

Article

Personalized Canine Diet Generation Using Machine Learning and Constraint Optimization

Aliya Kalykulova ^{1,*} , Kuanysh Bakirov ^{1,*} , Aruzhan Shoman ^{1,*}, Kadyrzhan Makangali ² 
and Gulzhan Tokysheva ²

¹ Research and Innovation Center “AgroTech”, Astana IT University, 010000 Astana, Kazakhstan

² Department of Technology of Food and Processing Industries, S. Seifullin Kazakh Agrotechnical Research University, 010000 Astana, Kazakhstan

* Correspondence: aliyakalykulova01@gmail.com (A.K.); bakir.kuanysh@gmail.com (K.B.); a.shoman@astanait.edu.kz (A.S.)

Abstract

The growing demand for customized pet diets highlights the shortcomings of commercial dog foods designed for all breeds, especially when it comes to addressing breed-specific diseases, metabolic disorders, and health risks. This research presents the development and evaluation of a hybrid system for formulating wet canine food recipes. The system combines data on ingredients, veterinary feeds, and breed-related diseases; the architecture includes a recommendation module for ingredient selection and a linear programming block for recipe optimization, considering veterinary nutrient restrictions. The evaluation of the system included automatic classification of foods by specialization, visual analysis of recipe clustering, and comparison of formulas obtained by different models. The average precision of label recovery was 85.4% for TF-IDF and 88.2% for the E5 model. A comparison of ingredient extraction methods showed that machine learning produces more stable recipes, while the statistical approach provides greater variability. The developed system demonstrates potential for automating recipe creation, filling in missing data, and developing veterinary decision support platforms aimed at personalized diet selection based on the physiological needs of animals.

Keywords: canine nutrition; recommender systems; personalized pet food; machine learning; veterinary informatics



Academic Editors: Daniel Riera Terrén, Angel A. Juan, Majsa Ammuriova and Laura Calvet

Received: 15 December 2025

Revised: 5 February 2026

Accepted: 21 February 2026

Published: 25 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

In the last few decades, the number of dogs around the world has steadily increased. In Europe alone, there are more than 90 million companion dogs and a pet food market worth EUR 29.1 billion [1]. This trend has made people pay more attention to the quality of pet care, with nutrition being a big part of that. The food a dog eats has a direct effect on its gastrointestinal health, skin condition, body weight, and behavior [2]. But the nutritional needs of dogs can be very different depending on their size, age, activity level, and whether they have any diseases. For instance, owners of small breeds may choose wet food because their dogs are more likely to get dental and urinary tract problems. Large-breed dogs require higher energy and protein intake to maintain muscle mass and meet metabolic demands. As a result, dry food is often preferred for large breeds because of its higher energy density and cost efficiency when fed in large amounts [3].

Even though commercial pet food has improved, many owners still feed their pets home-cooked or unusual diets. One survey found that 38% of respondents used these

diets [4]. These kinds of methods are often adopted without supervision, which can cause nutrient imbalances and health problems [5,6]. People are still worried about how accurate the energy calculations are in both homemade and store-bought diets. Studies have found that 22% of the metabolizable energy equations used by manufacturers are not clear, and measuring cups can make dosing errors of 80–152% [5]. Also, differences between the stated and measured levels of nutrients, like EPA/DHA levels being 25–50% lower than what was said, make people less likely to trust the product [6]. At the same time, sustainability has become a big problem in the making of pet food. Climate change is making it more important to lower the carbon footprint of diets, which means that formulations need to be more efficient and environmentally friendly [7]. The mentioned factors above make it clear that there is a need for scientifically sound, data-driven methods to help design diets that are safe, accurate, and long-lasting [8,9].

This paper proposes a model that combines text data processing, statistical analysis, and optimization algorithms for the automated calculation of specialized feed recipes. The developed system is based on several interrelated datasets, including ingredient characteristics, information about veterinary feed, and common dog diseases. Part of the data was collected using web scraping from a pet food e-commerce platform [10] as well as from websites providing information on dog breeds and diseases [11–13]. The dataset was further enriched through manual annotation and subsequently used to evaluate automatic annotation approaches based on TF-IDF and the E5 transformer model. These methods enable the recovery of missing information and support the consistent standardization of feed descriptions as the dataset is expanded.

The proposed model includes three key components. The first component involves data consolidation and cleaning, ensuring the connection between ingredients, diseases, and dietary recommendations. The second component is aimed at extracting relevant ingredients characteristic of various functional areas of feed. The third component represents recipe optimization using linear programming, considering nutrient standards and individual dog parameters. Based on the proposed model, a platform has been developed that allows us to calculate an individual diet considering the animal's characteristics.

The results show that automating the analysis of veterinary diets is possible and that text descriptions of ingredients contain sufficient information to determine the specialization of the feed. A comparison of ingredient extraction methods revealed differences: machine learning produces more stable recipes, while the statistical approach leads to greater variability. Analysis in a combined feature space and visualization using t-SNE confirmed that feeds of the same functional category form distinct, though not completely isolated, clusters, reflecting the diversity of the veterinary diet market.

Thus, the developed system is a comprehensive tool for the automated formation and analysis of specialized diets, combining machine learning, text analytics, and optimization methods. The model can serve as a basis for the further development of veterinary decision support systems, simplifying the individual selection of feeds, and a more in-depth study of structural patterns in functional diet recipes.

2. Related Work

Rational canine nutrition necessitates the evaluation of various factors that influence both the dietary nutritional value and its safety and digestibility. When making a diet, you need to think about the animal's nutrient needs, the estimated metabolizable energy, and the properties of the ingredients you have on hand, such as their biological value, digestibility, and possible risks. These parameters change based on things like age, how active the dog is, their health, any chronic diseases they have, and the physiology of their breed. Also, how tasty the food is and how easy it is to find in the local market are important

factors in owner compliance and practical adoption. Tables 1–4 present the methods used to calculate or recommend recipes.

One of the main techniques in making a dog's diet can be optimization-based methods, which use math models to find the right balance of nutrients in rations while keeping in mind cost, nutrition, or availability. Linear programming and its derivatives are commonly utilized in animal nutrition to reduce feed expenses while fulfilling energy and nutrient demands. In practice, though these methods are powerful, they often have problems with computational complexity and scalability.

The second option is to build a recommendation system that helps pet owners or veterinarians choose the right commercial diets. Traditional methods use content-based filtering or measures of similarity between nutrient profiles. More advanced models, on the other hand, use user feedback, text processing, and graph-based representations. For instance, nutrient-based recommendation algorithms (NRAs) analyze diets based on their nutrient content and compare them to dog profiles, which is better than k-nearest neighbors (KNN) and matrix factorization algorithms. Hybrid models that integrate clustering, embeddings, or knowledge graphs have been proposed in human nutrition and are progressively being adapted for veterinary applications.

Even with these improvements, there are still some problems. Optimization models often do not consider how tasty something is or how easy it is to find the right ingredients. Recommender systems, on the other hand, have problems with not having enough data and the cold-start problem. These gaps have led to the creation of hybrid frameworks that combine machine learning with optimization under veterinary nutritional constraints. This allows for both personalization and scientific validity in designing a dog's diet.

2.1. Feed Optimization Methods

Optimization techniques utilized in diet formulation rely on the mathematical modeling of ingredient ratios concerning their nutrient profiles and established limitations. Linear programming (LP) is the most common method. In LP, linear constraints show how the diet is made up, how much it costs, and how many ingredients are involved. LP has been extensively utilized in feed formulation across various species. For instance, Aguilar and Montoya [14] and Joban et al. [15] used LP to make cost-effective diets for dogs. Pavlova et al. [16] used the method to improve sheep nutrition and Metwally [17] showed how well it worked in poultry production, cutting costs at all stages of growth. Different versions of the basic LP model have also been investigated. Alqaisi and Schlecht [18] developed a multi-period linear programming model that integrated temporal price variations for optimizing cattle rations, facilitating adaptation to fluctuating economic conditions and ensuring long-term sustainability. LP is very popular, but it has certain disadvantages. One major issue is corner solutions, where the output is mostly made up of extreme values. In addition, LP does not give much diagnostic information when there is not a solution that works [19].

Nonlinear optimization methods let us build complex interactions between nutrients because they are more expressive. Li et al. [19] looked at how well deterministic global optimization (MINLP_DGO), iterative linear programming (IteLP), and sequential quadratic programming (SQP) worked for making diets for dairy cows. SQP was more accurate than IteLP, but it did not always give stable solutions. MINLP_DGO gave the most reliable and cost-effective results, but it required a lot more computing power.

Other methods have investigated using Bayesian optimization to make feed that meets multiple goals. Uribe-Guerra et al. [20] used multi-objective Bayesian optimization (MOBO) to find the best swine diets that had the most digestible energy and lysine while also being the least expensive. MOBO was accurate and adaptable, but it did not scale

well when more constraints were added. To solve this problem, the same authors later came up with MORBO (Multi-Objective Regionalized Bayesian Optimization), which combines spatial partitioning and parallel search [21]. MORBO was four to eight times more efficient than MOBO and produced more varied diet solutions, but it was harder to tune the hyperparameters.

In general, optimization methods work better or worse depending on the problem at hand. Linear methods are easy to use and work well for calculations, but they have trouble with nonlinear nutrient interactions. Nonlinear methods more accurately reflect biological and technological limitations, but they require a lot of computing power. Multi-objective Bayesian optimization methods show a lot of promise for balancing cost, nutrient quality, and sustainability, but they need careful parameterization and more testing. In practice, hybrid methods that combine the best parts of linear, nonlinear, and probabilistic frameworks may be the most reliable way to make strong feed formulations.

Table 1. Comparative overview of optimization and recommendation methods for diet formulation of dogs.

Study	Category	Dataset	Method	Advantages	Limitations
Aguilar & Montoya (2023) [14]	Cost-optimal dog diets	Ingredient DB with nutrient composition and prices (Central America)	Linear Programming	Diets 14.3% cheaper than 4 commercial foods; better nutrient levels	Nutrient changes after cooking; regional price variation; no practical validation
Joban et al. (2020) [15]	Cost optimization of homemade dog food	Danish Food Informatics, NRC nutrient tables	Linear Programming	Web tool for ingredient-based ration formulation	No experimental validation
Song & Kim (2020) [22]	Recommender for commercial dog food	838 products, 1007 reviews	Nutrient-based Recommendation Algorithm (cosine similarity on averaged nutrient vectors)	Outperforms k-NN (+20% AUC) and SVD (+9.7% AUC); no cold-start; no pretraining needed	Dependent on availability and quality of user reviews

2.2. Recommender Systems

Text mining, clustering, and similarity-based algorithms are used in recommender systems to personalize food choices. These systems usually depend on information about the nutritional value, ingredient composition, user feedback, and product descriptions. Their main goal is to suggest choices that fit with what the user wants, what their body needs, or what they can eat. One ongoing problem is the cold-start problem, which happens when there is not enough information about new users or products to make recommendations.

A significant example in canine nutrition is the Nutrient-based Recommendation Algorithm (NRA) created by Song and Kim [22]. This model sorts dog food products by their nutrients, considering the dog's breed and health. It makes an averaged nutrient vector from user reviews and then uses cosine similarity to compare it to other diet profiles. The NRA was very efficient and worked well even when it was cold-started. It beat k-nearest neighbors (k-NN) and singular value decomposition (SVD) by 20% and 9.7% in AUC, respectively.

Other methods focus on matching text. Putra et al. [23] used Term Frequency–Inverse Document Frequency (TF-IDF) and cosine similarity to match user searches with descriptions of dog food. This method, on the other hand, is sensitive to synonyms, misspellings,

and variations in word choice. Li et al. [24] talked about similar problems in recommending human food, where ingredient names were often not standardized. Their partial standardization cut the number of ingredient terms from 1863 to 158. A predictive model then filled in the gaps, making the system more accurate.

More advanced methods use graph neural networks (GNNs) that have been trained on different types of knowledge graphs. Ma et al. [25] put forward a GNN-based food recommendation model that shows how users, meals, and nutrients are related. This lets users personalize their recommendations based on both taste and nutrient needs at the same time. However, these models necessitate extensive training data and may inadequately consider individual constraints, such as allergies. Chen et al. [26] also came up with the FKGM model, which sorts recipes by how much salt, fat, sugar, and saturated fat they have, to find a balance between user satisfaction and health outcomes.

Time element has been incorporated as well. Rostami et al. [27] put forward a hybrid model that merges the clustering of taste preferences with time elements (such as meals throughout the day) and similarity searches utilizing ingredient embeddings. This method works, but it needs an active user community, which makes it hard to grow. Parameter tuning has also been examined and research [28] demonstrated that employing 10 calorie levels yielded the highest recommendation quality, whereas increased granularity diminished classification accuracy.

Generative approaches are coming up that go beyond traditional systems. Majumder et al. [29] utilized a BiGRU with attention mechanisms to produce customized recipe texts based on user preferences. This method was made for human food, but it could also be used to make dog food recipes. Sadovoy et al. [30] also talked about a predictive model that used decision trees and neural networks to figure out how many calories and nutrients a person needs each day and made suggestions for certain foods.

Table 2. Comparative overview of optimization and recommender methods for diet formulation in animals.

Study	Category	Dataset	Method	Advantages	Limitations
Pavlova et al. (2024) [16]	Sheep feed	Nutrient composition, requirements, ingredient prices	Linear Programming	Accounts for energy, protein, mineral, and cost constraints	No experimental validation
Metwally (2023) [17]	Poultry feed (Egypt)	9–10 regional ingredients with prices	Linear Programming	Cost reduced by 8.3–13.5% across growth stages	No validation
Alqaisi & Schlecht (2020) [18]	Dairy cattle feed	Ingredient prices, nutrient composition, availability scenarios	Multi-period LP with time series	Adapts to price fluctuations; reduced CH ₄ emissions by 21–53%	Corner solution problem
Li et al. (2022) [19]	Dairy cattle feed	NRC (2001) feed composition & nutrient requirements	Comparison of ItLP, SQP, MINLP_DGO	MINLP_DGO robust with complex constraints; SQP efficient	ItLP uneconomical; SQP may fail to converge; high computational cost; no practical validation

Table 2. Cont.

