JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

NutriCluster-India: Personalised Dietary Correction through K-Means Clustering of Indian Foods by Macronutrient Density

Aditya Saini[1] *B. Tech (AI&DS) 5th Sem*
Deepak Prajapati[2] *B. Tech (AI&DS) 5th Sem*
Ayush Azad[3] *B. Tech (AI&DS) 5th Sem*
Mr. Ritesh[4] *Assistant Professor*

Department of Artificial Intelligence & Data Science
Dr Akhilesh Das Gupta Institute of Professional Studies, New Delhi, India

Abstract*—*Background: **Unhealthy eating is a major contributor to the global burden of non-communicable disease, but the personalised nutrition tools in common use are almost all designed around Western food composition databases. In the Indian setting, where cuisine varies by region, has a distinct nutritional makeup and relies on ingredients that these tools cannot identify, the problem goes well beyond inconvenience: the tools are simply not relevant to what people eat.**

Methods: **This study presents NutriCluster-India, a system built on the Indian Food Nutrition dataset. Dishes are grouped by K-Means clustering applied to three per-calorie features: protein density, carbohydrate density and fat density. A custom quality score, which gives credit for protein and dietary fibre and subtracts for free sugar and excess fat, orders the foods inside each cluster. For every user, calorie and macronutrient targets come from the Mifflin–St Jeor BMR equation, adjusted for activity level and reduced by 20% to support weight loss. Reported intake is then compared against these targets to find which macronutrients fall short, and the system suggests particular foods from the cluster that best matches each shortfall, with portion sizes computed to cover the gap.**

Results: **The Elbow Method and the Silhouette Score both supported K = 4 as the number of clusters. The four profiles obtained (High Carbohydrate–Low Fat, Low Carbohydrate, High Protein–High Fat and Low Protein) match eating patterns that readers familiar with Indian food will readily recognise. In a case study based on a real user profile, the system detected the macronutrient deficiencies correctly and returned food suggestions with suitable serving sizes.**

Conclusion: **Grouping foods by macronutrient density offers a practical and easily interpreted way to stratify Indian dishes and to produce individualised dietary corrections. Because the system is modular, it can be adapted to other food databases or to other health goals with little effort.**

Index Terms*—*dietary pattern analysis, K-Means clustering, Indian food nutrition, macronutrient density, personalised diet correction, Mifflin–St Jeor, BMR, silhouette score, unsupervised learning, food recommendation

I. Introduction

Most nutrition apps assume a food environment very different from that of an Indian home. Everyday staples such as dal, sabzi, rice and roti have macronutrient profiles that differ considerably from those of the chicken salads and protein shakes around which Western health-tracking tools are built. If an app suggests foods that the user never actually encounters, the suggestion is ignored, and the personalisation effort fails before it begins.

The consequences for health are considerable. Imbalanced diets, whether low in protein, high in refined carbohydrate or poor in fat quality, rank among the leading modifiable risk factors for obesity, type-2 diabetes and cardiovascular disease [1]. The pressure is especially visible in India, where fast urbanisation and cheap processed foods have been displacing traditional meals based on legumes, whole grains and vegetables. The newer eating patterns are energy dense yet lack the nutrients that matter most [2]. Disease is accumulating faster than individual, face-to-face dietary counselling could ever be scaled to meet it.

Current digital diet tools can be divided into two groups, each with shortcomings. Meal plans written by experts are inflexible by design: they cannot respond to what a person really eats or to what that person's body needs. Machine-learning alternatives are more adaptable but nearly always need labelled data about users, for instance ratings, interaction logs or clinical outcomes, and such data are hard to obtain in most practical settings, India included [6]. A lighter and more transparent option is needed, one that derives structure from food composition data alone, couples it with standard physiological equations and gives personalised feedback without a labelled training set [3].

NutriCluster-India, the system described here, is intended to fill that need and has three components. It first clusters Indian dishes on their macronutrient-density profile, measuring protein, carbohydrate and fat per calorie rather than as absolute amounts. It then derives a user's calorie and macronutrient targets from weight, height, age, sex and activity level. Finally, it sets those targets against what the user has actually eaten, finds the shortfalls and proposes specific foods from the most suitable cluster, in portions chosen to make up each shortfall.

