Gaussian Mixture Models for Identifying Ultra-Processed Foods

Joshua Strickland[1], Murray Dare[2], and Wenjia Wang[1(✉)]

[1] University of East Anglia, Norwich, UK
{j.strickland,wenjia.wang}@uea.ac.uk
[2] Boone, Norwich Research Park, Norwich, UK
murray@myboone.app

Abstract. Ultra-processed foods (UPFs) are industrial formulations of multiple ingredients to make a highly palatable food product and account for nearly half of the average British diet and most developed countries. UPFs are strongly associated with obesity, chronic disease, and increased mortality. However, the identification of UPFs is currently still heavily relied on by human experts, making consistent and large-scale classification very difficult and resource intensive.

This research explores a semi-supervised pipeline in which available labels are leveraged during preprocessing and performance evaluation, but not used in unsupervised clustering tasks. We investigated and tested Gaussian Mixture Models (GMM) and few other commonly used clustering methods on real-world data, and then compared them with a baseline method k-means. This reflects the real-world scenario where food databases contain partially labelled data but full annotation remains impractical.

These results demonstrate that unsupervised learning offers a viable and scalable option for automated food categorization where labelled data is scarce.

Keywords: Unsupervised Machine Learning · Ultra-Processed Food · NOVA Classification Method

1 Introduction

1.1 Background and Motivation

Ultra-processed foods are industrially formulated products designed for convenience and hyper-palatability. They typically contain additives such as emulsifiers, artificial flavours, and colourings and very little whole foods.

UPFs make up to 48.6% of the average British diet despite the associated increased risk of non-communicable diseases, 79% increased risk of obesity, and links to higher all-cause mortality rates [6,13,15]. According to government reports, obesity and obesity-related illness alone cost the NHS £6.1 billion between 2014 and 2015 [12]. Thus, demonstrating the importance of reducing UPF consumption and the necessary improvement in the standard diet to improve public health.

E. Corchado et al. (Eds.): HAIS 2026, LNAI 16596, pp. 100–110, 2027.
https://doi.org/10.1007/978-3-032-29292-6_9

Recently, machine learning (ML) and artificial intelligence (AI) have been explored as tools to classify UPFs more efficiently and accurately. These technologies offer the potential to reduce reliance on expert opinion and to support recommendations for less processed food alternatives.

This study adopts a semi-supervised approach, utilising available labels to guide preprocessing steps such as class balancing and feature selection, whilst leaving the clustering task itself unsupervised. This reflects the practical reality of food databases, where a proportion of products carry NOVA labels but the vast majority do not. By leveraging partial labels in preprocessing rather than classification, this pipeline can scale to largely unlabelled datasets without requiring extensive expert annotation.

2 Related Work

2.1 Ultra-Processed Foods and the NOVA Classification System

Works by Monteiro et al. [10] describe UPFs as foods containing numerous ingredients undergoing industrial processes not typical in domestic cooking. Monteiro et al. [9] introduced the NOVA system, assigning foods to four groups ranging from unprocessed to ultra-processed based on the extent of industrial processing. Notably, NOVA classifies individual products rather than food categories; for example, bread may fall into Group 3 or Group 4 depending on whether it contains emulsifiers or preservatives [10].

Despite being the most widely used framework, NOVA has limitations. Expert subjectivity leads to inconsistent classifications; Braesco et al. [2] found unanimous agreement on only three of 120 branded products tested. Nevertheless, NOVA's relevance is well-supported: Suksatan et al. [15] linked a 10% increase in UPF calorie intake to a 15% increase in all-cause mortality, Askari et al. [1] associated UPF consumption with overweight and obesity across more than 2000 studies, and Lane et al. [7] reported a 20 to 81% increased risk of non-communicable diseases among high UPF consumers.

2.2 Unsupervised Clustering

Unsupervised learning approaches for clustering foods by processing level have received limited attention. However, some notable work has been carried out. The core idea involves representing nutritional information as vectors in a feature space, allowing similar items to be grouped without labels. Marcos et al. [8] applied a Gaussian Mixture Model (GMM) to a dataset containing 62 nutritional parameters and achieved a Gini coefficient of 0.5437, demonstrating moderate success in clustering foods with similar processing levels.