Study	Category	Dataset	Method	Advantages	Limitations
Uribe-Guerra et al. (2024a) [20]	Swine feed optimization	17 ingredients, 28 nutrient constraints	Multi-Objective Bayesian Optimization (MOBO)	Improved digestible energy (+10.7%), lysine (+14.8%), and cost (−3.8%) vs. stochastic programming	Poor scalability; limited nutrient coverage; no practical validation
Uribe-Guerra et al. (2024b) [21]	Swine feed	Ingredient proportions; 3 objectives (lysine, energy, cost)	MORBO (regionalized MOBO with parallel search)	4–8× faster than MOBO; more solution diversity	Lower hypervolume than MOBO; parameter tuning required; no validation
Saxena & Parasher (2019) [31]	Animal feed (general)	5000 synthetic samples	ANN model	High correlation with ground truth ($r = 0.99$); accuracy grows with hidden nodes	Requires large training data; no validation

Table 3. Comparative overview of optimization and recommender methods for diet formulation for humans.

Study	Category	Dataset	Method	Advantages	Limitations
Putra et al. (2024) [23]	Dog food text-based recommendation	10 product descriptions	TF-IDF + cosine similarity	Works without ratings/reviews	Sensitive to synonymy, typos, non-standard names
Ma et al. (2024) [25]	Food recommendation	Food Recommender dataset	GNN on heterogeneous knowledge graph	F1 = 0.81; better than FeaStNet, DeeperGCN, GraphSAGE, etc.	Requires detailed knowledge graph; high compute; possible missing nutrients
Chen et al. (2023) [26]	Health-aware food recommendation	Allrecipes + Freebase	FKGM (knowledge graph + multi-task learning)	Balances taste and health; outperforms CFKG, KGAT, BPRMF	Only 4 nutrients considered; no user validation
Rostami et al. (2022) [27]	Time-aware meal recommendation	Scraped Allrecipes dataset	Hybrid: clustering + embeddings + temporal factors	Outperforms LDA, HAFR, FGCN; higher accuracy	High computational load; sensitive to data quality
Song et al. (2022) [28]	Calorie-aware food recommendation	Allrecipes dataset	Heterogeneous graph (users, meals, nutrients, calories)	Outperforms HAFR and Cal-HAFR on AUC	Sensitive to data quality; cold-start problem

In general, nutrition recommender systems have come a long way from simple vector and text-based models to more advanced graph-based and generative architectures. The best systems for making dog food are those that consider physiological factors, nutrient profiles, and how well the dog can handle cold-start conditions. Using natural language processing and graph-based models from human nutrition opens up new ways to make accurate, personalized, and scalable diet recommendations for dogs.

The literature distinctly delineates a separation between optimization-based methodologies and recommender systems. For animals (Table 2), linear programming is still the most common method because it is simple and cheap, even though it does not always work well with nonlinear biological constraints and has not been tested in real life. To get around these problems, more advanced methods like nonlinear programming, Bayesian optimization, and neural models have been suggested. However, these methods make the calculations much more difficult and cause scaling issues. The field of human nutrition has moved toward knowledge graphs and deep learning methods, which allow for more personalized results but need large, high-quality datasets (Table 3). There are not many recommender systems for dogs (Table 1). The Nutrient-based Recommendation Algorithm (NRA) [22] is a rare example of a nutrient-driven, cold-start-resistant model. This gap underscores the necessity for hybrid frameworks that amalgamate optimization rigor with the adaptability of contemporary recommender systems, specifically customized for the veterinary sector.

Table 4. Summary of optimization and recommender methods applied in diet formulation.

Method	Description	Limitations	References
Linear Programming (LP)	Optimization technique that maximizes or minimizes an objective function subject to linear constraints. Used to determine optimal ingredient ratios under nutrient, cost, or composition restrictions.	Corner solution problem; cannot model nonlinear dependencies; fails to provide feasible solutions in some cases.	[14–18]
Multi-Objective Bayesian Optimization (MOBO)	Probabilistic global optimization method using surrogate models (e.g., Gaussian processes). Applied for multi-criteria feed formulation (e.g., digestible energy, lysine, cost).	Poor scalability with many variables; high computational cost; sensitive to kernel choice.	[20]
Nutrient-based Recommendation Algorithm (NRA)	Constructs an “ideal diet vector” from reviews (accounting for breed, disease, and physiology); recommends food using cosine similarity.	May face scalability issues with large diet diversity; dependent on review data.	[22]
TF-IDF (Text Mining)	Evaluates word importance (e.g., ingredient names in recipes) for similarity matching.	Ignores synonyms and morphology; sensitive to typos; limited semantic capture.	[23]
Cosine Similarity	Measures the angle between vectors representing nutrient composition or text embeddings. Used for identifying similar products or ingredients.	Does not consider vector magnitude; vectors with different scales but same direction may yield similarity = 1.	[22,23,27]
Graph Neural Networks (GNNs)	Model knowledge graphs linking users, meals, ingredients, and nutrients for personalized recommendations or recipe generation.	Require high-quality knowledge graph structure; high computational complexity; dependent on large datasets.	[24–28]
Sequential Quadratic Programming (SQP)	Iterative nonlinear optimization method solving quadratic approximations at each step. Suitable for nonlinear nutrient constraints.	Sensitive to initial guess; risk of local minima; high computational cost with many constraints.	[24]
MINLP with Deterministic Global Optimization (MINLP_DGO)	Combines discrete and continuous optimization for problems with categorical (e.g., ingredient inclusion) and nonlinear biological constraints.	Very high computational cost; requires precise definition of all constraints and relationships.	[24]

3. Materials and Methods

A model for formulating specialized canine diets was developed. The nutritional needs of domestic animals can vary significantly depending on age, physiological condition, and the presence of diseases. The proposed model includes three components that provide a consistent transition from data collection to the calculation of an optimized diet formulation (Figure 1).

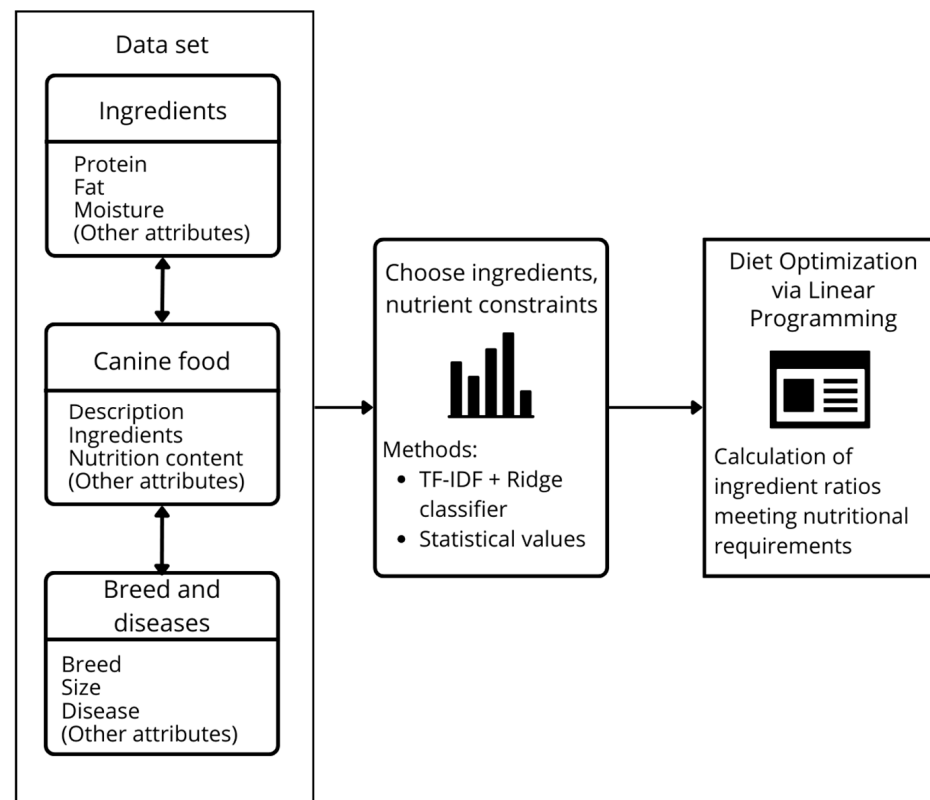


Figure 1. The general process of calculating a dog's diet.

The first component is a set of data required for calculating the recipe. It includes ingredient characteristics, information about commercial specialized foods, information about dog breeds and common diseases. Part of the data was obtained through web scraping from several sources (pet food e-commerce platform [10], websites providing information on dog breeds and diseases [11–13]). Certain information was incomplete and was additionally marked up manually. Since manual markup is a labor-intensive process, automatic annotation methods based on TF-IDF and the E5 transformer model were tested.

The second component of the model focuses on extracting and systematizing ingredients that are typical for specialized feeds. Feeds were grouped according to functional areas (such as Kidney Care, Heart Care, Aging Care, and other categories), and machine learning and statistical methods were used to identify the key ingredients typical for each area.

The third component optimizes the diet composition based on linear programming. The optimal ingredient ratio is determined, based on nutrient requirements and individual animal parameters. The optimization task is formulated as maximizing indicators that reflect the dog's needs for certain nutrients, given the selected category and constraints.

Based on the proposed model, a platform has been created that provides automatic calculation of the diet for a specific dog, considering its individual characteristics and the selected type of specialized nutrition.

3.1. Data Sources

The system is built upon several interconnected datasets that collectively form a comprehensive data ecosystem representing the nutritional domain of canine health. These datasets encompass detailed information on individual ingredients including nutrient composition, energy density, and ingredient class. Moreover, it contains canine diseases and disorders with corresponding dietary implications, contraindications, recommended nutrient adjustments, and veterinary diet formulations curated from commercial and research sources. The integration of these datasets, as shown in Figure 2 and Table 5, enables the system to establish meaningful relationships between ingredients, health conditions, and dietary recommendations, forming the foundation for both machine learning-based prediction and optimization-driven recipe generation processes.

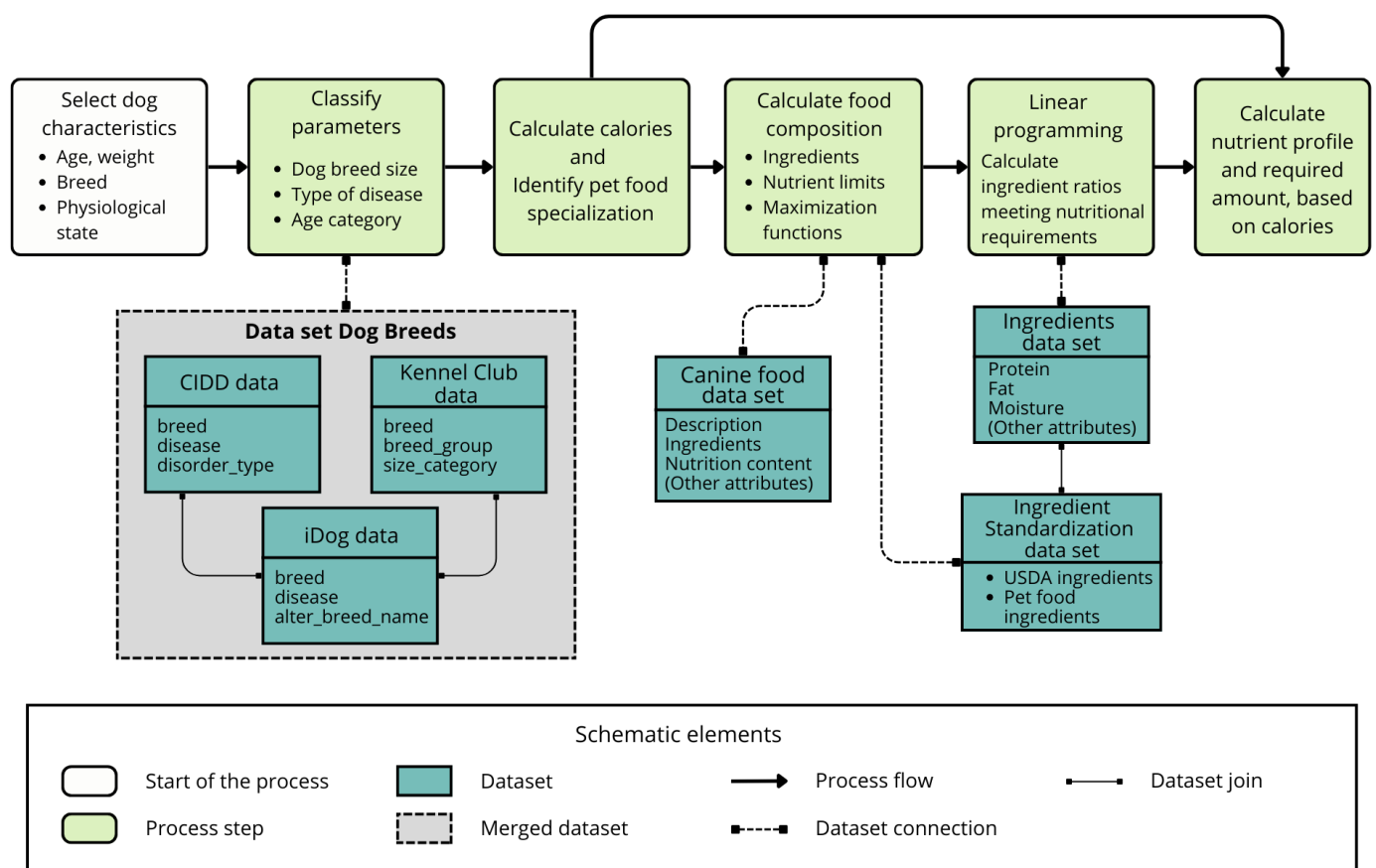


Figure 2. Integration of datasets in the feed recipe calculation process.