This paper makes three contributions:

- A K-Means pipeline for the Indian Food Nutrition dataset that works on density features, so that portion size no longer confounds the clustering.
- A composite nutritional quality score for ranking Indian dishes inside each cluster, favouring protein and fibre over dishes high in sugar or fat.
- A complete personalised correction workflow that begins with a user's body measurements and ends with specific foods and serving sizes in grams.

II. Related Work

A. Approaches to Dietary Pattern Analysis

Three families of method appear in the dietary pattern literature. Expert-defined indices, among them the Healthy Eating Index and the DASH diet score, are the first; they quantify how well someone's intake matches a fixed template [3]. They are transparent and clinically easy to explain, yet they can only recognise patterns that their authors had in mind, which makes them suited to testing hypotheses but not to generating them.

Data-driven dimension reduction, chiefly Principal Component Analysis (PCA) and Factor Analysis, forms the second family. These methods find patterns in the data itself with no prior assumptions, but they return continuous factor scores instead of discrete groups. A patient gains little from being told that they score 1.3 on a given factor, and such a number cannot be turned directly into advice about what to eat.

Cluster analysis is the third family. It places each subject, or in this work each food, in a distinct group that can be interpreted. Sauvageot et al. contrasted K-Means, K-Medians and Ward's hierarchical method over several numbers of clusters and judged reliability with three stability indices [4]; K-Means gave the most stable partition on every index. They also compared the outcome with PCA-derived factors and found that the K-Means clusters closely resembled the PCA patterns, which indicates that the groups capture real structure and are not products of the method.

B. K-Means Applied to Diet and Health Outcomes

Two recent studies bear directly on our approach. Kirti and Singh ran K-Means on dietary frequency data for 12,318 adolescents from the nationally representative CNNS survey (2016–18) in India [5]. Of the five clusters found, one was an obesogenic pattern that included 24% of the sample and was significantly linked with obesity and pre-diabetes. Their study is the nearest to ours, since both use K-Means on Indian dietary data and both choose K by combining the Hopkins statistic, the elbow method and the silhouette score. The difference lies in what is clustered: they group individuals by how often they eat certain foods, whereas we group foods by their macronutrient composition.

Park et al. used K-Means on the food intake records of 57,213 Korean adults in a population cohort and related cluster membership to the risk of chronic kidney disease (CKD) through Cox proportional-hazards regression [6]. People in the cluster marked by high carbohydrate intake and little dietary variety had a 59% greater CKD risk than those in the healthiest cluster (aHR 1.59, 95% CI 1.04–2.41). The result is relevant beyond kidney disease, because it shows that clusters derived only from nutritional composition can predict real health outcomes, which supports the biological meaning of such groupings.

C. The Gap This Work Fills

We found no earlier study, among these or elsewhere, that clusters Indian foods by macronutrient density, computes individual macronutrient targets from biometric equations and joins the two in an automatic recommendation engine that fills nutritional gaps. The present work is aimed at that gap.

III. Methods

A. Dataset

The Indian Food Nutrition dataset contains a broad selection of Indian dishes. Each entry gives values per 100 g for calories (kcal), carbohydrates (g), protein (g), total fat (g), free sugar (g), dietary fibre (g) and sodium (mg). Our analysis used the macronutrient variables only; micronutrients such as calcium and iron were left out. Their omission is not a statement about their importance. The system is built specifically to correct macronutrient intake, and adding micronutrient dimensions without a sound weighting scheme would introduce noise rather than useful signal into the clustering.

B. Preprocessing

Food composition records tend to be noisy in predictable ways, and clustering reacts strongly to the outliers they usually contain, so we cleaned the data in five steps.

