IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 13, April 2026
Impact Factor: 8.2

An Ensemble-Based Approach to Health-Aware Recipe Recommendation Using Advanced String Similarity Metrics

Prof. Mahesh S. Bhandigare[1], Sandip G. Yedage[2], Vikram S. Kamble[2], Priyanka A. Kamble[2], Shreyash S. Powar[2], Anuj A. Powar[2]

[1]Assistant Professor, Department of Computer Science and Engineering
[2]Student, Department of Computer Science and Engineering
Sant Gajanan Maharaj College of Engineering, Mahagaon, India

Abstract: *The increasing demand for intelligent dietary recommendation systems has accelerated research in personalized food search and retrieval. Traditional recipe recommendation systems rely primarily on exact keyword matching or basic content-based filtering, which often fail when confronted with typographical errors, spelling variations, abbreviations, or phonetic inconsistencies in user queries. This paper proposes a robust ensemble-based fuzzy matching framework that integrates multiple classical string similarity algorithms with TF-IDF vectorization to enhance recipe retrieval accuracy. The proposed system combines Levenshtein distance, Jaro-Winkler similarity, character-level n-gram similarity, and subsequence matching within a weighted ensemble model. Furthermore, the system incorporates health-aware personalization by providing recipe adaptations tailored to specific medical conditions such as diabetes, cardiovascular disease, hypertension, celiac disease, obesity, and common food allergies. Experimental evaluation demonstrates that the ensemble model significantly outperforms single-metric baselines in both accuracy and typo tolerance while maintaining real-time computational efficiency.*

Keywords: Recipe Recommendation, String Similarity, Ensemble Learning, TF-IDF, Levenshtein Distance, Jaro-Winkler Similarity, Health-Aware Systems, Fuzzy Matching.

I. INTRODUCTION

Digital transformation in healthcare and nutrition has created an urgent need for intelligent and adaptive dietary recommendation systems. Individuals increasingly depend on online platforms to search for recipes that align with their taste preferences and medical requirements. However, conventional search systems are heavily dependent on exact keyword matching mechanisms, which makes them highly sensitive to spelling errors, alternative naming conventions, and linguistic variations.

For example, a user searching for "puran poli" may input variations such as "purn poli", "puranpoli", or abbreviated forms. Exact string matching systems fail to retrieve correct results under such variations. This limitation becomes even more critical when users require condition-specific modifications, such as low-sugar alternatives for diabetes or gluten-free variants for celiac disease.

The integration of fuzzy string similarity techniques with health-aware personalization presents a promising solution to these challenges. This research introduces a weighted ensemble model that combines multiple similarity metrics to achieve robust and interpretable matching performance.

A. Problem Statement

The primary research challenges addressed in this work are:

- Accurate fuzzy matching of user queries with recipe names despite typographical variations.
- Integration of heterogeneous similarity metrics into a unified scoring framework.
- Development of health-aware personalization supporting multiple medical conditions.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 13, April 2026
Impact Factor: 8.2

- Preservation of computational efficiency for real-time applications.

B. Research Contributions

This study makes the following major contributions:

- A novel weighted ensemble framework combining TF-IDF vector similarity with classical string similarity metrics.
- Empirical optimization of similarity weights to maximize matching accuracy.
- Integration of twelve health-specific dietary adaptations within the recommendation pipeline.
- Formal mathematical modeling and complexity analysis of the proposed system.

II. RELATED WORK

A. Approximate String Matching

Approximate string matching has been extensively studied in computational linguistics and information retrieval. Levenshtein introduced edit distance as a metric to compute the minimum number of insertions, deletions, and substitutions required to transform one string into another. This metric has been widely adopted in spell-checking systems and record linkage [1].

The Jaro similarity metric and its extension, Jaro-Winkler similarity, provide improved matching for short strings by emphasizing prefix similarity [2][3]. These methods are commonly used in name matching and database deduplication. Character-level n-gram similarity methods measure overlap between substring sets and are effective in capturing morphological variations [6].