The first dataset is based on a list of ingredients and their chemical composition. These data were sourced from the open-access Food Nutrition Dataset [32], which was created using information published on the official website of the U.S. Department of Agriculture—USDA FoodData Central [33]. This resource represents one of the largest international databases on food products, their composition, and nutritional value. Since the original dataset covers a wide range of foods, many of which are not intended for canine consumption, only 126 out of 7413 ingredients were selected based on their safety and suitability for dogs. For analytical convenience, the selected ingredients were classified into seven functional categories: water and supplements, greens, grains, fats, meat, vegetables/fruits, and eggs/dairy products. The dataset contains average values of macro and micronutrient content, standardized per 100 g of edible portion, which ensures comparability and accuracy in diet modeling.

The second dataset covers information on dog breeds, diseases, and the relationships between them. It was created by combining data from several official sources. The primary information on breeds and associated diseases was obtained from the iDog platform [11], developed by the National Genomics Data Center (NGDC, China)—a leading research center specializing in the collection and systematization of genetic and health data on dogs. A key feature of this resource is its inclusion of both official and alternative breed names, which greatly facilitates the process of data standardization and integration across different sources. Additional information was incorporated from the Canine Inherited Disorders Database (CIDD) [12], maintained by the University of Guelph (Canada), and from The Kennel Club [13]. The CIDD provides structured data on hereditary canine diseases, including their classification by disorder type and the breeds most susceptible to each condition. The Kennel Club dataset offers details on breed nomenclature, group classifications (e.g., terriers, hounds, toy breeds), and size categories. After merging the data on breeds and diseases, a manual review and completion of missing information were conducted, along with the assignment of each disease to a corresponding disorder category.

The third dataset focuses on veterinary diets and their intended therapeutic purposes. Data on feed compositions were obtained from the official website of Hill's Pet Nutrition [10], an international manufacturer of specialized veterinary diets widely used in clinical practice. Both dry and wet food products were included, with detailed descriptions of their nutritional composition, such as protein, fat, and carbohydrate content. Moreover, information about each diet's specific health purpose was added, allowing for classification into categories such as skin and coat health support, weight management, joint health support, digestive health improvement, and other veterinary objectives. This dataset creates a direct link between diseases and corresponding dietary solutions.

Table 5. General description of the used datasets.

Source	Description	Data	Purpose
USDA FoodData Central [33]	Contains ingredients and their chemical composition: macro- and micronutrients, moisture, minerals, vitamins	7413 ingredients in the initial set; 126 ingredients after filtering and classification by category	Used to account for the nutrient composition of ingredients when calculating recipes and setting linear programming tasks
iDog platform [11]	Information about dog breeds: hereditary diseases, breed characteristics, alternative names	482 dog breeds	Used to determine breed characteristics, calculate calorie requirements, and select target feed purposes based on diseases; expand breed data
Canine Inherited Disorders Database (CIDD) [12]	Information about hereditary diseases in dogs: type of disease, classification, connection to breeds	147 diseases grouped into 11 categories; associations with 151 breeds	Used to compare breeds, disease types, and target feed purposes; supplementing dog health data
The Kennel Club UK [13]	Breed information: breed, breed group, size	277 dog breeds	Used to refine breed characteristics, calculate calorie requirements, and expand breed information

Table 5. Cont.

Source	Description	Data	Purpose
Hill's Pet Nutrition [10]	Commercial veterinary dog food: name, description, purpose, age, breed size, ingredients, nutritional composition	330 feeds	Used to determine specialized ingredients and nutrient restrictions depending on the feed's purpose
Ingredient coordination (manual processing)	Manual comparison of ingredient names from commercial feeds with ingredients from the USDA nutrient database	Pairs of "feed ingredient—ingredient from the database"	Used to combine recipe data with ingredient chemical composition information and generate an accurate list of recommended ingredients

The fourth dataset was constructed to standardize and harmonize ingredient names across the three datasets. Commercial feed compositions frequently contain generalized or non-specific ingredient descriptions, which complicate direct comparison and integration. To address this issue, ingredient normalization was performed using a frequency-based rule derived from the dataset itself. Specifically, generalized ingredient categories were replaced with the most frequently occurring specific ingredient within the corresponding category across all feed compositions. For example, the generic label "meat and animal by-products" was mapped to chicken meat, as chicken was the most commonly listed meat ingredient in the dataset. Likewise, the unspecific term "vegetables" was standardized to carrots, and "cereals" to rice, reflecting their highest occurrence frequencies within their respective categories.

The selection of these representative ingredients was based on their dominant presence in feed recipes, as demonstrated by the top 10 most frequently occurring ingredients in the dataset (Table 6). Normalization ensured consistency across datasets while preserving more realistic ingredient representations.

Table 6. Top 10 most frequently occurring ingredients in the pet food dataset.

Ingredient	Count
flaxseed	245
rice	204
corn	185
chicken	166
beet	162
carrot	159
chicken fat	137
fish oil	133
green pea	111
wheat	104

Collectively, these datasets combine information on ingredients, their composition, dog breeds, diseases, and specialized diets [34]. Based on them, dog diets were formulated with consideration of individual health conditions and nutritional needs.

3.2. Nutrient Requirement Modeling

The accurate determination of nutrient requirements is essential for generating personalized and physiologically balanced canine diets. In this study, nutritional requirements were derived from the FEDIAF Nutritional Guidelines for Complete and Complementary Pet Food for Cats and Dogs (2024 edition) [35], which provide recommended daily allowances (RDAs) and safe upper limits for macronutrients and micronutrients in dogs of different sizes, life stages, and physiological conditions. A custom computational was implemented to translate these recommendations into individualized nutritional targets, expressed as ranges against which optimized recipes could be validated.

The cornerstone of requirement modeling is the estimation of daily metabolizable energy, which determines the scale at which nutrients are supplied. ME was calculated using the metabolic body weight formula:

$$ME_{base} = \kappa \cdot \mathcal{W}^{0.75}, \quad (1)$$

where $\mathcal{W}$ is the body weight in kilograms, and the constant κ is defined by FEDIAF activity multipliers. Adjustment factors were incorporated for life stage and physiological status:

- Puppies: correction multipliers were applied by growth stage (weaning, 50% of adult weight, and near-maturity);
- Adults: multipliers distinguished sedentary, moderately active, and highly active dogs;
- Senior dogs: energy needs were reduced relative to adults to account for lower metabolic rates;
- Pregnancy: during the first 4 weeks of gestation, ME was increased modestly (typically 1.1–1.2× adult requirements), whereas during the last 5 weeks, energy demand rose significantly (up to 1.3–1.6×);
- Lactation: energy requirements were scaled both by week (weeks 1–4) and litter size, with maximum multipliers approaching 4–6× adult ME for large litters.

Thus, the final daily energy requirement was modeled as:

$$ME_{day} = f_{life-stage} \cdot f_{activity} \cdot \mathcal{W}^{0.75}, \quad (2)$$

where $f_{life-stage}$ and $f_{activity}$ are multiplicative correction factors derived from FEDIAF tables.

Protein, fat, and carbohydrate requirements were determined relative to either metabolic weight or daily energy intake. Protein requirements were scaled to metabolic body weight:

$$Protein_{req} = r_p \cdot \mathcal{W}^{0.75}, \quad (3)$$

where r_p is the stage and condition-specific protein coefficient. Adjustments were applied for growth, pregnancy, and lactation, reflecting increased anabolic demands. Fat and carbohydrate requirements were adopted directly from FEDIAF's recommendations as proportions of daily energy intake. Fat supplied a minimum proportion of daily calories to ensure the delivery of essential fatty acids and energy density, whereas carbohydrate requirements were formulated primarily as ranges that balanced energy provision and gastrointestinal health.

Micronutrient needs (minerals and vitamins) were computed according to FEDIAF standards, expressed either per kilogram of body weight or per 1000 kcal of metabolizable energy, depending on the nutrient category. For minerals such as calcium (Ca), phosphorus (P), magnesium (Mg), sodium (Na), and potassium (K), requirements were scaled to body weight:

$$Mineral_{req} = r_m \cdot \mathcal{W}, \quad (4)$$

where r_m is the recommended allowance (mg per kg body weight per day). For trace elements (iron, copper, zinc, manganese, selenium, iodine) and vitamins (A, D, E, B-complex, folate, choline), requirements were scaled per unit of energy intake:

$$Nutrient_{req} = r_n \cdot \frac{ME_{day}}{1000}, \quad (5)$$

where r_n is the recommended allowance per 1000 kcal. This approach ensures that nutrient densities remain adequate irrespective of overall energy requirement.

3.3. Data Preprocessing

The dataset obtained from Hill's Pet Nutrition [10] was initially characterized by significant heterogeneity in the structure and level of detail, which required standardization and supplementation. Missing data, the functional focus of the diet, the form of release (wet or dry food), the size of the target breed, the age category, and moisture content indicators for wet foods were added.

The processing of composition descriptions was aimed at extracting a list of ingredients and standardizing their presentation. Fragments containing a list of components were extracted from the descriptions, after which the text was cleaned of numerical designations and service markers. The text was then segmented and elements that did not correspond to food components (e.g., technical additives) were filtered out, followed by the normalization of lexical variants and standardization of names. Long-winded ingredients were combined into compound designations to ensure the uniformity of their presentation in subsequent stages of analysis.

The chemical composition indicators for pet food were converted to a comparable format. Nutrient values were normalized by removing units of measurement and converting them to numerical values. For wet foods, cases were considered where some of the indicators were presented in terms of dry matter, which led to a discrepancy in the total macronutrient content. In such situations, the values were converted back using the actual moisture content specified by the manufacturer.

3.4. Feed Classification by Functional Purpose

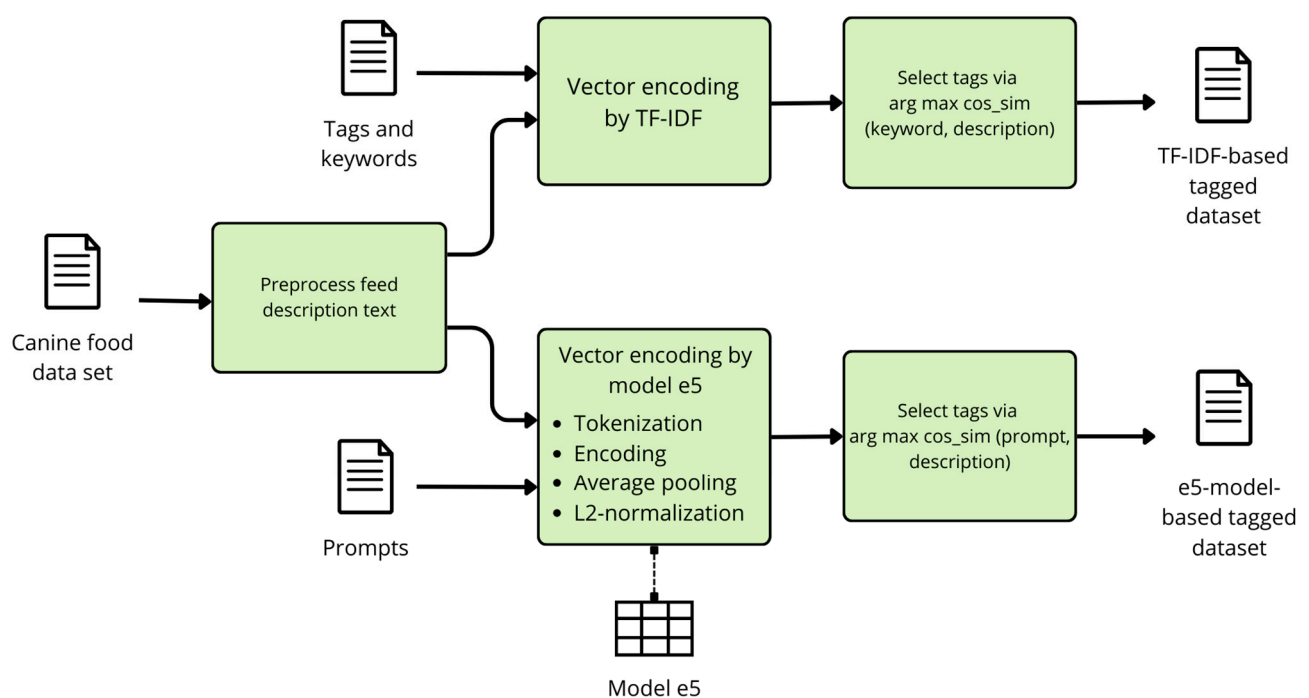
The initial dataset contained incomplete annotations: out of 330 feeds, 242 recipes were missing functional tags. These tags were originally assigned manually, which ensured acceptable accuracy but was labor-intensive and poorly scalable. To overcome this limitation, automatic restoration of missing categories was explored using information retrieval techniques and modern language models. In the future, automated tagging methods will enable rapid annotation when expanding the dataset with new recipes.

The functional categories used for annotation were derived from the classification scheme employed by the commercial website from which the data were scraped. Specifically, the taxonomy included 13 categories related to dietary correction or support under health conditions, as well as three general-purpose categories corresponding to immune support and complete balanced nutrition across three life stages: puppies, adult dogs, and senior dogs. To enable automatic labeling, a set of representative keywords for each category was defined manually, taking into account the terminology and phrasing commonly used in commercial feed descriptions. The resulting keyword lists are presented in Table 7 and serve as category descriptors in the subsequent tagging procedures.

