

Journal of Intelligent Decision Making and Information Science

www.jidmis.org
eISSN: 3079-0875



AIMRS: An Explainable Optimized Multi-Output Machine Learning Framework for Personalized Ayurvedic Formulation, Herb, Diet, and Yoga Recommendation

Krutika Lohiya¹, Anuradha Thakare¹, Santwana Gudadhe¹, Bal Virdee², Ashish Khanna³, Pratham Bokefode¹, Isha Thakur¹, Shamal Dighole¹, Lokesh Ajmera¹

¹ Department of Computer Science & Engineering (AI & ML), Pimpri Chinchwad College of Engineering, Savitribai Phule Pune University, Pune, India.

² Department of Communications Technology Research Centre, LMU.

³ Department of Computer Science and Engineering, Maharaja Agrasen Institute of Technology, New Delhi, India.

Emails: Krutika Lohiya: krutikabang@gmail.com, Anuradha Thakare: anuradha.thakare@pccoepune.org, Santwana Gudadhe:

s.santwana20@gmail.com, Bal Virdee: b.virdee@londonmet.ac.uk, Ashish Khanna: ashishkhanna@mait.ac.in, Pratham Bokefode:

pratham.bokefode@gmail.com, Isha Thakur: isha12906@gmail.com, Shamal Dighole: digholeshamal@gmail.com, Lokesh Ajmera:

lokeshajmera64@gmail.com

ARTICLE INFO

Article history:

Received 12 June 2026

Received in revised form 12 July 2026

Accepted 24 July 2026

ABSTRACT

Ayurveda is a comprehensive personalized healthcare approach where treatment recommendations are made not only based on disease diagnosis but also based on numerous individual-specific factors like symptoms, dosha, prakriti, diet, life style, sleep habit, stress, environmental exposure, medical history, and overall body condition of the patient. The traditional approach in Ayurveda depends largely on expert knowledge and manual analysis of interrelated factors, thus resulting in an expensive, subjective, and hard-to-standardize decision-making process. While there have been some efforts in recent times using advancements in artificial intelligence in disease prediction, prakriti classification, and recognition of medicinal plants in the literature, most of these efforts consider only one particular aspect of personalized Ayurvedic recommendation and are oblivious of the multi-dimensional nature of the problem. To bridge this gap, the Ayurvedic Intelligent Multi-output Recommendation System (AIMRS), a comprehensible and optimized machine learning model for personalized Ayurvedic recommendation support. The suggested framework predicts nine recommendation targets: Ayurvedic Herbs, Formulation, Diet and Lifestyle Recommendations, Yoga and Physical Therapy, Medical Intervention, Prevention, Prognosis, Complications, and Patient Recommendations, based on the Kaggle dataset with 446 preprocessed patient records and 19 input features and 9 target outputs. In order to address issues related to sparseness, imbalance, and length of the text-based recommendation labels, the framework implements label normalization, rare class processing, semantic clustering, target-wise modeling selection, hyperparameter tuning, Top-K recommendations, ablation analysis, and explainability. The experimental evaluation of the proposed model shows that it achieves the maximum Weighted F1-score of 0.7350 on the Medical Intervention target and significant improvements on other recommendation targets following the target-space optimization. The obtained results prove that the framework of optimized target-wise modeling is efficient and explainable enough for Ayurvedic multi-output recommendations. The suggested framework is

Keywords: Ayurveda; AI; multi-output learning; Machine Learning; Ayurvedic Formulation; Top-K recommendation; clinical decision support

aimed at academic research and practitioner-assisted decision making and is not intended for medical diagnoses or Ayurvedic treatment.

1. Introduction

Ayurveda is one of the oldest personalized healthcare systems, which does not rely on just a specific disease or symptom for treatment suggestions. There is consideration of several aspects about the patient like the patient's dosha, prakriti, diet, lifestyle, sleeping habit, stress, seasonal changes, environment, past history of the patient, risks factors and health condition of the patient before suggesting any herb, formulation, yoga and so on. It indicates that Ayurveda is based on a multidimensional decision making process not on prediction of a single label.

Due to the increasing access to the data and advanced techniques of machine learning, artificial intelligence is becoming increasingly popular in medical decision support systems. As far as Ayurveda is concerned, the recent literature has studied many topics including disease prediction, prakriti and dosha classification, nadi-pariksha, digital wellness platform, medicinal plants identification and personal healthcare based on ayurgenomics approach. These pieces of work indicate that AI technologies could help traditional medical systems to identify certain patterns in structured and unstructured data. Nevertheless, all these systems are aimed at solving a particular problem including prediction of diseases, classification of prakriti type, identification of medicinal plants from image data.

Unlike in the case of actual Ayurveda, recommending something does not stop at just producing one output. For instance, a patient description will require a number of relevant recommendations, ranging from herbs and formulations to food and lifestyle recommendations, yoga and physical therapy, interventions, prevention measures, prognosis, possible complications, and more. This means that the task of modeling Ayurveda recommendations will be impossible with just formulation-based classification or disease-based predictions. There is therefore a need for machine learning methods capable of modeling all of those aspects together in a meaningful manner while being interpretable and appropriate for practitioner review.

The other major problem in Ayurveda recommendation data is that the recommendations themselves tend to be long and sparse, and take different textual formats despite meaning the same thing. The result is a very large number of labels, most of which appear only once or twice, and thus suffer from very poor class imbalance and high target-cardinality. Under such circumstances, multiclass classification becomes an instable approach with misleading accuracy figures.

This paper proposed a novel explainable and optimized multi-output machine learning framework for personalized Ayurvedic recommendation assistance (AIMRS). This framework exploits clinical, demographic, lifestyle, environmental, dosha, and prakriti related features to predict nine recommendation targets: Ayurvedic Herbs, Formulation, Diet and Lifestyle Recommendations, Yoga and Physical Therapy, Medical Intervention, Prevention, Prognosis, Complications, and Patient Recommendations. The proposed framework employs TF-IDF based text representation, One Hot Encoding, target specific machine learning algorithms, label optimization, hyper parameter tuning,

Top-K Recommendation, ablation studies and feature importance based explainability.

Major contributions of this paper are as follows:

Multi-output machine learning framework for personalized Ayurvedic recommendation is developed instead of single output formulation/disease prediction

Kaggle based AyurGenixAI dataset comprising 446 cleaned records, 19 input/profile features, and 9 recommendation targets is used for academic ML experiments.

Label normalization, rare class handling, and conservative semantic groupings are performed to handle sparse and long text target labels.

Target specific models are developed and evaluated using imbalance aware metrics like Weighted F1, Macro F1, Balanced Accuracy, and Top-3 Accuracy.

Raw label and optimized label performances are compared to study the impact of target space optimization.

Ablation analysis is performed to understand the contribution of clinical, lifestyle, and Ayurveda-specific feature groups.

Explainability analysis is included to identify model-learned feature importance for practitioner-reviewable decision support.

The remainder of this paper is organized as follows. Section II provides an overview of related work in AI for Ayurveda, classification of prakriti and doshas, identification of medicinal plants, multi-output learning, imbalanced classification, and explainable healthcare AI. Section III states the research problems and contributions of this work. Section IV provides an overview of the data used in our experiments and exploratory data analysis. Section V explains the methodology proposed, which includes data preprocessing, label optimization, feature engineering, target-specific modeling, Top-K recommendation, and explainability. Section VI covers the experimental setting and reproducibility information. Section VII provides the definitions of the evaluation metrics used in our experiments.

