Intelligent Expert System to Optimize Nutritional Care: Analysis Based on Machine Learning and Clinical Rules

Miguel Angel Valles-Coral [1,*], Jorge Raul Navarro-Cabrera [1], Cristian García-Estrella [1], Jorge Valverde-Iparraguirre [1], Sarita Saavedra [2], and Luz Karen Quintanilla-Morales [2]

[1] Department of Systems and Computer Science, Faculty of Systems Engineering and Computer Science, National University of San Martín, Tarapoto, Peru
[2] Department of Human Medicine, Faculty of Human Medicine, National University of San Martín, Tarapoto, Peru
Email: mavalles@unsm.edu.pe (M.A.V.C.); jnavarroc@unsm.edu.pe (J.R.N.C.); cgarcia@unsm.edu.pe (C.G.E.); jdvalver@unsm.edu.pe (J.V.I.); sgsaavedrag@unsm.edu.pe (S.S.); lquintanilla@unsm.edu.pe (L.K.Q.M.)
*Corresponding author

***Abstract*—This study developed and validated an intelligent expert system based on machine learning and clinical rules to optimize nutritional care for patients in the San Martín region of Peru. The approach specifically addresses challenges posed by limited access to specialized dietary guidance. A dataset consisting of 615 patient records, each with 45 clinical, anthropometric, and lifestyle variables, was used to train and implement a hybrid architecture that combined unsupervised clustering (K-means) with a rule-based inference engine. The model achieved optimal segmentation into five nutritional profiles, supported by dimensionality reduction through Principal Component Analysis (PCA) and validated using the Silhouette (0.1229), Davies-Bouldin (1.3238), and Calinski-Harabasz (47.8740) indices. The expert system's dietary recommendations showed a global concordance of 77% with those of the clinical nutritionists, achieving up to 83% accuracy in standard metabolic clusters. These results confirm the system's ability to generate personalized and clinically coherent dietary plans, thereby improving the efficiency of nutritional services. Overall, the findings highlight the potential of intelligent systems to enhance healthcare delivery in resource-limited settings and pave the way for broader applications of AI in public health and clinical decision support.**

***Keywords*—clinical support, clustering algorithms, dietary recommendation, digital public health, nutritional profiling**

I. Introduction

Health is a fundamental aspect of human life [1]. Within this context, adopting a healthy lifestyle has become a key priority for overall well-being [2]. This requires the incorporation of appropriate eating habits–considering both the type and quantity of food consumed—to meet each individual's specific caloric and nutritional needs. These habits should be complemented by regular physical activity [3]. According to Tan [4], maintaining a healthy lifestyle not only optimizes physiological functions, but also significantly reduces the risk of developing chronic diseases such as diabetes, obesity, and other disorders associated with sedentary behavior and poor nutrition.

Within the healthcare context, a nutritious diet is essential for patients, as it strengthens the immune system, increases resistance to a range of diseases, and contributes to an improved quality of life [5, 6]. According to Werner and Osterbur [7], dietary plans for patients must ensure an appropriate balance of micronutrients and macronutrients, including vitamins, proteins, minerals, antioxidants, fiber, and healthy fats. However, implementing personalized plans faces significant challenges, such as high consultation costs and long waiting times for access to nutrition specialists [8]. In response to these barriers, technological solutions based on artificial intelligence and recommender systems have been developed to generate individualized diets, thereby optimizing the efficiency of nutritional care processes [9].

Reports from the Regional Health Directorate of the San Martín region indicate that one of the main causes of morbidity is linked to poor dietary habits, despite the region's vast food resources, to the extent that it is considered the most diverse food-producing area in Peru. Furthermore, according to the same source [10], the average number of patients per physician in the region is 1771, ranking second-lowest in the country, with only Cajamarca reporting a higher figure of 1939 patients per physician.

In practical terms, health service coverage indicators in the San Martín region are inadequately met [11]. Specifically, regarding nutrition professionals, the National Institute of Statistics and Informatics [12] reported that the region has approximately 900,000 inhabitants but only 22 nutritionists distributed throughout the territory. This creates a demand that is difficult to meet, as each professional attends an average of 15 patients per day in outpatient consultations, in addition to providing consultation services for other programs such as

Manuscript received May 29, 2025; revised August 11, 2025; accepted August 20, 2025; published November 21, 2025.

Tuberculosis (TB), Human Immunodeficiency Virus (HIV), and other infectious diseases. As a result, the average waiting time to obtain an appointment is approximately 10 days.

Once the consultation begins, the situation becomes even more critical. According to established parameters, the maximum allotted consultation time per patient is 20 min. However, due to the service's reliance on anamnesis, this time is insufficient. Part of the allotted time is underused because the nutritionist must manually analyze and cross-reference nutritional evaluation components using complex tables. This makes the process labor-intensive and tedious. Furthermore, inadequate identification, measurement, monitoring, and control of service quality parameters by nutritionists make it extremely difficult to improve care outcomes.

Our literature review indicates that machine learning-based nutritional recommendation systems have emerged as the most effective tools for helping users change their eating behavior through personalized suggestions [13, 14]. These systems are software applications that act as information filters, helping users navigate large sets of alternatives and identify options that match their preferences [15]. Tapia *et al.* [16] and Bondevik *et al.* [17] indicate that although these technologies still have limitations that need to be addressed, they have shown promising results in promoting health, with significant progress in the field of nutritional care.

This study aimed to recommend healthy diets to patients who were receiving nutritional services in the San Martín region of Peru. We applied unsupervised learning algorithms to classify patients based on their nutritional characteristics. Furthermore, we implemented an expert system that integrated supervised learning algorithms and nutritional evaluation components to generate personalized dietary recommendations. Finally, the outcomes of these recommendations were evaluated according to the clinical evolution of different patient categories, culminating in an analysis of the expert system's implementation.

II. Literature Review

A. Personalized Nutrition and AI-Driven Dietary Recommendation

Recent surveys have consistently shown a rapid shift from generic advice to Personalized Nutrition (PN) driven by Artificial Intelligence (AI) and Recommender Systems (RSs) for dietary guidance [13, 17]. Methodologically, three main families predominate: (i) supervised models that predict targets (e.g., glycemic response or diet class) from clinical and behavioral features; (ii) collaborative filtering models that leverage preference and consumption patterns; and (iii) knowledge-based or expert approaches that encode clinical guidelines. Hybridization is becoming increasingly prevalent to balance predictive power and clinical safety. For example, Internet of Medical Things (IoMT)-assisted models integrate streaming data from IoT devices with Machine Learning (ML) to tailor recommendations [18], whereas knowledge-centric frameworks like PROTEIN AI formalize expert rules and then optimize meal plans. Beyond technical accuracy, evidence from a Randomized Controlled Trial (RCT) using eNutri indicates that automated PN advice can improve diet quality and engagement in the short term [18], and usability studies (e.g., the SAlBi mobile nutrition education app, which provides self-monitoring and Mediterranean diet–based feedback) have highlighted its acceptability in real-world settings [19]. Large cohort analyses further suggest that Collaborative Filtering (CF)-style models can scale PN by exploiting population dietary patterns while retaining interpretability for major food groups [20].

B. Clinical Decision Support and Rule-based Safety in Nutrition Care

In clinical workflows, rule-based Clinical Decision Support Systems (CDSS) remain essential because they encode strict safety constraints (e.g., sodium restriction in hypertension, carbohydrate management in diabetes, avoidance of allergens), provide transparent rationales, and support auditable decision paths. Hybrid designs that prioritize explicit rules while using data-driven context as a probabilistic soft prior are recommended to mitigate risks from model uncertainty or population data drift. This study follows that paradigm: cluster context informs typical profiles, but diet prescriptions are ultimately governed by rule overrides in the presence of contraindications, which aligns with established best practices for safety-critical clinical decision support [21, 22].

C. Phenotyping by Clustering and Dimensionality Reduction

Unsupervised phenotyping (e.g., K-means, hierarchical clustering, Gaussian Mixture Model (GMM)) is widely used to uncover multimorbidity and metabolic profiles from mixed clinical-behavioral data. However, high dimensionality after feature encoding and cluster imbalance can reduce internal separation (e.g., moderate Silhouette scores) and create long-tail phenotypes. Standard practice combines Principal Component Analysis (PCA) for denoising and compactness, multi-index model selection (Silhouette, Davies–Bouldin Index (DBI), Calinski–Harabasz Index (CH), and stability checks (bootstrap Adjusted Rand Index (ARI)) to support robustness. Hybrid priors should be assigned lower weights for rare clusters with low statistical confidence. In nutrition informatics, clustering helps map actionable profiles (e.g., overweight with prediabetes vs. younger overweight with normal glycemia), which can be directly translated into diet rule templates—an approach that our system operationalizes [23].

