Georgia Institute of Technology

CSE 6242

Data and Visual Analytics

Final Report

Recipe Recommendation System

Team HealthNinja:
Wenxin Fang
Xingyu Liu
Victoria Wanjiang Zheng
Yujia Xie
Peicheng Hua
Yifei Wang

Supervisor:
Prof. Duen Horng Chau

April 26, 2017

Contents

1 Introduction

People always face same problem in their daily life: What to eat and how to eat healthily. What's more, people who love cooking usually don't know how to start when facing too many ingredient pairings. Our Recipe Recommendation System give users nutritious recipe choices which also satisfy their personal taste.

2 Problem Definiton

Suppose we have a dataset of recipes $\mathcal{A}$, in which there is each record of recipe $a \in \mathcal{A}$ along with its ingredients, and each ingredient along with its compounds. Another data the program has access to is the basic and flavor information of the customers. The goal of our program is to optimize over the customers' taste need with the constraint of nutrition requirement. In the end, our program offers a personalized optimal recommendation list of recipes that is in best consistent with the tastes of customers, as well as satisfying the EER nutritional constraints.

3 Literature Survey

The first challenge in our project is to measure the dietary intake which satisfies different users' daily nutritional goals. One well-known and widely-accepted method for the measurement of dietary intake is to use Estimated Energy Requirements(EER) equation [1] [2] based on the user's weight, height, age, sex and physical activity level. Physical activity level(PAL) [3] is an index related with the intensity and impact of various activities adults do in their daily life. According to the calorie level assessed, we could get daily macronutrients goals for specific users [4]. For those with high Body Mass Index(BMI), a reasonable deduction in calorie level in the first step will be made so as to help them lose weight [5]. By getting these statistics, we are allowed to recommend recipes which suits individual nutrition requirement.

Another challenge in our project is to obtain an ordered list of recipe recommendations that satisfy the nutritional constraints. Our first thought is constraint optimization algorithm, which requires to set up a utility function that measures the usefulness of items to users [6]. However, considering the magnitude of databases and the cost of coordinating databases, we believe it is a better idea to split the problem into selecting feasible set and building a recommender system [7]. First, we let users to choose what flavor they would like today, and shrink the dataset of recipes based on the compounds in the ingredients [8] [9]. Then, the dataset can be further reduced by adding nutritional constraints. Finally, we order the feasible set by hybrid recommending [10]. Like Cotter & Smyth (2000)[11], we plan to merge the content-based information in step one and results of collaborative filtering methods [7][12][13] to produce a final list. As for the cold start problem [14] that is usually faced by recommender system, since content-based approaches [15][16] do not rely on ratings from other users, they can be used to produce recommendations for all items, provided attributes of the items are

