Food Data Analysis using Multidimensional Visualizations based on Point Placement

Maria Eduarda M. de Holanda
Dep. de Ciência da Computação
Universidade de Brasília
Brasília, DF, Brazil
eduarda.holanda@aluno.unb.br

Bernardo Romão
Departamento de Nutrição
Universidade de Brasília
Brasília, DF, Brazil
bernardo.lima@aluno.unb.br

Raquel Braz Assunção Botelho
Departamento de Nutrição
Universidade de Brasília
Brasília, DF, Brazil
raquelbotelho@unb.br

Renata Puppin Zandonadi
Departamento de Nutrição
Universidade de Brasília
Brasília, DF, Brazil
renatapz@unb.br

Vinícius R. P. Borges
Dep. de Ciência da Computação
Universidade de Brasília
Brasília, DF, Brazil
viniciusrpb@unb.br

Abstract—**Food data comprise records regarding nutrients, ingredients, amounts of different vitamins and minerals that can be found in foods. The wide variety of food products that can be stored in large datasets makes the traditional analysis tasks unfeasible and time-consuming when conducted manually by the dietitians and related professionals. This paper describes a method for visualizing food data using point placement strategies to support specialists in tasks related to determining similar food products that can be replaced in specific diets. The proposed method generates a structured representation for food data to be used as input to some state-of-the-art and recent visualizations, such as PCA, t-SNE, UMAP and TriMap. Experiments were conducted to assess the quality of visualizations and the results reported that the nonlinear visualizations presented satisfactory discriminability regarding some food categories and better preservation of the data patterns. A case study based on a visual exploration process was also conducted and demonstrates the specialist successfully finding substitute food products for planning a vegan diet plan.**

Index Terms—**food composition data, data visualization, visual data mining, point placement strategies**

I. Introduction

The wide diversity of foods and its constituent compounds (nutrients and ingredients) bring several challenges to dietitians, nutrologist physicians and health-related professionals, here called the domain specialists. However, the large number of food products that may present several compounds makes the task of analyzing such amount of data infeasible when performed manually by the specialists. In this sense, approaches based on machine learning and natural language processing naturally appear as potential strategies for knowledge discovery on food data that might support the specialist in those tasks [1].

In the last two decades, several researches have employed data mining strategies for identifying relevant patterns and useful information from food datasets [2]. Morales-Garzón et al. [3] proposed an unsupervised algorithm for adapting ingredient recipes to user needs and preferences through a method based on word embeddings. As well, Matej Petkovi'c et al. [4] designed a AI workflow methodology named *DietHub* to annotate and classify recipes with the food concepts related to them, divided into representation learning and two predictive modeling tasks, classification and hierarchical multi-label classification.

Although the proposed methods were effective to support the specialists in several tasks, interpreting the results produced by machine learning techniques can not be simple and intuitive, as well as to understand the learning process of the associated mathematical models. This scenario is suitable to consider visualization techniques, which aims at generating intuitive graphical representations of data so that the specialists can interactively explore and gain insight of the underlying data by means of visual analysis [5].

To the best of our knowledge, there is little research regarding the use of point placement visualization techniques to analyze food data. This motivated us to proposed a method to study and explore point placement strategies for visualizing food composition data, in which also includes data preprocessing step for generating a structured representation of food data instances. One dataset were considered for the experiments, which employed well-known metrics for assessing the quality of the visualizations. Moreover, we demonstrate a case study regarding the identification of food substitutes based on their ingredients and nutrients.

The main contributions of this paper are enumerated as follows:

- An interactive approach that retains user control for visualizing food composition data;
- A structured representation of food data to be used as input to the visualizations based on point placement;
- A comparative study exploring state-of-the-art point placement strategies: Principal Component Analysis (PCA) [6], t-Distributed Stochastic Neighbor Embedding (t-SNE) [7], Uniform Manifold Approximation and Projection (UMAP) [8] and TriMap [9].

This paper is structured as follows. Section II details the proposed method and its constituting steps. Section III describes the experimental results that were conducted on a vegan food dataset and using qualitative evaluation metrics. Section IV concludes this paper and discusses possibilities for future work.

II. Proposed Method

The proposed method is divided into four steps: dataset definition, preprocessing, visualization based on point placement and evaluation of the visualizations quality as shown in Figure 1. Each step is detailed in the following subsections.