D. Missing Data, Interpretability, and External Validation

In clinical datasets with mixed data types, tree-based imputation methods (e.g., MissForest, Recursive Feature Elimination-MissForest (RFE-MF) variants) often achieve lower error rates than mean or mode, k-Nearest Neighbors (kNN), or Multiple Imputation by Chained Equations (MICE), but require rigorous nested validation to avoid

data leakage in unsupervised pipelines. Our conservative strategy is benchmarked, therefore, against these methods through sensitivity analyses.

Clinical datasets often exhibit complex patterns of missingness and mixed data types. While advanced imputers (e.g., k-Nearest Neighbors (kNN), Multiple Imputation by Chained Equations (MICE)) can be effective, they rely on assumptions—such as the reliability of distance metrics in high-dimensional one-hot encoded spaces or the Missing at Random (MAR) assumption for chained models—and may inflate correlations that distort PCA or clustering results. Conservative imputation (median/mode with missingness indicators) plus sensitivity analyses against kNN and MICE offer a reproducible alternative in unsupervised learning pipelines. Regarding interpretability, combining model-agnostic explanations (e.g., tree-based SHapley Additive exPlanations (SHAP) values) with feature importance from non-linear ensembles and classical effect-size tests improves clinical traceability. This is complementary to evidence from Randomized Controlled Trials (RCTs) and usability studies (eNutri, SAlBi), as well as expert concordance assessments used as external validation criteria [24–26].

E. Gap and Contribution

Despite notable advances in Recommender Systems and Precision Nutrition (RS/PN), few studies have reported hybrid rule-plus-clustering approaches that are validated against clinician judgment in resource-limited settings and incorporate safety overrides, imbalance handling, and stability checks. This study addresses that gap through four key contributions. First, it identifies nutritional phenotypes using PCA-aided clustering. Second, it implements a rule-first inference engine with adaptive integration of cluster priors. Third, it performs robustness analyses, including the gap statistic, stability assessment, and sensitivity to both imputation methods and rare clusters. Finally, it conducts external validation with clinical nutritionists, demonstrating that personalized, auditable recommendations can be delivered at scale within a public-health context.

III. Materials and Methods

A. Data Collection Instrument

For data collection, we used a structured form for clinical and nutritional assessment, specifically designed to capture relevant information within the nutrition service at Hospital II-2 in Tarapoto. The instrument comprised 45 variables divided into two main sections: sociodemographic and lifestyle data, and clinical and biochemical data.

The first section covered demographic characteristics (sex, age, educational level, employment status) and basic anthropometric parameters (weight and height). It also included lifestyle aspects, such as tobacco and alcohol use, average hours of sleep, screen time, and frequency of physical activity. Additionally, it recorded food preferences and aversions, allergies, number of daily meals, usual methods of food preparation, frequency of consumption by food groups, and adherence to specific diet plans. Finally, it incorporated a self-assessment of physical condition based on the perceived ability to engage in moderate- or high-intensity physical activity.

The second section addressed key clinical and biochemical variables for nutritional evaluation, including a history of hypertension, diabetes mellitus, and dyslipidemia, as well as the use of antihypertensive or hypoglycemic medications. The analytical components included glucose, total cholesterol, triglycerides, hemoglobin, and hematocrit levels. The tool was developed in coordination with clinical nutrition and public health professionals to ensure its relevance to the local context and its ability to comprehensively capture the determinants of nutritional status.

Before implementation, the structured instrument underwent validation through expert judgment and statistical analysis. The global Content Validity Index (CVI) was 0.91, indicating high item relevance. A pilot test enabled the evaluation of the internal structure through exploratory factor analysis, yielding a Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy of 0.82 and a significant Bartlett's sphericity test ($p < 0.001$). Five factors were identified, explaining 67.4% of the total variance. Internal consistency was high, with an overall Cronbach's alpha of 0.85. These results confirmed the validity and reliability of the instrument for use in similar clinical contexts.

B. Proposed Model

To integrate the automated processes for collecting, processing, and analyzing clinical–nutritional data, we designed a functional model that integrates several methodological stages aimed at providing personalized healthy diet recommendations. This model combines unsupervised machine learning techniques to identify nutritional patterns with an expert system based on clinical rules to generate individualized recommendations. The operational workflow comprises five main components: data collection, preprocessing, clustering, expert system implementation, and results validation. The interaction between these modules enables the development of an intelligent solution capable of assisting healthcare personnel in high-demand, resource-limited settings. Fig. 1 schematically presents the overall architecture of the proposed model.

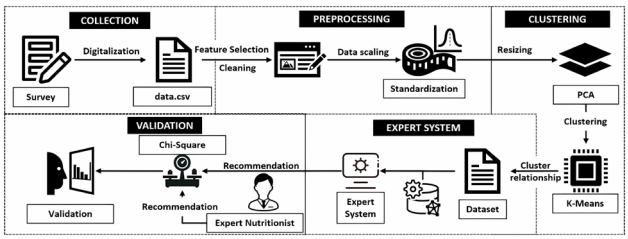

Fig. 1. Proposed model.

1) Phase 1: Data collection

Following the validation of the structured clinical and nutritional assessment instrument, we conducted

systematic field data collection, resulting in a database containing 623 records from individual patients. Each record included 45 variables covering sociodemographic, clinical, anthropometric, and lifestyle data. The information was initially digitized in Microsoft Excel (XLS) format, subjected to a consolidation and quality control process, and then converted to Comma-Separated Values (CSV) format for processing in the computational analysis modules. Each record was anonymized using a unique alphanumeric code to preserve participant confidentiality, in accordance with current ethical principles for clinical research. This process ensured data traceability without compromising subject identity, allowing for subsequent statistical and computational analyses.

The initial exploratory analysis of the dataset revealed substantial demographic heterogeneity, with balanced representation across different age groups, educational levels, and socioeconomic conditions, thus reinforcing the external validity of the proposed model. At the biochemical level, the observed values for glucose, cholesterol, and triglycerides were aligned with the expected epidemiological ranges for Latin American populations, indicating appropriate clinical consistency and data plausibility. Notably, the richness of the information regarding dietary habits and physical activity, representing approximately 40% of the total variables, allowed the identification of critical modifiable factors in the determination of nutritional status. This integration of clinical, metabolic, and behavioral data provides a robust foundation for identifying nutritional patterns and formulating personalized dietary recommendations.

From a data-structure perspective, the matrix exhibited a significant proportion of missing values: 40 of the 45 columns had at least one missing entry, which amounted to 8811 missing observations out of 28,035 possible, with no rows entirely complete. This characteristic presented substantial methodological challenges during the preprocessing phase, particularly regarding the imputation and differentiated handling of categorical and numerical variables. Table I summarizes the structural characteristics of the dataset.

TABLE I. PARTICIPANT PROFILE VALUES

Name	Values
Rows	623
Columns	45
Discrete Columns	10
Continuous Columns	35
Missing Columns	40
Missing Observations	8811
Complete Rows	0
Total Observations	28035

2) Phase 2: Data preprocessing

Data preprocessing began with an exploratory analysis aimed at evaluating the structure of the dataset, identifying inconsistencies, and quantifying the extent of missing values. A total of 8811 missing entries were identified, representing approximately 30% of all possible observations. A quality-control threshold was applied to exclude records with more than 50% missing data, resulting in the removal of eight records and yielding a final analytic sample of 615 patients.

Given the high proportion of missing values and the heterogeneous nature of the variables (clinical, anthropometric, and lifestyle), we adopted simple imputations based on variable mean/mode, avoiding advanced techniques such as MICE or KNN. While these methods can be powerful, they may introduce artificial correlations in clinical datasets with non-random missingness and low-frequency categories, potentially affecting interpretability and the rule-based inference process. Our approach prioritized preserving the original data structure and transparency in modeling, complemented by the use of clinical rules that mitigate the impact of imputed values on decision-making

This 1.3% reduction was considered an acceptable trade-off between sample size and data integrity, as the excluded records exhibited systematic data omissions caused by errors in data collection. The relationship between the exclusion threshold and the resulting sample size is shown in Fig. 2.

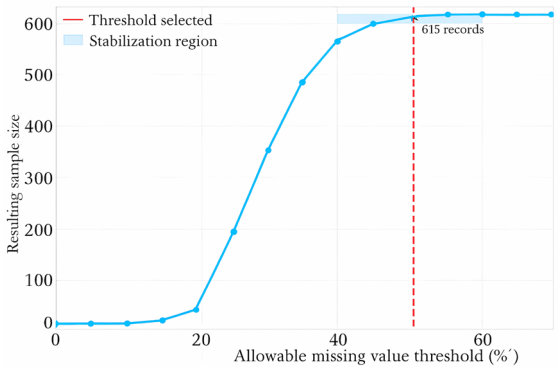

Fig. 2. Relationship between missing values threshold and sample size.