2. Related Work

Recent research shows growing interest in applying artificial intelligence to Ayurveda, especially in areas such as diagnostic assistance, wellness platforms, prakriti analysis, nadi-pariksha modernization, and Ayurvedic drug-discovery support [1], [2]. These studies show that computational methods can support traditional healthcare systems, but they also highlight common challenges such as limited dataset size, lack of clinical validation, and the absence of standardized Ayurvedic data formats.

Personalization is a central idea in Ayurveda, and earlier ayurgenomics work has connected prakriti-based stratification with personalized care and drug discovery [3]. Machine learning studies have also attempted to classify prakriti and dosha types using supervised and clustering-based methods [4]. However, most of these works focus on identifying constitution or diagnostic categories rather than generating a complete recommendation profile.

Another related direction is medicinal plant and herb recognition, where deep learning models are used to identify medicinal leaves or plants from images [5], [6]. While useful for herb identification, these systems do not address patient-specific recommendations across herbs, formulation, diet, yoga therapy, prevention, prognosis, complications, and patient guidance.

From the machine learning perspective, this work also connects with multi-output learning, imbalanced classification, and explainable healthcare AI. Multi-output learning is suitable when several recommendation targets are associated with one patient profile [7], while imbalanced learning explains why accuracy alone is unreliable when many classes have very low support [8]. Explainability methods such as LIME and SHAP support model auditing and interpretation [9], [10]. In healthcare decision-support, such transparency is important because AI-generated outputs must remain understandable, reviewable, and supervised by human experts [11], [12].

Table 1 presents a comparison between AIMRS and previous AI-assisted studies in Ayurveda. The existing literature primarily focuses on specific tasks, such as Prakriti or Dosha classification, identification of medicinal plants, well-being platforms, and general explainable health care AI. Conversely, AIMRS integrates personalized recommendation for nine different targets along with target-specific modeling, label optimization, Top-K recommendations, ablation analysis, and explainability. The primary drawback mentioned in Table 1 is that AIMRS still uses the research dataset and lacks clinical validation.

Table 1. Comparison with related AI-in-Ayurveda research.

Study/Area	Main Focus	Method/Approach	Limitation	Difference in AIMRS
AI in Ayurveda reviews [1]	Landscape of AI applications	Review/synthesis	Broad, not a reproducible model	Implements and evaluates a full pipeline
Prakriti/Dosha classification [3,4]	Constitution prediction	ML classifiers/clustering	Single-output diagnostic framing	Uses dosha/prakriti as input features for nine targets
Medicinal plant recognition [5,6]	Plant/herb image identification	CNN/transfer learning	Not patient-specific	Predicts patient-profile recommendation outputs
Digital Ayurveda wellness systems [2]	Conceptual decision support	Platform architecture	Limited target-level evaluation	Reports target-wise metrics and ablation
Explainable AI in healthcare [9 - 12]	Interpretability and trust	LIME/SHAP/feature attribution	Domain-general	Applies explainability to Ayurvedic recommendation
AIMRS (Proposed Work)	Personalized Ayurvedic recommendation	Optimized target-specific ML	Research dataset; no clinical validation	Combines optimization, Top-K, ablation, and explainability

3. Research Questions and Contributions

3.1 Research Questions

RQ1. Can structured patient-profile features support prediction of multiple Ayurvedic recommendation categories?

RQ2. Does label optimization and rare-class handling improve learnability under severe target sparsity?

RQ3. Which recommendation targets benefit most from optimized target-specific modelling?

RQ4. Which clinical, lifestyle, and Ayurveda-specific feature groups contribute most to different recommendation outputs?

RQ5. Can explainability analysis provide interpretable model-level insights for practitioner review?

RQ1 and RQ3 are addressed in the final optimized results. RQ2 is addressed by the before-vs-after optimization comparison. RQ4 is examined through ablation. RQ5 is addressed through feature-importance analysis.

3.2 Contributions

The main contribution is an optimized, explainable, target-specific ML pipeline for personalized Ayurvedic recommendation. It brings together patient-profile feature engineering, label normalization, rare-class handling, semantic grouping, imbalance-aware benchmarking, hyperparameter tuning, Top-K recommendation, ablation analysis, and explainability all while staying within a cautious decision-support framing rather than an autonomous one.

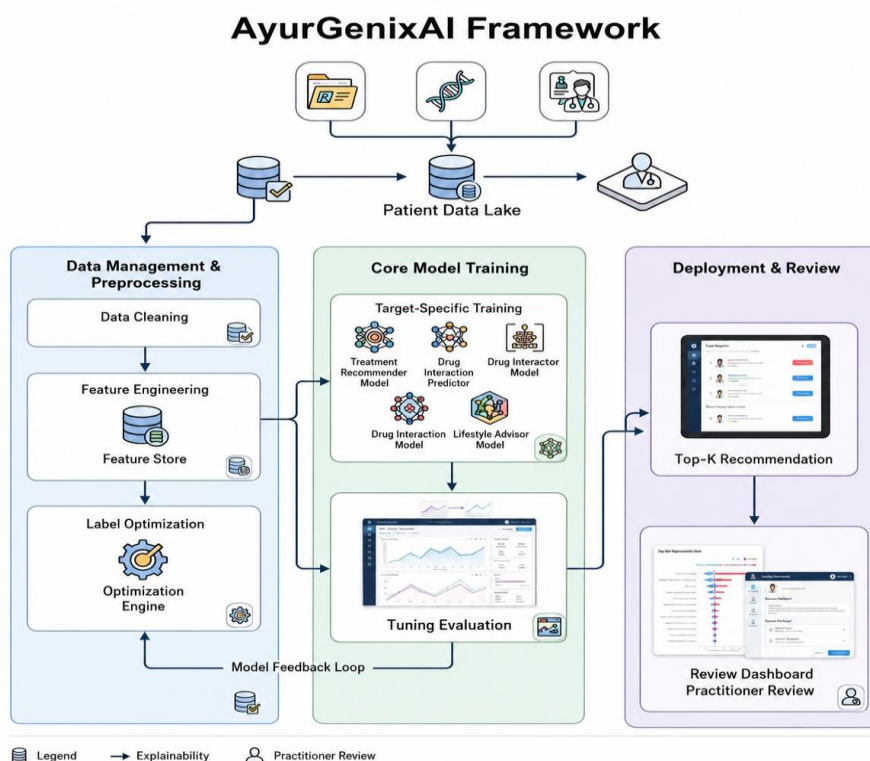


Figure 1. Proposed AIMRS framework for practitioner-reviewable recommendation generation.

Figure 1 shows the overall structure of the proposed Ayurvedic Intelligent Multi-output Recommendation System (AIMRS). The system starts with attributes such as demographic, clinical, life style, environmental, dosha and prakriti information about the patient's profile, and then proceeds to preprocessing and feature engineering. After that, independent machine learning models are trained for each recommendation type, which is done followed by Ayurvedic recommendations generation for the patient. The modular structure of the system allows us to optimize every target individually.