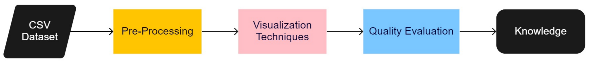


Fig. 1. Flowchart showing the steps illustrating the proposed method for visualizing food component data.

A. Dataset

To design the proposed method, we considered a Vegan Food Composition dataset that presents 277 data instances, described by 26 attributes as shown in Table I. This dataset is useful to analyze the nutrition compounds of vegan products in the Brazilian market, in which the target task is to identify the relations among ingredients to assist dietitians when preparing specific diets.

TABLE I
Description of the Vegan Food Composition Dataset's attributes.

Attributes	Type	Cardinality
Food category	Nominal	1
Product's name	Nominal	1
[Ingredients]	List of Nominals	1
{Nutrients}	Numerical	23

B. Preprocessing

As food datasets can present attributes from different types (numerical and nominal), a preprocessing is required to generate a new structured representation containing only numerical attributes to allow their input to the visualization techniques.

We employed the one-hot encoding approach to transform the attribute "Ingredients" to binary attributes once each food product can present a variable number of values (ingredients). The procedure consists of mapping each value of "Ingredients" to a new binary attribute indicating the presence or absence of the underlying characteristic. Furthermore, the remaining numerical attributes (nutrients) presenting real values were normalized, in which all obtained values for each attribute are in the range $[0, 1]$.

The preprocessing step produces a high dimensional feature vector for each data instance, in which the entire dataset defines a high dimensional space.

C. Visualization based on point placement

The goal of visualizations based on point placement is to map high dimensional objects to lower dimensional points in two-dimensional space in such a way that similar objects are placed by nearby points and dissimilar objects are placed by distant points in the visual space [10]. As a result, the visual analysis of the graphical representation (also known as layout) takes advantage from the human perception system to interpret and identify the global and local patterns of the underlying data according to their similarity relationships. The specialist can view the layout and identify groups of points as well as their visual properties, such as the shape (sparse, dense) and size of those groups.

Four state-of-the-art and recent visualization techniques were considered in the proposed method: Principal Component Analysis (PCA) [6], a statistical technique, reduces data dimensionality by performing a linear mapping of the data in a high dimensional space to a lower dimensional space by taking into account the variances of each attribute and their relations to generate the transformed data; t-Distributed Stochastic Neighbor Embedding (t-SNE) [7] approximates two probability distributions modeled over the pairwise dissimilarities of the objects in the high dimensional space (Gaussian distribution) and the corresponding data points in the visual space (t-Student distribution) using the Kullback–Leibler divergence; Uniform Manifold Approximation and Projection (UMAP) [8] computes a weighted graph based on the similarity relations for each pairwise distance, in which the graph's weights are determined in an optimization process that defines the low dimensional embedding; and TriMap [9] generates the visual space by preserving the distance properties regarding the nearest neighbors principle for each triplet of objects in such a way to retain the global and local structures.