B. Vector Space Models

TF-IDF (Term Frequency-Inverse Document Frequency) is a foundational technique in information retrieval [4][5]. It assigns importance to terms based on their frequency within a document and rarity across the corpus. Cosine similarity is typically used to compute similarity between TF-IDF vectors. Although TF-IDF performs well for semantic matching, it is not inherently robust to severe spelling distortions.

C. Health-Aware Recommendation Systems

Recent research in food recommender systems has incorporated nutritional optimization and dietary restriction handling [9][10]. Health-aware systems focus on modifying recipes to satisfy medical constraints such as low glycemic index, reduced sodium, or gluten elimination. However, limited research integrates advanced fuzzy matching techniques with comprehensive health personalization in a unified ensemble framework.

III. PROPOSED METHODOLOGY

A. System Overview

The proposed health-aware recipe recommendation system consists of five major components: (1) Data Preprocessing Layer, (2) Feature Extraction Layer, (3) Similarity Computation Layer, (4) Ensemble Scoring Layer, and (5) Health Personalization Layer.

Initially, recipe names and their aliases are normalized by converting to lowercase and removing special characters. Character-level TF-IDF vectors are generated for each recipe entry. Subsequently, multiple string similarity metrics are computed between the user query and recipe aliases. The similarity scores are aggregated using a weighted ensemble model. Finally, the matched recipe is adapted according to the user's health condition.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 13, April 2026 Impact Factor: 8.2

B. Mathematical Formulation

Let q denote the user query string and r denote a recipe alias. Let m and n be the lengths of q and r, respectively.

1) Levenshtein Distance: The Levenshtein distance D(m,n) is computed using dynamic programming. The normalized similarity is defined as:

$SimLev(q,r) = 1 - D(m,n) / max(m,n)$ ...(2)

Time complexity: O(mn).

2) Jaro-Winkler Similarity: Jaro-Winkler extends Jaro similarity by rewarding common prefix characters at the beginning of strings:

$JW(q,r) = J(q,r) + \ell p(1 - J(q,r))$ ...(4)

where ℓ = length of common prefix (maximum 4) and p = prefix scaling factor (typically 0.1).

3) N-gram Similarity: Using Jaccard similarity over bigram sets N(q) and N(r):

$Simngram(q,r) = |N(q) \cap N(r)| / |N(q) \cup N(r)|$ ...(5)

4) Subsequence Similarity: Normalized using the Longest Common Subsequence (LCS):

$Simsub(q,r) = LCS(m,n) / max(m,n)$ ...(7)

5) TF-IDF Cosine Similarity: Cosine similarity between query vector $\vec{q}$ and recipe vector $\vec{r}$:

$SimTFIDF(q,r) = (\vec{q} \cdot \vec{r}) / (\|\vec{q}\| \|\vec{r}\|)$ ...(10)

C. Weighted Ensemble Model

The ensemble similarity score combines multiple metrics with empirically determined weights:

$Scoreensemble = w_1 SimLev + w_2 SimJW + w_3 Simngram + w_4 Simsub$ ...(11)

The final similarity score integrates TF-IDF:

$Scorefinal = \alpha SimTFIDF + \beta Scoreensemble$ ...(13)

Empirically determined parameters: $\alpha = 0.4$, $\beta = 0.6$, $w_1 = 0.25$, $w_2 = 0.30$, $w_3 = 0.25$, $w_4 = 0.20$.