4. Dataset Description and Exploratory Analysis

After cleaning, the dataset came out to 446 records across 34 columns. Of these, 19 are patient/profile features that feed the model, while the remaining 9 serve as recommendation targets. Figure 2 gives a snapshot of the clinical, lifestyle, and Ayurveda-specific attributes that make up the profile side of the data.

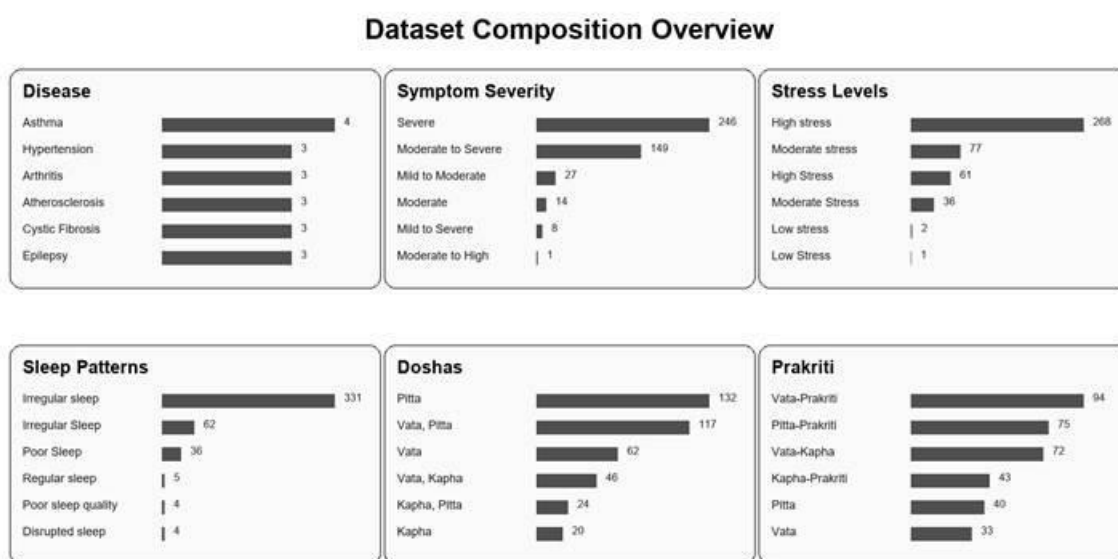
The AyurGenixAI dataset collected from Kaggle [13] and treated it purely as an academic research dataset a way to test whether explainable, multi-output Ayurvedic recommendation is even feasible, not as a clinically validated source of patient data. It includes patient-profile attributes alongside recommendation-oriented targets, but readers shouldn't mistake it for verified clinical records. Our focus here stays on the machine learning side: methodology, optimizing the target space, and building decision-support models that can explain themselves. If someone wanted to take this toward real clinical use, that would require validation from qualified Ayurvedic practitioners, testing against external datasets, and a proper ethics review. No patient-identifiable information appears anywhere in this paper.

Table 2 categorizes the patient profiles into six major categories: clinical parameters, patient's history and risks, lifestyle, environment, demographics, and Ayurveda-specific parameters. They are as follows: disease, symptoms, severity, medications, stress, sleep, diet, occupation, Dosha, and Prakriti. It is shown in Table 2 that AIMRS treats Ayurveda recommendation as a multi-dimensional task and not only based on disease or symptom. This wide range of features helps the framework study the impact of various features of patient profile on different recommendation targets.

Table 2. Input feature groups used in AIMRS.

Feature Group	Features	Role
Clinical	Disease; Symptoms; Diagnosis & Tests; Symptom Severity; Duration	Disease state and symptom profile
Patient History/Risk	Medical History; Current Medications; Risk Factors	Background and risk conditions

Lifestyle	Sleep Patterns; Stress Levels; Physical Activity; Diet; Lifestyle	Behavioral and routine factors
Environmental	Environmental Factors; Seasonal Variation	External exposure and seasonal influence
Demographic	Age Group; Gender; Occupation	Population context
Ayurveda-specific	Doshas; Prakriti	Constitutional attributes



Clinical, lifestyle, and Ayurveda-specific patient-profile attributes used by AyurGenixAI.

Figure 2. Dataset composition overview showing clinical, lifestyle, and Ayurveda-specific patient-profile attributes used by AIMRS.

Figure 2 depicts the dataset composition of AyurGenixAI after the preprocessing phase. This figure indicates the distribution of different patients' characteristics like patient demographics, disease category, dosha, prakriti, lifestyle characteristics, and recommendation target. This figure shows that this dataset is heterogeneous and has enough variety to evaluate personalized recommendations model. The other point depicted by the figure is that the natural distribution of some features is imbalanced; hence, there is a need for imbalance-aware learning.

5. Proposed Methodology

The introduced AIMRS framework converts patient-specific clinical, demographic, lifestyle, environmental, dosha, and prakriti data into personalized Ayurvedic recommendations using an efficient multi-output machine learning pipeline. The approach utilizes the steps of data

preprocessing, label optimization, model selection, hyperparameter optimization, Top-K recommendation, ablation study, and explanation in order to tackle the problem of sparse and imbalanced recommendation labels.

5.1 Data Preprocessing and Target Leakage Prevention

Before modeling, we normalized column names, checked for duplicate records, standardized missing text values as "Unknown," and cleaned up text fields generally. Recommendation target columns were kept out of the input features entirely no leakage there. Text vectorizers and encoders were fit only inside the modeling pipeline itself, raw-label and optimized-label experiments were stored separately, target-specific models were trained independently of one another, and feature importance was pulled after model selection rather than used to tune performance on the test set.

5.2 Label Optimization Procedure

Table 3. Label Optimization Procedure.

Step	Operation	Reproducibility Detail
1. Text normalization	Lowercase conversion, whitespace normalization, punctuation cleanup, and standardization of repeated phrases	Medically meaningful terms preserved
2. Raw class frequency analysis	Count labels for each target	Classes with fewer than 3 records treated as rare
3. Rare-class identification	Low-support classes identified using threshold < 3	Sparse classes flagged before modeling
4. Semantic grouping	Long-text targets grouped keyword-wise into broader interpretable categories	Category meaning preserved where possible
5. Formulation preservation	Formulation labels handled cautiously	Herb combinations, dosages, and carriers can carry meaning
6. Model comparison	Raw-label and optimized-label experiments compared	Improvements interpreted as target-space normalization

Optimized-label performance measures prediction of normalized recommendation categories, not exact reproduction of the original long text recommendation strings. This step was used to reduce target-space sparsity and improve learnability. It should not be interpreted as improving clinical correctness of the recommendations.

Table 3 illustrates the way this framework addresses the problem of sparseness and high number of recommendation labels. This procedure entails text normalization, analysis of the frequency of class occurrence, finding classes which occur less than 3 times, grouping of semantically similar long-text labels, preserving the formulation information, and comparing the raw and optimized label experiments. The purpose of this procedure is to minimize class fragmentation and make the

recommendation targets learnable. However, it should be noted that enhanced performance after optimization is a sign of improved predictions for normalized recommendation classes and should not serve as evidence of improved clinical correctness.

