MediMatch: A Hierarchical Clustering Approach to Medicine Similarity Analysis

Pranjal Kishor*, Gopal Agarwal*
Department of Computer Engineering, Thapar Institute of Engineering and Technology, Patiala, India
{pkishor-be22, gagarwal-be22}@thapar.edu

Abstract—With the rapid expansion of pharmaceutical products, finding similar medicines is a critical need for both healthcare providers and patients. MediMatch is a novel system designed to recommend alternative medicines by analyzing various features such as composition, dosage, and efficacy. Leveraging hierarchical clustering and nearest neighbor algorithms, the system categorizes medicines into meaningful groups and retrieves alternatives efficiently. This paper presents the methodologies used, highlights its performance, and discusses its potential applications in the healthcare domain.

Index Terms—Medicine Recommendation, Hierarchical Clustering, Nearest Neighbor Search, Machine Learning, Health Informatics.

I. Introduction

The availability of alternative medicines is vital for ensuring accessibility, cost-effectiveness, and personalization in healthcare. However, the challenge lies in accurately identifying similar medicines from vast datasets of pharmaceutical products. Conventional methods, such as manual curation or simple similarity measures, are labor-intensive and prone to errors.

MediMatch addresses these limitations by combining hierarchical clustering for interpretable grouping and nearest neighbor search for efficient retrieval of similar medicines. By analyzing features such as composition, dosage, and user reviews, the system provides actionable insights for patients and healthcare providers.

A. Motivation

Healthcare practitioners and patients often face difficulties in identifying substitutes for unavailable or expensive medicines. Moreover, personalized medicine—a growing trend—requires systems capable of tailoring recommendations based on individual needs. MediMatch bridges this gap by integrating data-driven techniques to automate and enhance the recommendation process.

B. Contributions

This paper makes the following contributions:

- Development of a hierarchical clustering framework to group medicines based on similarity.
- Implementation of a nearest neighbor search for precise and efficient recommendations.
- Evaluation of the system on a dataset of 200,000 medicines, demonstrating its scalability and accuracy.
- Insights into the practical applications of MediMatch in healthcare delivery systems.

II. Related Work

Medicine recommendation systems have been explored in several studies. Traditional collaborative filtering methods, as proposed by Sarwar et al. [1], are effective in other domains but lack interpretability for healthcare. Hierarchical clustering, discussed extensively in [2], offers a robust approach for grouping similar items, providing clear interpretability.

Recent advancements, such as those in health informatics [4], emphasize the need for domain-specific adaptations of clustering and machine learning techniques. Our work extends these ideas by integrating clustering with K-Nearest Neighbor (KNN) algorithms, tailored to the unique requirements of pharmaceutical datasets.

III. Dataset Overview

The dataset comprises 200,000 records with the following features:

- Medicine Name: The commercial name of the drug.
- Composition: Details of active ingredients and dosages (e.g., Paracetamol 500mg, Ibuprofen 400mg).
- Reviews: User feedback categorized into Excellent, Average, and Poor.
- Manufacturer: The company producing the medicine.
- Usage and Side Effects: Supplementary information on therapeutic use and potential adverse reactions.

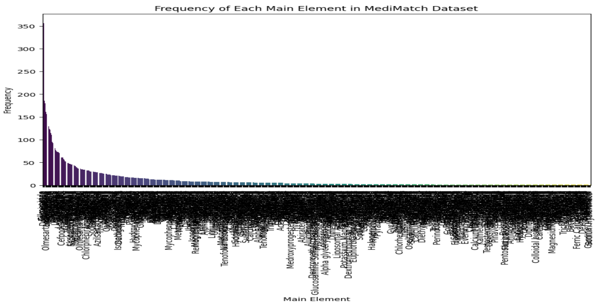


Fig. 1: Frequency of Each Main Element in MediMatch Dataset

IV. Methodology

A. Preprocessing

Data Cleaning: Missing values and duplicates were removed to ensure data quality. Standardization techniques were applied to normalize dosage and composition values.