The first step excluded items providing fewer than 50 kcal per 100 g, mainly garnishes, fresh herbs and watery drinks, which supply almost no macronutrients in realistic portions and whose exaggerated density ratios would skew the feature distributions. Next, items with under 1 g of protein were discarded, which removed confectionery and sweetened beverages that are of no use as dietary corrections. The third step filtered by keyword to exclude deep-fried snacks (pakora, vada, bhaji), sweets and desserts (halwa, laddu, barfi), condiments (chutney, pickle, sauce) and fatty garnishes (butter, cream, ghee). In the fourth step we removed items above 15 g of free sugar per 100 g, so that refined-sugar foods of poor nutritional value could not enter the High Carbohydrate cluster. The final step excluded items above 25 g of total fat per 100 g, because such outliers would drag the cluster centroids toward unrepresentative extremes.

Table I gives an overview of the dataset before and after these steps.

TABLE I
Dataset Summary Before and After Preprocessing

Attribute	Raw	Filtered
Features Used	7	3 (density-based)
Calorie Range (kcal)	0–>500	50–≈500
Sugar Ceiling (g)	Unrestricted	≤ 15
Fat Ceiling (g)	Unrestricted	≤ 25
Clusters Produced	–	4

C. Quality Score

Cleaning settles one issue and raises another, namely how to choose which foods to recommend from within a cluster. Foods in the same cluster differ in quality; for instance, some high-protein items also contain a lot of saturated fat or free sugar. We therefore defined a composite quality score Q as a measure of overall nutritional merit:

$$Q = 2P + 2F_b + 0.5C - 3S - 0.5F_t \quad (1)$$

in which P denotes protein (g), F_b dietary fibre (g), C carbohydrates (g), S free sugar (g) and F_t total fat (g), each per 100 g.

The weights were chosen for the following reasons. Protein is rewarded most strongly, as Indian diets most often fall short on it and it is the main factor in preserving lean mass while energy intake is restricted [2]. Fibre receives an equal reward, since it slows glucose absorption and benefits gut health and is likewise under-represented in urban Indian diets. A small positive weight on carbohydrate recognises it as a necessary source of energy. Free sugar carries the largest penalty, in line with the WHO guidance to keep it below 10% of daily energy [1]. Total fat is penalised mildly so that groups that are already rich in fat are not further filled with fatty options, a clustering-within-a-cluster artefact.

D. Feature Engineering

A basic problem arises when raw macronutrient values are clustered. A food with 20 g of protein and 100 kcal per 100 g

appears the same as one with 20 g of protein and 500 kcal, yet protein is a large part of the energy in the first and a minor part in the second. Clustering on absolute protein would therefore reflect energy density more than nutritional quality.

To avoid this, each macronutrient is expressed as a density, meaning the quantity supplied per calorie:

$$d_P = P/\text{kcal}, \quad d_C = C/\text{kcal}, \quad d_F = F_t/\text{kcal}. \tag{2}$$

These three features capture a food's macronutrient profile regardless of how energy dense it is. We then standardised them using z-scores, so that no feature has undue influence on the Euclidean distance used by K-Means.

E. Clustering

K-Means was applied to the normalised $n \times 3$ density matrix, with the objective of minimising the within-cluster sum of squares:

$$J = \Sigma_{k=1}^{K} \Sigma_{x \in *C*k} \|x - \mu_k\|^2. \tag{3}$$

The settings were k-means++ initialisation to avoid unfavourable starting points, *n_init* = 10 to lessen dependence on the random seed, and *random_state* = 42 so that results can be reproduced.

Two complementary diagnostics were used to choose K. The Elbow Method plots the within-cluster sum of squares (WCSS) for K between 2 and 9 and looks for the value after which extra clusters bring little further decrease in inertia. For a point i, the Silhouette Score is defined as

$$s(i) = [b(i) - a(i)] / \max(a(i), b(i)), \tag{4}$$

where $a(i)$ is the average distance from i to the other points in its own cluster and $b(i)$ is the average distance to the points of the closest other cluster. Values close to +1 mean that the point lies well apart from neighbouring clusters. For each candidate K we averaged this score over all points.