After handling the missing values, one-hot encoding was applied to the categorical variables [20, 25], thereby converting nominal categories into binary vectors. Consequently, the feature space expanded significantly, resulting in a matrix of 615 rows by 196 columns, thereby justifying the subsequent application of dimensionality reduction techniques. The numerical variables were then standardized using the StandardScaler method, which scales the data to have zero mean and unit variance [27]. This standardization prevented variables with larger numerical ranges from dominating the behavior of subsequent clustering algorithms.

As the final step, a correlation matrix was computed for the numerical variables to assess significant relationships between the clinical indicators. Relevant associations were identified, including positive correlations between Body Mass Index (BMI) and blood glucose, and between BMI and triglycerides ($r > 0.4, p < 0.01$), as well as between age and blood pressure ($r = 0.38, p < 0.01$), which is consistent with clinical patterns reported in the literature on metabolic syndrome in Latin American populations. These findings are presented graphically in Fig. 3 and serve as key inputs for clinical interpretation and for defining nutritional segmentation criteria in the subsequent phases of the study.

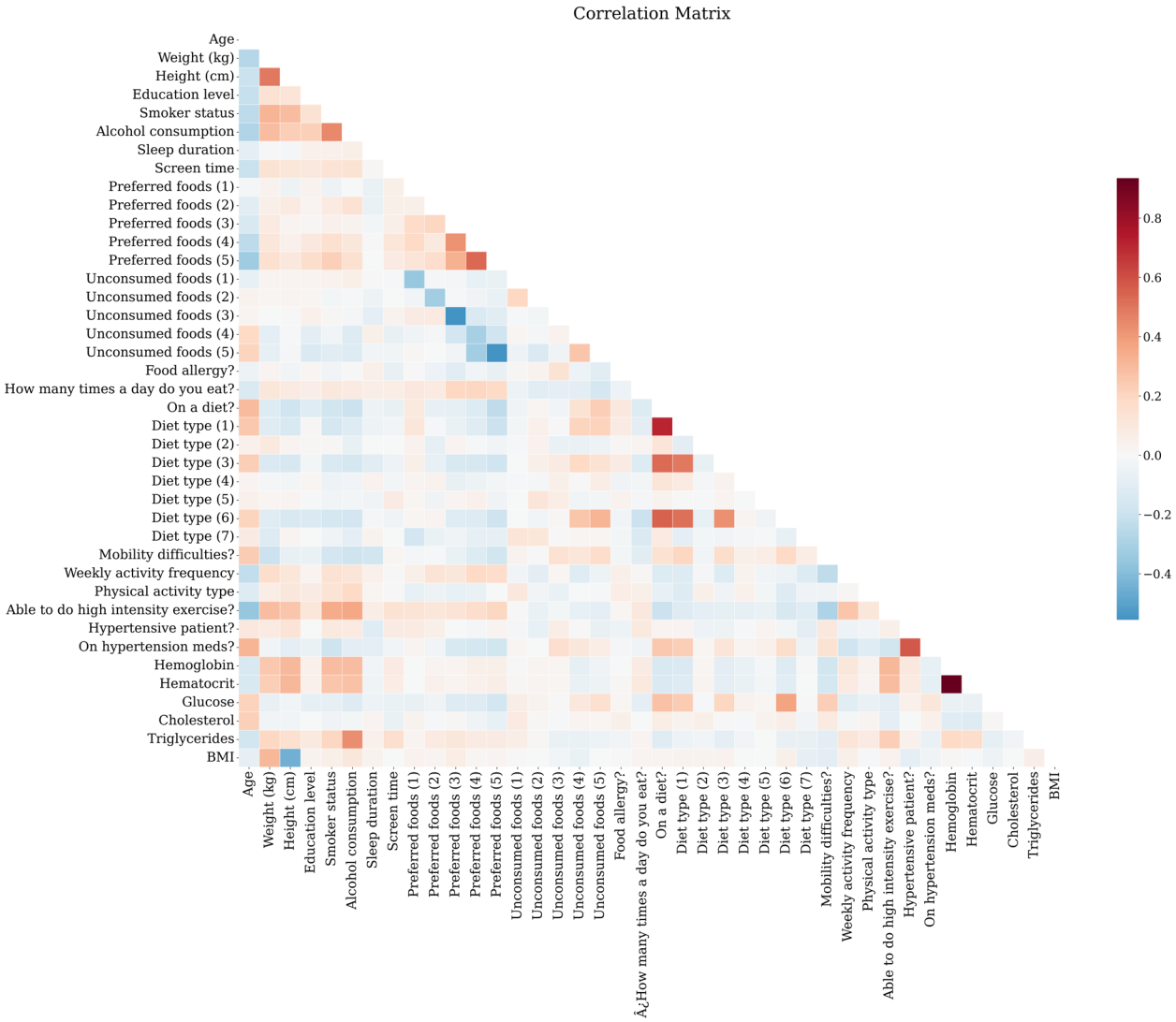

Fig. 3. Correlation matrix of numerical variables.

3) Phase 3: Clustering

The patient-clustering process was organized into three sequential methodological stages: dimensionality reduction, clustering analysis, and profile characterization.

In the first stage, Principal Component Analysis (PCA) was applied to reduce the high dimensionality, mainly caused by the one-hot encoding of categorical variables. This reduction minimized statistical noise and improved computational efficiency without compromising relevant information. The criterion for principal component retention was set to preserve 90% of the cumulative variance, thus ensuring preservation of the transformed dataset's information structure.

The retention of 23 principal components was determined by applying a $\geqslant 90\%$ cumulative variance threshold, identified from the explained variance curve. This criterion preserved most of the informative structure of the data (91.49%) while reducing dimensionality from 196 to 23 features, optimizing computational efficiency and lowering the risk of overfitting in the clustering analysis

Table A1 in Appendix lists PC1–PC23 along with their eigenvalues, explained variance percentages, and main variables ($|\text{loading}| \geq 0.30$). In summary, PC1 is dominated by categorical diet types (Types 1–7, all with negative loadings), PC2 by preferred foods (Types 1–5, positive loadings), and PC3 by foods not consumed (Types 1–5, negative loadings). PC4 is strongly associated with sex (male/female), hemoglobin, and hematocrit (all negative loadings), while PC5 captures anthropometric measures (weight, height) and lipid profile variables (cholesterol and triglycerides, negative loadings). Subsequent components (PC6–PC23) reflect patterns in lifestyle, comorbidities, physical activity, and metabolic indicators. The complete loading matrix is provided in Supplementary Data S1 to ensure reproducibility.

The second stage consisted of implementing and evaluating various unsupervised clustering algorithms to identify latent patterns within the dataset. Three approaches were compared: K-means, known for its computational efficiency and ability to produce well-defined partitions [28], Agglomerative Hierarchical Clustering, which identifies nested structures without assuming specific group shapes [29], and the Gaussian Mixture Model (GMM), which provides a flexible probabilistic framework capable of modeling clusters with

varying shapes and dispersions [30]. For each algorithm, cluster solutions ranging from k = 3 to k = 10 were explored based on practical criteria of clinical interpretability. Solutions with fewer than three clusters were excluded due to low discriminatory power, while those with more than ten clusters were avoided to maintain operational feasibility.

The selection of the optimal model was conducted using a multi-criteria approach employing three complementary evaluation metrics:

The Silhouette Index measures intra-cluster cohesion and inter-cluster separation [31], and can be formally expressed as shown in Eq. (1).

$$s(i) = \frac{(b(i)-a(i))}{Max\{a(i),b(i)\}} \quad (1)$$

where $a(i)$ is the average intra-cluster distance, and $b(i)$ is the average distance to the nearest neighboring cluster for each sample. Here, $s(i)$ represents the silhouette value of the i-th sample, indicating how appropriately the sample is assigned to its cluster. Values close to 1 denote well-clustered points, values near 0 indicate boundary cases, and negative values suggest possible misclassification

The Davies-Bouldin Index penalizes solutions with high internal dispersion and low intercluster separation [32], and we formally defined it in Eq. (2).

$$DB = \frac{1}{c}\sum_{i=1}^{c} Max_{i\neq j}\left\{\frac{d(X_i)+d(X_j)}{d(c_i,c_j)}\right\} \quad (2)$$

where c denotes the number of clusters, i and j are labeled clusters such that $d(X_i)$ and $d(X_j)$ represent all samples within clusters i and j with respect to their respective centroids, and $d(c_i, c_j)$ is the distance between these centroids. Here, DB is the overall Davies–Bouldin Index, which evaluates the clustering solution as a whole; lower values indicate more compact and well-separated clusters.