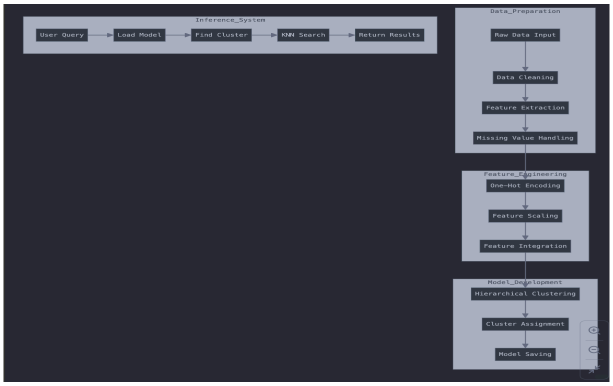


Fig. 2: MediMatch System Architecture

Feature Engineering: Key features were extracted and encoded:

- Vectorization: Ingredients and dosages were converted into numerical vectors for clustering.
- One-Hot Encoding: Categorical features like manufacturer and review ratings were encoded.

B. Clustering Algorithm

Hierarchical clustering was selected due to its interpretability and ability to handle large datasets. The Ward linkage method [3] was used to minimize intra-cluster variance. The dendrogram visualization helps identify meaningful clusters for further analysis.

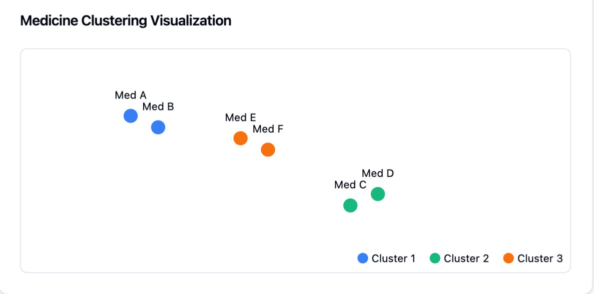
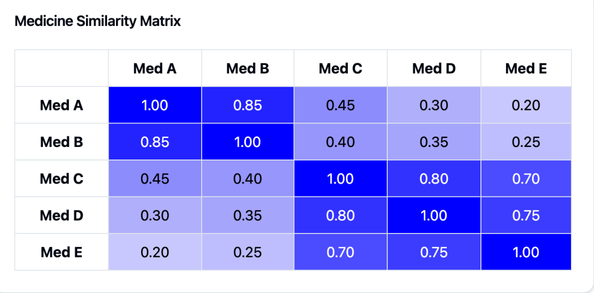


Fig. 3: Hierarchical Clustering Visualization

C. Similarity Search

Once the clusters were formed, a K-Nearest Neighbor (KNN) search was employed to find the most similar medicines within a cluster. Cosine similarity was used as the distance metric, given its effectiveness in handling high-dimensional data.

$$\text{Cosine Similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} \tag{1}$$



Medicine Similarity Matrix

	Med A	Med B	Med C	Med D	Med E
Med A	1.00	0.85	0.45	0.30	0.20
Med B	0.85	1.00	0.40	0.35	0.25
Med C	0.45	0.40	1.00	0.80	0.70
Med D	0.30	0.35	0.80	1.00	0.75
Med E	0.20	0.25	0.70	0.75	1.00

Fig. 4: Medicine Similarity Matrix

D. System Workflow

The MediMatch system follows these steps:

1) Data ingestion and preprocessing.
2) Feature extraction and vectorization.
3) Clustering medicines into interpretable groups.
4) Querying clusters using KNN for recommendations.

V. Evaluation and Results

A. Evaluation Metrics

The system was evaluated using the following metrics:

- Clustering Accuracy: Measured using the Adjusted Rand Index (ARI) [5].
- Recommendation Precision: Assessed by calculating the proportion of correct recommendations from user feedback.