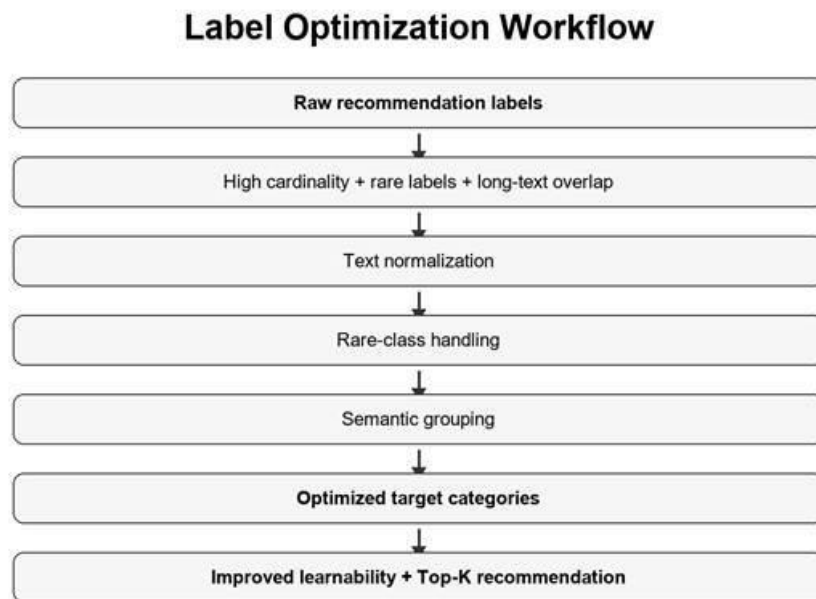


Figure 3. Label optimization workflow used to reduce target sparsity while preserving broad recommendation meaning.

As can be seen from Figure 3, the pipeline of label optimization is performed before the model training stage. The initial set of recommendation labels is first normalized and duplicates are eliminated; then, conservative semantic clusters and rare classes are processed. The resulting optimized labels are used to train target-specific models. Label sparsity is decreased without affecting the semantics of recommendations.

5.3 Feature Engineering and Modeling

For text-heavy fields we used TF-IDF, and categorical fields went through one-hot encoding [14,15]. Each of the nine targets got its own label encoder and, more importantly, its own model because the targets differ so much in class count, imbalance, and semantic granularity that forcing them into a single shared model didn't make sense. We ran the raw-label and optimized-label experiments as two separate tracks. Model selection came down to Weighted F1 and Macro F1 rather than plain accuracy, since accuracy alone would hide how badly the models handled rare classes. We also reported Top-3 accuracy to reflect how useful a recommendation list would actually be in practice, and used feature importance strictly as an auditing tool not as a stand-in for causal explanation.

The models we evaluated were LinearSVC [16], Logistic Regression [17], Random Forest [18], Extra Trees [19], Multinomial Naive Bayes [20], Complement Naive Bayes, and Gradient Boosting [21] when practical. XGBoost [22], LightGBM [23], and CatBoost [24] did not make the final run because their dependencies were not available in our environment. We want to emphasize that this is a limitation of our implementation, not a conclusion about the performance of those models.

5.4 Top-K Recommendation and Explainability

We provide both Top-1 and Top-3 candidate recommendations. Top-3 accuracy is more meaningful when considered alongside how many optimized classes a target has; fewer categories can make it easier to achieve a Top-3 hit, so the number shouldn't be viewed in isolation. That said, it serves as a useful signal, as AIMRS is designed to assist with tasks that may require ingredient-level breakdowns, an Ayurvedic classification system, formulation embeddings, or recommendation methods that extend beyond standard multiclass classification.

6. Experimental Setup and Reproducibility

Experiments were executed using `train_ayurgenixai_optimized.py`. Outputs were saved under `outputs/` as CSV tables, PNG figures, Markdown reports, JSON sample output, and joblib model artifacts.

Table 4 gives an overview of the entire experimental pipeline and steps which have been taken to increase reproducibility and avoid data leakage. It includes dataset cleaning, handling of missing values, TF-IDF transformation, One-Hot Encoding, optimization of labels, train-test splitting, selecting models for the targets, no hyperparameter tuning, evaluation, and artifacts saving. One crucial aspect of this pipeline is fitting of preprocessing parts such as vectorizers and encoders in the train pipeline but not on the whole dataset. Thus, this table proves that the results were achieved by means of a reproducible experiment.

Table 4. Reproducibility summary of the experimental pipeline.

Stage	Operation	Leakage Control / Reproducibility Detail
Dataset cleaning	Duplicate removal and text normalization	Cleaned CSV saved
Missing-value handling	Standardized unknown text/categorical values	Consistent `Unknown` token
Text transformation	TF-IDF on text-heavy fields	Fitted within train pipeline/splits
Categorical encoding	OneHotEncoder	Unknown categories handled safely
Label optimization	Rare threshold < 3 and semantic grouping	Raw and optimized labels saved separately
Train/test split	Stratified when feasible	Sparse-label limitations logged
Model selection	Target-specific classifiers	Best model selected by weighted/macro F1
Hyperparameter tuning	Limited target-specific tuning	Tuning results exported
Evaluation	Accuracy, balanced accuracy, macro F1, weighted F1, Top-3	CSV results exported

Artifact saving	Figures, reports, models	Saved under `outputs/`
-----------------	--------------------------	------------------------

7. Evaluation Metrics

Precision = $TP / (TP + FP)$

Recall = $TP / (TP + FN)$

F1 = $2 \times (Precision \times Recall) / (Precision + Recall)$

Weighted F1 = $\sum_c (support_c \times F1_c) / N$

Balanced Accuracy = average recall across classes

Top-K Accuracy = samples where the true label appears in top K predictions / total samples

Because class imbalance is severe, weighted F1, macro F1, balanced accuracy, and Top-3 accuracy [25,26] are more informative than accuracy alone.

8. Results and Discussion

Table 5. Summary of target-space optimization and final model performance.

Target Output	Raw Classes	Optimized Classes	Best Model	Weighted F1	Top-3 Accuracy	Main Challenge
Ayurvedic Herbs	75	17	Complement Naive Bayes	0.6491	0.8667	Sparse herb combinations
Formulation	210	30	LinearSVC Balanced	0.3480	0.6222	Fine-grained dosage/herb labels
Diet and Lifestyle Recommendations	445	6	Random Forest Balanced Tuned	0.5583	0.9111	Near-unique long text
Yoga & Physical Therapy	141	14	Logistic Regression Balanced	0.6828	0.9556	Overlapping therapy phrases
Medical Intervention	293	7	Random Forest Balanced	0.7350	0.9889	Rare intervention labels
Prevention	291	9	Extra Trees	0.6359	0.9222	Adjacent preventive guidance
Prognosis	99	15	LinearSVC Balanced	0.5162	0.7889	Moderate imbalance

Complications	382	16	LinearSVC Balanced	0.6520	0.8889	High- cardinality complications
Patient Recommendations	305	9	Logistic Regression Balanced	0.5329	0.8889	Long-text advice overlap

The table 5 illustrates how much the number of recommendation classes was reduced by optimization and contains the best results of models for each target. For instance, the number of classes of Ayurvedic Herbs was reduced from 75 to 17, Formulation – from 210 to 30, Medical Intervention – from 293 to 7, and Complications – from 382 to 16. The highest Weighted F1-score was obtained for the Medical Intervention (0.7350), while the lowest one was obtained for the Formulation (0.3480) due to detailed descriptions of herbs, dosages, and preparation procedures contained in labels. Thus, Table 5 demonstrates how label-space optimization increases the learnability, but its difficulty highly depends on recommendation targets.

