Composition-Based Medicine Recommendation Using Hybrid Machine Learning and Brand Mapping for Alternative Drug Suggestion

Madhan Raj P
Department of CSE,
Rajalakshmi Engineering College
Thandalam, Chennai, India
220701148@rajalakshmi.edu.in

***Abstract*—.** This research presents a Machine Learning–based Medicine Recommendation System designed to enhance accessibility and informed decision-making in pharmaceutical selection. The system recommends alternative medicines that possess the same chemical composition as a user-provided drug name but are produced under different brand labels. The recommendation engine leverages a structured drug composition dataset and utilizes a vectorized ingredient-matching algorithm for identifying equivalent formulations. Natural Language Processing (NLP) techniques are integrated to ensure accurate mapping between brand names and generic components. A hybrid filtering model, incorporating both content-based and rule-based mechanisms, is implemented to refine the suggested alternatives. The system is trained and evaluated on a traditional computing platform using Scikit-learn and Pandas libraries. It is optimized using the cosine similarity measure for matching compositions while minimizing redundancy in brand alternatives. Experimental validation confirms that the model effectively identifies relevant alternatives with a high degree of compositional fidelity, ensuring that users receive medically equivalent options. The proposed model contributes toward a scalable and interpretable framework for smart pharmaceutical recommendations in both clinical and retail domains.

Keywords—Medicine Recommendation, Drug Composition, Brand Alternatives, Machine Learning, NLP

I. Introduction

In recent years, the integration of artificial intelligence and machine learning in healthcare has seen rapid advancements, driven by the need to provide smarter, faster, and more accessible medical support systems. Among these, medicine recommendation systems have gained prominence due to their potential in addressing key challenges in drug availability, brand substitution, and prescription management. As healthcare providers and patients increasingly rely on technology for informed decision-making, machine learning models offer a scalable solution to navigate the complex landscape of pharmaceutical options. Traditional methods of medicine identification rely heavily on manual referencing and professional consultation, which may be time-consuming and limited by brand familiarity. The need for automated, intelligent systems capable of recommending therapeutically equivalent medicines with the same chemical composition but manufactured under different brand names is both timely and essential.

The proposed system addresses this gap by introducing a machine learning–based recommendation engine that, upon receiving the name of a specific medicine, suggests alternative drugs with identical active ingredients but produced by different pharmaceutical companies. This system leverages structured drug datasets, such as those containing details of generic names, brand labels, dosage forms, and therapeutic classes, to build a composition-aware recommendation pipeline. Techniques from Natural Language Processing (NLP) and similarity matching are employed to extract, vectorize, and compare the compositional profile of input drugs with others in the dataset. The system emphasizes pharmaceutical equivalency, ensuring that the therapeutic effect remains consistent across recommended alternatives.

In particular, cosine similarity is utilized to measure the closeness between composition vectors, while auxiliary filters consider dosage strength, form (e.g., tablet, syrup), and frequency of availability. This model is trained and evaluated using standard machine learning libraries including Scikit-learn and Pandas, ensuring compatibility with healthcare data structures. The implementation environment is built on a lightweight architecture to facilitate deployment in real-world applications such as online pharmacies, hospital management systems, and mobile health (mHealth) platforms.

Unlike traditional e-prescription systems that suggest alternatives based on cost or popularity, this system prioritizes molecular and compositional integrity, making it especially useful in resource-constrained regions or during shortages of popular brands. By enhancing accessibility and choice, the model contributes to a more patient-centric healthcare ecosystem. Furthermore, the system lays the foundation for future integration with clinical decision support tools and electronic health record (EHR) systems to provide holistic medication management solutions. Through rigorous testing and model evaluation, the system demonstrates high accuracy in mapping brands to generics and vice versa, with low divergence in recommended composition profiles.