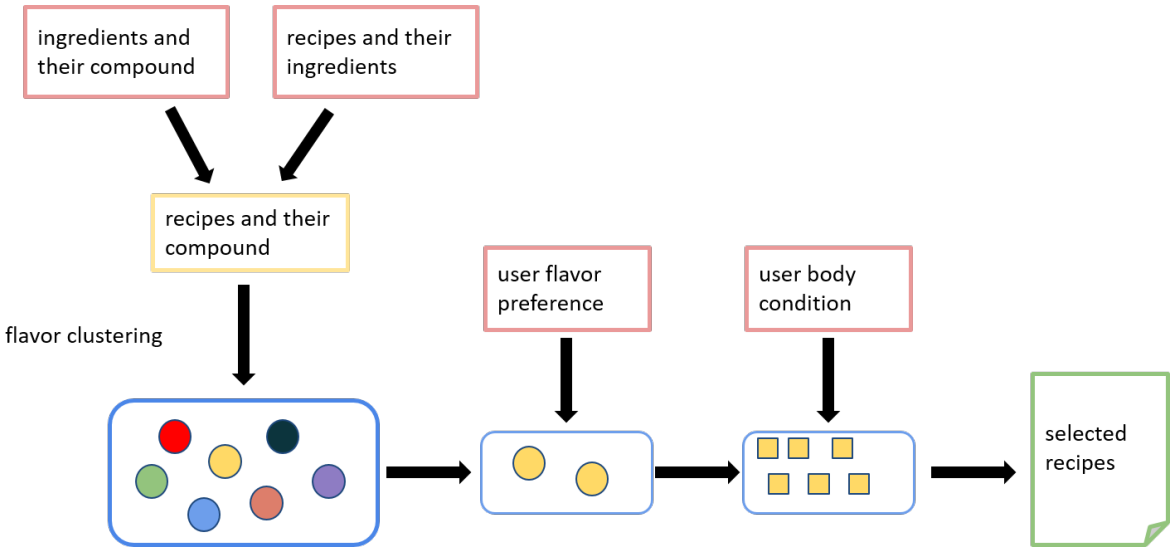


Figure 1: This is the complete architechture of our project. First, we parse our data into a table where recipes and their compounds are directly connected. Then, we further filter out ther recipe by flavor and nutrition constraints.

available. In fact, the content-based predictions of similar users can also be used to further improve predictions for the active user [17].

For our algorithm, recipe data will be collected and studied. For each recipe, we will parse out the information about the ingredients and compounds it contains. Then we will use flavour network [18] to project recipes with ingredients into our flavour space. In order to classify the recipes, different clusters among recipes will be created to select recipes from the cluster of flavor more close to user's preference. There are various ways to perform the clustering [19]. The traditional clustering method (K-means) performs better [20] than neural networks using Kohonen learning using simulated data with known cluster solutions [21]. A popular Lloyd's k-means clustering algorithm is also easy to implement and very efficient, requiring a kd-tree as the only major data structure [22]. In addition, one may consider employing hierarchical clustering, since they have the similar performance but potentially it might be easier to measure the dissimilarity among clusters.

However, our literature survey is far from being satisfactory. More work needs to be done on the clustering algorithms and recommender systems. We may also consider the requirement of users with special need such as pregnancy, lactation and living with diabetes in order to extend the range of recipe recommendation system application.

4 Proposed Methods

4.1 Intuition

The complete architecture of our prohect in shown in figure 1. Compared with existing recipe websites, our app has two advantages:

1. **Simple and user friendly.** We try to keep our app with a client-friendly interface, which would not require a lot of effort to use. Users only need to make one click and input two numbers, then they will get recipes based on their flavor and their nutritional requirements.

2. **Creative and surprising.** Our system is totally different from those websites which use keywords of recipe names or ingredients to search for recipes. We use clustering to decide which recipes are similar to each other; compound and flavor network are really convincing. We believe our result of cluster would include recipes with similar flavor from different countries and regions. Then users will find recipes which they are familiar with or not and hopefully they will feel surprising and surprisingly satisfied.

4.2 Data Processing

Data Collection: We collected recipe data using Yummly's API, which contains ingredients for each recipe. We project recipe and compound into one space by combining the table of recipe-ingredient and table of ingredient-compound. The first column of the result table is recipes' name while the remaining columns are boolean values which stand for whether such compound exist or not. As for nutrition data, we have multiple resources. One of them is from kaggle's Open Food Facts dataset. This dataset provide us with nutrition detail for each kind of food, such as fat, sugar. With ingredients' nutrition data known, we can calculate what and how much nutrition every recipe can provide. Further along the process, we collect data from restaurants and supermarkets databases to add to the missing data that could not be matched from the current record.

Data Cleaning: In order to reduce the dimension of the attribute of FoodFacts and to match nutrition attribute with the ingredient, we used multiple data cleaning techniques. For missing values, if it is numerical, listwise deletion, average imputation, regression substitution and educated guessing were used. For categorical missing value, classification method could also be used. For field matching, we use indexes to speed up the matching process.