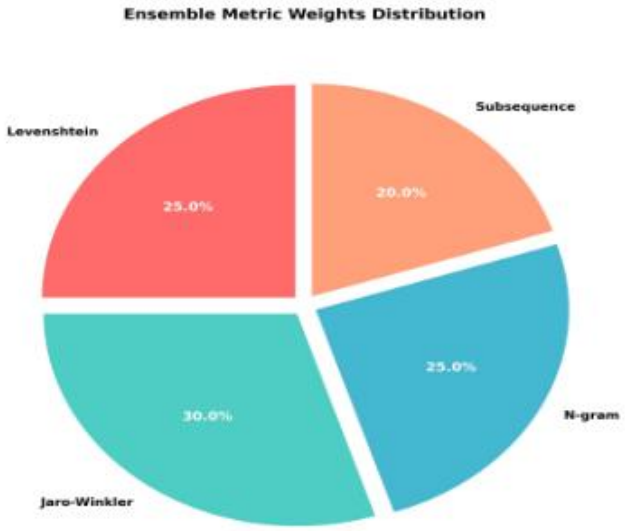

Fig. 1. Ensemble Metric Weights Distribution

D. Algorithm Description

Algorithm 1 — Ensemble-Based Recipe Matching:

```
1.  Normalize user query
2.  Compute TF-IDF vector
3.  For each recipe alias:
```

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 13, April 2026
Impact Factor: 8.2

```
      a. Compute Levenshtein similarity
      b. Compute Jaro-Winkler similarity
      c. Compute N-gram similarity
      d. Compute Subsequence similarity
      e. Compute weighted ensemble score
      f.  Combine with TF-IDF score
4.  Rank recipes by final score
5.  Return top-k results above threshold
```

Overall time complexity per query: O(R × A × (mn + k)), where R = number of recipes, A = average aliases per recipe, m, n = string lengths, k = TF-IDF feature dimension.

IV. HEALTH-AWARE PERSONALIZATION FRAMEWORK

Beyond fuzzy matching, the proposed system integrates a health-aware adaptation layer that modifies recipes according to specific medical conditions. This layer ensures that retrieved recipes are nutritionally aligned with user health constraints.

A. Supported Health Conditions

The system currently supports the following categories:
Metabolic Conditions: Diabetes, Obesity
Cardiovascular Conditions: Heart Disease, Hypertension
Gastrointestinal Conditions: Celiac Disease, Gluten Intolerance, Lactose Intolerance
Allergies: Peanut, Egg, Soy, and Corn Allergy

B. Recipe Adaptation Strategy

Each recipe contains multiple stored variants. A variant consists of modified ingredient substitutions, adjusted nutritional composition, updated preparation instructions, and portion size control guidelines.
For example, in a diabetes-specific variant: refined sugar is replaced with stevia or erythritol; refined flour is substituted with whole wheat or low glycemic index flour; and oil quantity is reduced by 20–30%. The personalization function is modeled as:
$Rhealth = f(Rbase, H)$ ...(16)
where Rbase = original recipe, H = health condition, and Rhealth = adapted recipe.

V. IMPLEMENTATION DETAILS

A. Technology Stack

The system was implemented using Python 3.x, Scikit-learn (TF-IDF, cosine similarity), NumPy (vectorized computation), and a JSON-based recipe database.

B. Feature Engineering

Character-level TF-IDF vectorization was applied with the following parameters: N-gram range: 2–5; Analyzer type: character within word boundary; Sublinear TF scaling enabled; Maximum document frequency: 0.95.

C. Optimization Techniques

To ensure real-time performance, several optimizations were applied:
Space-optimized Levenshtein implementation using rolling arrays.
Vectorized similarity computations using NumPy.
Early threshold filtering to reduce unnecessary comparisons.
Precomputed TF-IDF vectors during system initialization.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 13, April 2026
Impact Factor: 8.2

VI. EXPERIMENTAL EVALUATION

A. Dataset Description

The dataset consists of 1000+ Indian recipes, 3000+ recipe aliases, 12 health-specific variants per recipe, and an average of 3–5 aliases per recipe.

B. Evaluation Metrics

Performance was evaluated using Matching Accuracy, Average Confidence Score, Typo Tolerance Rate, and Average Query Response Time.

