *Erysipelotrichaceae incertae sedis* were risk factors. Future research directions may expand the analyses to human samples and/or include more microbial species in the study. It may be useful to handle the zero inflation commonly encountered in microbial features rather than removing sparse features, as done in this work. Other possible studies could be to analyze data collected on an experiment testing if an increase or decrease in some microbial genera affects the mice behaviors and examine whether exposure to cocaine would change the microbiome.

Acknowledgments

Anthony B Song would like to thank Drs. George Weinstock and Dong-Binh Tran for their guidance during his research and for providing him with the mouse data. Anthony also acknowledges the Academic Year Fellowship awarded by the Jackson Laboratory for Genomic Medicine during this research.

References

1. Consortium, H. M. P., Structure, function and diversity of the healthy human microbiome. *Nature* **2012**, *486* (7402), 207-214.
2. SAMHSA, S. A. a. M. H. S. A., Behavioral health trends in the United States: Results from the **2014** national survey on drug use and health. 2014, http://www.samhsa.gov/data/sites/default/files/NSDUH-FRR1-2014/NSDUH-FRR1-2014.pdf.
3. Kampman, K. M., The treatment of cocaine use disorder. *Science advances **2019***, *5* (10), eaax1532.
4. Goldstein, R. Z.; Volkow, N. D., Drug addiction and its underlying neurobiological basis: Neuroimaging evidence for the involvement of the frontal cortex. *American Journal of Psychiatry* **2002**, *159* (10), 1642-1652.
5. Li, C.-s. R.; Sinha, R., Inhibitory control and emotional stress regulation: Neuroimaging evidence for frontal–limbic dysfunction in psycho-stimulant addiction. *Neuroscience & Biobehavioral Reviews* **2008**, *32* (3), 581-597.
6. Meckel, K. R.; Kiraly, D. D., A potential role for the gut microbiome in substance use disorders. *Psychopharmacology* **2019**, *236* (5), 1513-1530.
7. Russell, J. T.; Zhou, Y.; Weinstock, G. M.; Bubier, J. A., The Gut Microbiome and Substance Use Disorder. *Frontiers in Neuroscience* **2021**, 1134.
8. Mayer, E. A.; Tillisch, K.; Gupta, A., Gut/brain axis and the microbiota. *The Journal of clinical investigation* **2015**, *125* (3), 926-938.
9. Babor, T. F.; Caetano, R., Subtypes of substance dependence and abuse: implications for diagnostic classification and empirical research. *Addiction* **2006**, *101* (Suppl. 1), 104-10.
10. Moloney, R. D.; Desbonnet, L.; Clarke, G.; Dinan, T. G.; Cryan, J. F., The microbiome: stress, health and disease. *Mammalian genome* **2014**, *25* (1), 49-74.
11. Nestler, E. J., The neurobiology of cocaine addiction. *Science & practice perspectives* **2005**, *3* (1), 4.
12. Volpe, G. E.; Ward, H.; Mwamburi, M.; Dinh, D.; Bhalchandra, S.; Wanke, C.; Kane, A. V., Associations of cocaine use and HIV infection with the intestinal microbiota, microbial translocation, and inflammation. *Journal of studies on alcohol and drugs* **2014**, *75 (2*), 347-357.
13. Xu, Y.; Xie, Z.; Wang, H.; Shen, Z.; Guo, Y.; Gao, Y.; Chen, X.; Wu, Q.; Li, X.; Wang, K., Bacterial diversity of intestinal microbiota in patients with substance use disorders revealed by 16S rRNA gene deep sequencing. *Scientific reports* **2017**, *7* (1), 1-9.
14. Kiraly, D. D.; Walker, D. M.; Calipari, E. S.; Labonte, B.; Issler, O.; Pena, C. J.; Ribeiro, E. A.; Russo, S. J.; Nestler, E. J., Alterations of the host microbiome affect behavioral responses to cocaine. *Scientific reports* **2016**, *6* (1), 1-12.
15. Sun, J.; Kranzler, H. R.; Gelernter, J.; Bi, J., A genome-wide association study of cocaine use disorder accounting for phenotypic heterogeneity and gene–environment interaction. *Journal of Psychiatry and Neuroscience* **2020**, *45* (1), 34-44. PMC6919916.
16. Morris, A.; Beck, J. M.; Schloss, P. D.; Campbell, T. B.; Crothers, K.; Curtis, J. L.; Flores, S. C.; Fontenot, A. P.; Ghedin, E.; Huang, L., Comparison of the respiratory microbiome in healthy nonsmokers and smokers. *American journal of respiratory and critical care medicine* **2013**, *187* (10), 1067-1075.