F. Personalised Target Computation

Four inputs from the user (weight, height, age and sex) are enough to compute the calorie and macronutrient targets. The basal metabolic rate (BMR) follows the Mifflin–St Jeor equations:

$$\text{BMR}_{\text{male}} = 10W + 6.25H - 5A + 5, \tag{5}$$

$$\text{BMR}_{\text{female}} = 10W + 6.25H - 5A - 161, \tag{6}$$

with W the weight in kg, H the height in cm and A the age in years. Multiplying BMR by an activity factor (AF) gives the total daily energy expenditure:

$$\text{TDEE} = \text{BMR} \times \text{AF}, \tag{7}$$

The activity factors are 1.2 for sedentary, 1.55 for moderately active and 1.75 for highly active individuals. For weight loss the target is set 20% below expenditure:

$$E_{\text{target}} = 0.8 \times \text{TDEE}. \tag{8}$$

To keep the prescription clinically safe, the target is never allowed to fall below 1,500 kcal for males or 1,200 kcal for females. The macronutrient targets are then calculated as follows:

$$\text{Protein}_{\text{req}} = 1.6 \times W \ \text{(g/day)}, \tag{9}$$

$$\text{Fat}_{\text{req}} = 0.8 \times W \ \text{(g/day)}, \tag{10}$$

$$\text{Carbs}_{\text{req}} = [E_{\text{target}} - 4\,\text{Protein} - 9\,\text{Fat}] / 4 \ \text{(g/day)}. \tag{11}$$

The 1.6 g/kg protein figure is at the top of the range supported by evidence for retaining lean mass during an energy deficit [2]. Carbohydrate is given whatever energy remains once the protein and fat allowances are subtracted.

G. Gap Analysis and Recommendation

With the targets set, the system measures the shortfall against reported intake. For each macronutrient m, the difference is

$$\Delta_m = m_{\text{intake}} - m_{\text{req}}. \tag{12}$$

A negative difference indicates intake below the requirement. Shortfalls are graded as mild when they lie between −15 and −40 g for protein or fat, or between −30 and −80 g for carbohydrates, and as severe when they exceed those limits.

For every deficient macronutrient, the engine picks the cluster whose mean density of that nutrient is highest. It ranks the foods in that cluster by their quality score Q and returns the three best that still fit within the user's remaining calorie allowance. The portion that bridges the gap is calculated as

$$s = \min(150, |\Delta_m| / d_m \times 100) \text{ g}, \tag{13}$$

with a ceiling of 150 g per suggestion so that the quantities remain realistic.

If the user has already reached the calorie target, the system enters *Replacement Mode*. In this mode its suggestions are flagged as swaps, that is, foods to substitute for others and not to add to the day's intake.

IV. Results and Discussion

A. Exploratory Data Analysis

The feature distributions were inspected before clustering. In keeping with food composition data generally, raw macronutrient values were right-skewed, with most dishes at the low end and a long tail of nutrient-rich items. The correlation heatmap (Fig. 1) showed a notable pattern: calories correlate strongly with carbohydrates ($r \approx 0.80$) and moderately with fat, whereas protein is only weakly related to calories ($r \approx 0.35$). This implies that clustering on raw protein is to a large extent clustering on energy density. Using density features removes that confound so that the macronutrient profile can be seen by itself.

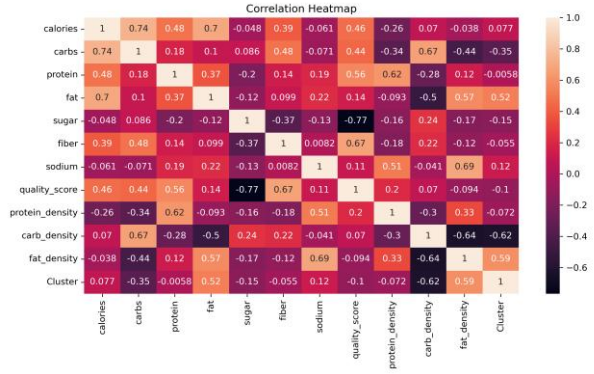


Fig. 1. Pearson correlation heatmap of the raw nutritional features. The strong calories–carbohydrates correlation ($r \approx 0.80$) is the reason for clustering on per-calorie density features.