II. Literature Review

The growing demand for personalized healthcare and efficient drug management has led to significant advancements in medicine recommendation systems powered by machine learning techniques. These systems assist in identifying alternative drugs with similar active pharmaceutical ingredients (APIs), thereby improving accessibility and reducing treatment costs. Several studies have explored various algorithmic approaches to address these challenges through the integration of drug datasets, similarity matching, and deep learning architectures.In [1], the authors present a recommendation engine based on content-based filtering that suggests alternative drugs by comparing drug compositions and molecular structures. The system uses cosine similarity over a vectorized representation of drug attributes such as composition, strength, and form. The approach, while effective in identifying exact chemical matches, is limited in its ability to handle multi-ingredient drugs or brand-specific formulations.A hybrid collaborative filtering framework is proposed in [2], which combines user preferences (e.g., user prescriptions or historical drug usage) with drug features such as therapeutic class and chemical composition. By applying matrix factorization techniques, the model effectively predicts similar medicines from different brands. However, the reliance on historical user data poses a challenge for cold-start scenarios where user history is absent.The work in [3] introduces a machine learning-based classification model to group medicines based on the Anatomical Therapeutic Chemical (ATC) classification system. Random Forest and Support Vector Machine (SVM) classifiers are used to categorize drugs into therapeutic classes, enabling a more structured drug substitution process. While the model achieves high accuracy, it does not directly offer brand-level substitution, which is crucial for real-world prescription filling.In [4], Natural Language Processing (NLP) is employed to parse unstructured drug descriptions from online pharmacy databases. Named Entity Recognition (NER) techniques are utilized to extract drug names, compositions, and usage contexts. The study highlights the significance of combining structured and unstructured data to build comprehensive medicine recommendation systems, especially in multilingual or inconsistent data environments.A deep learning-based approach using autoencoders is introduced in [5] to model drug composition similarity. The encoder compresses the medicine's features into latent space, and the decoder reconstructs similar compositions. The approach allows the model to recommend substitutes based on compositional proximity. While promising, the model's performance heavily depends on a sufficiently large and diverse training dataset.In [6], a graph-based model is used to represent drug relationships where nodes represent medicines and edges encode compositional similarity or therapeutic equivalence. Graph Convolutional Networks (GCNs) are applied to propagate similarity across the graph structure, thus enabling effective recommendation even in sparse data settings. This method is well-suited for large-scale drug networks but requires careful normalization of the graph.A reinforcement learning-based recommendation framework is explored in [7] for medicine substitution in chronic treatment pathways. The system learns to recommend drugs over multiple time steps, considering drug effectiveness, side effects, and cost. The model demonstrates improved adherence outcomes but demands extensive longitudinal health records for training.In [8], the authors propose a federated learning model that allows different healthcare institutions to collaboratively train a drug recommendation model without sharing sensitive patient data. The study emphasizes privacy-preserving learning while still achieving high accuracy in brand substitution and dosage optimization.The application of Explainable AI (XAI) in drug recommendation is detailed in [9], where models are built with interpretability in mind. SHAP (SHapley Additive exPlanations) values are used to understand which features most influence the selection of a particular drug alternative. This improves trust among healthcare providers and patients but introduces computational overhead.In [10], a deep neural network is trained using prescription and drug interaction datasets to avoid suggesting alternatives with known adverse interactions. The system integrates interaction databases such as DrugBank to enhance safety. This approach addresses a critical gap in conventional systems that focus solely on compositional similarity.Study [11] proposes an ontology-driven approach where drug properties and relationships are modeled in a formal semantic network. By using SPARQL queries on ontologies like SNOMED CT and RxNorm, the system ensures semantic consistency in recommending medicines. However, the implementation complexity is high due to the need for robust knowledge engineering.A cloud-based medicine recommendation engine is developed in [12] to support low-resource clinical setups. The system integrates RESTful APIs, cloud databases, and ML models to deliver medicine suggestions with low latency. The study demonstrates practical deployment scenarios but requires consistent internet connectivity for real-time functioning.In [13], a cross-lingual medicine recommendation system is introduced using multilingual word embeddings. This system is particularly useful in multilingual nations where drug labels or prescriptions may be in different languages. The model bridges the language barrier and improves accessibility but may struggle with rare or region-specific brand names.These studies collectively highlight the diverse methodologies employed in building intelligent medicine recommendation systems. Despite their respective contributions, challenges such as data privacy, multilingual support, interpretability, and scalability remain open areas of research. The proposed system aims to address these challenges by combining compositional matching, NLP, and lightweight ML models for efficient, brand-level medicine recommendation.

III. Proposed Methodology

This section outlines the proposed framework for a machine learning-based Medicine Recommendation System that identifies and recommends alternative medicines with the same chemical composition but from different brands. The system leverages both predictive modeling and explainable artificial intelligence (XAI) techniques to ensure recommendations are accurate and transparent. Key phases of the methodology include data collection, preprocessing, feature engineering, model selection and training, integration

of XAI tools like SHAP and LIME, and performance evaluation. The framework is designed to support decision-making in healthcare by making medicine substitution safer and more comprehensible for users.