Raw classes refer to original target labels before normalization. Optimized classes indicate normalized recommendation categories after handling rare classes and grouping similar concepts. Therefore, optimized-label performance measures category-level recommendation prediction instead of precisely reproducing the original long-text recommendation strings.

Table 6. Detailed final optimized model results.

Target	Best Model	Accuracy	Balanced Accuracy	Macro F1	Weighted F1	Top-3 Accuracy	Classes
Ayurvedic Herbs	Complement Naive Bayes	0.6889	0.2761	0.2336	0.6491	0.8667	17
Formulation	LinearSVC Balanced	0.3778	0.1799	0.1455	0.3480	0.6222	30
Diet and Lifestyle Recommendati ons	Random Forest Balanced Tuned	0.5556	0.4825	0.4408	0.5583	0.9111	6
Yoga & Physical Therapy	Logistic Regression Balanced	0.7111	0.3465	0.2773	0.6828	0.9556	14
Medical Intervention	Random Forest Balanced	0.7667	0.3111	0.3278	0.7350	0.9889	7
Prevention	Extra Trees	0.6556	0.5217	0.5296	0.6359	0.9222	9
Prognosis	LinearSVC Balanced	0.5222	0.3134	0.2997	0.5162	0.7889	15

Complications	LinearSVC Balanced	0.6778	0.4423	0.3920	0.6520	0.8889	16
Patient Recommendations	Logistic Regression Balanced	0.5444	0.5685	0.4578	0.5329	0.8889	9

Table 6 presents the detailed evaluation of target-specific models for Accuracy, Balanced Accuracy, Macro F1, Weighted F1, Top-3 Accuracy, and number of classes. In this case, the best target can be considered the Medical Intervention, where accuracy is 0.7667, Weighted F1 – 0.7350, and Top-3 accuracy – 0.9889. However, the Macro F1 is very low (0.3278) in comparison with other metrics and means that performance is much higher for common classes. Such difference in values of Weighted F1 and Macro F1 can be noticed for other targets too.

Raw classes refer to original target labels before normalization. Optimized classes indicate normalized recommendation categories after handling rare classes and grouping similar concepts. Therefore, optimized-label performance measures category-level recommendation prediction instead of precisely reproducing the original long-text recommendation strings.

Medical Intervention performed the best, with a Weighted F1 of 0.7350, accuracy of 0.7667, and Top-3 accuracy of 0.9889. Yoga & Physical Therapy, Complications, Ayurvedic Herbs, and Prevention also achieved solid weighted F1 scores. Together, these results address RQ1 and RQ3.

Patient profile features can support several recommendation targets, but the effectiveness varies significantly between outputs.

While the weighted F1 improved considerably after optimization, macro F1 remained lower for several targets. In other words, the optimized models became much better at common recommendation categories but did not fully improve on the less common ones. It is more accurate to describe this framework as a reproducible baseline for common recommendation patterns a starting point, not a complete solution for rare Ayurvedic cases.

Formulation remained the most challenging of the major targets. This is largely due to formulation labels containing detailed therapeutic information such as herb combinations, dosage phrasing, and preparation methods. In its raw state, this target included 210 classes, 178 of which had fewer than three records each. Even after optimization, it still retained 30 classes, making it notably more complex than broader targets like Medical Intervention or Diet and Lifestyle Recommendations. This suggests that accurately predicting formulations might eventually require ingredient-level breakdown, an Ayurvedic ontology, formulation embeddings, or retrieval-based recommendations, rather than relying solely on standard multiclass classification.

8.1 Before-vs-After Optimization

Table 7. Before-vs-after weighted F1 comparison.

Target	Raw	Optimized	Improvement	Optimized
--------	-----	-----------	-------------	-----------

	Weighted F1	Weighted F1		Top-3
Ayurvedic Herbs	0.4654	0.6491	0.1836	0.8667
Formulation	0.1717	0.3480	0.1763	0.6222
Diet and Lifestyle Recommendations	0.0000	0.5583	0.5583	0.9111
Yoga & Physical Therapy	0.4589	0.6828	0.2239	0.9556
Medical Intervention	0.1289	0.7350	0.6061	0.9889
Prevention	0.1086	0.6359	0.5272	0.9222
Prognosis	0.4075	0.5162	0.1087	0.7889
Complications	0.1333	0.6520	0.5186	0.8889
Patient Recommendations	0.1137	0.5329	0.4192	0.8889

The direct effects of label optimization on the Weighted F1-scores are shown in Table 7. The greatest improvements have been observed for the Medical Intervention from 0.1289 to 0.7350, Diet and Lifestyle from 0.0000 to 0.5583, Prevention from 0.1086 to 0.6359, and Complications from 0.1333 to 0.6520. The reason for those improvements is related to the high level of fragmentation and sparsity in the initial targets. Thus, it is safe to assume that a key problem with the performance of raw models was target-space fragmentation and sparsity rather than just the classifier used.

The biggest improvements appeared for targets with particularly sparse raw labels: Medical Intervention, Prevention, Complications, Patient Recommendations, and Diet and Lifestyle Recommendations. This pattern supports our earlier argument that the poor performance of raw labels stemmed mainly from target-space sparsity, not just from selecting weaker models.

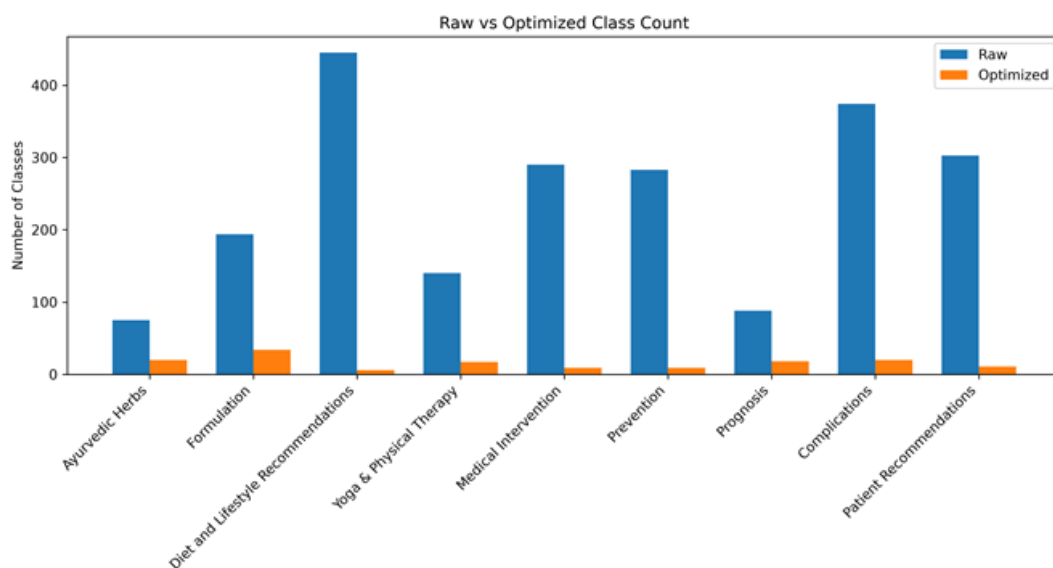


Figure 4. Raw versus optimized class count after label normalization and rare-class handling.

