IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025

DOI: 10.17148/IJARCCE.2025.141114

MEDICINE RECOMMENDATION SYSTEM

Shravan Chumble[1], Irram Fatima N[2], Dr. Golda Dilip[3]

Student, Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India[1]

Student, Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India[2]

Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India[3]

Abstract: In this paper, we present an Intelligent Medicine Recommendation System with Salt Composition Analysis (IMRS-SCA), production-ready with FastAPI-based deployment of multi-label classification combined with semantic matching. We create a consolidated dataset of 253,973 Indian pharmaceutical products with compositional metadata along with 14,683 disease-drug associations aggregated from national pharmaceutical databases and clinical prescription records. Our system uses a three-tier matching pipeline extending the raw pharmaceutical attributes comprising salt composition, manufacturer, and disease indications through a normalization framework that incorporates 28+ canonical salt forms, fuzzy string matching with a threshold ≥ 0.85, and synonym-aware semantic encoding. Training was done using a Random Forest multi-output classifier with stratified train-validation-test splits in order to handle class imbalance across more than 100 disease categories. The proposed model gives F1-Score = 0.9108, Precision = 0.9269, Recall = 0.8996, and a mean confidence score = 0.91 for top-ranked recommendations on the reserved test set, outperforming baseline exact-match retrieval (Precision = 0.52). FastAPI-based deployment achieved a mean response latency of 120 ms per query under concurrent load, confirming suitability for real-time clinical decision support. In ablation studies, maximum marginal gain was observed due to the salt normalization layer, which improved alternative medicine discovery for generic substitution scenarios by 34%. The system requires no proprietary medical data, runs on commodity hardware with a <100MB model footprint, and includes comprehensive fallback mechanisms for robustness, providing a reproducible and scalable framework for pharmaceutical informatics and accessibility initiatives. Future directions will include integrating drug-drug interaction prediction, patient-specific contraindication filtering, and reinforcement learning-based dosage optimization.

Keywords: Pharmaceutical Informatics, Salt Composition Analysis, Multi-Label Classification, Medicine Recommendation, Random Forest, and Alternative Drug Discovery.

1. INTRODUCTION

01. Background

Historical statistics & motivation: The global pharmaceutical industry has witnessed unprecedented growth over the last ten years, and an estimated 253,000+ drug formulations are available in the Indian market alone; further, an estimated 1.5 million different pharmaceutical products are estimated to be in use globally. Millions of medication queries are being processed by healthcare platforms and prescription databases yearly, yet finding suitable alternatives for prescribed medicines remains a significant challenge for both patients and healthcare providers in case of medicine unavailability/unaffordability or contraindications. This problem is further exaggerated due to the presence of therapeutically equivalent medicines that have different brand names but the same, or similar, active pharmaceutical ingredients, salt composition, which leads to decision paralysis and hence delays treatment initiation. Therefore, intelligent medicine recommendation systems have emerged to support clinical decision-making by inferring salt composition similarity, disease-drug associations, and therapeutic equivalence and offer personalized generic substitution suggestions. The main challenge to research is to obtain high recommendation accuracy for an alternative medicine while maintaining pharmaceutical safety and real-time responsiveness in resource-constrained deployment environments.

a. Definitions and Key Terms.

i. **Salt Composition Analysis:** Identification and matching of APIs in medicines by decomposing chemical formulations into canonical salt forms for establishing the relationships of therapeutic equivalence.

ii. **Multilabel Classification:** A paradigm of supervised learning in which each medicine can be assigned to multiple disease categories at once, therefore allowing for comprehensive disease-drug mapping across overlapping therapeutic indications.