The Calinski-Harabasz Index, which evaluates the ratio of between-cluster variance to within-cluster variance, is useful for quantifying the compactness and separation of clusters [33], and we formally expressed it in Eq. (3).

$$CH = \frac{trace(S_B)}{trace(S_w)} \cdot \frac{n_p-1}{n_p-k} \quad (3)$$

where S_B denotes the between-cluster dispersion matrix, S_w the within-cluster dispersion matrix, n_p the number of clustered samples, and k the number of clusters. Here, CH represents the overall Calinski–Harabasz Index; higher values indicate clustering solutions with better separation and compactness.

Beyond internal indices (Silhouette, Davies–Bouldin, Calinski–Harabasz), we statistically justified the choice of k using three complementary procedures. First, we computed the Gap Statistic, comparing the within-cluster dispersion of the observed data against a reference null distribution, and identified a local maximum at k = 5 (Gap = 0.041, SE = 0.006), indicating non-random clustering structure. Second, we evaluated clustering stability via bootstrap resampling (100 replicates, 80% subsampling) and quantified agreement with the Adjusted Rand Index (ARI). The mean ARI peaked at k = 5 (0.732), exceeding that of neighboring solutions k = 4 (0.684) and k = 6 (0.701). Third, we estimated Prediction Strength under a 5-fold cross-validation scheme; only k = 5 reached a strength above the commonly accepted threshold (0.83). To address size imbalance, we imposed a minimum cluster size criterion of 1% of N to flag extremely small clusters as rare phenotypes/outliers. These were reported descriptively but did not drive global rule design in the expert system. Collectively, these analyses support k = 5 as the most stable and clinically interpretable solution. Clusters containing fewer than 1% of N (Clusters 3 and 4, with one patient each) were treated as rare phenotypes/outliers. They are reported descriptively but excluded from generating or calibrating global rules in the expert system, ensuring that imbalance does not compromise the robustness and interpretability of recommendations.

The Silhouette = 0.1229 is moderate, as expected with high-dimensional data, imbalanced clusters, and overlapping clinical phenotypes. Since this index can underestimate structure in anisotropic groups, we reinforced internal validity through complementary evidence: stability (ARI), Prediction Strength, and an external criterion (agreement with nutritionists). In the expert system, clusters act as phenotypic context, while patient-level clinical rules determine the final recommendation; small clusters are treated as rare phenotypes

Finally, the model that maximized between-cluster separation while minimizing inter-cluster overlap, as indicated by internal validation metrics, was retained for subsequent analysis.

In the third stage, a multivariate characterization of the identified clusters was conducted. Measures of central tendency and dispersion were calculated for numerical variables, while frequency distributions were generated for categorical variables. To identify the most relevant attributes for group differentiation, feature importance analysis was conducted using the Random Forest algorithm. This technique was chosen for its robustness in modeling nonlinear relationships and its capacity to handle correlations among predictors without compromising ranking stability. The analysis revealed the most influential variables in the segmentation process, enabling a more precise and clinically grounded interpretation of the emerging profiles.

Random Forest was selected for feature importance analysis due to its ability to handle non-linear relationships and high collinearity among predictors without strict parametric assumptions. As the aim was to obtain a global view of the most relevant attributes for differentiating profiles, this method provided a robust and stable approach for heterogeneous clinical, anthropometric, and one-hot encoded data. While methods such as LASSO regression or SHAP values offer interpretability advantages, they are primarily suited to supervised learning contexts or scenarios requiring formal statistical testing, which was not the main objective of this exploratory stage.

4) Phase 4: Expert system implementation

The expert system was implemented using a rule-based clinical decision-making approach structured around the nutritional profiles identified in the clustering phase. A structured knowledge base was defined, comprising 11 clinical dietary types: normocaloric, hypocaloric, hypercaloric, hyperproteic, hypoglycemic, hypolipidemic, hypoproteic, soft, gluten-free, lactose-free, and hypoallergenic. Each dietary regimen was linked to specific applicability criteria, formulated based on clinical thresholds for variables such as BMI, glucose, cholesterol, triglycerides, and glycemic status, as well as individual characteristics including food allergies, physical activity level, and pre-existing diagnoses. This encoding formalizes expert knowledge in a computational format that can be interpreted by the inference engine.

Based on this structure, the inference engine of the system is designed as a deterministic algorithm using conditional evaluation logic. This engine compares individual patient clinical data with a predefined rule set for each diet type, following a hierarchical logic that prioritizes personal attributes over the generalizations of the assigned cluster. This hybrid architecture provides accurate, personalized, and clinically coherent recommendations. The output of the system is a structured object for each patient, including an anonymous identifier, the assigned cluster, relevant clinical variables (age, sex, BMI, and biochemical parameters), and a list of suggested diets accompanied by technical descriptions and justifications based on the rules triggered during inference. In cases where no specific criteria for specialized diets are met, the system defaults on a balanced normocaloric diet, which is considered the baseline standard for nutritional maintenance. This phase established a scalable and adaptable clinical support tool aimed at improving the nutritional care efficiency through automated and reproducible processes.

In a safety-critical setting, we prioritize individual clinical attributes (explicit rules) over cluster-level context (phenotypic prior) to prevent contraindicated prescriptions in borderline or atypical cases. We formalized the hybrid decision as a weighted score per diet d for a patient with features x, as shown in Eq. (4).

$$S(d|x) = \big(1-\alpha(x)\big)R_{rules}(x,d) + \alpha(x)R_{cluster}(x,d) \quad (4)$$

where $S(d|x)$ is the hybrid score for assigning diet d to a patient with profile x; $\alpha(x)$ is an adaptive weight balancing clinical rules against cluster-based priors; and (x,d) indicates that both components are computed jointly from patient features and the candidate diet. Here, $R_{rules}(x,d)$ encodes the contribution of explicit clinical rules (e.g., thresholds for BMI, glucose, or allergies), while $R_{cluster}(x,d)$ encodes the cluster-based prior, i.e., the probability of diet d given the nutritional phenotype inferred from unsupervised clustering.

Rule layer. We model rules as (crisp/fuzzy) activations $r_j(x,d) \in [0, 1]$ with clinical weights $w_j \geq 0$ and hard contraindication masks $c_j(x,d) \in \{0, 1\}$ (0 = forbidden). We defined the cluster-based score in Eq. (5)

$$R_{cluster}(x,d) = \big(\sum_j w_j r_j(x,d) \times \prod_j c_j(x,d)\big) \quad (5)$$

where $r_j(x,d)$ is the activation of rule j for patient features x under candidate diet d; w_j is its corresponding clinical weight; and the product $w_j r_j(x,d)$ represents the weighted contribution of each rule to the final score. The term $c_j(x,d)$ is a hard safety mask: it takes the value 1 when the diet is admissible, and 0 when a contraindication is triggered, in which case the entire contribution is nullified. This ensures that contraindicated diets are strictly excluded.

Cluster prior. Let $k(x)$ be the soft assignment to clusters from k-means (distance-based) and n_{kd} the count of diet d observed/validated in cluster k. We compute a smoothed prior as shown in Eq. (6):

$$R_{cluster}(x,d) = \sum_k \pi_k(x) \frac{n_{kd}+\lambda}{n_k+\lambda|D|'} \quad (6)$$

where n_k is the total number of patients in cluster k, $|D|'$ denotes the total number of possible diets, and $\lambda > 0$ is a Laplace smoothing factor to prevent zero probabilities when a diet has not been observed in a cluster. Here, $\pi_k(x)$ indicates the soft membership of patient x in cluster k, obtained by normalizing inverse squared distances to the centroids (or alternatively using probabilistic assignments from a Gaussian Mixture Model). The fraction $(n_{kd}+\lambda)/(n_k+\lambda|D|')$ thus represents the smoothed probability of diet d within cluster k, ensuring robustness against sparsity and rare events.

Adaptive weight (x). To down-weight cluster influence under uncertainty, imbalance, or missingness, we define the adaptive weight as shown in Eq. (7).

$$\alpha(x) = \alpha_{max} \times c(x) \times m(x) \times c(1-k_x) \quad (7)$$

where $c(x) \in [0, 1]$ is cluster-confidence (e.g., rescaled point-silhouette or normalized inverse distance to the nearest centroid), $m(x) \in [0, 1]$ is data completeness (fraction of non-missing predictors used by the rules), and $c(1-\kappa(x)) \in [0, 1]$ is a conflict penalty that increases with the number/severity of triggered contraindications. We capped α_{max} at 0.20 (tuned by inner validation to maximize agreement with nutritionists), enforced $(x) = 0$ for rare clusters ($n_k < \tau$, with τ = 1% of N) or low confidence $c(x) < \delta$, and used rules-only when any hard contraindication fired for diet d.

Decision. The recommended diet(s) are arg $\max_d S(d \mid x)$. We return the top-k options with their scores and an explanation listing (i) triggered rules and thresholds, (ii) any contraindications applied, and (iii) the residual cluster contribution.