As seen in Figure 4, the effect of applying label optimization on the number of unique recommendation classes is evaluated. There are many sparsely occurring and semantically similar labels in the original dataset, which leads to class fragmentation and poor learnability of classes. As a result of the normalization of labels and the cautious merging of rare classes, the number of classes decreases but at the same time, their clinical meaning is retained.

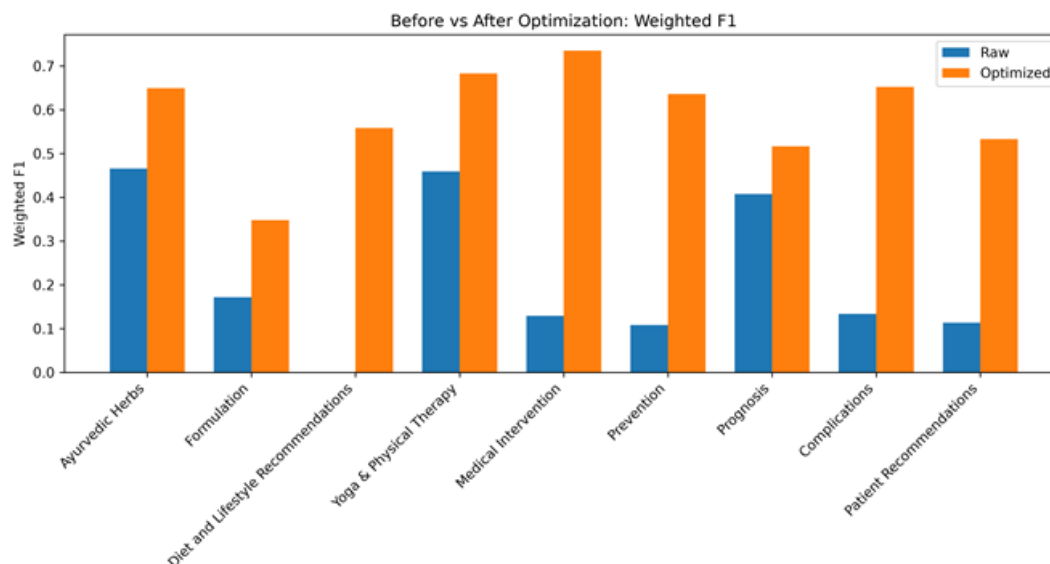


Figure 5. Before-vs-after optimization comparison for weighted F1-score.

Figure 5 shows the Weighted F1-score for each recommendation target prior to and post application of the proposed optimization pipeline. As seen from the figure, target-specific pre-processing, label optimization and hyperparameter tuning lead to an increase in predictive performance in most recommendation types. The highest gain can be achieved for Medical Intervention, and other recommendation targets like Yoga and Physical Therapy, Ayurvedic Herbs, and Prevention have made some noticeable improvements too.

9. Ablation Study

Table 8. Ablation summary by weighted F1.

Target	Best Feature Setting	Best Weighted F1	All-Features Weighted F1
Ayurvedic Herbs	E6_All features	0.6491	0.6491
Complications	E6_All features	0.6520	0.6520
Formulation	E3_Clinical Lifestyle	0.3801	0.3480
Prognosis	E6_All features	0.5162	0.5162
Yoga & Physical Therapy	E6_All features	0.6828	0.6828

Table 8 shows the effects of different types of input features on the performance of selected recommendation targets. The all-features configuration delivers the best or comparable results in terms of Weighted F1-score for Ayurvedic Herbs, Complications, Prognosis, and Yoga & Physical Therapy. Nevertheless, the combination of Clinical and Lifestyle features is more beneficial for the

Formulation target, providing a Weighted F1-score of 0.3801 as opposed to 0.3480 of all-features one. Thus, target-specific modeling is justified since different recommendation outputs require different input feature combinations.

All features were either best or competitive for Herbs, Complications, Prognosis, and Yoga & Physical Therapy, while Clinical and Lifestyle scored best for Formulation. This indicates that different recommendation targets rely on different feature groups and supports target-specific modeling. The ablation study backs up the main design choice of the framework. Ayurvedic recommendations should not be modeled as a single output, since different targets react differently to clinical, lifestyle, and Ayurveda-specific features.

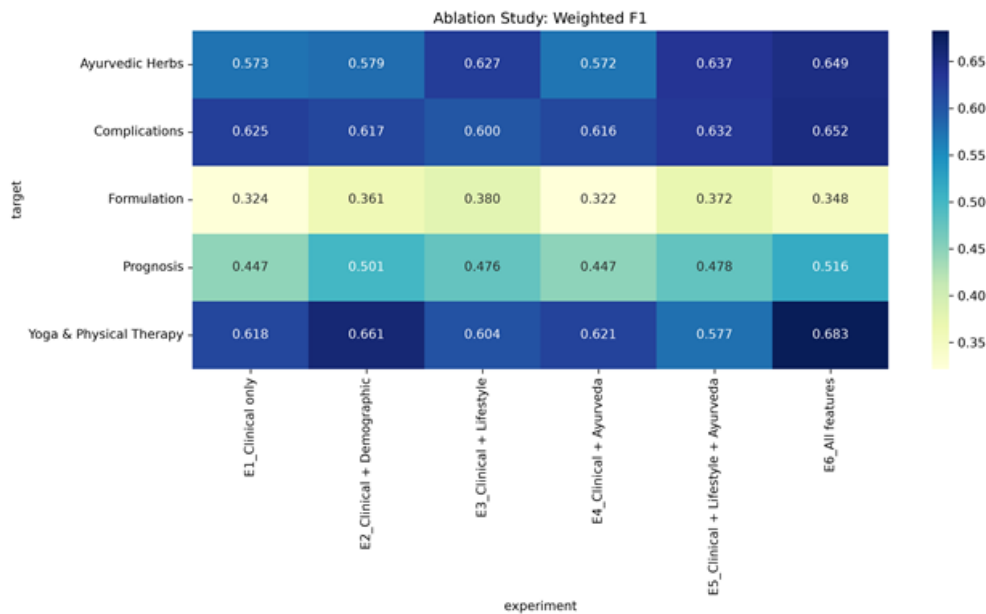


Figure 6. Ablation study weighted F1 heatmap across feature groups and selected recommendation targets.

The effects of various sets of features on model performance have been studied through ablation experiments shown in Figure 6. In each experiment, particular feature types were excluded or kept isolated to test their individual influence on prediction quality. Clinical features have proven to be the best predictors, whereas dosha, prakriti, lifestyle, and environmental features add complementary information for some of the recommendation tasks. Thus, it has been demonstrated that a personalized Ayurveda-based recommendation is more effective when combined clinical and Ayurveda-related features are used together.

10. Explainability Analysis

Table 9. Top explainability features for selected targets.