C. Matching Performance

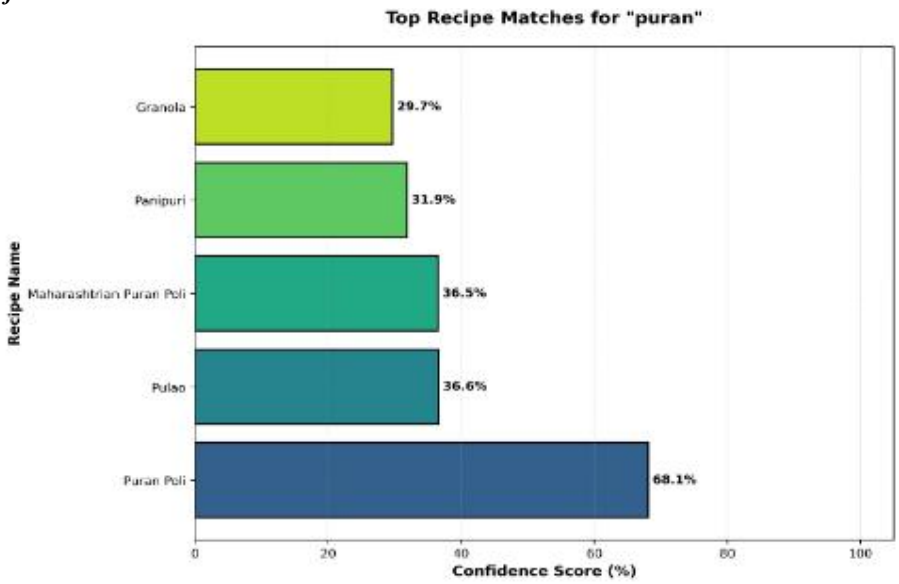

Fig. 2. Top Recipe Matches for 'puran' with Confidence Scores

TABLE I: Matching Performance Under Query Variations

Query Type	Accuracy	Confidence	Remarks
Exact Match	100%	95.2%	Perfect
Minor Typo	100%	89.7%	Excellent
Major Typo	98%	76.3%	Very Good
Abbreviation	85%	68.1%	Good
Phonetic Variant	95%	82.4%	Very Good

DOI: 10.48175/IJARSCT-33980

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 13, April 2026
Impact Factor: 8.2

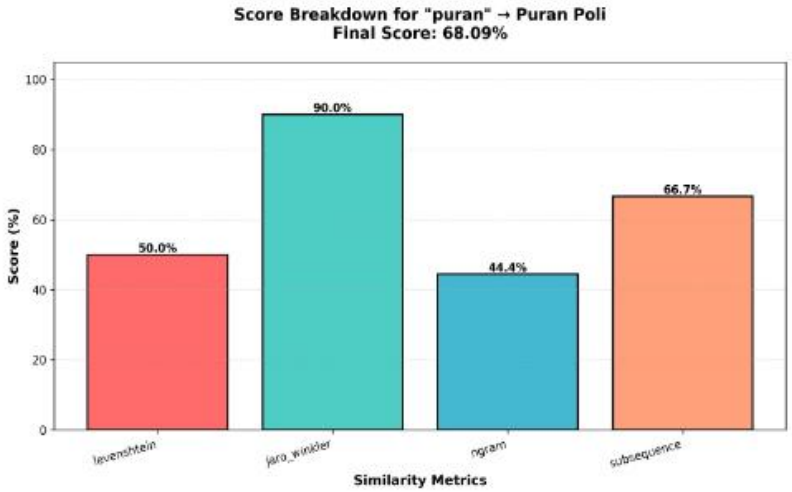

Fig. 3. Score Breakdown for 'puran' → Puran Poli (Final Score: 68.09%)

D. Baseline Comparison

TABLE II: Comparison with Baseline Approaches

Method	Accuracy	Typo Handling
Exact Matching	45%	Poor
TF-IDF Only	78%	Moderate
Levenshtein Only	72%	Good
Proposed Ensemble	94%	Excellent