A. Data Collection and Synthesis
Due to the absence of a public, structured dataset containing brand-wise medicine composition, a synthetic dataset was curated. This dataset simulates real-world medicine catalogs and includes brand names, medicine names, combinations of active ingredients, dosage forms, strengths, pricing, manufacturer information, and therapeutic class. Each record also captures patient-relevant attributes like preferred brand, price sensitivity, and known allergies. The target outcome is a list of suitable alternative medicines that share the exact chemical composition but differ in branding or pricing.

B. Data Preprocessing
To ensure high data quality, preprocessing steps included handling missing values through median/mode imputation and standardizing names using string normalization techniques (lowercasing, stemming, abbreviation resolution). Multi-ingredient compositions were parsed and sorted alphabetically to enable accurate matching. Categorical variables such as dosage form and manufacturer were label encoded, and numerical fields like price and strength were normalized using Min-Max scaling. Duplicate entries and inconsistencies in spelling were resolved through fuzzy matching algorithms.

C. Feature Engineering
In addition to the base fields, the following derived features were created:

- **Ingredient Hash**: A unique identifier created by hashing sorted ingredient lists to group all chemically equivalent medicines.
- **Price Tier**: A discretized feature categorizing medicines as low, medium, or high cost.
- **Brand Popularity Score**: Estimated using frequency of brand occurrence in the dataset.
- **Patient Match Score**: A composite score reflecting alignment with patient preferences (e.g., low price + no allergen match).

These engineered features enable the model to recommend alternatives not only based on chemistry but also user-specific considerations.

D. Model Choice and Training
Three machine learning models were explored for alternative medicine recommendation:

- **Random Forest**: Chosen for its robustness and interpretability, especially in handling categorical and structured data.
- **XGBoost**: Selected for its efficiency and accuracy in ranking-based recommendation systems.
- **Logistic Regression**: Used as a baseline due to its simplicity and ease of interpretation.

Training was conducted using a supervised learning setup, where input features described a primary medicine and user profile, and the label indicated a suitable alternative match. Grid search with 5-fold cross-validation was applied for hyperparameter tuning, optimizing for precision and recall.

E. Explainability Integration
To ensure recommendations are interpretable:

- **SHAP** was used with tree-based models (RF, XGBoost) to explain feature contributions both globally and at the instance level. Summary and force plots helped visualize which features—such as matching ingredient hash or price tier—most influenced a recommendation.
- **LIME** was applied with Logistic Regression to generate understandable local explanations for simpler use cases, making it accessible to pharmacists and patients.

F. Evaluation Metrics
The models were evaluated using the following criteria:

- **Accuracy Metrics**: Precision, Recall, and F1-score to assess the correctness of recommended alternatives.
- **Ranking Metrics**: Mean Reciprocal Rank (MRR) and Normalized Discounted Cumulative Gain (nDCG) to measure how well true alternatives were ranked among suggestions.
- **Explainability Metrics**: Human clarity score (1–5) based on feedback from pharmacists assessing the transparency of the explanations.

G. Model Comparison Framework
Each model was evaluated based on:

- **Prediction Quality**: Top-5 alternative match accuracy and ranking consistency.
- **Interpretability**: Visual clarity, explanation consistency, and pharmacist rating.
- **Computational Cost**: Training time and explanation latency.

Random Forest with SHAP struck the best balance, offering high performance and meaningful insights. XGBoost had higher accuracy but was slightly less interpretable. Logistic Regression offered fastest inference and clearer explanations, though less precise in complex cases.

H. System Implementation Pipeline
The system architecture was modular and included:

1. **Data Loader** – Prepares structured medicine data and user input.
2. **Feature Constructor** – Builds features such as ingredient hash and patient match score.
3. **Model Inference Engine** – Uses trained models to suggest alternatives.

4. **XAI Module** – Generates SHAP or LIME visual explanations.
5. **Recommendation Dashboard** – Displays recommended alternatives, their similarity score, and explanation visuals in a web interface.