Target	Top Positive Features
Ayurvedic Herbs	Sleep Patterns (0.0501); Stress Levels (0.0488); Current Medications (0.0480); Environmental Factors (0.0444); Symptoms (0.0439)

Complications	Gender: Both genders (0.2079); Sleep Patterns: Irregular sleep (0.1316); Doshas: Pitta (0.1302); Seasonal Variation: All seasons (0.1171); Stress Levels: High Stress (0.1120)
Formulation	Gender: Both genders (0.2160); Doshas: Vata, Pitta (0.1371); Sleep Patterns: Irregular sleep (0.1300); Symptom Severity: Moderate to Severe (0.1241); Doshas: Pitta (0.1188)
Prognosis	Gender: Both genders (0.2086); Symptom Severity: Severe (0.1492); Sleep Patterns: Irregular sleep (0.1381); Stress Levels: High Stress (0.1335); Stress Levels: Moderate stress (0.1311)
Yoga & Physical Therapy	Age Group: Adults (0.3411); Stress Levels: High stress (0.2883); Sleep Patterns: Irregular sleep (0.2649); Doshas: Vata (0.2630); Age Group: All ages (0.2617)

Table 10 demonstrates how AIMRS behaves when it encounters a recommendation profile belonging to low-support or highly imbalanced classes. Instead of relying entirely on a single prediction, the system provides a Top-1 recommendation together with three ranked candidates and a prediction score. Several targets return a Rare/Expert Review Category, indicating that insufficient training evidence exists and the case should be reviewed by a practitioner rather than automatically accepted. For example, Yoga and Physical Therapy produces a strong score of 0.9476 for Anulom Vilom and Pranayama, whereas Prognosis has a much weaker Top-1 score of 0.1052, demonstrating that confidence varies significantly across outputs. The table therefore illustrates both the Top-K recommendation mechanism and the framework's intended practitioner-review safety behaviour.

These relationships should be interpreted as model-learned associations in the dataset, not as causal medical conclusions. Some high-importance categorical values such as All ages or Both genders may reflect dataset construction patterns rather than patient-specific clinical effects. Therefore, feature importance is used for model auditing and hypothesis generation only, not as causal medical evidence.

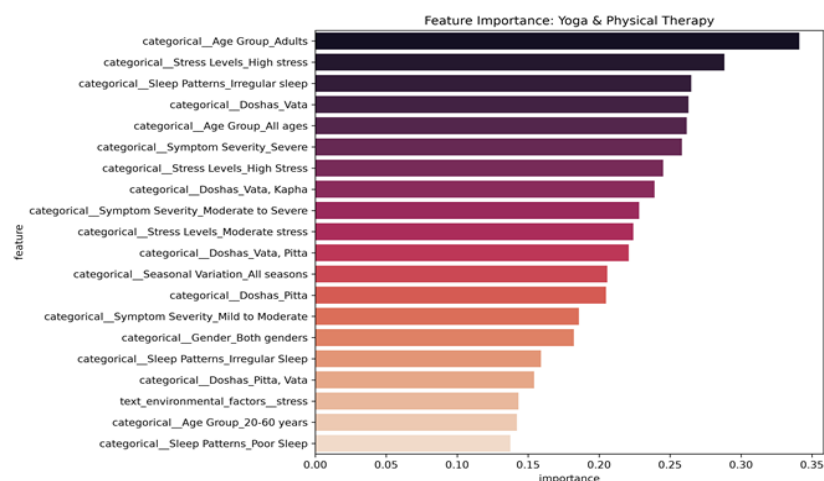


Figure 7. Feature importance for Yoga & Physical Therapy, shown as an example of model-level explainability.

Figure 7 shows the feature importance analysis of the recommendation model for yoga and physical therapy, which highlights the importance of various attributes of the patient for predicting the results. Clinical features like disease type and symptoms turn out to be the most important predictors in the model, followed by other features like lifestyle, dosha, prakriti, and environment. This analysis is a demonstration of the explainability of the proposed model.

11. Error Analysis and Sample Output

Errors arise from rare classes, high-cardinality labels, overlapping recommendations, and long-text label noise. Rare outputs have limited training evidence, while formulation labels may differ by herb combination, dosage expression, or preparation medium. Broader categories such as Medical Intervention and Prevention benefited more from label optimization.

The following example is a rare-case recommendation profile. It is intentionally retained to demonstrate the framework's safety behavior when a profile falls into low-support recommendation regions.

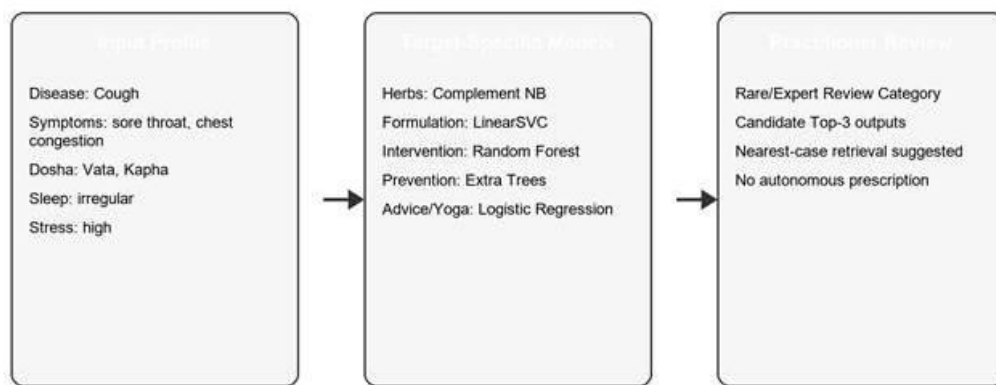
Table 10 shows the behavior of AIMRS upon encountering the recommendation profile of a target class which falls under low-support and highly imbalanced. The system offers not only a single prediction but gives Top-1 recommendation as well as three other ranked candidates along with their scores. Many targets fall under a category known as "Rare/Expert Review" where there is little evidence in the training process and needs review by an expert instead of accepting it automatically. Some examples include that of Yoga and Physical Therapy where there is a high score of 0.9476 for Anulom Vilom and Pranayama while Prognosis has a very low Top-1 score of 0.1052.

Table 10. Rare-case Top-K recommendation example.

Target	Top-1 Recommendation	Top-3 Candidates	Top-1 Score
Ayurvedic Herbs	Rare/Expert Review Category	1.Rare/Expert Review Category; 2. tulsi, ginger; 3. ashwagandha, guggul	0.9425
Formulation	ginger-honey based formulation	1.ginger-honey based formulation; 2.Rare/Expert Review Category; 3.turmeric-milk based formulation	0.1842
Diet and Lifestyle Recommendations	avoid oily spicy cold foods	1.avoid oily spicy cold foods; 2.hydration and warm fluids; 3.Rare/Expert Review Category	0.4070

Yoga & Physical Therapy	anulom pranayama vilom,	1.anulom vilom, pranayama; 2.Rare/Expert Review Category; 3. breathing exercises	0.9476
Medical Intervention	Rare/Expert Review Category	1.Rare/Expert Review Category; 2. medication support; 3.vaccination, supportive care	0.7625
Prevention	avoid triggers	1.avoid triggers; 2.stress management; 3.Unknown	1.0000
Prognosis	usually mild	1.usually mild; 2. chronic; 2.can be managed with treatment	0.1052
Complications	Rare/Expert Review Category	1. Rare/Expert Review Category; 2.respiratory complications; 3.mental stress related complications	0.1805
Patient Recommendations	Rare/Expert Review Category	1.Rare/Expert Review Category; 2.avoid triggers; 3. monitor symptoms	0.6850

The Rare/Expert Review Category is not clinical output; it indicates that the model assigned the case to a low-frequency recommendations group and should trigger practitioner review or nearest-neighbour evidence retrieval.



This example demonstrates safety behavior for low-support recommendation regions.