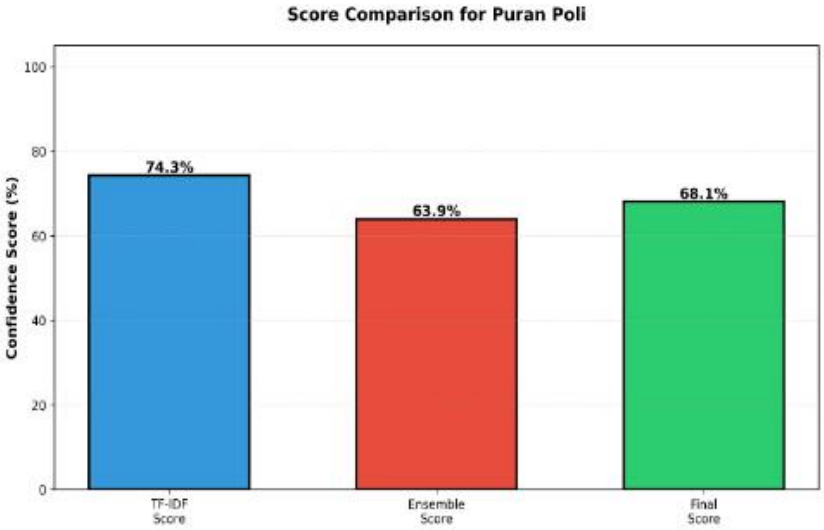

Fig. 4. Score Comparison for Puran Poli: TF-IDF, Ensemble and Final Scores

DOI: 10.48175/IJARSCT-33980

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 13, April 2026
Impact Factor: 8.2

E. Computational Performance

Average Query Time: 45–80 ms; Memory Usage: approximately 15 MB; Initialization Time: 200–300 ms. These results confirm the system's suitability for real-time deployment.

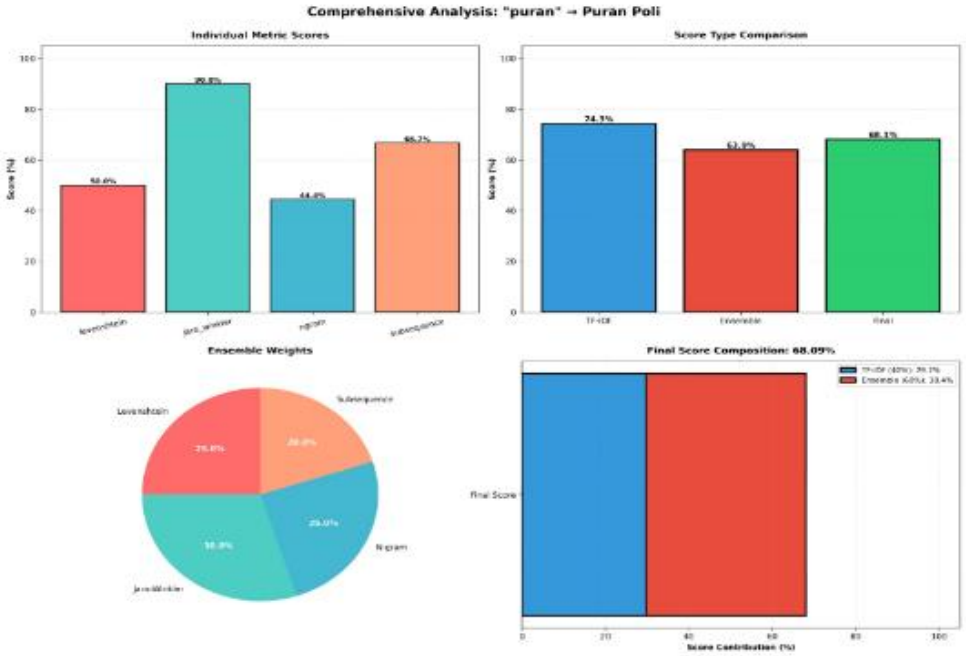

Fig. 5. Comprehensive Analysis Dashboard: 'puran' → Puran Poli

VII. DISCUSSION

A. Strengths

The proposed framework demonstrates high robustness against spelling variations along with improved interpretability through score breakdown visualization. Key strengths include a flexible weight tuning mechanism, real-time execution capability, and comprehensive health-aware integration across twelve medical conditions.