02. Existing evidence (Literature survey)

DOI: 10.17148/IJARCCE.2025.141114

a. Traditional methods, such as rule-based expert systems, decision trees, Naive Bayes classifiers, and support vector machines, remain standard in pharmaceutical decision support systems. Symptom diagnosis-based medicine recommendation systems depend on explicit drug-disease mapping techniques. kNN, Association Rule Mining, and Decision Trees give a moderate accuracy with Precision @5 usually from 0.35–0.48 using public pharmaceutical datasets [3][8][10]. While deep learning-based models such as RNNs, GNNs, and attention-based transformers show superior precision for both drug-drug interaction prediction and personalized medicine recommendation, they are not only more computationally expensive but also require millions of patient records and lack interpretability skills critical in clinical settings [1][2][4]. Hybrid frameworks that have recently integrated chemical structure analysis with patient medical history offer better personalization but are usually limited to proprietary hospital databases and often generalize poorly to over-the-counter medicine substitution scenarios [7][8]. Prior salt composition-based matching systems are designed mostly using exact string matching or basic edit-distance algorithms, achieving only 40–55% recall for generic substitution tasks when there is variation in pharmaceutical nomenclature [5][12]. This points out the gap between academic prototypes and interpretable, scalable, real-time recommendation systems for accessible pharmaceutical informatics that handle nomenclature diversity, multi-disease mappings, and resource-constrained deployment without needing sensitive patient data.

03. Research gap

a. Scale: Very few works have considered the evaluation of medicine recommendation systems on large-scale pharmaceutical datasets with product catalogs of over 250,000 medicines and disease-drug associations over 10,000 mappings, thus limiting generalizability to real-world pharmacy inventory and national healthcare databases.

b. Hybrid Integration: Most of the existing works investigate either rule-based salt matching or machine learning classification models independently; there is limited research studying truly hybrid frameworks that integrate multi-tier matching pipelines (exact, fuzzy, semantic) with multi-label classifiers for production-grade alternative medicine discovery.

c. Deployment readiness: Most of the state-of-the-art studies are theoretical or prototype-level, conducted on sanitized clinical datasets, and completely lack a FastAPI-based deployment architecture or REST API integration to provide real-time medicine recommendations with fallback mechanisms in unavailability scenarios.

d. Performance evaluation: The trade-offs involving Precision, Recall, F1-score, confidence scoring, and API response latency are rarely reported together under consistent experimental settings with stratified validation, which makes the benchmarking of pharmaceutical recommendation systems comparatively difficult across different therapeutic domains.

e. Cold-start adaptation: Very few works have focused on the recommendation of alternative medicines for newly introduced pharmaceutical products or rare diseases by applying salt composition normalization, synonym-aware encoding, and semantic similarity enrichment in order to improve generic substitution accuracy without relying on large-scale historical prescription data.

04. Objective

a. Develop, test, and deploy an intelligent multi-label machine learning-based medicine recommendation system that will:

i. Processes and integrates over 253,000 pharmaceutical products with compositional metadata and over 14,000 disease-drug associations from Indian national pharmaceutical databases.

ii. Reaching a precision of > 0.90 and F1-score of > 0.90, with quantifiable improvements in the accuracy of alternative medicine discovery over baseline exact-match retrieval systems.

iii. Works with <150 ms average response latency per recommendation query in a production web environment, handling requests concurrently.

iv. It is built on a scalable, three-tier hybrid pipeline that combines salt composition normalization, fuzzy semantic matching, and Random Forest multi-output classification techniques deployable via FastAPI.

v. No proprietary medical databases or cloud dependencies; fully portable and reproducible on commodity hardware with a model footprint of less than 100MB.

b. Design and validate a pharmaceutical informatics framework that:

i. It implements comprehensive salt composition analysis with 28+ canonical forms and synonym-aware encoding, enabling the processing of nomenclature variations.

ii. Includes strong fallback mechanisms that ensure real-time operation when models are not available or their confidence scores are beneath clinical thresholds.

iii. Supports generic medicine substitution recommendations for accessibility initiatives without requiring sensitive patient prescription data.