Table 7. Keyword sets for identifying specialized feeds using TF-IDF search.

Feed Specialization	Keywords
Kidney Care	kidney kd
Liver Care	liver ld removal of toxic
Heart Care	heart hd
Dental Care	dental teeth
Urinary Care	urinary bladder stones
Cancer Patient Support	cancer patient support oncare onc
Urgent Care	urgent multiorgan recovering surgery illness injury
Digestive Care	digestive digestion stool
Food Sensitivity	food sensitivity hypoallergenic stomach
Skin Health	skin coat allergy
Musculoskeletal and Joint Care	muscle joint bone cartilage jd
Weight Management	weight management overweight
Nervous System Care and Stress	nervous system stress disrupted sleep brain
Aging Care	aging senior mature
Puppy Care	puppy grow start
Adult Care	adult immune optimal delicious

Since the functional purpose of a feed is typically explicitly or implicitly reflected in its textual description, missing tags were restored by analyzing product descriptions. To determine the functional affiliation of diets, two independent unsupervised approaches were applied, as illustrated in Figure 3: a method based on TF-IDF text representations and a method relying on sentence embeddings obtained using the E5 model.

**Figure 3.** The process of labeling a dataset using TF-IDF and the E5 model-based methods.

In the first approach, product descriptions were lemmatized and then vectorized using the TF-IDF (Term Frequency–Inverse Document Frequency) method. This method allows a collection of texts to be converted into a numerical representation. For each word t in the feed description d , TF-IDF is defined as the product of the word frequency and its inverse document frequency:

$$tfidf(t, d) = tf(t, d) \cdot \log\left(\frac{N}{df(t)}\right), \quad (6)$$

where $tf(t, d)$ is the frequency of word t in the feed description d , N is the total number of feeds, and $df(t)$ is the number of feeds containing word t . After constructing the matrix:

$$T = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{bmatrix}, \quad (7)$$

where each row v_i is the TF-IDF vector of the corresponding recipe, a search was performed for the most relevant recipes by keywords presented in Table 7. The search query—a set of key tokens—was also converted into a TF-IDF vector:

$$q = tfidf(query). \quad (8)$$

To assess the semantic proximity between the query and each recipe, the cosine similarity was calculated:

$$\cos(q, v_i) = \frac{q \cdot v_i}{\|q\| \|v_i\|}, \quad (9)$$

where q is the TF-IDF vector of the query, and v_i is the TF-IDF vector of the recipe description.

The similarity values obtained were added to the data frame as new features. For each recipe, the most relevant class (food category) corresponding to the key query with the maximum cosine similarity was selected. If we denote the set of search queries as $K = \{k_1, k_2, \dots, k_m\}$, then for feed d , the label was defined as:

$$label(d) = \operatorname{argmax}_{k \in K} \cos(q_k, v_d). \quad (10)$$

If there were several classes with the same similarity value, a set of all labels satisfying the threshold was used:

$$\cos(q_k, v_d) \geq \max_j \cos(q_j, v_d). \quad (11)$$

This method allowed recipes to be automatically classified according to categories typical for specialized feeds, and the results to be used for subsequent recipe development.

In the second approach, the E5 (Enhanced Embedding Encoder) language model was used to encode text descriptions of feed, forming high-dimensional semantic representations of texts [36]. Unlike the TF-IDF bag-of-words model, the E5 model extracts contextual features and allows for the evaluation of semantic proximity between texts even in the absence of direct term matches.

For each functional category, a specialized search query was formed, differing only in the name of the purpose. All queries had a common structure:

$$\text{Query}(c) = \text{"Instruct: Find a recipe for a given request that matches the specified purpose. Query : Find dog food related to } \{c\} \text{."},$$

where c is the name of the functional area (e.g., Kidney Care, Heart Care, Digestive Support).

Each query was encoded in embedding $q_c \in \mathbb{R}^d$, where d is the dimension of the model's hidden space. Similarly, a semantic representation of each text description of the recipe was formed:

$$v_i = E5(description). \quad (12)$$

To obtain a stable feature representation, average pooling of hidden states masked by the attention matrix was used:

$$v_i = \frac{1}{\sum_j m_{ij}} \sum_j h_{ij} \cdot m_{ij}, \quad (13)$$

where h_{ij} is the hidden state of the j -th token, and m_{ij} is the corresponding attention mask, which takes the value 0 for padding and 1 for real tokens. Next, all vectors were normalized using the L2 norm:

$$\tilde{v}_i = \frac{v_i}{\|v_i\|}, \quad (14)$$

$$\tilde{q}_c = \frac{q_c}{\|q_c\|}. \quad (15)$$

After obtaining normalized embeddings, the cosine similarity between each recipe and all formed queries was evaluated using Formula (9). For each recipe, the category with the maximum similarity value was selected (Formulas (10) and (11)).

Markup quality assessment: For both approaches—TF-IDF search and the E5 model-based method—a quantitative assessment of the quality of automatic functional category determination was performed. Standard information retrieval metrics were used: precision, recall, and $F1$ score.

$$Precision = \frac{TP}{TP + FP}, \quad (16)$$

$$Recall = \frac{TP}{TP + FN}, \quad (17)$$

$$F1 = \frac{Precision * Recall}{Precision + Recall}. \quad (18)$$

To evaluate the correctness of the annotation and identify the categories determined by the models, the true lists of feeds and the lists obtained automatically were compared for each specialty.

3.5. Analysis of a Specialized Feed Dataset

The composition and numerical data of nutrients were used to analyze the feed. The ingredient composition was considered as textual features, while the nutrient indicators (macronutrient and moisture content) were considered as numerical characteristics. Both groups of features were combined into a general representation.

Text data (ingredients) were converted into a vector representation using TF-IDF. The TF-IDF value for ingredient t in recipe d is determined by Formula (6). Where $tf(t, d)$ is the frequency of ingredient t in document d , N is the total number of recipes, and $df(t)$ is the number of recipes containing ingredient t . The resulting TF-IDF vectors were combined with numerical chemical composition features into a single combined vector:

$$X_{combined} = [X_{tfidf}, X_{numeric}], \quad (19)$$

where X_{tfidf} are ingredient vectors, and $X_{numeric}$ is a chemical composition vector. To reduce the dimensionality and visualize the combined feature matrix $X_{combined}$, t-SNE is applied, allowing us to obtain a 2D embedding (y_1, y_2) that preserves local similarities:

$$(y_1, y_2) = t - SNE(X_{combined}). \quad (20)$$

Functional categories of feed (e.g., support for the liver, cardiovascular system, immune function, etc.) were displayed in the visualization using color markers.

3.6. Ingredient Extraction

The model for automatically formulating specialized feed recipes involved two key stages: extracting relevant ingredients for a specific functional purpose and building optimized recipes using linear programming. The first stage involved identifying candidate ingredients that were most characteristic of a given disease category, while the second stage involved creating a balanced feed composition, considering nutritional value restrictions.

The ingredients characteristic of specialized feeds were identified based on machine learning, statistical and frequency analysis of the distribution of ingredients in functional categories. Two approaches were considered: a machine learning-based feature extraction method and a statistical method including the χ^2 criterion, Bayesian/PMI, and Log-Likelihood Ratio.

A list of candidate ingredients was formed, after which one component representing four key functional groups was selected from it: protein sources, fats, cereal components, and fruits/vegetables. Water was included in each recipe as a mandatory component since the modeling was performed for wet foods.

3.6.1. Machine Learning-Based Method

A hybrid machine learning framework was developed to predict nutrient content and rank candidate ingredients for canine-specific dietary profiles. The system integrates textual product information with structured categorical features and employs both regression and classification models to generate recommendations (Figure 4).

All textual fields from the commercial dataset, including product titles, descriptions, claims, and related metadata, were concatenated into a single document representation per product. These documents were transformed into numerical features using the term frequency-inverse document frequency (TF-IDF) scheme. For a given term t in document d , the TF-IDF weight was computed as

$$tfidf(t, d) = \frac{f(t, d)}{\sum_{\omega \in \vartheta} f(\omega, d)} \cdot \log \left(\frac{M}{1 + \{d' : t \in d'\}} \right), \quad (21)$$

where $f(t, d)$ is the frequency of term t in d , ϑ denotes the vocabulary, and M is the number of documents in the corpus. To reduce sparsity and capture latent semantic structure, truncated singular value decomposition (SVD) was applied, yielding a dense embedding space of 100 latent dimensions:

$$X_{tfidf} \approx U_K \Sigma_K V_K^T, Z = U_K \Sigma_K \in \mathbb{R}^{M \times 100}. \quad (22)$$

The final input representation was constructed as the horizontal concatenation of textual embeddings. To infer the likelihood of ingredient presence, Ridge classifiers were trained for all ingredient tokens. For each candidate ingredient g a binary label was assigned, indicating its presence or absence in product i . The Ridge classifier then learned a linear decision function:

$$s^g(\chi) = \omega_g^T \chi + b_g, \quad (23)$$

which was used to rank ingredients by predicted relevance. Unlike probabilistic classifiers, the Ridge classifier provides raw decision scores, which were directly employed to construct top-ranked ingredient lists.

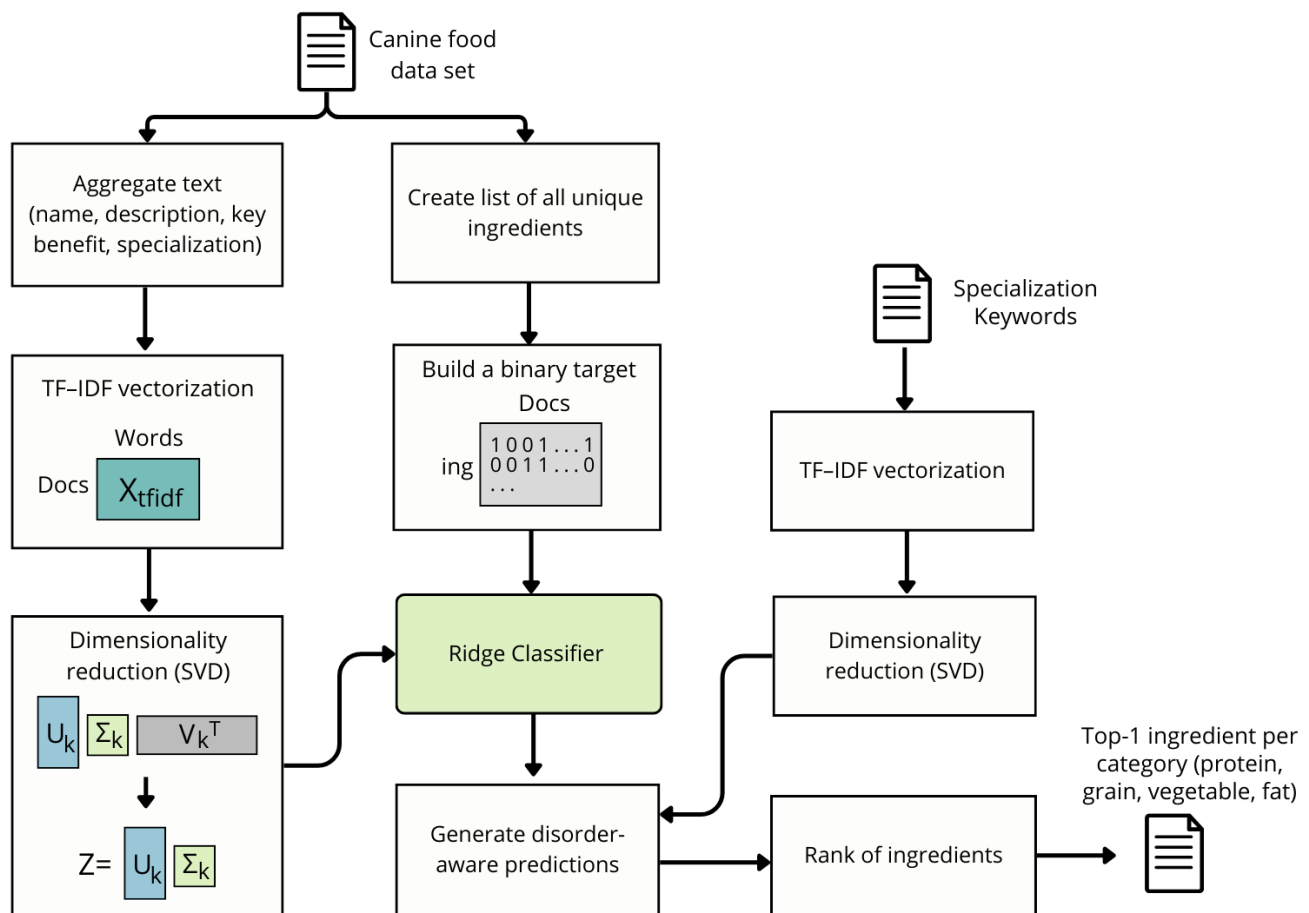


Figure 4. The process of selecting ingredients based on machine learning.

To incorporate disorder-specific context, the keyword dictionary from Table 7 was used. These queries were vectorized using the same TF-IDF and SVD pipeline, producing a disorder embedding z_c . This representation was passed through the ingredient classifiers to generate disorder-aware predictions, while ingredient ranking scores were given by:

$$s^g(Z_c) = \omega_g^T Z_c + b_g. \quad (24)$$

The resulting disorder-aware ranked ingredient lists were subsequently integrated into the linear programming optimization module, which enforced nutritional and categorical constraints to produce feasible wet food recipes.

3.6.2. Statistical Value-Based Method

In the second approach, ingredients were extracted based on the X^2 criterion, Bayesian/PMI, and Log-Likelihood Ratio. In the first stage, X^2 criterion values were calculated for each ingredient in relation to functional categories. The X^2 statistic was determined using the formula:

$$X^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \quad (25)$$

where O_{ij} is the observed number of recipes with/without an ingredient for a given category, and E_{ij} is the expected number of recipes under the hypothesis of independence. High X^2 values indicated a statistically significant relationship between the ingredient and the category. Preference was given to this criterion, as it allows the significance of associations to be confirmed.