5) Phase 5: Expert system validation

The expert system validation was designed to assess the consistency of its dietary recommendations against clinical expert judgment. A stratified sampling strategy, stratified by cluster membership, ensured proportional

representation of the different profiles generated during the unsupervised clustering phase. This approach enabled the inclusion of cases exhibiting sufficient phenotypic variability, incorporating both patients with predominant cluster features and those with borderline or atypical phenotypes. The expert system was then executed on this sample, generating automated dietary recommendations for each patient based on their clinical characteristics and the structured knowledge base encoded in the inference engine.

In parallel, three certified clinical nutritionists independently evaluated the same cases using a blinded, standardized protocol. The experts were not provided with any information regarding the system's recommendations or the patients' cluster assignments. To ensure a fair comparison, only the variables utilized by the system were made available to the experts and were presented in a unified format. Recommendations from both approaches were subsequently harmonized through a standardized dietary taxonomy, enabling direct comparative analysis. To statistically assess overall agreement between the two decision sources, Pearson's Chi-square test was applied, given its suitability for evaluating independent categorical distributions. This methodological choice was justified by the nominal nature of the proposed diets, the independence of the decision sources, and the requirement to detect statistically significant discrepancy patterns. A significance level of $\alpha = 0.05$ was adopted, establishing a robust threshold for validating the system's reliability as a complementary tool in clinical decision-making.

We selected $\alpha_{max} \in \{0.0, 0.1, 0.2, 0.3\}$, confidence threshold $\delta \in [0.1, 0.3]$ and minimum cluster size $\tau \in \{1\%, 2\%\}$ via inner cross-validation to maximize concordance with expert nutritionists subject to safety overrides. The chosen configuration ($\alpha max = 0.20$, $\delta = [\ldots]$, $\tau = 1\%$) prioritized rule-based safety while preserving informative cluster context

IV. Result and Discussion

The application of Principal Component Analysis (PCA) enabled a substantial reduction in the dataset's dimensionality, decreasing the original 196 variables to 23 principal components, which together accounted for 91.49% of the total variance. This reduction mitigated redundancy and reduced statistical noise without compromising the structural integrity of the data, thereby optimizing the computational performance for the subsequent analytical stages. Examination of the factor loadings revealed that biochemical parameters—particularly hematocrit and hemoglobin—along with anthropometric variables, notably BMI, contributed significantly to the explained variance. These findings align with those of previous studies, highlighting their clinical relevance as key indicators in nutritional status assessments. This transformation also facilitates the three-dimensional visualization of the clustering patterns generated by the applied algorithms, as illustrated in the three-dimensional projection shown in Fig. 4.

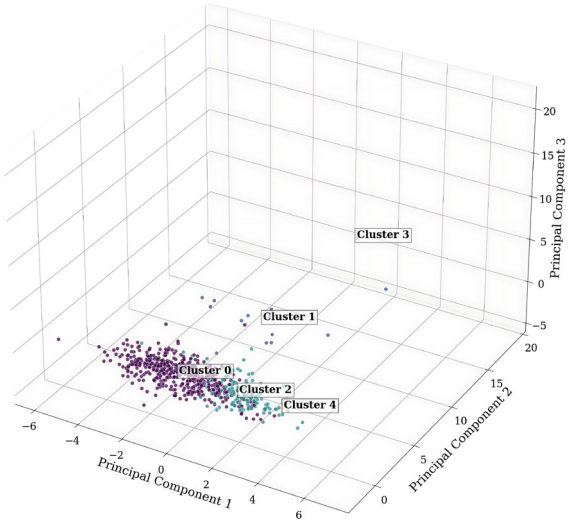

Fig. 4. 3D visualization of clusters using K-means.

A systematic comparison of three clustering algorithms—K-means, Agglomerative Hierarchical Clustering, and Gaussian Mixture Models (GMM)—was conducted by exploring solutions with k values ranging from 3 to 10. Performance was evaluated using three complementary metrics: the Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index. The K-means algorithm demonstrated the best overall performance at $k = 5$, achieving a Silhouette Score of 0.1229, a Davies–Bouldin Index of 1.3238, and a Calinski–Harabasz Index of 47.8740. Although Hierarchical Clustering and GMM obtained their best metric values at $k = 10$, their Silhouette Scores were noticeably lower (0.0833 and 0.0820, respectively), indicating weaker internal cohesion among the clusters formed.

It is worth noting that the Calinski-Harabasz Index reached its maximum value with k-means at $k = 4$ (48.3475); however, the solution with $k = 5$ provided a better overall balance when considering all three evaluation metrics simultaneously. These results support the selection of $k = 5$ as the optimal clustering solution, offering the best trade-off between separation, compactness, and clinical interpretability.

Consistent with the internal indices, the Gap Statistic favored $k = 5$ over adjacent values. The bootstrap stability analysis yielded the highest mean ARI at $k = 5$ (0.732), and the Prediction Strength was also maximal at $k = 5$ (0.83). Importantly, using an external criterion relevant to clinical utility—the agreement between the expert system and clinical nutritionists—the concordance observed with $k = 5$ (77% global; Clusters 1 and 0: 75% and 83%) exceeded that obtained with $k = 4$ (74%) and $k = 6$ (75%). Thus, both stability and task-oriented external performance support retaining $k = 5$.

The moderate Silhouette = 0.1229 reflects the expected overlap among real-world clinical phenotypes in our dataset. While it indicates less distinct boundaries between clusters, complementary evidence from stability analysis, Prediction Strength, and external agreement with nutritionists confirmed the robustness of the segmentation.

In practice, the hybrid expert system uses these clusters only as contextual information, relying on patient-level clinical rules for recommendations. This approach maintained clinical interpretability and supported the high agreement observed in homogeneous clusters (83% in Cluster 1 and 75% in Cluster 0).

Given the highly skewed cluster sizes, we applied a minimum size threshold of [1% of N] (rounded up, ≥7 patients) to flag Clusters 3 and 4 as rare phenotypes/outliers. These clusters were reported descriptively and excluded from global rule calibration, while the primary clinical interpretation and system rules were anchored on the dominant, well-defined profiles (Clusters 1 and 2).

The evolution of the Silhouette Score as a function of the number of evaluated clusters is presented in Fig. 5.

	3	4	5	6	7	8	9	10
GMM	0.1205	0.0781	0.1183	0.0807	0.5532	0.0806	0.023	0.082
Hierarcical Clusterin	0.0872	0.0767	0.0662	0.0701	0.0775	0.0806	0.0804	0.0833
K-Means	0.1167	0.0978	0.1229	0.0943	0.0891	0.1114	0.0853	0.0911

Fig. 5. Evolution of the silhouette metric as a function of the number of clusters.

Based on the comparative evaluations described above, the K-means solution with k = 5 was selected as the optimal configuration, prioritizing the Silhouette Score as the primary selection criterion while also considering alignment with the Davies–Bouldin and Calinski–Harabasz indices. The resulting segmentation exhibited a markedly asymmetric distribution of patients across clusters: Cluster 0, 13 patients (2.1%); Cluster 1, 349 patients (56.7%); Cluster 2, 251 patients (40.8%); and Clusters 3 and 4, only one patient each (0.2%). This pattern indicates the presence of clinical outliers—effectively isolated by the algorithm—while the majority of cases were concentrated within two dominant profiles represented by Clusters 1 and 2.

To clinically interpret the groupings, a feature importance analysis was performed using the Random Forest algorithm, enabling the identification of the most influential predictors for distinguishing between clusters. The five top-ranked attributes were hematocrit (0.1114), hemoglobin (0.0936), ability to perform high-intensity exercise (0.0825), height (0.0623), and age (0.0589). This ranking underscores the pivotal role of hematological and functional parameters in defining the nutritional profiles generated by the model, in contrast to traditional approaches that often emphasize biochemical markers such as glucose or lipids. The distribution of metabolic variables by cluster is shown in Fig. 6, providing a clear visualization of the structural differences among the identified groups.

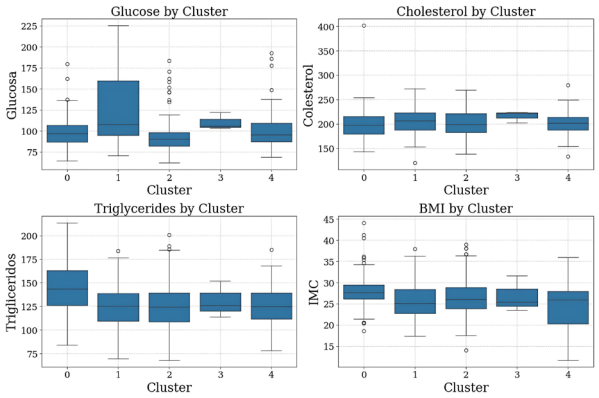

Fig. 6. Distribution of metabolic variables by cluster.