B. Choosing the Number of Clusters

In the WCSS curve (Fig. 2), inertia drops quickly between $K = 2$ and $K = 4$ and then levels off. The bend at $K = 4$ is clear; a fifth cluster lowers inertia by very little and just splits an existing group instead of revealing a new one.

Silhouette scores (Fig. 3) are highest for $K = 3$ and $K = 4$. We preferred $K = 4$ because the extra cluster distinguishes the Low Protein profile from the Low Carbohydrate profile, and these call for quite different dietary corrections. The two solutions differ in silhouette score by less than 0.02, so the loss in statistical terms is trivial, while the gain in interpretability is substantial.

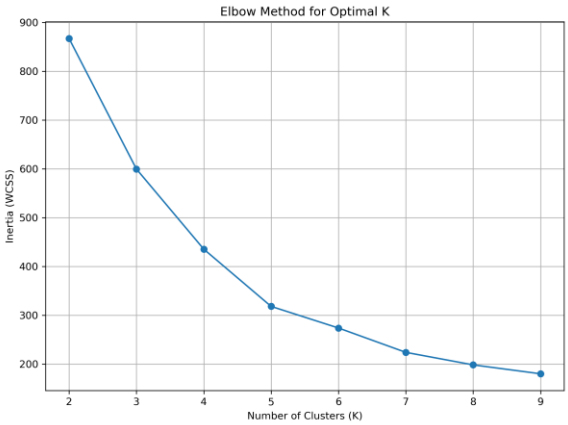


Fig. 2. Elbow Method: within-cluster sum of squares for each value of *K*, with a clear bend at *K* = 4.

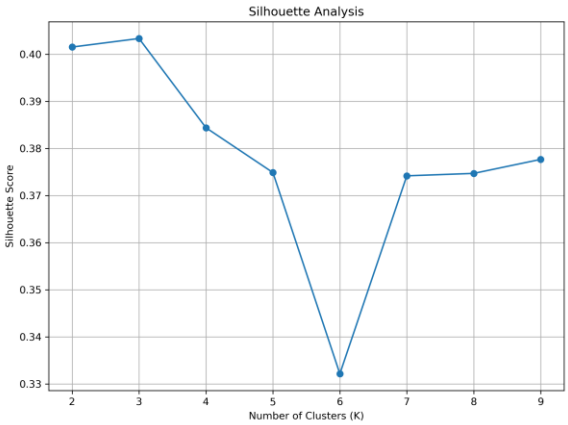


Fig. 3. Mean Silhouette Score for each value of *K*. The scores for *K* = 3 and *K* = 4 are similar, and *K* = 4 is chosen because it is easier to interpret.

C. Cluster Profiles

Table II lists the four clusters with the relative density of each macronutrient.

Cluster 0 is the biggest group. It holds cereal- and pulse-based preparations such as dal varieties, rice dishes, khichdi and roti combinations, where carbohydrate supplies most of the energy and fat is comparatively low. Together they make up the core of a typical Indian meal.

Protein-rich foods are concentrated in Cluster 2, which includes thick lentil soups (dal makhani, chana dal), egg dishes and paneer preparations. That high protein occurs together with high fat is to be expected, as both legumes and dairy contain a good deal of fat besides protein.

TABLE II
K-MEANS CLUSTER PROFILES (K = 4). DENSITY LABELS ASSIGNED BY RANKING CLUSTER CENTROIDS.

Cluster	Label	d_P	d_C	d_F
0	High Carb, Low Fat	Low	High	Low
1	Low Carb	Medium	Low	Medium
2	High Protein, High Fat	High	Medium	High
3	Low Protein	Low	Medium	Medium

Cluster 1 consists of non-starchy vegetable dishes, including dry sabzis, raitas and mixed salads. These are low in total calories and so act as bulk foods for users who have nearly used up their energy allowance.