However, given the limited size of the dataset, the number of significant ingredients was often insufficient to form a complete recipe database. In such cases, alternative measures of association were used: Bayesian/PMI and Log-Likelihood Ratio. Bayesian/PMI was calculated using the formula:

$$PMI(w, c) = \log \frac{P(w, c)}{P(w)P(c)}, \quad (26)$$

where $P(w, c)$ is the joint probability of ingredient w and category c appearing, and $P(w)$ and $P(c)$ are their marginal probabilities.

Ingredients with PMI values greater than 0 were ranked from highest to lowest and added to the candidate list. This criterion was used as a milder measure of association, allowing us to identify ingredients that may not be frequent enough to pass the X^2 threshold but are still informative in terms of joint occurrence.

When there was a shortage of ingredients after applying X^2 and PMI, the Log-Likelihood Ratio (LLR) was used, which is well suited for analyzing small samples and rare events. Its value was calculated using the formula:

$$LLR = 2 \left[k \ln \frac{k}{E_1} + (n - k) \ln \frac{n - k}{E_2} \right], \quad (27)$$

where k is the number of recipes containing an ingredient in each category, n is the total number of recipes, and E_1 and E_2 are the expected frequency values under the independence hypothesis. Ingredients were selected in descending order of LLR, which allowed us to identify the most likely associations between rare ingredients and categories.

This approach ensures the identification of both statistically significant and informationally related ingredients, which was critical given the limited amount of data and the need to form a complete set of recipe components.

3.7. Diet Optimization via Linear Programming

Linear programming (LP) was used to optimize the ingredient ratios in specialized feeds, solved via the Python module. The main goal was to develop recipes that meet nutritional requirements and maximize the content of key macronutrients characteristic of specific functional feed categories. The following data were used as input: the nutrient profile of the ingredients, restrictions on their quantity in the recipe, and nutrient values for the target feed categories. The problem was formulated as follows:

Let n denote the number of candidate ingredients selected for formulation. The decision variable for each ingredient i is the fraction x_i (per 100 g recipe), constrained to be non-negative and within category-specific bounds:

$$l_i \leq x_i \leq u_i, \quad i = 1, \dots, n, \quad (28)$$

where l_i and u_i are user-defined lower and upper limits (e.g., oils 1–10%). The sum of ingredient fractions must equal 100 g:

$$\sum_{i=1}^n x_i = 1. \quad (29)$$

This ensures that the output represents a complete 100 g recipe. To maintain dietary diversity and nutritional plausibility, category-level constraints were imposed. For example,

$$\begin{aligned} 0.3 &\leq \sum_{i \in P} x_i \leq 0.60 \text{ (proteins)}, 0.1 \leq \sum_{i \in C} x_i \leq 0.35 \text{ (cereals)}, \\ 0.1 &\leq \sum_{i \in V} x_i \leq 0.25 \text{ (vegetables)}, \\ 0.01 &\leq \sum_{i \in O} x_i \leq 0.10 \text{ (oils)}, \\ 0.00 &\leq \sum_{i \in W} x_i \leq 0.30 \text{ (water/other)}, \end{aligned} \quad (30)$$

Here, P, C, V, O, W denote ingredient sets belonging to protein-rich foods, cereals, vegetables, oils, and water/other categories, respectively. Each ingredient i contributes a known nutrient profile, where (a_{ij}) is the amount of nutrient j per gram of ingredient i . The nutrient level in the recipe is

$$y_j = \sum_{i=1}^n a_{ij} x_i \cdot 100. \quad (31)$$

where y_j is the content of nutrient j in the recipe.

Determination of nutrient ranges. For each macronutrient (moisture, protein, fat, carbohydrates), minimum and maximum values were determined based on existing specialized feeds. To do this, the minimum and maximum values for wet feeds were calculated separately for each purpose. These values formed the lower and upper limits of the constraints in the linear programming model:

$$y_j^{\min} \leq y_j \leq y_j^{\max}, \quad j \in \{\text{moisture, protein, fat, carbohydrate}\}. \quad (32)$$

The optimization objective was defined as maximizing the sum of selected nutrients (default: protein and moisture). The objective function was formed based on a comparative analysis of the average values of macronutrients in specialized and all feeds. If the average value of a nutrient in specialized feeds exceeded the average value in the total sample, this nutrient was included in the maximization function. The selection was carried out separately for dry and wet feeds, after which the results were combined. Formally, the objective function had the following form:

$$\max \sum_{j \in S} y_j. \quad (33)$$

In LP form, this is equivalent to minimizing their negative contributions:

$$\min_x f^T x, \quad f_i = - \sum_{j \in S} a_{ij}, \quad (34)$$

where S is the set of nutrients chosen for maximization.

Figure 5 shows an example illustrating the process of finding an optimal solution using linear programming for a case with two ingredients and two nutrients. The feasible region is defined by constraints, and the optimal point is determined by an objective function aimed at maximizing the total amount of nutrients. This example demonstrates how linear programming is applied to determine the most effective combination of ingredients to achieve the desired nutrient balance.

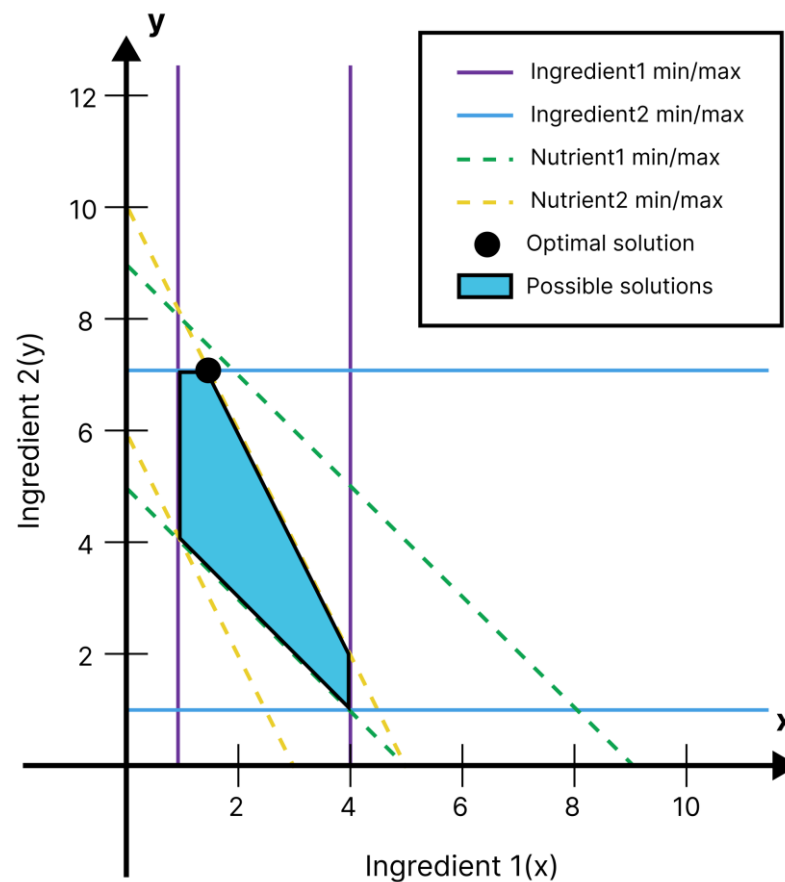


Figure 5. Geometric interpretation of a linear programming problem with two variables, where the objective function aims to maximize the total amount of nutrients.

If the solver failed to converge, a brute-force combinatorial search was executed to identify feasible solutions with minimum penalty scores for nutrient deviations. In this fallback method, candidate recipes were generated on a discretized grid (percentages summing to 100 g). Each candidate was scored with a penalty function measuring deviation from target nutrient ranges:

$$P(x) = \sum_j (\max\{0, L_j - Y_j\} + \max\{0, Y_j - U_j\}). \quad (35)$$

The feasible recipe with the lowest penalty was selected as the final solution.

3.8. Formulated Recipe Analysis

The Jaccard coefficient was used to quantify structural similarity between generated recipes based on their ingredient sets and to analyze the internal consistency of the recipe generation process. Importantly, this metric was not used as an indicator of nutritional quality, nor to assess similarity to commercial pet foods.

The analysis was conducted separately for each ingredient extraction method. For each method, a set of generated recipes was obtained, and pairwise comparisons were performed only within the same method. This approach made it possible to evaluate the degree of variability introduced by a given method and to identify potential tendencies toward unification or heterogeneity of ingredient composition. The metric was determined by the formula:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}, \quad (36)$$

where A and B are the sets of ingredients of two recipes. The values of the coefficient range from 0 to 1: values close to 0 indicate high recipe diversity, while values close to 1 indicate almost identical composition.

For each recipe R_i , the average value of the Jaccard coefficient was calculated relative to all other recipes R_j within the same method:

$$\bar{J}(R_i) = \frac{1}{N-1} \sum_{j \neq i} J(R_i, R_j), \quad (37)$$

where N is the total number of recipes in the method.

This analysis allows for the assessment of whether ingredient extraction methods produce sufficiently diverse recipes and whether the resulting ingredient sets generated by different methods exhibit comparable levels of internal consistency.

For a comparative analysis of the recipes formed between the two models, a method based on the vector representation of the ingredient and nutrient composition of each recipe was used.

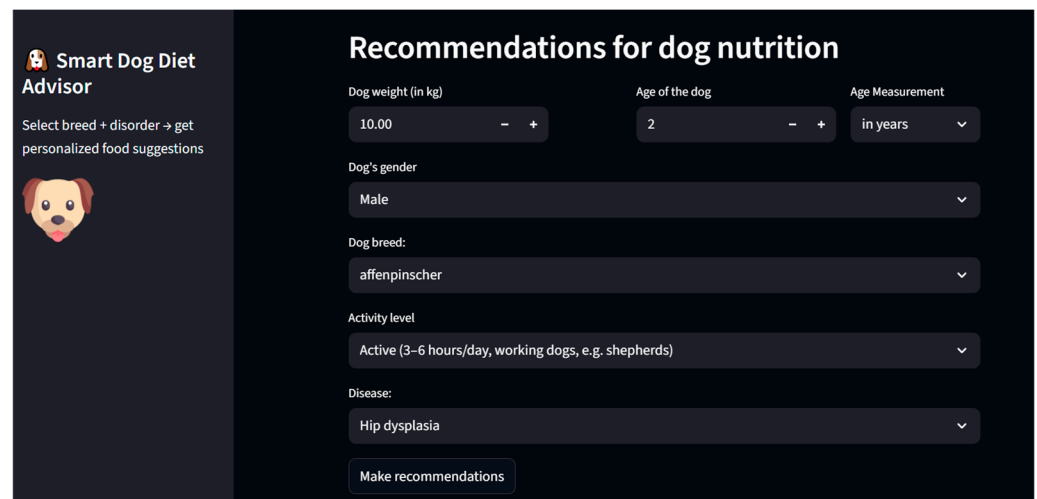
The textual description of the ingredient composition was converted into a vector representation using TF-IDF. Numerical nutrient indicators, including moisture, protein, carbohydrate, and fat content, were normalized using standard scaling. After normalization, the nutrient features were combined with the TF-IDF matrix into a single multidimensional feature space. The t-SNE dimension reduction technique was used for visual analysis of the distribution of recipes in multidimensional space.

Recipes calculated by two methods and related to the same functional purpose were compared using cosine similarity. The obtained cosine similarity values made it possible to quantitatively assess the consistency of approaches to recipe formation and identify categories for which the methods give similar or, conversely, significantly different solutions.

3.9. Implementation Details

The proposed system was built in Python 3.10 and used common scientific and machine learning libraries like Streamlit, pandas, NumPy, SciPy, scikit-learn, and Matplotlib. The Streamlit framework was used to make the application an interactive web interface as illustrated in Figure 6. This lets both experts in the field and pet owners use the recommendation pipeline in real time. The user interface shows people how to do things step by step:

1. Profile specification: The user gives important information like breed, body weight, age, activity level, sex, and reproductive status (for example, pregnancy or lactation). Additionally, curated disease taxonomies can be used to find breed-linked predispositions to inherited disorders.
2. Requirement calculation: The system figures out the metabolizable energy (ME) and nutrient needs based on the input profile, following FEDIAF guidelines and making physiological changes for the dog's age and health.
3. Recommendation retrieval: A hybrid ML pipeline (TF-IDF, dimensionality reduction, Ridge regression/classification) finds food products and ingredient vectors that are related to the disorder profile that was given.
4. Recipe optimization: The app uses a constrained linear programming problem to find wet food formulations that meet both nutritional ranges and ingredient category limits. If direct optimization does not work, a combinatorial fallback method is used.
5. Interactive refinement: Users can change ingredients and the amounts of each ingredient using sliders, and see right away how those changes affect how well the recipe meets nutrient ranges. Bar charts and caloric equivalence calculations are all clear ways to show how well a diet works.



Smart Dog Diet Advisor

Select breed + disorder → get personalized food suggestions



Recommendations for dog nutrition

Dog weight (in kg): 10.00 — +

Age of the dog: 2 — +

Age Measurement: in years ▼

Dog's gender: Male ▼

Dog breed: affenpinscher ▼

Activity level: Active (3–6 hours/day, working dogs, e.g. shepherds) ▼

Disease: Hip dysplasia ▼

Make recommendations

Figure 6. Main page with recommendations for dog nutrition generated based on input parameters.

This design makes sure that the system strikes a good balance between being hard to use and being easy to use. One side of the underlying architecture uses advanced machine learning algorithms and linear optimization models that strictly follow veterinarian nutritional rules. This makes sure that the formulations that are made are scientifically accurate and can be repeated. The user interface, on the other hand, was designed to be clear and interactive, so it turns these complicated calculations into easy-to-understand graphs and suggestions.