4.3 Algorithms

Machine Learning: With recipe-compound data on hand, we use two machine learning algorithms, hierarchical clustering and DBSCAN, to cluster flavor of recipes based on their compound. There are two considerations when we select clustering algorithm. Firstly, both algorithms can accept hamming distance as parameter to judge how

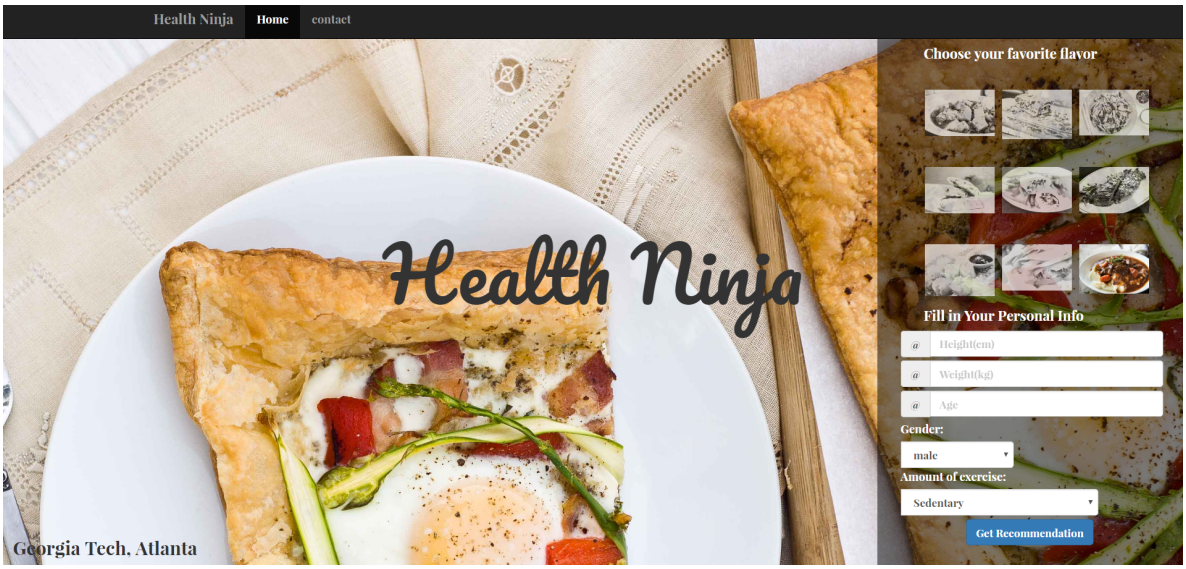


Figure 2: This is the first page of user-friendly interface of our recommendation system. On the right side of the website, the user can choose one dish from nine and our system will judge user's personal flavor by it. The user can also fill his body condition such as weight, height, age and activity level in the forms.

close two recipes are, which is suitable if we treat each column (compound) as one feature. Second, since the number of compound exceeds 1000, we definitely need scalable algorithms. It is hard to say what is the correct label for each flavor group, so we use one typical recipe for every flavor group.

Nutritional constraint: To enforce our result meet the nutritional need of users, we use information of user's body condition to further filter out the recipe to recommend. Estimated Energy Requirements equations are used to measure individual needs for dietary intake.

4.4 User Interfaces

As for user interfaces, we are designing a web application. In general, the structure could be separated into two parts. The first part presents our clustered flavors as visualized nodes with simple descriptions about different flavors, which are generated by clustering results from previous study. Users can click on nodes to choose the flavor in their preference. There will also be following options to collect their user information (height, weight and age). We try to keep this step as simple as possible since it might reduce the user satisfaction. Based the previous information collected, our app will redirect to the second part which recommend recipes. The result will be presented as pictures of the food. Once users click on picture, a semi transparent container would appear where users can see details of cuisine like ingredients, cooking time and nutrition. Also, there is a button at the bottom of container which can direct to the recipe instructions once clicked on. Figure 2 and 3 is the final result of the user interface.