Figure 8. Sample Top-K recommendation flow for a low-support recommendation profile.

The Top-K recommendation scheme employed by AIMRS is illustrated in Figure 8 in which a patient profile corresponds to a low support or an imbalanced recommendation class. Rather than generating a single prediction, the system sorts a number of candidate recommendations by the likelihoods with which they would occur. The approach can be especially useful in personalized Ayurvedic decision support where there might exist several potential therapies and human expertise plays an important role in decision making.

12. Ethical and Clinical Considerations

This framework is meant strictly for academic research and practitioner-reviewed decision support — nothing beyond that. It doesn't diagnose disease, it doesn't prescribe Ayurvedic medicine on its own, and it hasn't been clinically validated. Any recommendation it generates needs to be checked by a qualified Ayurvedic or medical professional before it's used in practice. The dataset was downloaded from Kaggle and used purely as an academic research dataset. The class imbalance and rare labels remain major constraints throughout. The Feature-importance patterns may simply be reflecting bias baked into the dataset.

13. Conclusion

This paper presented the **Ayurvedic Intelligent Multi-output Recommendation System (AIMRS)**, an explainable and optimized machine learning framework to support personalized Ayurveda recommendations. The proposed framework solves the problem of multi-dimensional character of Ayurveda decisions making through prediction of nine connected recommendation categories based on patient-specific clinical, demographic, lifestyle, environmental, dosha, and prakriti attributes. To handle problems caused by sparsity and imbalance of the recommendation labels, the framework uses data preprocessing, label normalization, rare class treatment, target-specific model selection, hyperparameter optimization, Top-K recommendation, ablation analysis, and explainability through feature importance. The experimental results show that the optimization techniques significantly boost the performance of predictions in multiple recommendation categories, where Medical Intervention reaches the highest Weighted F1-score 0.7350, followed by Yoga and Physical Therapy, Complications, Ayurvedic Herbs, and Prevention. The before-and-after optimization analysis confirms the efficiency of the proposed methodology, while the ablation analysis and explainability analysis increase the interpretability and reproducibility of the models. While highly sparse targets like Formulation pose difficulties, the proposed framework lays down the foundation of the explainable AI-driven Ayurveda recommendation. Further works will concentrate on developing the system through larger clinically validated datasets, expert-reviewed recommendations, advanced machine learning techniques, and clinical evaluation.

References

- [1] S. Shaumya and R. Kumar, "Artificial Intelligence in Ayurveda: A Systematic Review (2020–2025)," *Journal of Ayurveda and Naturopathy*, vol. 2, no. 2, pp. 05–19, 2025, doi: 10.33545/ayurveda.2025.v2.i2.A.24.
- [2] V. Meena and S. Mandloi, "AI-Driven Digital Healthcare and Wellness System with Ayurveda," *International Journal of Creative and Open Research in Engineering and Management*, vol. 2, no. 5, 2026, doi: 10.55041/ijcope.v2i5.801.
- [3] Z. Huang et al., "An Ayurgenomics Approach: Prakriti-Based Drug Discovery and Development for Personalized

- Care," *Frontiers in Pharmacology*, vol. 13, Art. no. 866827, 2022, doi: 10.3389/fphar.2022.866827.
- [4] P. Bidve, S. Mishra, and A. J., "Enhancing Ayurvedic Diagnosis using Multinomial Naive Bayes and K-modes Clustering: An Investigation into Prakriti Types and Dosha Overlapping," arXiv:2310.02920, 2023, doi: 10.48550/arXiv.2310.02920.
- [5] D. Chetia, S. K. Kalita, P. P. Baruah, D. Dutta, and T. Akhter, "Identification of Traditional Medicinal Plant Leaves Using an Effective Deep Learning Model and Self-Curated Dataset," arXiv:2501.09363, 2025, doi: 10.48550/arXiv.2501.09363.
- [6] M. S. I. Musyaffa, N. Yudistira, M. A. Rahman, and J. Batoro, "IndoHerb: Indonesia Medicinal Plants Recognition using Transfer Learning and Deep Learning," arXiv:2308.01604, 2023, doi: 10.48550/arXiv.2308.01604.
- [7] G. Tsoumakas and I. Katakis, "Multi-label Classification: An Overview," *International Journal of Data Warehousing and Mining*, vol. 3, no. 3, pp. 1–13, 2007.
- [8] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.
- [9] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
- [10] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [11] R. T. Sutton et al., "An Overview of Clinical Decision Support Systems: Benefits, Risks, and Strategies for Success," *npj Digital Medicine*, vol. 3, Art. no. 17, 2020, doi: 10.1038/s41746-020-0221-y.
- [12] World Health Organization, *Ethics and Governance of Artificial Intelligence for Health*, Geneva, Switzerland: WHO, 2021. [Online] . Available: <https://www.who.int/publications/i/item/9789240029200>. Accessed: 13 July 2026.
- [13] "AyurGenixAI: Ayurvedic Dataset," Kaggle. [Online] . Available: <https://www.kaggle.com/datasets/kagglekirti123/ayurgenixai-ayurvedic-dataset>. Accessed: 13 July 2026.
- [14] G. Salton and C. Buckley, "Term-weighting Approaches in Automatic Text Retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.
- [15] Zeng, G. (2023). On the analytical properties of category encodings in logistic regression. *Communications in Statistics - Theory and Methods*, 52(6), 1870–1887. <https://doi.org/10.1080/03610926.2021.1939382>
- [16] T. Joachims, "Training linear SVMs in linear time," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 217–226, 2006, doi: 10.1145/1150402.1150429.
- [17] E. Y. Boateng and D. A. Abaye, "A review of the logistic regression model with emphasis on medical research," *Journal of Data Analysis and Information Processing*, vol. 7, no. 4, pp. 209–219, 2019, doi: 10.4236/jdaip.2019.74012.
- [18] Breiman, L. Random Forests. *Machine Learning* 45, 5–32 (2001). <https://doi.org/10.1023/A:1010933404324>
- [19] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006, doi: 10.1007/s10994-006-6226-1
- [20] L. Jiang, S. Wang, C. Li and L. Zhang, "Structure extended multinomial naive Bayes," *Information Sciences*, vol. 329, 2016, pp. 346-356, doi: 10.1016/j.ins.2015.09.037
- [21] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451
- [22] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016, doi: 10.1145/2939672.2939785
- [23] J. Zhang, D. Mucs, U. Norinder, and F. Svensson, "LightGBM: An effective and scalable algorithm for prediction of chemical toxicity—Application to the Tox21 and Mutagenicity Data Sets," *Journal of Chemical Information and Modeling*, vol. 59, no. 10, pp. 4150–4158, 2019, doi: 10.1021/acs.jcim.9b00633
- [24] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush and A. Gulin, "CatBoost: Unbiased Boosting with Categorical Features," *Advances in Neural Information Processing Systems*, 2018, pp. 6639-6650, doi:

10.48550/arXiv.1706.09516.

- [25] J. Opitz and S. Burst, "Macro F1 and Macro F1," 2019, doi: 10.48550/arXiv.1911.03347
- [26] J. Opitz, "A closer look at classification evaluation metrics and a critical reflection of common evaluation practice," Transactions of the Association for Computational Linguistics, vol. 12, pp. 820–836, 2024, doi: 10.1162/tacl_a_00675