DOI: 10.36838/v5i5.4

17. Lin, S. W.; Freedman, N. D.; Shi, J.; Gail, M. H.; Vogtmann, E. ; Yu, G.; Klepac-Ceraj, V.; Paster, B. J.; Dye, B. A.; Wang, G. Q., Beta-diversity metrics of the upper digestive tract microbiome are associated with body mass index. *Obesity* **2015**, *23* (4), 862-869.
18. Sharma, D.; Xu, W., phyLoSTM: a novel deep learning model on disease prediction from longitudinal microbiome data. *Bioinformatics* **2021**, *37* (21), 3707-3714.
19. Namkung, J., Machine learning methods for microbiome studies. *Journal of Microbiology* **2020**, *58* (3), 206-216.
20. Oh, M.; Zhang, L., DeepMicro: deep representation learning for disease prediction based on microbiome data. *Scientific reports* **2020**, *10* (1), 1-9.
21. Marcos-Zambrano, L. J.; Karaduzovic-Hadziabdic, K.; Loncar Turukalo, T.; Przymus, P.; Trajkovik, V.; Aasmets, O.; Berland, M.; Gruca, A.; Hasic, J.; Hron, K., Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. *Frontiers in microbiology* **2021**, 313.
22. The Jackson Laboratory, Standard Operating Procedure-Intravenous Self-administration Paradigm (CSNA IVSA v2.0). *Center for Systems Neurogenetics of Addiction - Behavioral* **2019**, https://www.jax.org/-/media/jaxweb/files/research-and-faculty/tools-and-resources/system-neurogenetics/intravenous-self-administration-ivsa-paradigm.pdf?la=en&hash=FA64135F219C7DF65937A1CF9270301B0E771836.
23. Goodfellow, I.; Bengio, Y.; Courville, A., *Deep learning*. MIT press: 2016.
24. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. In Learning deep features for discriminative localization, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016; pp 2921-2929.
25. Dickson, P. E.; Ndukum, J.; Wilcox, T.; Clark, J.; Roy, B.; Zhang, L.; Li, Y.; Lin, D.-T.; Chesler, E. J., Association of novelty-related behaviors and intravenous cocaine self-administration in Diversity Outbred mice. *Psychopharmacology* **2015**, *232* (6), 1011-1024.
26. Soriano-Lerma, A.; Pérez-Carrasco, V.; Sánchez-Marañón, M.; Ortiz-González, M.; Sánchez-Martín, V.; Gijón, J.; Navarro-Mari, J. M.; García-Salcedo, J. A.; Soriano, M., Influence of 16S rRNA target region on the outcome of microbiome studies in soil and saliva samples. *Scientific reports* **2020**, *10* (1), 1-13.
27. Chakravorty, S.; Helb, D.; Burday, M.; Connell, N.; Alland, D., A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *Journal of microbiological methods* **2007**, *69* (2), 330-339.
28. Servick, K., Of mice and microbes. *Science* **2016**, *353* (6301), 741-743.
29. Tran, T. D. B.; Nguyen, H.; Sodergren, E.; Dickson, P. E.; Wright, S. N.; Philip, V. M.; Weinstock, G. M.; Chesler, E. J.; Zhou, Y.; Bubier, J. A., Microbial glutamate metabolism predicts intravenous cocaine self-administration in Diversity Outbred mice. *Neuropharmacology* **2022**, 109409.
30. Beals, E. W., Bray-Curtis ordination: an effective strategy for analysis of multivariate ecological data. *In Advances in ecological research*, Elsevier: 1984; Vol. 14, pp 1-55.
31. Bro, R.; Smilde, A. K., Principal component analysis. *Analytical methods* **2014**, *6* (9), 2812-2831.
32. Armstrong, G.; Martino, C.; Rahman, G.; Gonzalez, A.; Vázquez-Baeza, Y.; Mishne, G.; Knight, R., Uniform manifold approximation and projection (UMAP) reveals composite patterns and resolves visualization artifacts in microbiome data. *Msystems* **2021**, *6* (5), e00691-21.
33. McInnes, L.; Healy, J.; Melville, J., Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv*:1802.03426 **2018**.
34. Shahapure, K. R.; Nicholas, C. *In Cluster quality analysis using silhouette score*, **2020** IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), IEEE: 2020; pp 747-748.
35. LeCun, Y.; Bengio, Y.; Hinton, G., Deep learning. *nature* **2015**, *521* (7553), 436-444.
36. Refaeilzadeh, P.; Tang, L.; Liu, H., Cross-validation. *Encyclopedia of database systems* **2009**, *5*, 532-538.
37. Fawcett, T., An introduction to ROC analysis. *Pattern Recognition Letters* **2006**, *27* (8), 861-874.
38. Zhu, L.; She, Q.; Chen, Q.; Meng, X.; Geng, M.; Jin, L.; Jiang, Z.; Qiu, B.; You, Y.; Zhang, Y., Background-aware Classification Activation Map for Weakly Supervised Object Localization. *arXiv preprint arXiv*:2112.14379 **2021**.
39. Hogg, R. V.; Tanis, E. A.; Zimmerman, D. L., *Probability and statistical inference*. Pearson/Prentice Hall Upper Saddle River, NJ,USA: **2010**.
40. Ostertagova, E.; Ostertag, O.; Kováč, J. *In Methodology and application of the Kruskal-Wallis test*, Applied mechanics and materials, Trans Tech Publ: **2014**; pp 115-120.

Author

Anthony Song was a senior at Edwin O Smith High School. He was the founder of the coding club in his school. His research proposal to the Beamline for School competition was ranked top 2 among the US teams in 2021. His work on autonomous vehicle was selected for presentation at an ACM/IEEE conference in 2022. He is passionate about computer science, and will pursue a college education in machine learning and its biomedical or engineering applications.