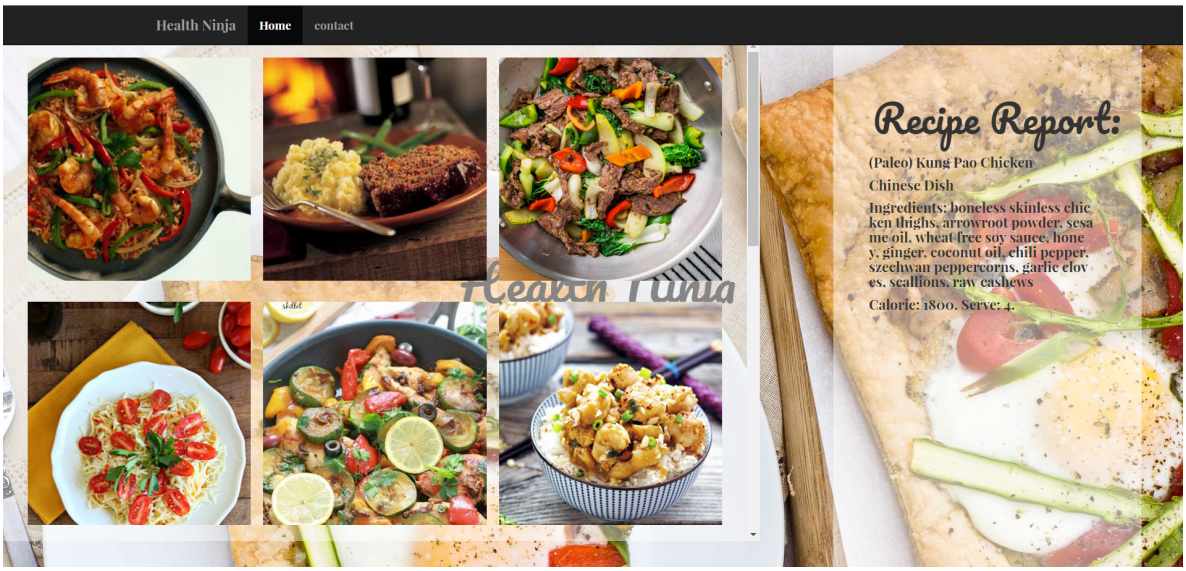


Figure 3: This is the second page of user-friendly interface of our recommendation system. After clicking "get recommandation" icon, the browser will jump into this website with list of dishes meeting user's personal taste and nutrition need at the same time. Each row here represent a set of recommendation.

5 Experiments/Evaluation

Figure 4 is the result of clustering. In figure 4, each node represent one ingredient in the food and the color of the node represent the category it is belong to. The diameter of node is a scale of the number of one ingredient in our dataset. This figure demonstrates the relation of recipes by their compounds, i.e. their flavors. The result seems good.

The result of UI is shown in figure 2 and 3. For example, if your input of 2 is **the left bottom recipe image**, **165**cm in height, **50**kg in weight, **22** years old, activity level as **sedentary**, and you are a **female**, you will get an ordered list of recipe set, each contain three recipes:
1. Ginger Beef & Broccoli, Mrs. Brogan's Curry Gyoza, Three Bean Chipotle Chili
2. Chinese Pork and Vegetable Hot Pot, Chinese Roast Pork, Shredded Mexican Beef
and so on...

Each set of the recipe is to recommend your meals for a day. Our recommended recipes can not only meet your nutritional need, but also satisfy your taste.

For algorithm evaluation, we took both intuitive approach and mathematical approach.

Intuitive approach: we randomly pick 10 recipes with the same flavor group to see if they are the same flavor from human's view. And we use two cluster model (DBSCAN,Hierarchy cluster) to compare flavor groups. This result was validated by user interview. Based on the feedback we think the results are good.

Mathematical approach: ideally, there are multiple ways to evaluate the clustering results. Internal index, external index and similarity/dissimilarity matrix. From calculation of correlation, we get result of 0.65, which is not ideal but comparing to randomly