4. Results

4.1. Feed Classification Evaluation

The results of automated recipe annotation are presented in Figures 7 and 8. The TF-IDF and E5 methods demonstrated comparably high overall classification precision: the average values were 85.4% and 88.2%, respectively. Both approaches allow for the fairly accurate assignment of functional labels to prescriptions and provide an acceptable quality of initial automated annotation.

When analyzing the distribution of errors in the TF-IDF model, it was found that the precision of the Adult Care category is significantly lower, while recall, on the contrary, is higher. For Skin Health, the opposite situation is observed: recall is significantly reduced, and precision is high. These results may indicate a weak differentiation between the mentioned categories in the recipe description texts. Many specialized feeds are intended for adult animals, so the model could have generated biased predictions, assigning some of the Skin Health recipes to Adult Care.

Low recall values were also recorded for the Aging Care, Musculoskeletal and Joint Care, Urgent Care, and Weight Management categories, probably due to greater variability in their text descriptions or less pronounced unique key terms.

Categories with a narrow physiological focus, such as Heart Care, Urinary Care, and Cancer Patient Support, demonstrate high precision and recall scores, indicating clearer and more structurally consistent verbal descriptions. It should also be noted that the TF-IDF approach assigns a label based solely on the maximum cosine similarity value; therefore, additional or secondary specializations of the prescription found in the source data are less well captured by the model.

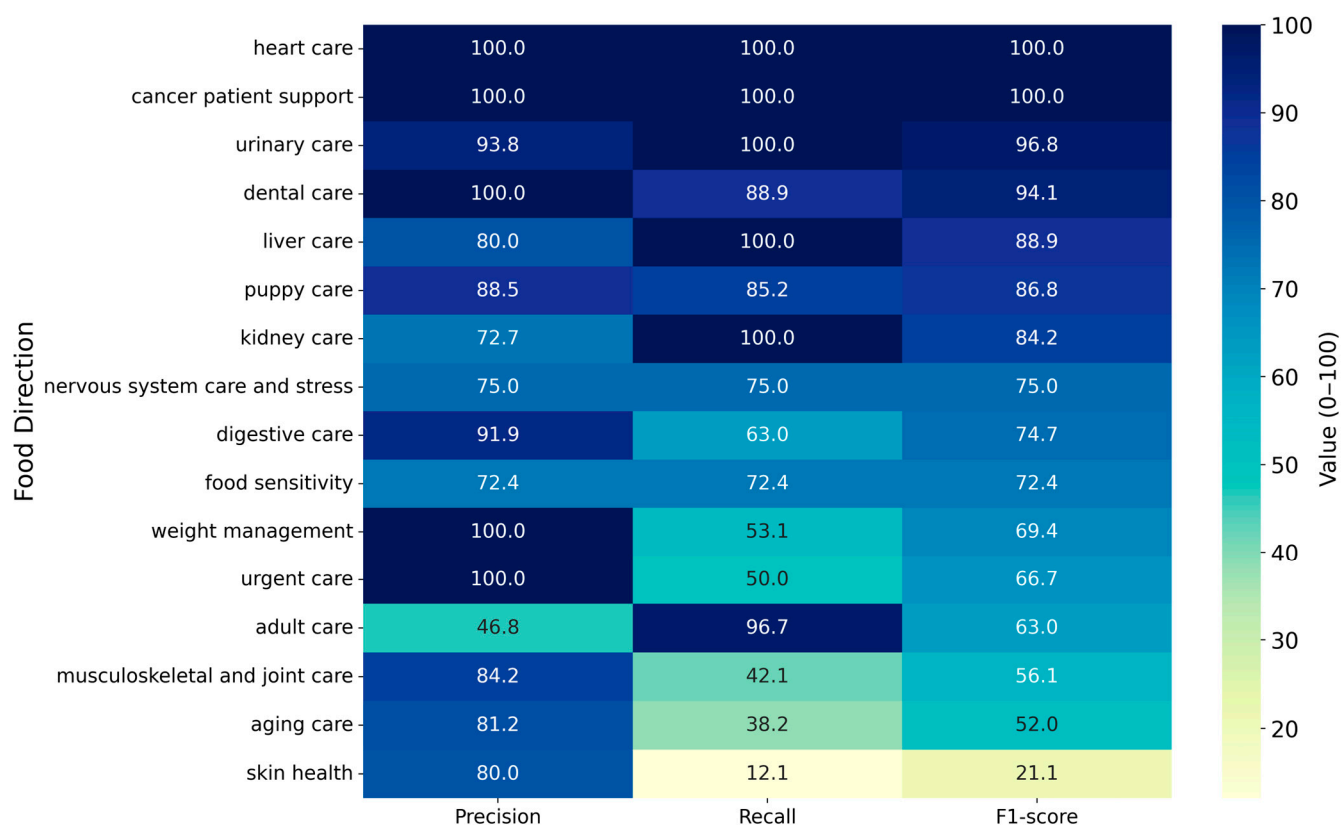


Figure 7. Precision, recall, and F1-measure metrics for automated recipe tagging using a TF-IDF-based method.

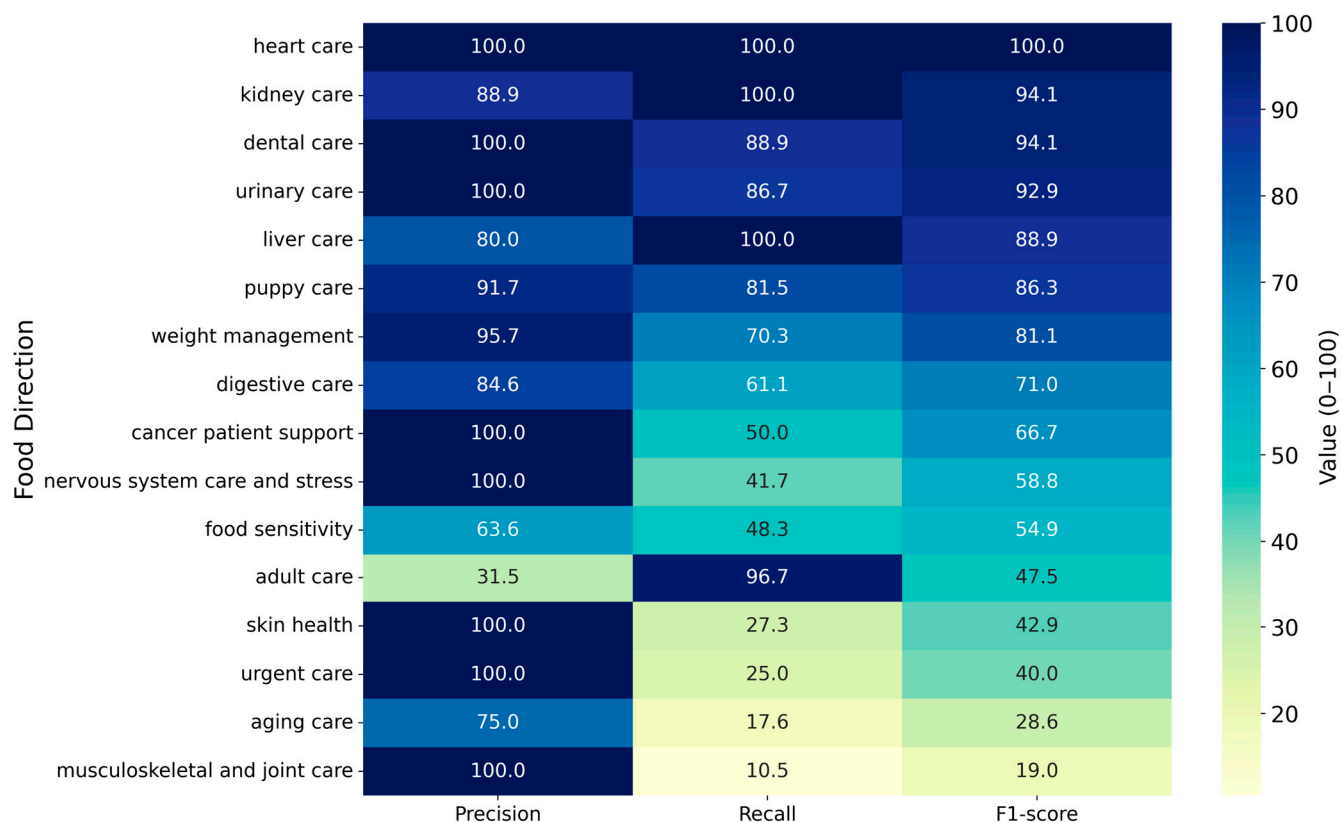


Figure 8. Precision, recall, and F1-measure metrics for automated recipe classification by the E5 model.

The E5 model (Figure 8), which uses contextual representations, showed higher precision for most categories. F1 scores exceeded 90% for Heart Care, Dental Care, Kidney Care, and Urinary Care, highlighting the language model's ability to consider semantic relationships between ingredients and functional characteristics.

However, recall for several categories remains low, indicating that the model is not sensitive enough to less explicit or ambiguous recipe descriptions. Particularly low F1 scores were observed for the categories Musculoskeletal and Joint Care (19%), Urgent Care (40%), and Aging Care (28.6%). This is likely since recipe texts in these groups often include general wording related to maintaining overall health, as well as references to the intended use of the feed for adult animals. These elements may dominate the description and reduce the possibility of unambiguously assigning the recipe to a narrow-specialized category.

A comparison of the two methods shows that E5 provides higher precision and better handles semantically complex and branched categories, where the interpretation of contextual relationships is important, not just the coincidence of terms. At the same time, TF-IDF is characterized by higher recall: the average recall value was 73.5%, while for the E5 model it was 62.8%. This makes TF-IDF preferable for tasks that require maximum data coverage, such as building large samples for subsequent analysis or model training. In contrast, E5 is a more reliable tool for tasks where annotation precision is critical and minimizing misclassifications is required.

Overall, both models demonstrate high efficiency when working with categories based on clearly defined physiological characteristics (heart, kidneys, oral cavity), while categories with more general or overlapping functional descriptions cause difficulties for both approaches. Joint analysis shows that TF-IDF provides better coverage, while E5 provides more accurate semantic interpretation, allowing the approach to be selected depending on the analytical goals of the study.

4.2. Specialized Feed Dataset Analysis

To understand the similarities and differences between foods based on ingredients and nutritional composition, data obtained from Hill's Pet Nutrition [32/12] was visualized. Figure 9 shows several distinct clusters, within which foods of different functional categories are located. This indicates structural differences between recipes. At the same time, some of the clusters overlap, reflecting the similarity of recipes and the use of similar basic components in feeds for different purposes.

Two large areas are particularly noticeable: the left, where dry feeds are mainly concentrated, and the right, corresponding to wet recipes. Dry feeds are more numerous and are characterized by greater variability in composition, which makes the functional differences between them more pronounced (Figure 10).

Some categories show a high degree of intra-group homogeneity and form more compact clusters. These include Puppy Care, Kidney Care, and Weight Management. This is due to the narrower specialization of these recipes and strict nutritional requirements: restrictions on protein and minerals (for Kidney Care), increased energy density (Puppy Care), or reduced calorie content and weight control (Weight Management).

Many categories overlap significantly, and there are virtually no completely isolated groups. The functional purpose is not always determined by a unique formula: general care foods—Adult, Aging, Digestive, Skin Health—are distributed more widely, reflecting the diversity of their ingredient profiles and the absence of strict nutrient restrictions.

Adult Care is often grouped closer to the center of the clusters. Aging Care shifts toward the top and bottom, which may reflect the variability of recipes for older dogs. Digestive Care is concentrated mainly on the left side of the clusters. Skin Health and Food Sensitivity foods are often located together, which is consistent with their intended use:

both types are designed for animals with allergies and intolerances, which often manifest as skin reactions.

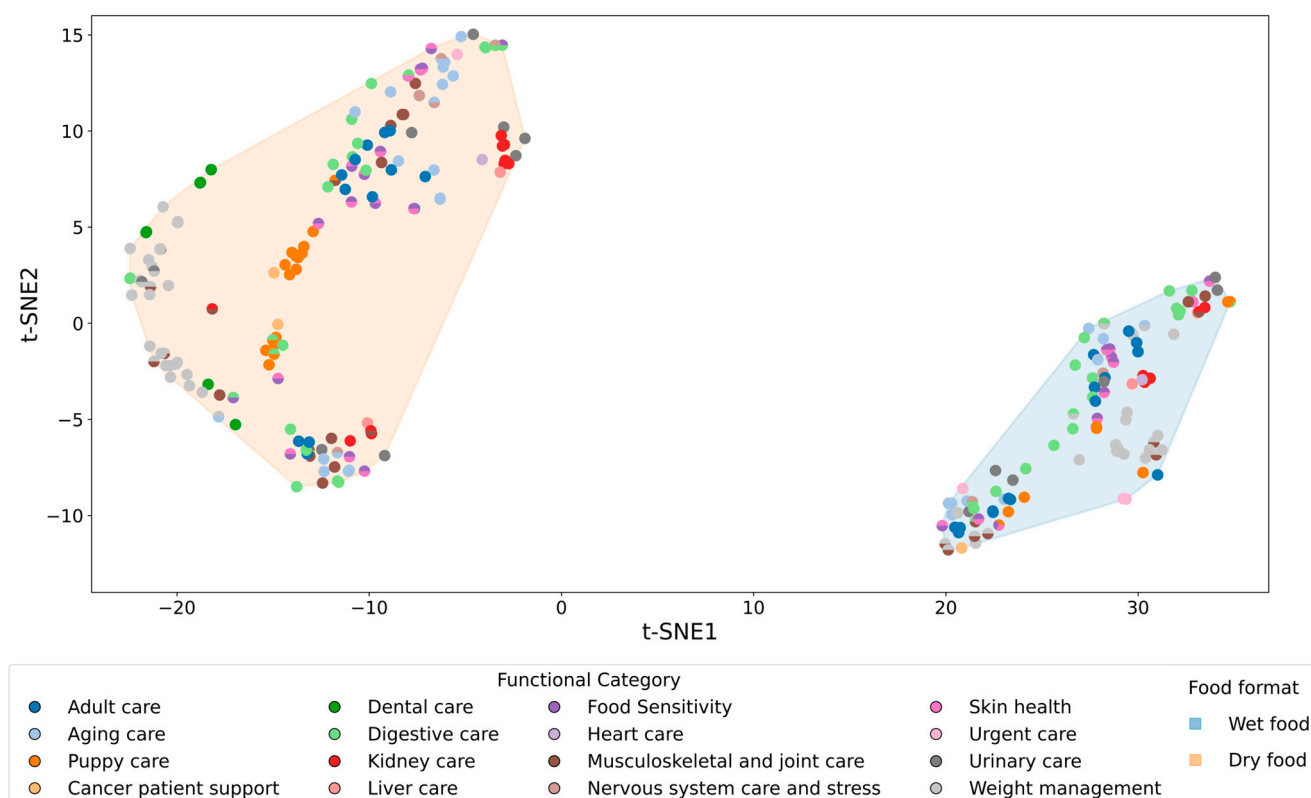


Figure 9. Feed visualization by ingredient and chemical composition.