Clinical characterization of the profiles grouped by K-means revealed differentiated patterns with relevance for the design of personalized dietary strategies. Each cluster presented distinct demographic, anthropometric, and metabolic configurations, supporting its clinical interpretation.

Cluster 0 (n = 13) was distinguished by a predominantly female composition (69.2%) and a mean age of 47.0 years (±14.2). This group exhibited elevated BMI values (27.2), glucose levels in the prediabetic range (109.7 mg/dL), a high prevalence of hypertension (84.6%), and a low frequency of physical activity (2.2 days per week). This profile corresponds to patients with evident metabolic alterations requiring nutritional interventions focused on weight reduction, glycemic control, and cardiovascular risk management.

Cluster 1, which included the largest proportion of patients (n = 349; 56.7%), was also predominantly female (76.5%) and older (mean age 59.2 ± 25.1 years). It was characterized by overweight status (average BMI 26.5), slightly elevated glucose (107.5 mg/dL), high total cholesterol levels (205.7 mg/dL), and a high prevalence of hypertension (90.0%). This group represents a typical profile of patients with multiple cardiometabolic risk factors requiring a comprehensive dietary approach focused on lipid, glucose, and blood pressure control.

In contrast, Cluster 2 (n = 251; 40.8%) consisted mostly of males (74.5%) and had a younger average age (45.2 ± 12.8 years). Although overweight was also observed (average BMI 28.3), this group displayed more favorable metabolic parameters, including normal glucose (96.9 mg/dL) and cholesterol levels (196.1 mg/dL) levels, albeit with a high prevalence of hypertension (90.8%). This profile suggests an early stage cardiometabolic risk group in which preventive dietary interventions could be particularly effective.

Clusters 3 and 4 (n = 1 each) represented atypical cases with unique profiles. Both exhibited above-average physical activity levels (4.0 days per week) and distinctive metabolic parameters that justified their separate classification by the algorithm, despite the minimal sample size. These cases may represent clinical outliers or less common metabolic subtypes that are useful for exploratory studies or individualized analyses.

Detailed values for glucose, cholesterol, triglyceride, and BMI for each cluster are presented in Table II, allowing for a quantitative comparison of the metabolic characteristics across the groups.

TABLE II. SUMMARY OF METABOLIC VARIABLES BY CLUSTER

Indicator	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Glucose-Mean	109.75	107.49	96.94	120.3	120.8
Glucose-Standard Deviation	22.17	33.07	17.65		
Cholesterol-Mean	195.7	205.69	196.07	185.3	196.3
Cholesterol-Standard Deviation	23.97	28.29	25.63		
Triglycerides-Mean	133.95	121.75	140.79	148.4	132.2
Triglycerides-Standard Deviation	21.37	23.05	25.23		
BMI-Mean	27.2	26.52	28.27	27.68	27.1

A. Results of the Expert System Implementation

The dietary recommendation expert system was implemented using a hybrid architecture that combined a rule-based clinical inference module with a contextualization layer informed by the clustering-derived patterns. The knowledge base was organized as a hierarchical matrix of conditional rules covering 11 clinical diet types. Each dietary category was linked to a defined set of applicability conditions expressed as numerical ranges or categorical value lists. This structure supports the simultaneous evaluation of multiple constraints related to metabolic, anthropometric, and pathological parameters. The inference engine was built following a forward-chaining approach, processing a 196-dimensional feature vector for each patient. During this process, the activated rules generate a vector of weighted matches, which form the basis for the final dietary recommendation.

As part of the decision-making mechanism, a dual-weighting strategy was adopted: individual patient data were given the highest priority, while average cluster characteristics were used only as a complementary source in cases of incomplete data or outlier profiles. This integration enhanced the system's robustness in the face of clinical data heterogeneity. The profile-based contextualization module generates synthetic feature vectors that merge individual-level information with the representative centroids of each cluster, thereby improving the system's adaptability to diverse population profiles. Fig. 7 depicts the complete architecture of the expert system, illustrating the data flow from input components through inferential processing and contextualization to the final recommendation output.

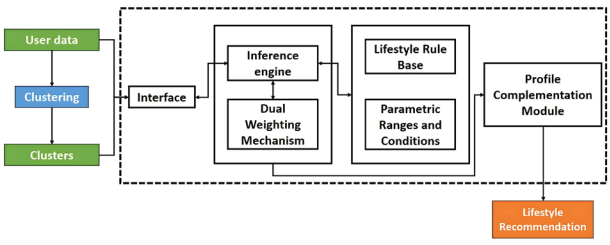

Fig. 7. Architecture of the expert dietary recommendation system.

B. Expert System Validation

The validation of the expert system yielded an overall agreement rate of 77% between the automated dietary recommendations and those provided by clinical specialists, reflecting a high level of concordance between the two decision-making approaches. Agreement was particularly strong in Cluster 1 (83%) and Cluster 0 (75%), which comprised patients with more standardized clinical and metabolic profiles. In contrast, Clusters 2 and 4 showed the highest discrepancy rates, primarily involving borderline cases straddling two dietary categories or patients presenting atypical clinical characteristics. Table III summarizes the distribution of dietary recommendations issued by both the expert system and clinical specialists for each evaluated diet type.

TABLE III. CONCORDANCE OF DIETARY RECOMMENDATIONS

Diet Type	Expert System (*n*)	Nutritionist (*n*)
Balanced Normocaloric Diet	12	15
Low-Calorie Diet	22	19
High-Calorie Diet	2	3
High-Protein Diet	5	6
Low-Protein Diet	1	2
Low-Carbohydrate Diet	8	7
Low-Liquid / Heart-Healthy Diet	7	6
Soft or Gastric Protection Diet	1	1
Gluten-Free Diet	0	0
Lactose-Free Diet	1	0
Hypoallergenic Diet	1	1
Total	60	60

For statistical comparison, Pearson's Chi-square test (χ^2) was applied using a significance level of $\alpha = 0.05$. Table IV presents the detailed confusion matrix by diet type, distinguishing between matching and non-matching cases. The total number of comparisons exceeds 60 patients because some individuals received more than one dietary recommendation from both sources.

TABLE IV. CONFUSION MATRIX OF DIETARY RECOMMENDATIONS: EXPERT SYSTEM VS. EXPERT NUTRITIONIST

Diet Type	Match	No Match	Total
Balanced Normocaloric Diet	10	7	17
Low-Calorie Diet	16	9	25
High-Calorie Diet	2	1	3
High-Protein Diet	4	3	7
Low-Protein Diet	1	1	2
Low-Carbohydrate Diet	6	3	9
Low-Liquid / Heart-Healthy Diet	5	3	8
Soft or Gastric Protection Diet	1	0	1
Gluten-Free Diet	0	0	0
Lactose-Free Diet	0	1	1
Hypoallergenic Diet	1	0	1
Total	46	28	74*

*The total exceeds 60 patients because some individuals received more than one dietary recommendation from both the system and the nutritionist.

Table V presents a comparative analysis by cluster profile, showing the concordance rate within each group and the main discrepancies observed. Clusters with greater clinical homogeneity (Clusters 0 and 1) exhibited the highest levels of agreement, while clusters with smaller sample sizes or atypical clinical profiles (Clusters 2 and 4) showed more frequent discrepancies.

The results of the chi-square test are presented in Table VI, showing no statistically significant differences ($p > 0.05$) between the distribution of recommendations, supporting the hypothesis that the system's decisions are clinically consistent with human judgment. However, qualitative analysis identified mild to moderate discrepancies in 37% of the non-matching cases, with the most frequent being substitutions between hypocaloric and hypoglycemic diets, as well as between normocaloric and heart-healthy diets.

TABLE V. COMPARATIVE ANALYSIS BY CLUSTER PROFILE

Cluster	No Total	No Sample	Agreement (%)	Main Discrepancy
0	13	8	75%	Low-Calorie vs. Low-Carbohydrate Diet
1	349	35	83%	Low-Choice vs. Normo-Calorie Diet
2	251	15	73%	Low-Calorie vs. Normo-Calorie Diet
3	1	1	100%	No Discrepancies
4	1	1	0%	Low-Carbohydrate vs. Hypoallergenic Diet
Total	615	60	77%	

TABLE VI. ARE TEST RESULTS

Statistic	Values
Chi-Square	1.654
Degrees of Freedom	8
p-value	0.990
Significant ($\alpha = 0.05$)	No

The implementation of the developed expert system achieved an overall agreement rate of 77% with the clinical judgment of nutritionists, aligning with previous findings such as those reported by Stefanidis *et al.* [34], whose PROTEIN AI system reached 92% accuracy in generating expert-validated dietary plans. Although our model did not surpass this benchmark, the high concordance observed in clinically homogeneous clusters (83% in Cluster 1 and 75% in Cluster 0) underscores its capacity to emulate expert reasoning in standardized contexts. The main discrepancies were concentrated in atypical profiles, a phenomenon previously noted by González-Ramírez *et al.* [26], who emphasized the need for adaptive feedback mechanisms to improve personalization in dietary recommendations.