I. Application in Practical Scenario

In a simulated user session, a patient inputs the name of a prescribed medicine. The system extracts its composition, price, and dosage, and then recommends equivalent alternatives from different brands. SHAP visualization highlights that price tier and matching ingredient hash were the most influential factors, helping both the patient and pharmacist understand the reason behind the recommendation. If the patient has allergies, the system excludes matching medicines and explains the decision.

J. Limitations Addressed

While the system is highly functional, a few limitations were noted:

- **Synthetic Data**: May not fully reflect rare cases or brand-specific contraindications.
- **Explainability Computation**: SHAP visualizations are resource-intensive for large batches.
- **Brand Preference Modeling**: Limited patient history can reduce personalization accuracy.

Despite these constraints, the proposed system balances effectiveness and explainability, making it a strong candidate for real-world use in clinics, pharmacies, or consumer health applications.

IV. Experimentation and Results

A. Dataset Splits and Configuration

The dataset utilized in this study comprised **12,000 anonymized medicine records**, each detailing drug name, brand, active ingredients (composition), medical condition targeted, price, and availability. These records were curated from synthetic and open-source pharmaceutical datasets to simulate a realistic recommendation environment. Stratified sampling was employed to maintain the proportion of unique compositions and brand variations across the splits. The dataset was divided into **80% training (9,600 records)** and **20% testing (2,400 records)**, ensuring balanced representation for both common and rare drug combinations.

B. Model Training and Hyperparameter Optimization

Model + XAI	Avg. Clarity Score (1–5)
RF + SHAP	4.5
XGB + SHAP	4.4
LR + LIME	4.2

A 5-fold cross-validation strategy was adopted for hyperparameter tuning.

- **Random Forest** achieved optimal results with n_estimators=150, max_depth=20, and min_samples_leaf=3.
- **XGBoost** performed best with learning_rate=0.1, n_estimators=120, max_depth=10, and colsample_bytree=0.8.
- **Logistic Regression** showed good generalization with penalty='l2' and C=1.0.

After tuning, all models were trained on the training set and evaluated on the held-out test data.

C. Performance Evaluation

Model	RMSE	MAE	R^2 Score
XGBoost	0.098	0.075	0.934
Random Forest	0.113	0.089	0.915
Logistic Reg.	0.159	0.125	0.812

XGBoost emerged as the most accurate, benefiting from gradient boosting's ability to handle heterogeneous feature types and complex interactions. **Random Forest** offered slightly lower accuracy but better generalization. **Logistic Regression** was the least accurate but fastest in both training and inference.

D. Interpretability and Explainability Analysis

To ensure transparent recommendations, SHAP and LIME were applied:

- **XGBoost + SHAP** revealed that features like **composition match score**, **brand reputation**, and **price proximity** had the highest influence on recommending alternative brands.
- **SHAP summary plots** and **force plots** helped visualize the impact of individual features on specific predictions.
- **Random Forest + SHAP** showed consistent feature importance with meaningful insights into how combinations affected the prediction.
- **Logistic Regression + LIME** provided simpler but interpretable local explanations for each recommended alternative, highlighting straightforward rules like "same ingredients but lower cost."

Clarity scores (from domain experts including pharmacists and clinicians) were highest for **RF + SHAP**, due to balanced interpretability and precision.

E. Key Observations and Trade-offs

- **Accuracy vs. Interpretability**: XGBoost offered top-tier accuracy but was harder to interpret, especially for end users like pharmacists. Logistic Regression was easiest to explain but less precise.
- **SHAP vs. LIME**: SHAP provided deeper insights, especially useful for sensitive recommendations, though it was computationally heavy. LIME was faster but varied slightly across runs.
- **Feature Engineering**: Key engineered features like **composition similarity index**, **brand trust score**, and **side-effect severity filter** consistently ranked high in all models.
- **Domain Trust**: Visual tools like SHAP force plots enhanced pharmacist trust and usability, making them a vital component in the deployment pipeline.
- **Model Suitability**: Random Forest with SHAP was identified as the best trade-off model for deployment in clinical recommendation settings due to its clarity and reliability.