3 Methodology

This study adopts a semi-supervised preprocessing strategy, in which available NOVA group labels are used to improve data quality and feature selection prior

to clustering. The clustering models themselves operate without label supervision, ensuring the approach remains applicable to datasets where labels are scarce or absent.

3.1 Data and Pre-processing

This research relied exclusively on publicly available crowdsourced data from Open Food Facts [11], a comprehensive database of food products from around the world. The initial dataset provided detailed information on a vast amount of products, including their nutritional composition (e.g., energy, macronutrients, sodium, fibre) and full ingredient lists. The dataset was further refined by filtering for products available in the United Kingdom, maintaining a consistent geographic focus aligned with the regional scope of this study.

The crowdsourced nature of this data introduced potential challenges, including inconsistencies in user-submitted entries and a high proportion of missing values for certain nutrients and ingredients. These issues were addressed through a rigorous data cleaning and preprocessing pipeline, detailed in Sect. 4, to ensure the robustness of the models.

Following this process, the final curated dataset consisted of 43,116 samples and 24,400 features, providing a large scale comprehensive dataset for analysis.

To address class imbalance in the classification task, multiple sampling techniques such as Synthetic Minority Over-sampling Technique (SMOTE), random oversampling(ROS), Edited Nearest Neighbour (ENN), and random undersampling(RUS) were applied to the training data, [5]. Each method was evaluated independently to determine its influence on model performance.

3.2 Normalisation and Standardisation

Each sampled dataset was processed using different scaling methods including Standardisation (Z-score normalisation), Min-Max Scaling, Maximum Absolute Scaling, and Robust Scaling. This ensures a thorough understanding of how feature scaling affects both clustering and classification models.

3.3 Dimensionality Reduction

Following scaling, we evaluated whether dimensionality reduction techniques could improve clustering accuracy and cohesion. Principal Component Analysis (PCA) provides a linear baseline that maximises variance.

3.4 Model Selection

Four clustering models were evaluated. k-means serves as the baseline, its simplicity providing a clear performance floor for more advanced techniques. A Gaussian Mixture Model (GMM) was selected to address k-means' assumption of spherical clusters; its probabilistic framework allows for elliptical clusters and

soft assignments, offering a more flexible model for the likely distribution of foods in the feature space. Hierarchical Clustering was included to explore whether a natural hierarchy exists between processing levels, with a dendrogram potentially revealing meaningful sub-groupings within NOVA categories. Finally, BIRCH was selected for its scalability with large datasets [17], making it suitable for real-world applications where food products are continuously added.

3.5 Evaluation

Each technique is evaluated 10 times using different random seeds, with the mean score taken forward to ensure robustness and reproducibility. Two metrics are used: ARI and silhouette score. ARI measures alignment between predicted clusters and ground truth labels, adjusting for chance, with scores ranging from -1 (worse than random) to 1 (perfect agreement) [16]. Silhouette score is an intrinsic metric requiring no labels, measuring how well each sample fits its assigned cluster relative to neighbouring clusters, also ranging from -1 to 1 [14]. Together these metrics provide both an external and internal assessment of clustering quality.

4 Data Analysis

4.1 Exploratory Data Analysis

Missing values within the dataset were explored. We found that 100% of rows were missing values with 99.8% of columns were also missing values. Overall 97.03% of the dataset was found incomplete, indicating a requirement for careful preprocessing for this dataset.

To address the substantial amount of missing data, imputation appeared to be impossible for ensuring robustness and accuracy. Instead, any sample (row) with missing values in the following key features was removed: NOVA Group, NutriScore, NutriScore Grade, energy per 100 g, protein per 100 g, saturated fat per 100 g, salt per 100 g, fibre per 100 g, sugar per 100 g, ingredient tags, additives tag, and categories. These features were selected based on their demonstrated importance in the classification of NOVA groups, as highlighted in previous research, such as the work by Monteiro et al. [10]. NOVA Group was used solely to ensure complete ground truth labels were available for evaluation purposes; it was not included as an input feature to the clustering models. Finally, after the rows with important missing data was removed, we removed any further columns still missing values.