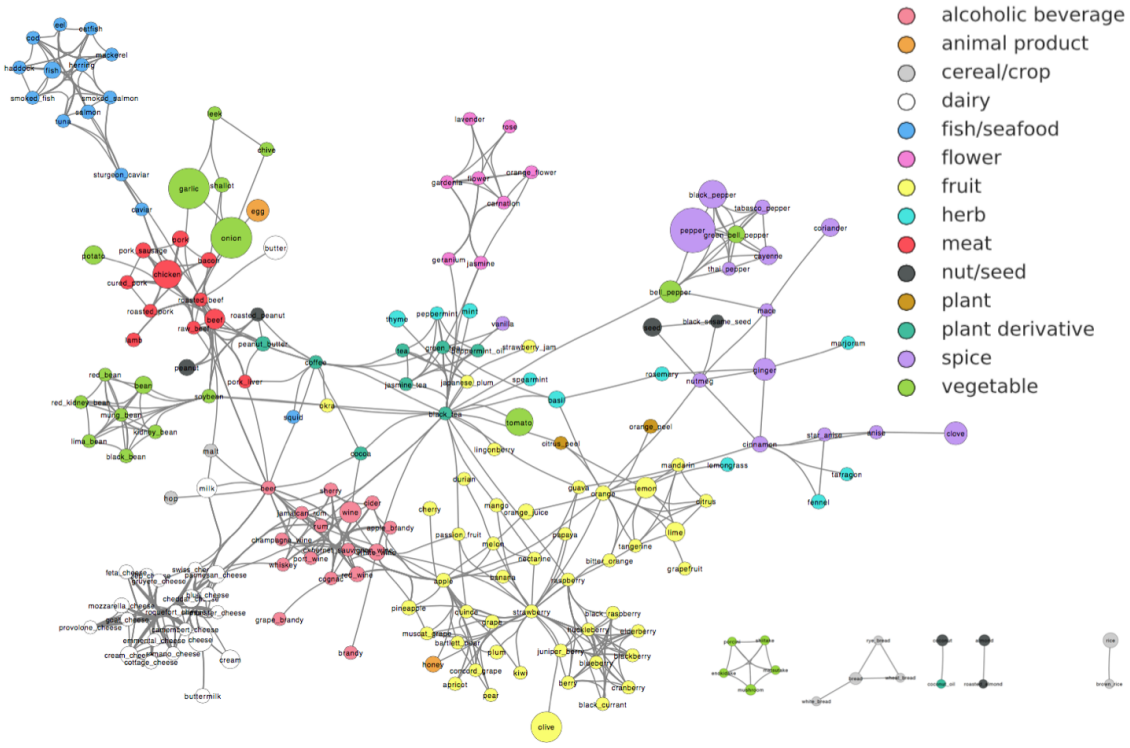
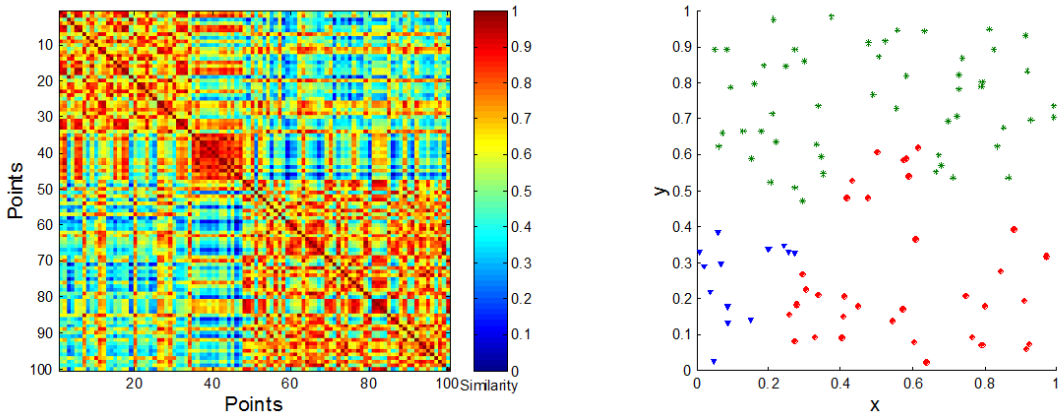


Figure 4: This is a graph where each node represent one ingredient in the food and the color of the node represent the category it is belong to. The diameter of node is a scale of the number of one ingredient in our dataset.

generated data, which is below 0.5, the result is relatively good.

From the visualized cluster below we can see that the edges are not crisp in this clustering. Further improvement could be used for improving the results.



(a) Similarity Matrix (b) Cluster Plot

Figure 5: This is the visualization of some typical recipes, where different colors stand for different flavors. We normalize the distance of all nodes and thus they can be presented better.

6 Conclusions and discussion

To conclude, the main innovation in our approach is to combine flavor of recipes with people's health. As a recipe recommendation system, we value not only users' preference but also their health. Besides, most recipe recommendation systems or apps recommend recipes based on users' rating and comments. But in our implementation, clustering flavors based on ingredients' compound makes our result more objective. The nutritional constraints make our result healthy. We believe our project will help people to live a more healthy and tasty life.