From the perspective of clustering model performance, the K-means algorithm outperformed both Hierarchical Clustering and the Gaussian Mixture Model, achieving a Silhouette Index of 0.1229, a Davies–Bouldin Index of 1.3238, and a Calinski–Harabasz Index of 47.874. These results surpass those reported by Puraram *et al.* [35], who obtained a 19% improvement in the Davies–Bouldin Index using a hybrid approach that combined PSO with K-means. Although the Silhouette Index in both studies remained moderate, our findings confirm the value of multicluster analysis for segmenting nutritional profiles and provide a robust basis for subsequent automated decision-making processes.

Variable importance analysis using Random Forest identified hematocrit (0.1114) and hemoglobin (0.0936) as the most influential predictors, which is consistent with the findings of Iwendi *et al.* [22], who emphasized the importance of caloric content and clinical parameters in personalized diet prediction. This concurrence validates the clinical-metabolic approach adopted by our system, particularly in incorporating hematological and functional variables, such as the ability to perform high-intensity exercise (0.0825). Compared to the work of Silva *et al.* [19], whose collaborative filtering system achieved a precision of 91% based on dietary consumption patterns, our inclusion of objective clinical variables further enriched the interpretive context beyond self-reported eating habits alone.

Finally, the evaluation of the expert system's hybrid architecture revealed a clear operational advantage by integrating codified clinical rules with clustering-based contextualization. This approach allowed the system to effectively manage data heterogeneity, particularly in cases with missing or atypical values, thereby delivering robust, personalized recommendations. This outcome is consistent with the findings of Franco *et al.* [18], who reported that 64% of eNutri platform users adhered to the advice received, with motivation to improve dietary habits being a key determinant. Although adherence was not directly measured in our study, the implemented inferential framework exhibited strong statistical consistency ($\chi^2 = 1.654$, $p = 0.990$), reinforcing its potential as a complementary tool in nutrition services operating with limited personnel and high patient demand, such as those in the San Martín region.

V. CONCLUSION

The objective of this study was to develop and evaluate an intelligent expert system capable of generating personalized dietary recommendations for patients based on clinical, anthropometric, and biochemical data, with application in a regional context characterized by high healthcare demand. The results demonstrated that combining unsupervised clustering with a rule-based inference engine enabled identification of clinically coherent nutritional profiles, achieving an overall agreement rate of 77% with expert dietitians. Notably, the highest concordance was observed in the standard clinical groups (Clusters 0 and 1), suggesting that the system operates robustly in scenarios with well-defined metabolic patterns. The integration of variables such as hematocrit, hemoglobin, and functional capacity into inferential logic enhanced the system's clinical accuracy, in line with previous evidence on the importance of comprehensive metabolic profiling in nutritional care.

Future research should validate the system using a larger and more diverse sample, incorporating longitudinal follow-up to assess adherence, clinical outcomes, and

behavioral changes induced by the recommendations. Additionally, the expert system architecture can be expanded using reinforcement learning components or Explainable AI (XAI) modules to improve its transparency and adaptability. From a policy perspective, the implementation of such a system in public hospitals could significantly reduce waiting times and enhance the efficiency of nutritional services, especially in regions with limited access to specialized care. The proposed model represents a scalable and reproducible digital health intervention with the potential to strengthen clinical decision-making and promote equitable access to personalized nutrition in underserved populations.

APPENDIX A: VARIANCE CONTRIBUTIONS AND VARIABLE LOADINGS FOR PCS

TABLE AI. VARIANCE CONTRIBUTIONS AND LEADING VARIABLE LOADINGS FOR PRINCIPAL COMPONENTS (PC1–PC23)

Principal Component	Eigen value	Explained Variance (%)	Cumulative Variance (%)	Main Variables (|loading| ≥ 0.30)
PC1	6.5046	15.1	15.1	Diet Type (1) (−0.38), Diet Type (2) (−0.37), Diet Type (3) (−0.36), Diet Type (4) (−0.37), Diet Type (5) (−0.36), Diet Type (6) (−0.37), Diet Type (7) (−0.37)
PC2	4.8606	11.29	26.39	Preferred Foods (1) (+0.35), Preferred Foods (2) (+0.34), Preferred Foods (3) (+0.34), Preferred Foods (4) (+0.34), Preferred Foods (5) (+0.34)
PC3	4.5628	10.59	36.98	Foods Not Consumed (1) (−0.36), Foods Not Consumed (2) (−0.34), Foods Not Consumed (3) (−0.35), Foods Not Consumed (4) (−0.34), Foods Not Consumed (5) (−0.35)
PC4	2.3428	5.44	42.42	Sex (Male/Female) (−0.41), Hemoglobin (−0.60), Hematocrit (−0.60)
PC5	2.0089	4.66	47.09	Weight (kg) (−0.41), Height (cm) (+0.39), Cholesterol (−0.54), Triglycerides (−0.52)
PC6	1.3198	3.06	50.15	Smoker Status (−0.31), Do you have difficulty moving or need assistance? (+0.53), Are you capable of high-intensity exercises? (−0.38)
PC7	1.2576	2.92	53.07	Smoker Status (−0.35), Do you follow a specific diet? (+0.31), Glycemic Status (−0.47), Glucose (+0.47)
PC8	1.2199	2.83	55.9	Weight (kg) (−0.38), Height (cm) (−0.40), Education Level (+0.33), Sleep Duration (−0.44)
PC9	1.2173	2.83	58.73	Smoker Status (+0.35), Sleep Duration (+0.36), Do you suffer from any food allergy? (−0.49)
PC10	1.1624	2.7	61.43	Age (−0.34), Alcohol Consumption (+0.35), Hypertension Medications Consumed (−0.51)
PC11	1.1479	2.67	64.09	Weekly Activity Frequency (−0.50), Are you a hypertensive patient? (+0.37), Hypertension Medications Consumed (−0.33)
PC12	1.1222	2.61	66.7	Screen Time (TV/Mobile) (−0.49), Type of Physical Activity (−0.46), Are you a hypertensive patient? (−0.31)
PC13	1.0516	2.44	69.14	How many times a day do you consume food? (−0.31), Do you follow a specific diet? (+0.57), Weekly Activity Frequency (+0.32)
PC14	1.0193	2.37	71.51	Age (+0.41), Alcohol Consumption (−0.32), Employment Status (+0.54), Are you capable of high-intensity exercises? (−0.32)
PC15	0.9992	2.32	73.83	Education Level (+0.31), How many times a day do you consume food? (+0.50), Type of Physical Activity (+0.39), Are you currently taking medication for hypertension? (−0.31)
PC16	0.9754	2.26	76.09	Age (+0.33), Alcohol Consumption (+0.31), Screen Time (TV/Mobile) (+0.43), How many times a day do you consume food? (+0.32), Are you currently taking medication for hypertension? (+0.34)
PC17	0.9568	2.22	78.31	Are you a hypertensive patient? (−0.45), Are you currently taking medication for hypertension? (+0.60)
PC18	0.9355	2.17	80.48	Education Level (+0.44), Alcohol Consumption (−0.45)
PC19	0.9219	2.14	82.62	Education Level (−0.39), Smoker Status (+0.35), Are you capable of high-intensity exercises? (−0.43)
PC20	0.8581	1.99	84.62	Do you suffer from any food allergy? (+0.49), Are you a hypertensive patient? (−0.34), Hypertension Medications Consumed (−0.32), Glucose (+0.35)
PC21	0.8493	1.97	86.59	Employment Status (−0.36), Weekly Activity Frequency (+0.46)
PC22	0.8098	1.88	88.47	Smoker Status (−0.31), Screen Time (TV/Mobile) (−0.33), Do you follow a specific diet? (−0.50), Glucose (+0.34)
PC23	0.7913	1.84	90.31	Age (−0.44), Employment Status (+0.33), How many times a day do you consume food? (+0.36)

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This study did not involve any experimental intervention or clinical treatment. Only previously collected clinical and nutritional data were used, with prior written informed consent from all participants authorizing their anonymized use for research purposes. All procedures complied with the principles of the Declaration of Helsinki and applicable local data protection regulations. Data were de-identified before analysis and handled under secure access controls to ensure confidentiality.

DATA PROTECTION AND PRIVACY

Personal identifiers were not collected; records were irreversibly de-identified and stored on secured servers with role-based access. Data handling complied with [applicable law/policy, e.g., Peru's Law 29733 on personal data protection] and institutional data governance policies.