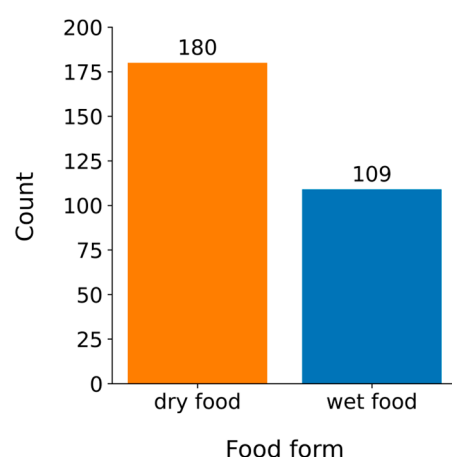


Figure 10. Distribution of dry and wet pet foods within the dataset.

Weight Management and Musculoskeletal Care foods are mainly located in the wet recipe zone, closer to the bottom of the cluster. These products are used for overweight animals, which puts extra strain on their joints, as well as to support large breeds. Based on this, combined uses may have similar recipes and form similar groups. However, the data presented is insufficient for a complete analysis of such foods.

The t-SNE visualization demonstrates that functional categories of feed do not form completely isolated clusters. This may be since, for some categories, it is important to consider the composition of micronutrients, technical additives, and ingredient ratios. Nevertheless, groups based on strict nutrient standards show a more pronounced structure and form noticeable clusters.

4.3. Formulated Recipe Evaluation

To analyze the consistency and diversity of the developed recipes, an assessment of the ingredient composition similarity was conducted. Table 8 presents the average Jaccard coefficient values calculated for each recipe based on its similarity to other recipes within the same method.

Table 8. Jaccard similarity scores for each approach, and cosine similarity of recipes across methods within each category.

Feed Specialization	Machine Learning- Based Method	Statistical Values- Based Method	Cosine Similarity
Kidney care	0.17	0.2	0.09
Liver care	0.28	0.18	0.91
Heart care	0.48	0.17	0.56
Dental care	0.52	0.15	0.89
Urinary care	0.38	0.2	0.79
Cancer patient support	0.34	0.24	0.80
Urgent care	0.5	0.21	0.83
Digestive care	0.49	0.17	0.82
Food sensitivity	0.4	0.22	0.60
Skin health	0.48	0.2	0.53
Musculoskeletal and joint care	0.5	0.19	−0.69
Weight management	0.5	0.22	0.56
Nervous system care and stress	0.5	0.17	−0.13
Aging care	0.38	0.2	0.43
Puppy care	0.5	0.2	0.84
Adult care	0.49	0.2	0.38
Average value	0.43	0.19	-

The machine learning method showed an average Jaccard coefficient of 0.43, indicating more pronounced similarities in composition between recipes. In contrast, the statistical method showed a significantly lower average value of 0.19, indicating noticeably more diverse and less overlapping ingredient compositions.

In the statistical approach, almost all recipes have a unique set of components (range 0.17–0.24), while in the machine learning-based method, the differences are more pronounced (range 0.17–0.52). More unique recipes within the ML-based method are observed for the categories Kidney Care (0.17), Liver Care (0.28), and Cancer Patient Support (0.34), reflecting the need for specialized ingredients for these areas.

The differences between the methods are related to the fact that the ML-based approach forms denser and semantically consistent sets of ingredients compared to statistical association measures. Since key components often appear in feed descriptions, TF-IDF + SVD representations capture the hidden semantic structure of texts, and the ridge classifier additionally provides continuous significance scores for all candidate ingredients. This allows the model to transfer contextual information between functional categories and generate more overlapping recommendation lists. In contrast, the statistical approach identifies ingredients characteristic of a particular category based solely on frequency patterns, resulting in greater recipe variability and less overlap between sets. Thus, the

TF-IDF-based method is more effective in tasks where the textual proximity of uses and ingredients is important, while the statistical approach scales better and emphasizes statistically significant components. These observations characterize the internal behavior of the recipe generation methods and do not imply the superiority or nutritional adequacy of any specific formulation.

To assess the consistency between the two approaches, the cosine similarity between pairs of prescriptions generated by each method for the same functional category was calculated. The final values show significant variability, as reflected in Figure 11 and Table 8. High cosine similarity values are observed in the Liver Care (0.91) and Dental Care (0.89) categories. In these cases, both methods agree on the set of ingredients and generate recipes with a similar nutrient profile. This degree of similarity indicates that the requirements of these categories are reflected quite clearly in the texts and statistics. A similar situation is observed in the categories Puppy Care, Urgent Care, Digestive Care, and Cancer Patient Support, where the cosine similarity exceeds 0.80. Here, the consistency of approaches is also high, which may indicate the presence of a pronounced and well-structured pattern in the composition of feeds characteristic of these specific purposes.

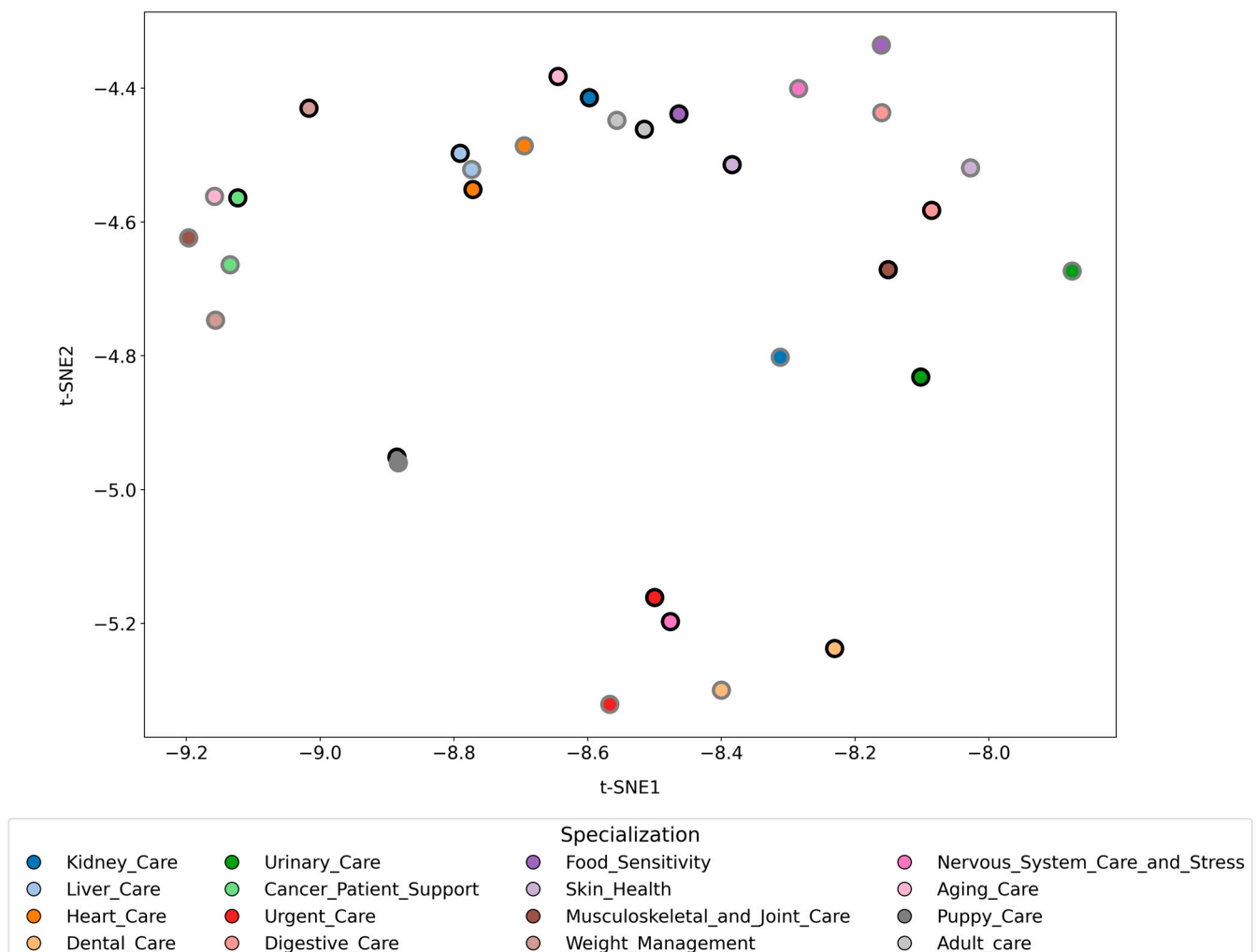


Figure 11. Visualization of developed recipes obtained using machine learning and statistical values. Points with gray borders indicate recipes calculated by ML-based method, while black borders indicate recipes formed by the statistical approach.

In contrast, categories with low cosine similarity values exhibit substantial differences in both ingredient composition and nutrient profiles. These include Kidney Care,

Musculoskeletal and Joint Care, and Nervous System Care and Stress. Such variability may be partially attributed to greater heterogeneity in formulation strategies within these categories.

Negative cosine similarity values indicate that recipe vectors are oriented in opposite directions within the joint ingredient–nutrient feature space, suggesting fundamentally different formulation strategies. In particular, for the Musculoskeletal and Joint Care category, the negative similarity value (−0.69) reflects a pronounced divergence in both ingredient selection and nutrient emphasis between the two methods.

Overall, high cosine similarity values can be interpreted as an indicator of the stability and reproducibility of approaches: if two different methods arrive at a similar recipe, then such a recipe probably reflects the functional requirements of the category well. Low values, on the contrary, emphasize the sensitivity of models to data structure and indicate areas where the wording of descriptions, the distribution of ingredients, or the specificity of the category itself are less clear, requiring a further refinement of criteria or expansion of the dataset.

5. Discussion

The results demonstrate the potential for automating the calculation of specialized recipes for dogs and confirm the possibility of filling in missing data using text markup. The analysis of ingredient profiles showed that foods with the same functional purpose can be similar in composition and nutrient profile. A comparison of recipes constructed using two methods also revealed both areas of overlap and significant variability.

The high precision of the TF-IDF and E5 models confirms that text descriptions contain sufficient information for the correct classification of specialization. The categories with clear physiological requirements, such as Kidney Care, Heart Care, Dental Care, and Urinary Care, are most clearly defined. The low recall for Aging Care, Musculoskeletal and Joint Care, and Urgent Care indicates vague wording and a lack of unique terms, which reduces the sensitivity of the models.

The TF-IDF and E5 approaches demonstrate different specifics. E5 is more accurate due to context consideration, while TF-IDF achieves greater completeness by responding to terminological matches and capturing more relevant prescriptions. Therefore, TF-IDF is preferable for tasks requiring maximum data coverage, and E5 is preferable when critical precision and the need to minimize false annotations are important. At the same time, TF-IDF remains dependent on manual keyword selection.

T-SNE visualization confirms that the recipes have a distinct but not completely isolated cluster structure: many feeds use similar ingredients, and differences between categories may be determined by micronutrient supplements or their proportions. More homogeneous clusters are observed in Puppy Care, Kidney Care, and Weight Management, while Adult, Aging, Digestive, and Skin Health are characterized by greater variability.

The ingredient recommendation process is primarily governed by the user-selected dietary specialization, which defines the target nutrient profile and corresponding constraints within the linear programming framework. Energy requirements are computed based on individual dog characteristics, including age, body weight, activity level, and reproductive status. These parameters affect the total caloric demand but do not alter the structural composition of the optimized recipe. As a result, moderate changes in input parameters lead to proportional adjustments in feeding quantity rather than abrupt shifts in ingredient composition, indicating a structural stability of the optimization model.

Despite these advantages, several limitations should be acknowledged. First, the quality and stability of the ingredient recommendations depend on the size and diversity of the labeled specialized feed dataset. In this study, the number of available specialized products

was relatively limited, which may constrain the generalizability of the learned ingredient associations. Nevertheless, the use of TF-IDF representations combined with truncated singular value decomposition (SVD) helped reduce feature sparsity and dimensionality, thereby mitigating overfitting risks. In addition, ingredient ranking based on linear Ridge classifiers further improved robustness due to regularization. Importantly, the proposed framework is scalable, and its performance is expected to improve as more annotated data become available.

Other limitations are associated with the fact that the recipes were formed based on a database that includes mainly conventional food ingredients. This excludes specialized components—hydrolysates, substrates, functional additives—and makes it difficult to model categories where they play a key role. The absence of substitutes for such ingredients leads to an incomplete reproduction of professional recipes.