7 Distribution of team member effort

Project work:
Project Framework: Xingyu Liu
Data collection: Yujia Xie
Data clean: Wenxin Fang, Wanjiang Zheng, Yifei Wang, Peicheng Hua
Clustering: Xingyu Liu
Nutritional Contraints: Yujia Xie, Peicheng Hua
User interface: Wenxin Fang, Wanjiang Zheng, Yifei Wang, Xingyu Liu
Report work:
Proposal: All
Presentation: Wanjiang Zheng
Progress report: All
Poster Making:Wanjiang Zheng
Final report: All

References

[1] A. Bergman and J. Distler, "Wormholes in maximal supergravity," 2007.

[2] B. E. Millen, S. Abrams, L. Adams-Campbell, C. A. Anderson, J. T. Brenna, W. W. Campbell, S. Clinton, F. Hu, M. Nelson, M. L. Neuhouser, *et al.*, "The 2015 dietary guidelines advisory committee scientific report: development and major conclusions," *Advances in Nutrition: An International Review Journal*, vol. 7, no. 3, pp. 438–444, 2016.

[3] I. of Medicine, *Dietary reference intakes for energy, carbohydrate, fiber, fat, fatty acids, cholesterol, protein and amino acids*. Washington (DC):National Academies Press, 2002.

[4] I. of Medicine., *Dietary reference intakes: the essential guide to nutrient requirements*. Washington (DC):National Academies Press, 2006.

[5] M. D. Jensen, D. H. Ryan, C. M. Apovian, J. D. Ard, A. G. Comuzzie, K. A. Donato, F. B. Hu, V. S. Hubbard, J. M. Jakicic, R. F. Kushner, *et al.*, "2013 aha/acc/tos guideline for the management of overweight and obesity in adults," *Circulation*, vol. 129, no. 25 suppl 2, pp. S102–S138, 2014.

[6] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE transactions on knowledge and data engineering*, vol. 17, no. 6, pp. 734–749, 2005.

[7] G. Linden, B. Smith, and J. York, "Amazon. com recommendations: Item-to-item collaborative filtering," *IEEE Internet computing*, vol. 7, no. 1, pp. 76–80, 2003.

[8] Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, and A.-L. Barabási, "Flavor network and the principles of food pairing," *Scientific reports*, vol. 1, 2011.

[9] C.-Y. Teng, Y.-R. Lin, and L. A. Adamic, "Recipe recommendation using ingredient networks," in *Proceedings of the 4th Annual ACM Web Science Conference*, pp. 298–307, ACM, 2012.

[10] P. Melville and V. Sindhwani, "Recommender systems," in *Encyclopedia of machine learning*, pp. 829–838, Springer, 2011.

[11] P. Cotter and B. Smyth, "Ptv: Intelligent personalised tv guides," in *AAAI/IAAI*, pp. 957–964, 2000.

[12] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Communications of the ACM*, vol. 35, no. 12, pp. 61–70, 1992.

[13] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "Grouplens: an open architecture for collaborative filtering of netnews," in *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pp. 175–186, ACM, 1994.

[14] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock, "Methods and metrics for cold-start recommendations," in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 253–260, ACM, 2002.

[15] R. J. Mooney and L. Roy, "Content-based book recommending using learning for text categorization," in *Proceedings of the fifth ACM conference on Digital libraries*, pp. 195–204, ACM, 2000.

[16] M. Pazzani and D. Billsus, "Learning and revising user profiles: The identification of interesting web sites," *Machine learning*, vol. 27, no. 3, pp. 313–331, 1997.

[17] P. Melville, R. J. Mooney, and R. Nagarajan, "Content-boosted collaborative filtering for improved recommendations," in *Aaai/iaai*, pp. 187–192, 2002.

[18] Y.-Y. Ahn and S. Ahnert, "The flavor network," *Leonardo*, vol. 46, no. 3, pp. 272–273, 2013.

[19] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[20] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[21] P. S. Balakrishnan, M. C. Cooper, V. S. Jacob, and P. A. Lewis, "A study of the classification capabilities of neural networks using unsupervised learning: A comparison withk-means clustering," *Psychometrika*, vol. 59, no. 4, pp. 509–525, 1994.

[22] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 881–892, 2002.