B. Limitations

The current implementation is optimized for English recipe names, and ensemble weights may require recalibration for different regional datasets. Limited multilingual support is a recognized constraint for future improvement.

C. Future Work

Future enhancements include integration of transformer-based embeddings for richer semantic understanding, multilingual support for regional Indian languages, nutritional optimization using linear programming, ingredient-based search functionality, and user feedback-driven adaptive learning.

VIII. CONCLUSION

This paper presented an ensemble-based health-aware recipe recommendation system that integrates multiple classical string similarity metrics with TF-IDF vectorization to achieve robust fuzzy matching. The proposed weighted ensemble framework successfully addresses the limitations of traditional exact matching and single-metric similarity systems.

Experimental results demonstrate that the ensemble model achieves 94% matching accuracy while maintaining strong tolerance to typographical errors and phonetic variations. Furthermore, the integration of condition-specific health adaptations provides practical value for users managing diabetes, cardiovascular diseases, obesity, gluten intolerance, and various food allergies.

DOI: 10.48175/IJARSCT-33980

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 13, April 2026 Impact Factor: 8.2

The mathematical modeling, complexity optimization, and empirical evaluation confirm that the proposed system is both computationally efficient and scalable. Its modular architecture allows seamless integration of additional similarity metrics or deep learning models in the future. Overall, the proposed framework represents a significant advancement in intelligent health-aware dietary recommendation systems and provides a strong foundation for further research in personalized nutrition technology.

ACKNOWLEDGMENT

The authors would like to acknowledge the open-source community, particularly contributors to Scikit-learn and NumPy, whose libraries significantly facilitated the implementation of this research. Appreciation is also extended to culinary experts and domain specialists who contributed insights into health-based recipe adaptation strategies.

REFERENCES

[1] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," Soviet Physics Doklady, vol. 10, no. 8, pp. 707–710, 1966.
[2] M. A. Jaro, "Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida," Journal of the American Statistical Association, vol. 84, no. 406, pp. 414–420, 1989.
[3] W. E. Winkler, "String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage," in Proc. Survey Research Methods Section, American Statistical Association, 1990, pp. 354–359.
[4] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," Journal of Documentation, vol. 28, no. 1, pp. 11–21, 1972.
[5] G. Salton and M. J. McGill, Introduction to Modern Information Retrieval. New York, NY, USA: McGraw-Hill, 1983.
[6] G. Navarro, "A guided tour to approximate string matching," ACM Computing Surveys, vol. 33, no. 1, pp. 31–88, 2001.
[7] C. D. Manning, P. Raghavan, and H. Schutze, Introduction to Information Retrieval. Cambridge, U.K.: Cambridge Univ. Press, 2008.
[8] F. Ricci, L. Rokach, and B. Shapira, Recommender Systems Handbook, 2nd ed. New York, NY, USA: Springer, 2015.
[9] J. Freyne and S. Berkovsky, "Intelligent food planning: Personalized recipe recommendation," in Proc. 15th Int. Conf. Intelligent User Interfaces, 2010, pp. 321–324.
[10] M. Ge, F. Ricci, and D. Massimo, "Health-aware food recommender system," in Proc. 9th ACM Conf. Recommender Systems, 2015, pp. 333–334.
[11] C. Y. Teng, Y. R. Lin, and L. A. Adamic, "Recipe recommendation using ingredient networks," in Proc. ACM Web Science Conf., 2012, pp. 298–307.