The most varied group is Cluster 3, which collects preparations that fit none of the other three clusters well: their protein content is modest compared with their energy, and they have no distinct carbohydrate or fat signature.

Fig. 4 shows that Clusters 0 and 3 contain most of the dishes. This is hardly surprising, since it reflects the carbohydrate-rich, protein-poor nature of mainstream Indian eating, a shortfall repeatedly highlighted by the ICMR as a public-health concern [2].

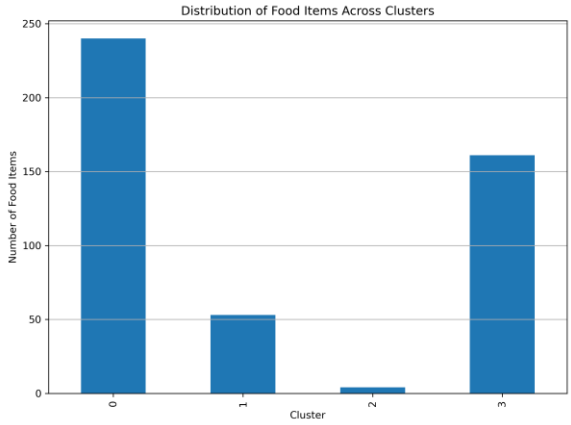


Fig. 4. Count of food items per cluster. Clusters 0 and 3 are the largest, in line with the carbohydrate-rich Indian food environment.

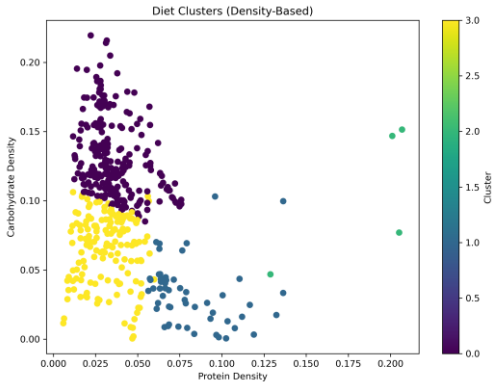


Fig. 5. Protein density against carbohydrate density, coloured by K-Means cluster. Clusters 0 and 2 are clearly separated.

D. Case Study: End-to-End Validation

To check that the whole pipeline works from input to output, we applied it to a typical profile from our own group: a 19-year-old moderately active male, 85 kg and 175 cm.

The targets it generated were BMR = 1,854 kcal, TDEE = 2,873 kcal, a calorie target of 2,299 kcal after the 20% deficit, protein = 136 g, fat = 68 g and carbohydrates = 286 g. On the day in question the user had eaten exactly 2,299 kcal, but the distribution across macronutrients was far from the targets, at protein = 100 g, fat = 40 g and carbohydrates = 190 g.

The gap analysis detected three shortfalls at once: $\Delta_P = -36$ g (a severe protein gap), $\Delta_C = -96$ g (a large carbohydrate deficit) and $\Delta_F = -28$ g (a mild fat deficit). With the calorie budget already used up, the system appropriately entered Replacement Mode. A representative part of its output appears in Table III.

TABLE III
SAMPLE RECOMMENDATIONS GENERATED (REPLACEMENT MODE)

Gap	Food	Qty (g)	Macro Provided	kcal
Protein	Millet Soup	150	16.9 g protein	84
Protein	Mulligatawny Soup	150	16.9 g protein	82
Carbs	Maa Chaane Ki Dal	150	64.7 g carbs	517
Fat	Millet Soup	150	18.9 g fat	84
Fat	Mulligatawny Soup	69	9.1 g fat	37

Millet soup is on both the protein list and the fat list. This is correct and not a fault: the soup is in Cluster 2 (High Protein, High Fat), and within that cluster it earns a high quality score thanks to its relatively high fibre and low free sugar. A deployed version should remove duplicates so that an item is not displayed twice, but the logic itself is valid, because the algorithm found the single best food for two deficiencies at once.