After this targeted removal of rows and subsequent feature cleaning, the resulting dataset had a shape of 20488 rows by 205 columns.

Finally, the distribution of NOVA categories was analysed. There is a clear imbalance and over-representation of NOVA group 4 foods, counting for 61.6% of the data, and an under-representation of group 2 foods (1.5%). The remaining groups, Group 1 and 3 make up 11.8% and 25.1% of the data respectively. This may be due to the nature of the dataset, which likely includes more packaged

and branded products, common in ultra-processed categories. Additionally, as the database is open source and user-submitted, there may be inherent biases in the types of products that are logged more frequently. This indicates that class balancing methods will need to be employed when building the model to prevent bias towards over-represented categories and ensure fair performance across all NOVA groups.

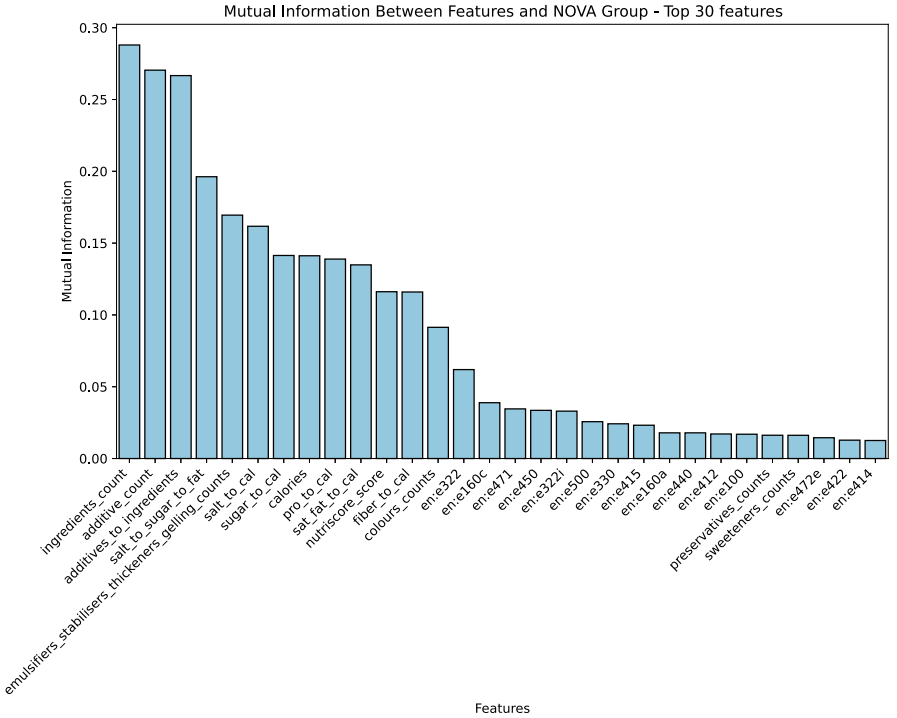

Fig. 1. Features with the highest Mutual Information to NOVA group.

4.2 Feature Engineering and Selection

Nutrient-to-energy ratios (protein, salt, saturated fat, fibre, sugar), a hyper-palatability composite capturing combined salt/sugar/fat intensity [3], additive category counts (colourings, preservatives, sweeteners, emulsifiers) [4], ingredient and additive totals, and one-hot encoded additive presence were computed, yielding 337 features across 20,488 samples.

Mutual information was calculated between each feature and NOVA group to identify the most discriminative features, as shown in Fig. 1. Ingredient count, additive count, and engineered nutritional ratios such as sugar-to-calorie scored highest, consistent with expectations for processed foods. Features with

zero mutual information, predominantly rare or highly specific additives, were dropped to reduce noise and computational complexity.

Table 1. Results from different sampling methods with k-means.

	RUS	ENN	SMOTE	ROS	Base
Mean ARI Score	0.1515	0.0472	0.1530	0.1512	0.0070
Mean Silhouette Score	0.5655	0.4853	0.5718	0.5718	0.4914

Table 2. Results from differing normalisation and standardisation techniques with k-means.