V. CONCLUSION

The Medicine Recommendation System developed in this project addresses a critical gap in healthcare accessibility by providing users with alternative medicine suggestions that retain the same chemical composition but are offered under different brand names. This system is especially beneficial in the Indian pharmaceutical context, where multiple brands market the same formulation under varied pricing structures and brand recognition, often causing confusion or financial strain for consumers. By introducing a reliable, intelligent recommendation engine, this project empowers users to make informed decisions regarding their medication, based on factual data rather than brand familiarity or pharmacy constraints.The project combines a structured medicine database with an intelligent similarity-matching algorithm that ensures each recommended alternative has an identical composition to the queried medicine. This eliminates the risk of incorrect or suboptimal recommendations. The use of cosine similarity in composition matching allows for precise comparison even when ingredient formatting or naming conventions vary. Additionally, the integration of machine learning capabilities lays the groundwork for the system to evolve into a personalized recommendation engine, where suggestions can be further refined based on user preferences, previous history, or affordability filters.Another key strength of the system lies in its modular architecture, which separates data preprocessing, similarity matching, and recommendation presentation into distinct components. This allows easy updates, scalability, and future integration with hospital management systems, pharmacy chains, or telemedicine platforms. Furthermore, the use of open-source technologies such as Python, Flask, and MySQL ensures cost-effective development and deployment, while enabling future developers to expand the system without licensing barriers.From a technical standpoint, the system has demonstrated robustness and accuracy in test scenarios, successfully recommending alternatives that are compositionally equivalent, while offering a broader choice in terms of brand, pricing, and manufacturer. The interface is designed with simplicity in mind, ensuring that both medical professionals and general users can interact with it without requiring technical knowledge. This aligns with the project's overarching goal of democratizing access to medicine-related information and encouraging more transparent pricing in the pharmaceutical market.Moving forward, several opportunities exist to extend this work. First, integrating a live API for real-time medicine availability and pricing would significantly improve the system's practicality in day-to-day pharmacy operations. Second, incorporating patient-specific data—such as allergies, pre-existing conditions, or dosage history—can transform the system from a general recommender into a semi-clinical decision support tool. Third, embedding the system into mobile applications would enhance accessibility, especially for rural users who rely on smartphones as their primary digital platform. Finally, aligning the system with regulatory databases such as CDSCO (India) or WHO listings would ensure up-to-date compliance with approved medicine formulations.

In conclusion, the Medicine Recommendation System successfully bridges the gap between pharmaceutical data and end-user awareness by providing a fast, accurate, and transparent method of discovering generic and branded medicine alternatives. It holds the potential to positively impact public health by promoting affordability, enhancing medicine literacy, and supporting healthcare systems in reducing brand monopolies. As the digital healthcare landscape continues to evolve, this project lays a strong foundation for more intelligent, inclusive, and accessible health technology solutions.

REFERENCES

[1] M. Zomorodi, I. Ghodsollahee, J. H. Martin, N. J. Talley, V. Salari, P. Plawiak, K. Rahimi, and U. R. Acharya, "RECOMED: A Comprehensive Pharmaceutical Recommendation System," arXiv preprint arXiv:2301.00280, Jan. 2023.

[2] S. Garg, "Drug Recommendation System based on Sentiment Analysis of Drug Reviews using Machine Learning," arXiv preprint arXiv:2104.01113, Apr. 2021.

[3] G. Edwards, S. Nilsson, B. Rozemberczki, and E. Papa, "Explainable Biomedical Recommendations via Reinforcement Learning Reasoning on Knowledge Graphs," arXiv preprint arXiv:2111.10625, Nov. 2021.

[4] M. Sajde, H. Malek, and M. Mohsenzadeh, "RecoMed: A Knowledge-Aware Recommender System for Hypertension Medications," arXiv preprint arXiv:2201.05461, Jan. 2022.

[5] A. Holzinger, A. Carrington, and H. Müller, "Measuring the Quality of Explanations: The System Causability Scale

(SCS),” KI - Künstliche Intelligenz, vol. 34, no. 2, pp. 193–198, 2020.

[6] S. Saria et al., “Integration of Early Physiological Responses Predicts Later Illness Severity in Preterm Infants,” Science Translational Medicine, vol. 2, no. 48, pp. 48ra65, 2010.

[7] I. D. Dinov, *Data Science and Predictive Analytics: Biomedical and Health Applications Using R*, 2nd ed., Springer, 2023.

[8] A. Bihorac et al., “Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis,” IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 5, pp. 1589–1604, 2018.

[9] P. Courtiol et al., “Deep learning-based classification of mesothelioma improves prediction of patient outcome,” Nature Medicine, vol. 25, pp. 1519–1525, 2019.