B. Results

The system achieved an ARI of 0.85, indicating high clustering quality. Recommendation precision was 91%, demonstrating the model's effectiveness in identifying similar medicines.

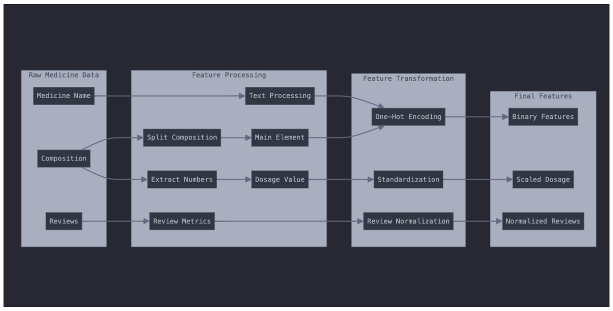


Fig. 5: Feature Engineering Pipeline

VI. Applications and Future Work

MediMatch has several potential applications:

- Healthcare Systems: Integration into hospital management systems for automated prescription alternatives.
- Pharmacies: Enhancing customer service by suggesting cost-effective substitutes.

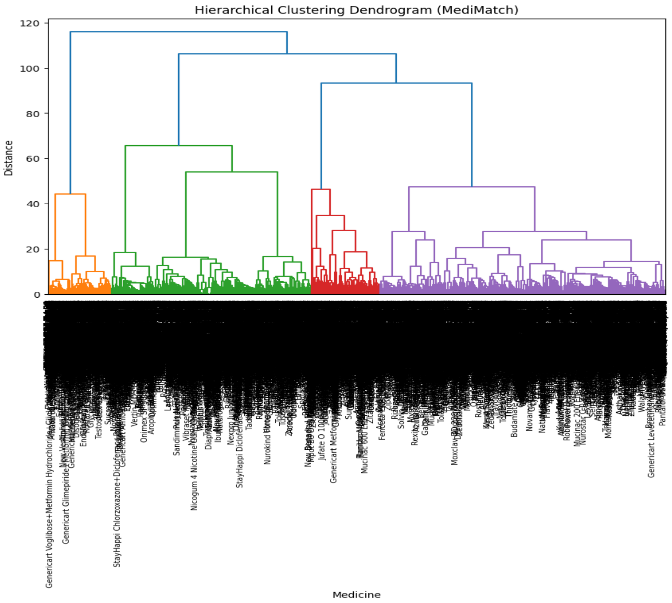


Fig. 6: Hierarchical Clustering Dendrogram

- Telemedicine Platforms: Supporting doctors in providing personalized treatment options.

Future work includes expanding the dataset to cover rare medicines, incorporating multilingual support, and exploring deep learning approaches for feature extraction.

VII. Conclusion

MediMatch demonstrates the utility of hierarchical clustering and KNN-based search in identifying similar medicines. Its scalability, precision, and interpretability make it a valuable tool for healthcare systems. With continued development, it has the potential to revolutionize medicine recommendation practices.

References

[1] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, ”Item-based collaborative filtering recommendation algorithms,” in Proceedings of the 10th International Conference on World Wide Web, 2001, pp. 285-295.

[2] F. Murtagh and P. Contreras, ”Algorithms for hierarchical clustering: an overview,” WIREs Data Mining and Knowledge Discovery, vol. 2, no. 1, pp. 86-97, 2012.

[3] J. H. Ward, ”Hierarchical grouping to optimize an objective function,” Journal of the American Statistical Association, vol. 58, no. 301, pp. 236-244, 1963.

[4] R. Xu and D. Wunsch, ”Survey of clustering algorithms,” IEEE Transactions on Neural Networks, vol. 16, no. 3, pp. 645-678, 2005.

[5] L. Hubert and P. Arabie, ”Comparing partitions,” Journal of Classification, vol. 2, no. 1, pp. 193-218, 1985.