CLINICAL USE DISCLAIMER

The developed system is a clinical decision-support prototype intended to assist qualified nutrition professionals. It does not provide medical diagnosis or

prescriptions, and it must not replace clinical judgment. Final treatment decisions remain the responsibility of licensed clinicians. The system is not intended for emergency use or unsupervised patient self-management, and it has not been cleared or approved by a medical device regulatory authority. Recommendations should always be interpreted within the patient's comprehensive clinical context.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Conceptualization, M.A.V.C., J.R.N.C., C.G.E., J.V.I., S.S. and L.K.Q.M.; Methodology, M.A.V.C., J.R.N.C., and C.G.E.; Software, M.A.V.C., J.R.N.C., C.G.E. and J.V.I.; Validation, M.A.V.C., J.R.N.C., S.S. and L.K.Q.M.; Formal Analysis, M.A.V.C., J.R.N.C., C.G.E., and L.K.Q.M.; Investigation, M.A.V.C., J.R.N.C., C.G.E., J.V.I., S.S. and L.K.Q.M.; Resources, M.A.V.C., J.R.N.C., C.G.E., J.V.I., S.S. and L.K.Q.M.; Data Curation, M.A.V.C., and J.R.N.C.; Writing-Original Draft Preparation, M.A.V.C., J.R.N.C., C.G.E., J.V.I., S.S. and L.K.Q.M.; Writing-Review & Editing, M.A.V.C., and J.R.N.C.; Visualization, M.A.V.C., J.R.N.C., J.V.I., and L.K.Q.M.; Supervision, M.A.V.C., J.R.N.C., C.G.E., J.V.I., S.S. and L.K.Q.M.; Project Administration, M.A.V.C., J.R.N.C., C.G.E., J.V.I., S.S. and L.K.Q.M.; Funding Acquisition, M.A.V.C., J.R.N.C., C.G.E., J.V.I., S.S. and L.K.QM.. All authors had approved the final version.

FUNDING

This research was funded by the Universidad Nacional de San Martín (UNSM) under the project titled "Improving Efficiency and Personalization in Nutritional Consultations through an Expert System Integrating Nutritional Evaluation Components and Advanced Deep Learning Models", supported by University Council Resolution No. 541-2024-UNSM/CU-R.

REFERENCES

[1] M. V. Uribe *et al.*, "Realizing the right to health in Latin America, equitably," *Int. J. Equity Health*, vol. 20, 34, 2021.
[2] K. A. Dyer, "Daily healthy habits to reduce stress and increase longevity," *J. Interprofessional Educ. Pract.*, vol. 30, 100593, 2023.
[3] L. I. Reyes *et al.*, "Actions in global nutrition initiatives to promote sustainable healthy diets," *Glob. Food Sec.*, vol. 31, 100585, 2021.
[4] X. Q. Tan, "The role of healthy lifestyles in preventing chronic disease among adults," *Am. J. Med. Sci.*, vol. 364, no. 3, pp. 309–315, 2022.
[5] P. Machado *et al.*, "Measuring adherence to sustainable healthy diets: A scoping review of dietary metrics," *Adv. Nutr.*, vol. 14, no. 1, pp. 147–160, 2022.
[6] T. Wilson and A. Bendich, "Nutrition guidelines for improved clinical care," *Med. Clin. North Am.*, vol. 106, no. 5, pp. 819–836, 2022.
[7] C. Werner and E. Osterbur, "Decoding plant-based and other popular diets: Ensuring patients are meeting their nutrient needs," *Physician Assist. Clin.*, vol. 7, no. 4, pp. 615–628, 2022.
[8] G. Salloum and J. Tekli, "Automated and personalized nutrition health assessment, recommendation, and progress evaluation using fuzzy reasoning," *Int. J. Hum. Comput. Stud.*, vol. 151, 102610, 2021.
[9] J. Lopez-Barreiro *et al.*, "Artificial intelligence-powered recommender systems for promoting healthy habits and active aging: A systematic review," *Appl. Sci.*, vol. 14, no. 22, 10220, 2024.
[10] Ministry of Health. (August 2023). Analysis of the health situation of Peru 2021. [Online]. Available: https://www.gob.pe/institucion/ensap/informes-publicaciones/4509305-analisis-de-situacion-de-salud-asis-2021 (in Spanish)
[11] Ministry of Health. (January 2021). Diagnosis of infrastructure and equipment gaps in the health sector. [Online]. Available: https://www.gob.pe/institucion/minsa/informes-publicaciones/1848369-diagnostico-de-brechas-de- (in Spanish)
[12] National Institute of Statistics and Informatics. (October 2018). Final results for the department of San Martín. [Online]. Available: https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1573/ (in Spanish)
[13] D. Tsolakidis, L. P. Gymnopoulos, and K. Dimitropoulos, "Artificial intelligence and machine learning technologies for personalized nutrition: A review," *Informatics*, vol. 11, no. 3, 62, 2024.
[14] C. Nutakki *et al.*, "Predicting markers of cognitive decline within small population samples of daily life activities," *Int. J. Online Biomed. Eng.*, vol. 21, no. 6, pp. 111–123, 2025.
[15] C. Song *et al.*, "A hybrid recommendation approach for viral food based on online reviews," *Foods* , vol. 10, no. 8, 1801, 2021.
[16] E. C. O. Tapia *et al.*, "Framing thematic trends around computational methods research in health sciences," *Rev. Cient. Sist. Inform.*, vol. 5, no. 1, e913, 2025.
[17] J. N. Bondevik *et al.*, "A systematic review on food recommender systems," *Expert Syst. Appl.*, vol. 238, 122166, 2024.
[18] R. Z. Franco *et al.*, "Effectiveness of web-based personalized nutrition advice for adults using the enutri web app: Evidence from the eatwelluk randomized controlled trial," *J. Med. Internet Res.*, vol. 24, no. 4, e29088, 2022.
[19] V. C. Silva *et al.*, "Recommender system based on collaborative filtering for personalized dietary advice: A cross-sectional analysis of the ELSA-brasil study," *Int. J. Environ. Res. Public Health*, vol. 19, no. 22, 14934, 2022.
[20] L. M. Matos *et al.*, "Categorical attribute transformation environment (CANE): A python module for categorical to numeric data preprocessing," *Softw. Impacts*, vol. 13, 100359, 2022.
[21] Y. F. Chen *et al.*, "Design of clinical support systems using integrated genetic algorithm and support vector machine," in *Proc. International Conf. on Computer Analysis of Images and Patterns*, 2009, pp. 791–798.
[22] C. Iwendi *et al.*, "Realizing an efficient IoMT-assisted patient diet recommendation system through machine learning model," *IEEE Access*, vol. 8, pp. 28462–28474, 2020.
[23] L. Yin *et al.*, "A fusion decision system to identify and grade malnutrition in cancer patients: Machine learning reveals feasible workflow from representative real-world data," *Clin. Nutr.*, vol. 40, no. 8, pp. 4958–4970, 2021.
[24] K. H. Hart *et al.*, "The suitability of dietary recommendations suggested by artificial intelligence technology via a novel personalised nutrition mobile application," in *Proc. Nutr. Soc.*, 2022, E37.
[25] F. Bolikulov *et al.*, "Effective methods of categorical data encoding for artificial intelligence algorithms," *Mathematics*, vol. 12, no. 16, 2553, 2024.
[26] M. González-Ramírez *et al.*, "SAlBi educa (Tailored nutrition app for improving dietary habits): Initial evaluation of usability," *Front. Nutr.*, vol. 9, 782430, 2022.
[27] L. B. V. d. Amorim, G. D. C. Cavalcanti, and R. M. O. Cruz, "The choice of scaling technique matters for classification performance," *Appl. Soft Comput.*, vol. 133, 109924, 2023.
[28] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, 1982.
[29] E. K. Tokuda, C. H. Comin, and L. F. Costa, "Revisiting agglomerative clustering," *Phys. A: Stat. Mech. its Appl.*, vol. 585, 126433, 2022.
[30] Y. Fu *et al.*, "Gaussian mixture model with feature selection: An embedded approach," *Comput. Ind. Eng.*, vol. 152, 107000, 2021.
[31] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.

[32] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227,1979.
[33] T. Caliński and J. Harabasz, "A dendrite method foe cluster analysis," *Commun. Stat.*, vol. 3, no. 1, pp. 1–27, 1974.
[34] K. Stefanidis *et al.*, "PROTEIN AI advisor: A knowledge-based recommendation framework using expert-validated meals for healthy diets," *Nutr.*, vol. 14, no. 20, 4435, 2022.
[35] T. Puraram *et al.*, "Thai food recommendation system using hybrid of particle swarm optimization and K-means algorithm," in *Proc. 2021 6th International Conf. on Machine Learning Technologies*, 2021, pp. 90–95.