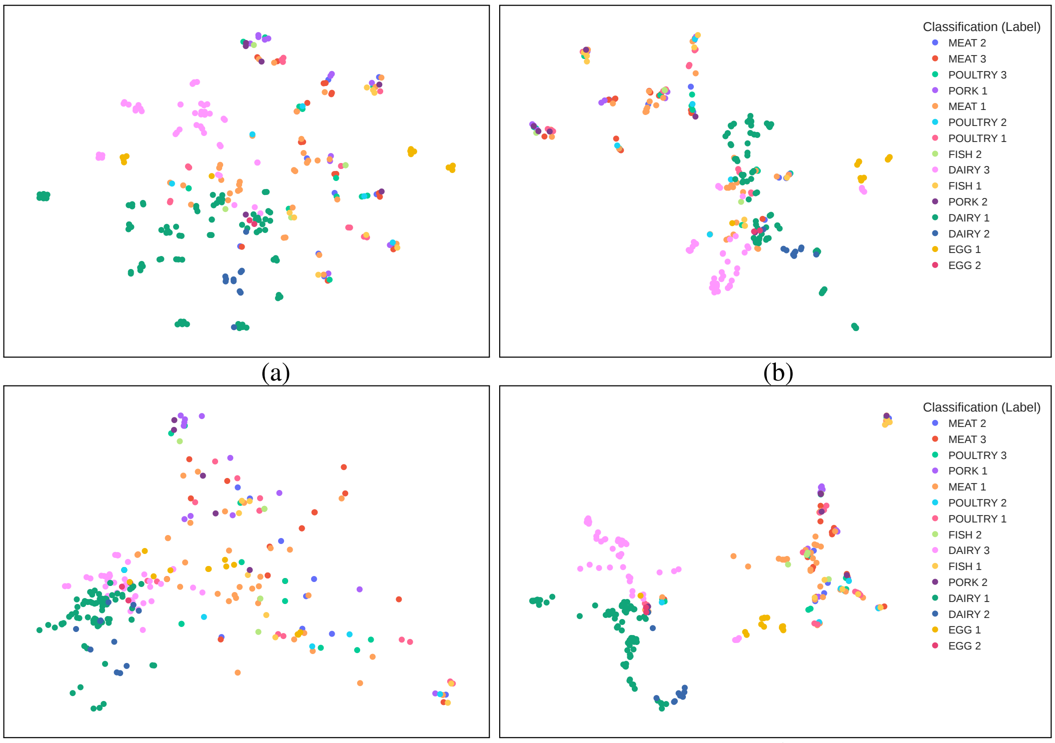
D. Evaluation of the Visualizations' Quality

The specialist's performance using the proposed method highly depends on the quality of the produced layouts. Thus, the visualizations are evaluated using two state-of-the-art metrics to measure the neighborhood preservation, the separability between groups of points and the similarity preservation among instances. These metrics were selected since we are following the quality assessment strategy presented in related researches [11].

The separability between groups of points in the layout regardless the data labels is evaluated using the silhouette coefficient (SC). For visualization purposes, SC is appropriate for only computing the separability between groups of points, which are determined using a clustering algorithm in the visual space [12]. For an instance x, the cohesion $a(x)$ is computed according to the mean intra-cluster distances, while the separation $b(x)$ is obtained by the minimum distance between x to any other instance belonging to another cluster.

In order to compute SC, the points in the visual space must be clustered using an algorithm. Here, we chose K-Medoids algorithm to determine K clusters since medoid centers are more appropriate representatives than mean centers due to the presence of categorical values in the dataset. The Euclidean distance is considered for K-Medoids and for computing the silhouette coefficient. The analysis of the clustering results was performed by running K-Medoids and varying K within the range $[2, 30]$, in which the silhouette coefficient is computed for each obtained clustering. Finally, we compute the mean and standard deviation from those values.

In order to measure the preservation of the local structure and similarity relations between neighbor data instances, we



(a) (b)

(c) (d)

Fig. 2. Layouts produced by the visualizations based on point placement by considering both ingredients and nutritional values of the Vegan Food Dataset: (a) t-SNE; (b) UMAP; (c) PCA; (d) TriMap.

compute the trustworthiness for each visualization [13]. For each object in the high dimensional space, the proportion of its k-nearest neighbor is computed and compared in relation to the k-neighbor points in the visual space for the corresponding 2D point. Finally, we compute this neighborhood preservation rate for a neighborhood value k by averaging the precision for all data instances. This rate value lies in the range $[0, 1]$, in which higher values are related to a better preservation of the neighborhood structure of data instances in the visual space.

III. Experimental results

In this section we performed experiments aiming to evaluate the quality of the employed visualization techniques considered in Section II-C. The development of the proposed method was based on Python 3.8 alongside with pandas, scikit-learn, Plotly, umap and trimap libraries for data processing and the multidimensional visualization techniques. The source code is available in a GitHub repository [1].

In order to evaluate the quality of the selected visualization techniques, we followed two steps: finding the best choices of hyperparameters in each method and then comparing the results with the previous mentioned evaluation metrics. To choose the best hyperparameters of each visualization technique, we performed a brute force search. For that purpose, we varied the values that are typically in a specific range. For instance, the learning rate for t-SNE is usually in the range $[10.0, 1000.0]$, which can be found in the library documentation. Following, for each combination, we calculated the trustworthiness value and the silhouette coefficient. After that, we plotted the point placement visualization layouts and then compared the evaluation metrics for each technique so that we selected the combination of values that yielded the best results.