	Standardisation	MinMax	Robust	Maximum Absolute	Base
ARI	0.0382	0.0665	0.0000	0.1144	0.1530
Silhouette	0.7004	0.4682	0.9973	0.3597	0.5718

5 Unsupervised Clustering Results

5.1 Sampling/Class Imbalance

SMOTE's performance was particularly notable, achieving both the highest ARI (0.1530) and joint-highest silhouette score (0.5718), as shown in Table 1. This demonstrates SMOTE's effectiveness at generating synthetic samples that preserve cluster structure while maintaining clear cluster boundaries. These results highlight the importance of input data quality for model performance and demonstrate how resampling methods can significantly enhance unsupervised learning outcomes. Consequently, SMOTE was used in subsequent tests.

5.2 Standardisation/Normalisation

The experiments examining different normalisation and standardisation techniques demonstrated that the base (unscaled) data achieved the highest ARI score of 0.1530, as shown in Table 2. Whilst Standardisation and Robust scaling achieved higher silhouette scores of 0.7004 and 0.9973 respectively, these did not correspond to meaningful cluster alignment with NOVA groups.

The superior performance of the base data indicates that the original feature space already exhibits appropriate scale relationships for clustering. The extremely high silhouette score (0.997) but zero ARI observed with Robust scaling and high silhouette score (0.7) but low ARI score (0.0382) of Standardisation demonstrates that geometric cluster separation does not necessarily correspond to meaningful data groupings. Consequently, the base data therefore represented the best balance between external validity and internal cohesion, and was selected for subsequent experiments.

Table 3. PCA results across different Principal Component counts (averaged over 10 runs) for unsupervised learning.

Explained Variance	Number of Components	Mean ARI	STD ARI	Mean Silhouette	STD Silhouette
0.980	1.00	0.1524	0.0026	0.6071	0.0232
0.990	2.00	0.1525	0.0027	0.5746	0.0271
0.995	3.00	0.1528	0.0024	0.5693	0.0183
0.999	4.00	0.1529	0.0023	0.5660	0.0182
0.9999	4.30	0.1528	0.0023	0.5660	0.0182
0.9999	10.90	0.1528	0.0024	0.5661	0.0173

5.3 Dimensionality Reduction

The dimensionality reduction experiments using PCA have been summarised in Table 3. PCA demonstrated consistent performance across all tested variance thresholds (98–99.99%), with ARI scores ranging from 0.1524 to 0.1529 and silhouette scores similarly stable between 0.5661 and 0.6071.

Whilst PCA with 1 component achieved a marginally higher silhouette score (0.6071), it retained only 98% of variance and collapsed the feature space into a single dimension, likely discarding meaningful structure relevant to food processing levels. PCA with 4 components was selected as it retained 99.9% of variance whilst keeping dimensionality low, offering a more complete representation of the underlying feature space. Given that ARI scores were effectively identical across all configurations (differing only in the 4th decimal place), the selection was made on the basis of representational completeness rather than marginal metric differences.

Note that fractional component counts (e.g., 4.30, 10.90) arise because SMOTE oversampling was applied prior to PCA independently across 10 random seeds; the reported values represent the mean number of components selected across seeds at each variance threshold.

5.4 Model Evaluation

Evaluation of clustering models on the base (unscaled), SMOTE sampled data reduced to 4 PCA components showed distinct performance differences. As seen in Table 4, the GMM achieved the highest ARI of 0.2026, outperforming other algorithms considerably. This reflects GMM's probabilistic nature and flexibility in modelling elliptical clusters with varying shapes and orientations, which allowed it to better uncover structure aligned with the ground truth labels.

k-means achieved the second-highest ARI of 0.1529 with the highest silhouette score of 0.5660, indicating compact and well-separated clusters. However, its assumption of spherical clusters with equal variance likely limited its ability to capture the true underlying data structure, as evidenced by its lower ARI compared to GMM.