These findings illustrate a fundamental trade-off between semantic coherence and statistical specificity. The machine learning-based method integrates contextual cues from product descriptions, which improves the interpretability and consistency of ingredient sets but also increases overlap between categories. Statistical association measures, on the other hand, isolate category-distinctive ingredients more aggressively, yet their outputs are more sensitive to data sparsity and category imbalance. This contrast suggests that the two approaches capture different dimensions of the underlying structure: one emphasizes semantic proximity, while the other reflects distributional distinctiveness.

The cosine similarity between the methods shows the degree of their consistency. High values for Liver Care, Dental Care, Puppy Care, Digestive Care, and Cancer Patient Support indicate virtually identical results from two independent methods. Such recipes can be considered more reliable: the coincidence of approaches reflects clearly defined category characteristics, a stable ingredient profile, and consistency between text data and actual formulas.

Low values indicate significant discrepancies due to ambiguous descriptions, high ingredient variability, or weak category formalization. In such cases, it is necessary to refine the criteria or expand the data.

An additional limitation is the lack of information on the ratio of ingredients in commercial feeds. The methods identify characteristic components but do not consider proportions, which means that the simulated recipes may deviate from production standards or veterinary requirements. In addition, the simulated recipes do not include micronutrients and technological additives that are important for the balance of industrial products.

The combined results show that the effectiveness of automated annotation and recipe generation depends on the structural homogeneity of the category, the quality of text descriptions, and the severity of nutrient restrictions. The use of two complementary methods allows for a deeper understanding of the nature of recipe variability and the identification of categories where additional formalization or improvement of descriptions is required. This combination of approaches provides a more reliable basis for the development of automated feed analysis systems and the subsequent optimization of recipes, considering the functional needs of animals.

6. Conclusions

A system for the automated creation of specialized diets for dogs was developed and tested, combining methods of text analytics, statistics, and linear programming. The integration of diverse data sources and the use of TF-IDF and the E5 model showed that text descriptions of feed contain sufficient information to reliably identify the functional focus of recipes and fill in missing data.

A comparative analysis of ingredient extraction methods revealed their complementary characteristics: machine learning approaches provide more stable and consistent results, while statistical methods capture greater variability in the generated recipes. The combined use of these approaches improves the robustness of key ingredient identification and helps highlight ingredient categories that require more precise descriptions or expanded datasets.

Current limitations related to the absence of explicit ingredient proportions, incomplete micronutrient information, and the heterogeneity of measurement units constrain the accuracy of nutritional optimization. Future research will focus on extending the model to incorporate micronutrients (vitamins and minerals) and technological additives, which are essential for ensuring complete nutritional balance in industrial dog feeds. Addressing these aspects will require the integration of more detailed compositional datasets and the development of normalization strategies for heterogeneous nutritional units.

Thus, the proposed system can serve as a basis for creating a more complete and more accurate platform for supporting veterinary decisions, providing automated analysis of recipes, formation of individual diets, and subsequent optimization of compositions considering the needs of animals.

Author Contributions: Conceptualization, A.K. and K.B.; methodology, A.K. and K.B.; software, A.K. and K.B.; validation, A.K., K.B., A.S. and K.M.; formal analysis, A.K. and K.B.; investigation, A.K. and A.S.; resources, K.M. and G.T.; data curation, A.S.; writing—original draft preparation, A.K. and K.B.; writing—review and editing, A.K., K.B. and G.T.; visualization, A.K. and K.B.; supervision, A.S.; project administration, A.S.; funding acquisition, A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Science and Higher Education of the Republic of Kazakhstan, grant number AP23488904.

Data Availability Statement: The original data presented in the study are openly available in FigShare at <https://doi.org/10.6084/m9.figshare.30344017> (accessed on 13 October 2025).

Acknowledgments: The authors express their sincere gratitude to all colleagues and research staff from Astana IT University and S. Seifullin Kazakh Agrotechnical Research University for their valuable assistance and constructive feedback throughout this study. The authors also acknowledge the administrative and technical support provided by collaborators, which greatly contributed to the successful completion of this work.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

EPA	Eicosatetraenoic acid
DHA	Docosahexaenoic acid
NRA	Nutrient-based recommendation algorithm
KNN	k-Nearest Neighbors
AUC	Area Under the Curve
SVD	Singular Value Decomposition
DB	Database
NRC	National Research Council
LP	Linear Programming
MINLP_DGO	Mixed-Integer Nonlinear Programming with Deterministic Global Optimization
IteLP	Iterative Linear Programming

SQP	Sequential Quadratic Programming
MOBO	Multi-Objective Bayesian Optimization
MORBO	Multi-Objective Regionalized Bayesian Optimization
ANN	Artificial Neural Network
TF-IDF	Term Frequency–Inverse Document Frequency
GNN	Graph Neural Network
FKGM	Food Knowledge Graph Model
BiGRU	Bidirectional Gated Recurrent Unit
CFKG	Collaborative Filtering with Knowledge Graph
KGAT	Knowledge Graph Attention Network
BPRMF	Bayesian Personalized Ranking Matrix Factorization
LDA	Latent Dirichlet Allocation
HAFR	Hierarchical Attention-based Food Recommender
FGCN	Food Graph Convolutional Network
USDA	United States Department of Agriculture
NGDC	National Genomics Data Center
CIDD	Canine Inherited Disorders Database
FEDIAF	Fédération Européenne de l’Industrie des Aliments pour Animaux Familiers (European Pet Food Industry Federation)
RDA	Recommended Daily Allowance
ME	Metabolizable Energy
ML	Machine Learning
LLR	Log-Likelihood Ratio
PMI	Pointwise Mutual Information
t-SNE	t-Distributed Stochastic Neighbor Embedding

References

1. FEDIAF. *Facts & Figures Based on Aggregated Data from 2023*; European Pet Food Industry Federation: Brussels, Belgium, 2025; Available online: <https://europeanpetfood.org/about/statistics/> (accessed on 10 September 2025).
2. Tapsell, L.C.; Neale, E.P.; Satija, A.; Hu, F.B. Foods, nutrients, and dietary patterns: Interconnections and implications for dietary guidelines. *Adv. Nutr.* **2016**, *7*, 445–454. [\[CrossRef\]](#)
3. O’Brien, J.S.; Tolbert, M.K.; Ruple, A. Dog and owner demographics impact dietary choices in Dog Aging Project cohort. *J. Am. Vet. Med. Assoc.* **2024**, *262*, 1676–1685. [\[CrossRef\]](#)
4. Hoummady, S.; Fantinati, M.; Maso, D.; Bynens, A.; Banuls, D.; Santos, N.R.; Roche, M.; Priymenko, N. Comparison of canine owner profile according to food choice: An online preliminary survey in France. *BMC Vet. Res.* **2022**, *18*, 163. [\[CrossRef\]](#)
5. Ionică, C.; Daina, S.; Pop, R.; Macri, A. Home-prepared dog food: Benefits and downsides. *Front. Anim. Sci.* **2025**, *6*, 1506003. [\[CrossRef\]](#)
6. Morelli, G.; Bastianello, S.; Catellani, P.; Ricci, R. Raw meat-based diets for dogs: Survey of owners’ motivations, attitudes and practices. *BMC Vet. Res.* **2019**, *15*, 74. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Pomar, C.; Andretta, I.; Remus, A. Feeding strategies to reduce nutrient losses and improve the sustainability of growing pigs. *Front. Vet. Sci.* **2021**, *8*, 742220. [\[CrossRef\]](#)
8. Akintan, O.; Gebremedhin, K.G.; Uyeh, D.D. Animal feed formulation—Connecting technologies to build a resilient and sustainable system. *Animals* **2024**, *14*, 1497. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Kumar, R.; Sharma, A. Machine learning in pet food: A comprehensive review of applications, challenges, and future directions. *J. Almaty Technol. Univ.* **2025**, *147*, 55–63. [\[CrossRef\]](#)
10. National Genomics Data Center. Dog Breed Information. NGDC. Available online: <https://ngdc.cnbc.ac.cn/idog/> (accessed on 24 September 2025).
11. Canine Inherited Disorders Database (CIDD). Disorder Types. CIDD. Available online: <https://cidd.discoveryspace.ca/disorder-types.html> (accessed on 24 September 2025).
12. The Kennel Club. Breeds. Available online: <https://www.thekennelclub.org.uk/search/breeds-a-to-z/> (accessed on 24 September 2025).
13. Hill’s Pet Nutrition. Dog Food. Available online: <https://www.hillspet.co.uk/products/dog-food> (accessed on 24 September 2025).
14. Aguilar, F.M.R.; Montoya, I.A.A. Development of a cost-optimal nutritionally balanced homemade canine diet through linear programming. *LACCEI* **2023**, *1*, 6–7. [\[CrossRef\]](#)

15. Joban, H.A.O.; Şahman, M.A.; İnal, F. Cost optimization of homemade diet for dogs. *Int. J. Appl. Math. Electron. Comput.* **2020**, *8*, 236–240. [CrossRef]
16. Pavlova, K.; Trichkova-Kashamova, E.; Dimitrov, S. Applying a mathematical model for calculating the ideal nutrition for sheep. *Mathematics* **2024**, *12*, 1270. [CrossRef]
17. Metwally, M.M. Broiler chicken feeds cost optimization using linear programming technique under Egyptian conditions. *Acta Sci. Vet. Sci.* **2023**, *5*, 53–63. [CrossRef]
18. Alqaisi, O.; Schlecht, E. Feeding models to optimize dairy feed rations in view of feed availability, feed prices and milk production scenarios. *Sustainability* **2020**, *13*, 215. [CrossRef]
19. Li, J.; Kebreab, E.; You, F.; Fadel, J.; Hansen, T.; VanKerkhove, C.; Reed, K. The application of nonlinear programming on ration formulation for dairy cattle. *J. Dairy Sci. Dairy* **2022**, *105*, 2180–2189. [CrossRef]
20. Uribe-Guerra, G.D.; Múnera-Ramírez, D.A.; Arias-Londoño, J.D. Feed formulation using multi-objective Bayesian optimization. *Comput. Electron. Agric.* **2024**, *224*, 109173. [CrossRef]
21. Uribe-Guerra, G.D.; Múnera-Ramírez, D.A.; Arias-Londoño, J.D. Swine diet design using multi-objective regionalized Bayesian optimization. *arXiv* **2024**, arXiv:2409.12919. [CrossRef]
22. Song, H.S.; Kim, Y.A. A dog food recommendation system based on nutrient suitability. *Expert Syst.* **2020**, *38*, e12623. [CrossRef]
23. Putra, N.A.S.; Widyaningsih, N.P.; Sumarlinda, N.S. Animal nutrition selection recommendation system with content-based filtering method. *J. Inotera* **2024**, *9*, 363–369. [CrossRef]
24. Li, K.; Jiang, Z.; Wang, H.; Liu, X. Healthy diet recommendation via food-nutrition-recipe graph mining. *Proc. Assoc. Inf. Sci. Technol.* **2020**, *57*, e319. [CrossRef]
25. Ma, W.; Li, M.; Dai, J.; Ding, J.; Chu, Z.; Chen, H. Nutrition-related knowledge graph neural network for food recommendation. *Foods* **2024**, *13*, 2144. [CrossRef]
26. Chen, Y.; Guo, Y.; Fan, Q.; Zhang, Q.; Dong, Y. Health-aware food recommendation based on knowledge graph and multi-task learning. *Foods* **2023**, *12*, 2079. [CrossRef]
27. Rostami, M.; Oussalah, M.; Farrahi, V. A novel time-aware food recommender system based on deep learning and graph clustering. *IEEE Access* **2022**, *10*, 52508–52524. [CrossRef]
28. Song, Y.; Yang, X.; Xu, C. Self-supervised calorie-aware heterogeneous graph networks for food recommendation. *ACM Trans. Multimed. Comput. Commun. Appl.* **2022**, *19*, 1–23. [CrossRef]
29. Majumder, B.P.; Li, S.; Ni, J.; McAuley, J. Generating personalized recipes from historical user preferences. *arXiv* **2019**, arXiv:1909.00105. [CrossRef]
30. Sadovoy, V.V.; Voblikova, T.V.; Permyakov, A.V.; Shchedrina, T.V.; Morgunova, A.V.; Stolyarova, V.V. Modeling the recipe composition of food products and recommendations for their use in individual nutrition. *IOP Conf. Ser. Earth Environ. Sci.* **2020**, *613*, 012127. [CrossRef]
31. Saxena, P.; Parasher, Y. Application of artificial neural network (ANN) for animal diet formulation modeling. *Procedia Comput. Sci.* **2019**, *152*, 261–266. [CrossRef]
32. Food Nutrition Dataset. Kaggle. 2021. Available online: <https://www.kaggle.com/datasets/shrutisaxena/food-nutrition-dataset/data> (accessed on 11 October 2025).
33. U.S. Department of Agriculture. FoodData Central. Available online: <https://fdc.nal.usda.gov/> (accessed on 24 September 2025).
34. Kuanysh, B. Canine Diet Datasets. Figshare. Dataset. 2025. Available online: https://figshare.com/articles/dataset/Canine_diet_datasets/30344017/1 (accessed on 13 October 2025).
35. FEDIAF. *Nutritional Guidelines for Complete and Complementary Pet Food for Cats and Dogs*; FEDIAF: Brussels, Belgium, 2024; Available online: https://europeanpetfood.org/wp-content/uploads/2024/09/FEDIAF-Nutritional-Guidelines_2024.pdf (accessed on 24 September 2025).
36. Wang, L.; Yang, N.; Huang, X.; Yang, L.; Majumder, R.; Wei, F. Multilingual E5 text embeddings: A technical report. *arXiv* **2024**, arXiv:2402.05672. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.