A. Evaluation based on Visual Analysis

The goal of the evaluation based on visual analysis consists of interpreting the layouts by taking into account the points' positioning so that we can identify global and local data patterns. In order to enrich the visual analysis, the points in the layout were colored according to the "Food category" attribute for each dataset. The formation of groups of points can be associated with the points' colors (labels) in such a way that the specialist can find similar food products sharing similar ingredients and nutrients.

Figure 2 illustrates the layouts obtained by PCA, t-SNE, UMAP and TriMAP for the Vegan Food Dataset. The visual analysis in Figure 2(a) shows that t-SNE displays small groups of points, meaning that the associated food products are similar regarding their ingredient and nutrients. Some groups presents points that belongs to the same category, such as "Dairy 1", "Dairy 2" and "Eggs 1", while other groups shows products from different categories, indicating that they can be substitutes from each other when planning a diet. In Figure 2(b) the UMAP's layout presents some dense groups of points regarding the labels "Dairy 1", "Dairy 2" indicating that the vegan products are way dissimilar from the others. Figure 2(c) depicts the layout obtained from PCA, which does not show satisfactory classes segregation, nor visual perceptible groups of points since their colors are mixed. The TriMap's layout, shown in Figure 2(d), presents

[1] github.com/dudaholandah/sibgrapi2022_wuw

three visible larger regions of points of the same categories: “Dairy 1”, “Dairy 2” and “Dairy 3”.

B. Evaluation of the Visualizations’ Quality

We now compare the visualization techniques regarding the preservation of the global and local structure of the data in the high dimensional space in the layouts. Figure 3 presents the results of the trustworthiness for the considered multidimensional visualizations. It suggests that PCA and TriMap present better preservation of original neighbors in the low dimensional space for the Vegan Food Dataset, indicating that the local structure and similarity relations were better retained than UMAP and t-SNE.

Table II describes the obtained SCs for the visualization techniques on the Vegan Food Dataset. TriMap and UMAP achieved the higher silhouette scores, while PCA and t-SNE obtained similar scores. These results indicate better overall separability between the generated clusters in the visual space for TriMap and UMAP. It is worth noting that SC is independent regarding the label “Category”, thus emphasizing that the focus is to analyze the separability between the groups of points in the layout.

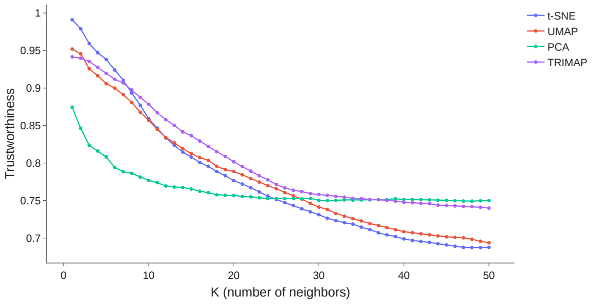


Fig. 3. Neighborhood preservation based on trustworthiness.

TABLE II
EVALUATION METRICS FOR VEGAN FOOD COMPOSITION DATASET

Visualization Techniques	Silhouette Coefficient	
	Mean	***Standard Deviation***
t-SNE	0.428	0.059
UMAP	0.523	0.059
PCA	0.417	0.046
TriMap	0.541	0.037

It can be seen that the nonlinear point placement strategies (t-SNE, UMAP and TriMap) presented better results regarding the visual analysis and the preservation of the patterns of the data according to the aforementioned metrics. As we are seeking visualization techniques that can group points related to similar food products, it is recommended to select those which minimize as much as possible the similarity relations regarding the ingredients and nutrients of food products.

The formation of small groups of points in the layouts suggest the possibility of using the proposed method on unlabeled food datasets, in which the specialist can assign categories to the food products sharing similar ingredients and nutrients. For that purpose, interaction tools and statistical analysis can be employed to support the specialist in an exploratory approach for knowledge discovery on food data.

The next subsection describes a case study in order to simulate the specialist visualizing the Vegan Food Dataset based on the visual exploration proposed by Daniel Keim [5].

C. Case study

We demonstrate the proposed method in a case study regarding the finding of substitute products for a diet plan in the context of vegan products. The idea relies on the use of different ingredients in preparations that mimic animal products commonly results in similar products from the sensory point of view, but divergent with regard to nutritional composition [14]. In this situation, suppose, for instance, a diet plan consisting of the categories pork, fish and poultry. If the specialist’s interest consists of finding appropriate substitutes for pork, we can perform the following steps:

1. Global view of data: Here, the multidimensional visualization receives as input the Vegan Food Dataset and displays the layout. We chose TriMap for that purpose because, predominantly, it has shown the best evaluation metrics results.