GMM's silhouette score of 0.2882 was notably lower than k-means (0.5660) and Hierarchical clustering (0.5362), indicating weaker geometric cluster cohesion and separation. This discrepancy highlights a fundamental trade-off between external clustering accuracy and internal cluster quality. GMM's probabilistic soft-assignment approach allows it to capture overlapping and irregularly shaped clusters that better reflect the actual data groupings, even if these clusters are less geometrically distinct in the reduced feature space.

These results demonstrate that accurate identification of meaningful clusters (as measured by ARI) does not necessarily correspond to geometrically compact and well-separated clusters (as measured by silhouette score). For unsupervised learning where the goal is to discover ground truth structure, external validity metrics like ARI are more important than internal cohesion metrics. Despite forming less geometrically compact clusters, GMM was selected as the best-performing model due to its superior ability to identify clusters that align with the actual data groupings.

Table 4. Clustering results of four models on the test data.

	K Means	GMM	BIRCH	Hierarchical
Adjusted Rand Index (ARI)	0.1529	0.2026	0.1403	0.1459
ARI STD	0.0023	0.0060	0.011	0.0054
Silhouette Score	0.5660	0.2882	0.4920	0.5362
Silhouette STD	0.0182	0.0073	0.0245	0.0261

5.5 Hyperparameter Tuning

A Grid search was conducted over covariance type (`diag`, `tied`, `full`), initialisation method (`k-means`, `k-means++`, `random`), convergence tolerance (1e-1 to 1e-6), and maximum iterations (500, 750, 1000), with regularisation covariance fixed at 0.1 for numerical stability.

Results revealed a clear divergence driven primarily by covariance structure. Top ARI models (Table 5) consistently used diagonal covariance with k-means++ initialisation and tolerance 0.01, achieving ARI of 0.2191 with moderate silhouette scores of 0.2762. Top silhouette models (Table 6) exclusively used tied covariance with k-means initialisation and looser tolerance 0.1, achieving silhouette of 0.5515 but lower ARI of 0.1543. This reflects a fundamental modelling difference: diagonal covariance allows each cluster to adapt its shape independently, better capturing true data groupings at the cost of geometric separation, while tied covariance enforces uniform cluster shapes, producing compact but potentially misaligned clusters. Both configurations were robust to iteration count, confirming convergence is achieved well before 500 iterations. Given that the objective is to uncover true underlying structure, diagonal covariance with k-means++ initialisation represents the superior configuration.

Table 5. Top 5 GMM Models by ARI Score

Model	ARI	Silhouette	Covariance	Init	Tol	Max iterations
4	0.2191	0.2762	diag	k-means++	0.01	1000
5	0.2191	0.2762	diag	k-means++	0.01	500
6	0.2191	0.2762	diag	k-means++	0.01	750
14	0.2190	0.2809	diag	k-means	0.01	500
15	0.2190	0.2809	diag	k-means	0.01	750

Table 6. Top 5 GMM Models by Silhouette Score

Model	Silhouette	ARI	Covariance	Init	Tol	Max iterations
71	0.5515	0.1543	tied	k-means	0.1	500
70	0.5515	0.1543	tied	k-means	0.1	1000
72	0.5515	0.1543	tied	k-means	0.1	750
69	0.5476	0.1544	tied	k-means	0.01	750
67	0.5476	0.1544	tied	k-means	0.01	1000

6 Evaluation and Discussion

6.1 Evaluation

The best-performing model, a Gaussian Mixture Model (GMM), achieved a maximum ARI of 0.22, consistent with Marcos et al. [8], who applied a GMM to a smaller dataset and reported moderate success. While low in absolute terms, this confirms that some meaningful cluster structure related to processing level exists and can be partially identified without labels. However, these moderate results underscore a fundamental challenge: the nutritional feature space does not contain natural clusters that perfectly align with expert-defined NOVA categories. This likely stems from the open-source nature of the dataset; as Braesco et al. [2] noted, experts frequently disagree on NOVA classifications, meaning differing labelling rationales produce conflicting patterns in the feature space. Without label guidance, what is intended as a four-cluster problem effectively becomes a far more complex one, with clusters reflecting combinations of expert rationale and processing level rather than processing level alone.