The carbohydrate suggestion, Maa Chaane Ki Dal at 517 kcal, may look odd for Replacement Mode considering how many calories it carries. It is deliberate, however. Replacement Mode does not screen on calorie cost; it selects the best macronutrient match and uses the replacement flag to inform the user that something else must be removed to make space. One obvious refinement would be to rank candidates by the macronutrient gap closed per calorie displaced, giving a better balance between benefit and energy cost.

Fig. 5 plots the four clusters against protein density and carbohydrate density. The separation between Cluster 0 (High Carb) and Cluster 2 (High Protein) is especially clean, which shows that macronutrient density discriminates well between Indian foods.

E. Limitations

The principal limitation is that the system considers macronutrients only. Shortages of iron, vitamin B12, vitamin D and calcium, which are common in Indian vegetarian diets, cannot be seen by the pipeline at all. The most valuable next step towards clinical use is to extend the quality score and the gap analysis to micronutrient density.

Because intake is self-reported, every recommendation is subject to recall bias and to errors in estimating portions. Removing this problem would require an image-based food-logging tool or a verified tracking application.

A further assumption of K-Means is that clusters are roughly spherical and of similar size. Fig. 4 shows that Clusters 0 and 3 are clearly bigger than Clusters 1 and 2. Gaussian Mixture Models, which estimate a separate covariance structure for each cluster, might cope better with this imbalance, and future work should compare them with K-Means.

Last, only weight loss is supported as a goal. For muscle gain or weight maintenance, different macronutrient targets are required, along with a calorie surplus or balance in place of a deficit.

V. CONCLUSION

The aim here was a dietary correction tool that genuinely suits Indian food, starting from what people in India eat and not from a Western food environment adapted as an afterthought. NutriCluster-India does so by clustering Indian dishes on macronutrient-density features and linking the clusters to a personalised target engine based on the Mifflin–St Jeor equation.

The four clusters found at $K = 4$, which both the Elbow Method and the Silhouette Score support, are more than statistical artefacts. Each corresponds to a recognisable dietary profile: staple dal and roti, low-calorie vegetable sides, protein-rich lentil soups and a starchy mixed group in between, all of which an Indian nutritionist would recognise at once. The dominance of Clusters 0 and 3 matches the ICMR's documented findings on protein deficiency in Indian diets, so the clustering seems to mirror real nutritional conditions as well as being technically sound.

The case study showed that the pipeline performs correctly from start to finish. It found three macronutrient deficiencies at the same time, switched to the proper mode once the calorie budget was spent and returned specific recommendations with portion sizes from the relevant cluster, which goes beyond what most diet apps now offer Indian users.

Future work should prioritise three things: extending the gap analysis to micronutrients, tracking users over time so that patterns can be detected and not just single-day snapshots, and introducing a proper macronutrient-per-calorie scoring rule for Replacement Mode. Comparing Gaussian Mixture Models with K-Means on the same features would also strengthen the methodology. A study involving real users and a qualified dietitian is indispensable before any clinical use.

REFERENCES

[1] World Health Organization, "Healthy diet," WHO Fact Sheet, Apr. 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/healthy-diet

[2] Indian Council of Medical Research, "Dietary Guidelines for Indians," 2nd ed., NIN, Hyderabad, 2020.

[3] J. Zhao et al., "A review of statistical methods for dietary pattern analysis," *Nutrition Journal*, vol. 20, no. 37, 2021, doi: 10.1186/s12937-021-00692-7.

[4] N. Sauvageot et al., "Stability-based validation of dietary patterns obtained by cluster analysis," *Nutrition Journal*, vol. 16, no. 4, 2017, doi: 10.1186/s12937-017-0226-9.

[5] K. Kirti and S. K. Singh, "Obesogenic diet and metabolic syndrome among adolescents in India: data-driven cluster analysis," *BMC Cardiovascular Disorders*, vol. 23, no. 393, 2023, doi: 10.1186/s12872-023-03429-y.

[6] D. Park et al., "Dietary patterns associated with the new onset of chronic kidney disease using clustering algorithm," *Scientific Reports*, vol. 15, no. 33569, 2025, doi: 10.1038/s41598-025-18627-1.