2. Zoom and filter: Assuming, for instance, the specialist’s interest consists of finding appropriate substitutes for pork, specially for “Pork 2”. The idea is to find groups of points that present related products. For that purpose, using the points’ color as a guide, the specialist can find groups of points that includes “Pork 2” category, represented by the dark purple points in Figure 4.

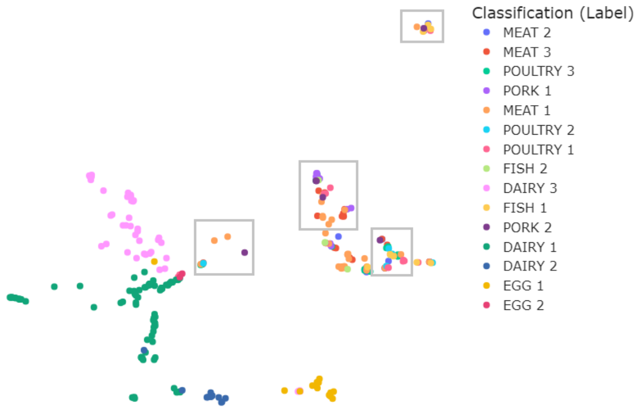
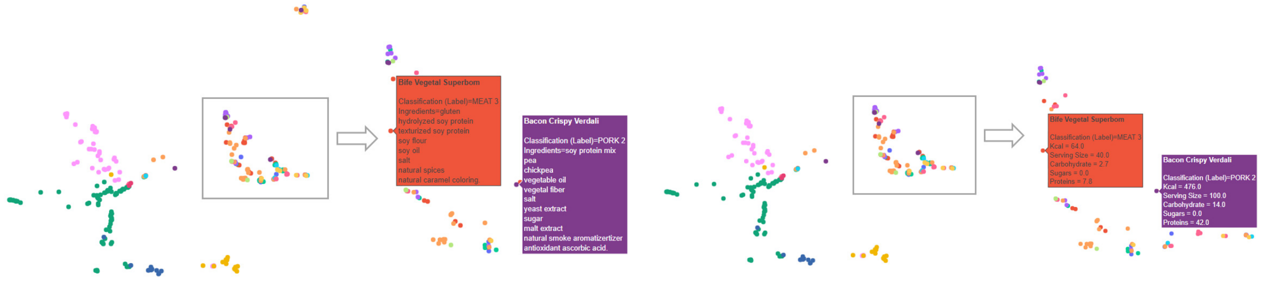


Fig. 4. TriMap Layout in which the rectangles present “Pork 2”-related products.

3. Details on demand: The specialist can use interactive techniques so that he can explore original information such as the nutrients and ingredients according to his goal. After zooming and filtering the layout by the requested category, it is possible to select similar products in order to visualize nutritional information or associated ingredients and evaluate whether suitable similar products are found. For instance, if the specialist zoomed in the group 1 referenced in Figure 4, he would have obtained, depending on his goal, either ingredients or nutrient values related to the products explored, as it can be seen in Figures 5(a) and 5(b), respectively.

IV. CONCLUSION

This paper explored the use of visualizations based on point placement for analyzing food composition data that


Bife Vegetal Superbom

Classification (Label)=MEAT 3
Ingredients=gluten
hydrolyzed soy protein
texturized soy protein
soy flour
soy oil
salt
natural spices
natural caramel coloring.

Bacon Crispy Verdali

Classification (Label)=PORK 2
Ingredients=soy protein mix
pea
chickpea
vegetable oil
vegetal fiber
salt
yeast extract
sugar
malt extract
natural smoke aromatizertizer
antioxidant ascorbic acid.