6.2 Discussion of Limitations

Future work should explore model generalisability across international food databases, this warrants investigation to ensure findings are not limited by regional dietary patterns. Most importantly, this work highlights the case for a continuous processing score rather than rigid NOVA categories, which would reduce labelling subjectivity and provide a more reliable foundation for dietary guidance tools.

7 Conclusion

A GMM was the strongest unsupervised performer, achieving an ARI of 0.20 and silhouette score of 0.29 before tuning. Hyperparameter tuning revealed a fundamental trade-off: optimising for silhouette yielded substantial gains (0.5515) but reduced ARI (0.1543), while optimising for ARI produced only marginal improvement (0.2191). Unsupervised methods offer a scalable, low-cost approach but currently fall short of practical accuracy requirements, particularly given that expert label disagreement obscures natural cluster structure. Nevertheless, they remain valuable where labelled data is scarce, and the integration of machine learning offers a promising path toward more data-driven, objective food classification.

References

1. Askari, M., Heshmati, J., Shahinfar, H., Tripathi, N., Daneshzad, E.: Ultra-processed food and the risk of overweight and obesity: a systematic review and meta-analysis of observational studies. Int. J. Obes. **44**(10) (2020). https://doi.org/10.1038/S41366-020-00650-Z, https://pubmed.ncbi.nlm.nih.gov/32796919/
2. Braesco, V., et al.: Ultra-processed foods: how functional is the nova system? Eur. J. Clin. Nutr. **76**(9), 1245–1253 (2022)
3. Fazzino, T.L., Rohde, K., Sullivan, D.K.: Hyper-palatable foods: development of a quantitative definition and application to the us food system database. Obesity **27**(11), 1761–1768 (2019)
4. Food Standards Agency: Approved additives and e numbers (2025). https://www.food.gov.uk/business-guidance/approved-additives-and-e-numbers
5. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. **21**(9), 1263–1284 (2009)
6. Jayedi, A., Soltani, S., Abdolshahi, A., Shab-Bidar, S.: Healthy and unhealthy dietary patterns and the risk of chronic disease: an umbrella review of meta-analyses of prospective cohort studies. Br. J. Nutr. **124**(11), 1133–1144 (2020). https://doi.org/10.1017/S0007114520002330
7. Lane, M.M., et al.: Ultraprocessed food and chronic noncommunicable diseases: a systematic review and meta-analysis of 43 observational studies. Obes. Rev. **22**(3), e13146 (2021)
8. Marcos, I.F.V., Bordel, B., Cira, C.I., Alcarria, R.: A methodology based on unsupervised learning techniques to identify the degree of food processing. In: 2022 17th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1–6. (2022). https://doi.org/10.23919/CISTI54924.2022.9820513
9. Monteiro, C.A., et al.: Nova. the star shines bright. World Nutr. **7**(1–3), 28–38 (2016)
10. Monteiro, C.A., et al.: Ultra-processed foods: what they are and how to identify them. Public Health Nutr. **22**(5), 936–941 (2019)
11. Open Food Facts contributors: Open food facts (2025). https://world.openfoodfacts.org. Accessed 28 May 2025
12. Public Health England: Health matters: Obesity and the food environment (2017). https://www.gov.uk/government/publications/health-matters-obesity-and-the-food-environment/health-matters-obesity-and-the-food-environment--2

13. Rauber, F., et al.: Ultra-processed food consumption and risk of obesity: a prospective cohort study of UK biobank. Eur. J. Nutr. **60**, 2169–2180 (2021)
14. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)
15. Suksatan, W., et al.: Ultra-processed food consumption and adult mortality risk: a systematic review and dose–response meta-analysis of 207,291 participants. Nutrients **14**(1), 174 (2021)
16. Wagner, S., Wagner, D.: Comparing clusterings: an overview (2007)
17. Zhang, T., Ramakrishnan, R., Livny, M.: Birch: an efficient data clustering method for very large databases. ACM SIGMOD Rec. **25**(2), 103–114 (1996)