Bife Vegetal Superbom

Classification (Label)=MEAT 3
Kcal = 64.0
Serving Size = 40.0
Carbohydrate = 2.7
Sugars = 0.0
Proteins = 7.8

Bacon Crispy Verdali

Classification (Label)=PORK 2
Kcal = 476.0
Serving Size = 100.0
Carbohydrate = 14.0
Sugars = 0.0
Proteins = 42.0

(a) (b)

Fig. 5. TriMap Layout detailing (a) ingredients associated and (b) nutritional information about "Pork 2" product and "Meat 3" similar product.

might support specialists in knowledge discovery tasks. For that purpose, we proposed a method that comprised data pre-processing for generating an appropriate representation of the high dimensional food data prior the visualization techniques. Here, we chose PCA, t-SNE, UMAP and TriMap as point placement visualizations and compared their performances regarding the quality of generated layouts.

The experimental results showed that the layouts obtained by the nonlinear visualization techniques (PCA, UMAP and TriMap) presented better separability between clusters in the visual space according to the silhouette coefficients as well as better neighborhood preservation, indicating that the main structures of the data were retained in the visual space. The case study showed the possibility to employ point placement-based visualizations to support the specialists is tasks related to finding food products from different categories sharing similar ingredients and nutrients.

Future work can be guided to incorporate interactive tools to enable users to obtain statistical information from selected points in the layout. Furthermore, the exploration of other tasks on food composition data that can lead to solutions based on classification and clustering techniques is also a potential possibility.

Acknowledgment

Maria Eduarda M. de Holanda and Vinícius R. P. Borges would like to thank *Fundação de Apoio a Pesquisa do Distrito Federal* (FAPDF) and ProIC/UnB for the grant that supported this research, process number 00193-00001288/2021-16. Raquel Braz Assunção Botelho and Renata Puppin Zandonadi acknowledge *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq) for the grants related to the process numbers 307822/2021-4 and 302602/2021-6, respectively.

References

[1] M. Van Erp, C. Reynolds, D. Maynard, A. Starke, R. Ibáñez Martín, F. Andres, M. C. Leite, D. Alvarez de Toledo, X. Schmidt Rivera, C. Trattner *et al.*, "Using natural language processing and artificial intelligence to explore the nutrition and sustainability of recipes and food," *Frontiers in Artificial Intelligence*, p. 115, 2021.

[2] W. Min, S. Jiang, L. Liu, Y. Rui, and R. Jain, "A survey on food computing," *ACM Computing Surveys (CSUR)*, vol. 52, no. 5, pp. 1–36, 2019.

[3] A. Morales-Garzón, J. Gómez-Romero, and M. J. Martin-Bautista, "A word embedding-based method for unsupervised adaptation of cooking recipes," *IEEE Access*, vol. 9, pp. 27 389–27 404, 2021.

[4] M. Petković, G. Popovski, B. K. Seljak, D. Kocev, and T. Eftimov, "Diethub: Dietary habits analysis through understanding the content of recipes," *Trends in Food Science & Technology*, vol. 107, pp. 183–194, 2021.

[5] D. A. Keim, "Information visualization and visual data mining," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, no. 1, pp. 1–8, 2002.

[6] I. T. Jolliffe, "Principal component analysis and factor analysis," *Principal Component Analysis*, pp. 150–166, 1986.

[7] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.

[8] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[9] E. Amid and M. K. Warmuth, "TriMap: Large-scale Dimensionality Reduction Using Triplets," *arXiv preprint arXiv:1910.00204*, 2019.

[10] F. V. Paulovich and R. Minghim, "Hipp: A novel hierarchical point placement strategy and its application to the exploration of document collections," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1229–1236, 2008.

[11] R. Motta, R. Minghim, A. de Andrade Lopes, and M. C. F. Oliveira, "Graph-based measures to assist user assessment of multidimensional projections," *Neurocomputing*, vol. 150, pp. 583–598, 2015.

[12] J. G. S. Paiva, W. R. Schwartz, H. Pedrini, and R. Minghim, "Semi-supervised dimensionality reduction based on partial least squares for visual analysis of high dimensional data," in *Computer Graphics Forum*, vol. 31, no. 3pt4. Wiley Online Library, 2012, pp. 1345–1354.

[13] L. Van Der Maaten, "Learning a parametric embedding by preserving local structure," in *Artificial Intelligence and Statistics*. PMLR, 2009, pp. 384–391.

[14] C. T. Gallagher, P. Hanley, and K. E. Lane, "Pattern analysis of vegan eating reveals healthy and unhealthy patterns within the vegan diet," *Public Health Nutrition*, pp. 1–11, 2021.