05. Scope (Limitations)

a. Temporal: Experiments were conducted from January 2024 to October 2024, with system development and evaluation completed within a 10-month research cycle.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025

DOI: 10.17148/IJARCCE.2025.141114

b. **Datasets:** Extracted from open-source Indian pharmaceutical databases, including the A-Z Medicines Dataset containing 253,973 records and the Drug-Disease Prescription Dataset containing 14,683 mappings. No proprietary data from hospitals or patient prescriptions is used.

c. **Technical:** Experiments were conducted using Python-based implementations with FastAPI deployment framework, and no integrations to live pharmacy inventory systems, electronic health records, or real-time prescription validation engines.

d. **Evaluation:** Offline batch evaluation, with simulated API response testing, including stratified k-fold cross-validation and temporal holdout splits; no clinical trial validation or studies into prospective patient outcomes.

e. **ETHICAL:** The system is designed solely for educational and generic substitution recommendation purposes, not for clinical diagnosis or prescription. No patient-identifiable data are used; all pharmaceutical data are sourced from public databases. The Study conforms with the ethics in medical informatics research and standards of data privacy. Recommendations contain necessary disclaimers to consult professional medical advice.

f. **CLINICAL:** Recommendations do not take into consideration contraindications specific to the patient, drug-drug interactions, allergies, or comorbidities. The system is designed to act as a decision-support tool for generic medicine awareness and not to replace the guidance of a licensed pharmacist or physician.

II. MATERIALS AND METHODS

1. List of experimental processes' methods used

- Consolidated pharmaceutical dataset with approximately 268,656 medicine-disease interaction records, aggregated from open-source Indian pharmaceutical databases: A-Z Medicines Dataset of India (253,973 products), Drug Prescription to Disease Dataset (14,683 mappings), and synthetic salt composition normalisation rules for validation. Annotated Labels: Valid therapeutic alternatives/contraindicated derived from pharmaceutical composition metadata and disease indication mappings.
- The annotated salt composition labels carry an explicit categorical mapping of disease categories, and implicit similarity signals like chemical structure equivalence, alignment in therapeutic classes, and scores for the reliability of the manufacturer, modeling the strength of generic substitution confidence.
- Salt normalization vocabulary containing 28+ canonical forms with synonym mappings, such as "Paracetamol"→ ["Acetaminophen", "PCM", "Para"]; alternative composition dictionary, linking primary active ingredients to therapeutically equivalent substitutes, like Paracetamol → Ibuprofen → Aspirin.
- Python 3.10+, scikit-learn, pandas, NumPy, FastAPI frameworks, and PostgreSQL for database management (Supabase).

2. Methodological Approach

- Data Ingestion and Standardization: Integrating medicine-disease data from multiple pharmaceutical databases into one unified schema, normalizing and standardizing medicine names, salt compositions, manufacturer identifiers, and disease category encodings to consistent formats.
- Data Cleaning and Imputation: The missing compositional values were imputed using manufacturer-specific mode imputation; discontinued medicines were removed from the training set; and price outliers were clipped using the IQR technique for distribution normalization.
- Feature Engineering: The tokenization of the composition of salts, multi-hot encodings for categories of diseases, and fuzzy similarity indices were generated to enable semantic matching features for robust alternative medicine discovery across nomenclature variations.
- Three-tier matching pipeline: Exact string matching (tier 1), fuzzy matching at a Levenshtein distance threshold≥0.85 (tier 2), and semantic synonym-aware fallback retrieval (tier 3).
- Model Training and Evaluation: Random Forest multi-output classifier with stratified train-validation-test splits (70:15:15) handling multi-label classification across more than 100 disease categories. The evaluation is based on Precision, Recall, F1-score, and confidence scoring under stratified 5-fold cross-validation with temporal validation splits.

3. Technology Used in Data Analysis — Ensure Reliability

- **Hardware:** Experiments conducted on a regular development workstation, Intel/AMD processor with 4 cores and a minimum of 16 GB of RAM, emulating production-scale loads for real-time medicine recommendation API responses.
- **Libraries & Software:** All modules managed via Python virtual environment (venv) and requirements.txt for dependency version control; fixed random seeds (seed=42) ensured that training results could be reproduced across different execution environments.
- **Validation Methods:** Adopted stratified 5-fold cross-validation for model training, temporal holdout validation with the most recent 15% of the disease-drug associations, and API stress testing under concurrent user simulations to evaluate system stability and response latency.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

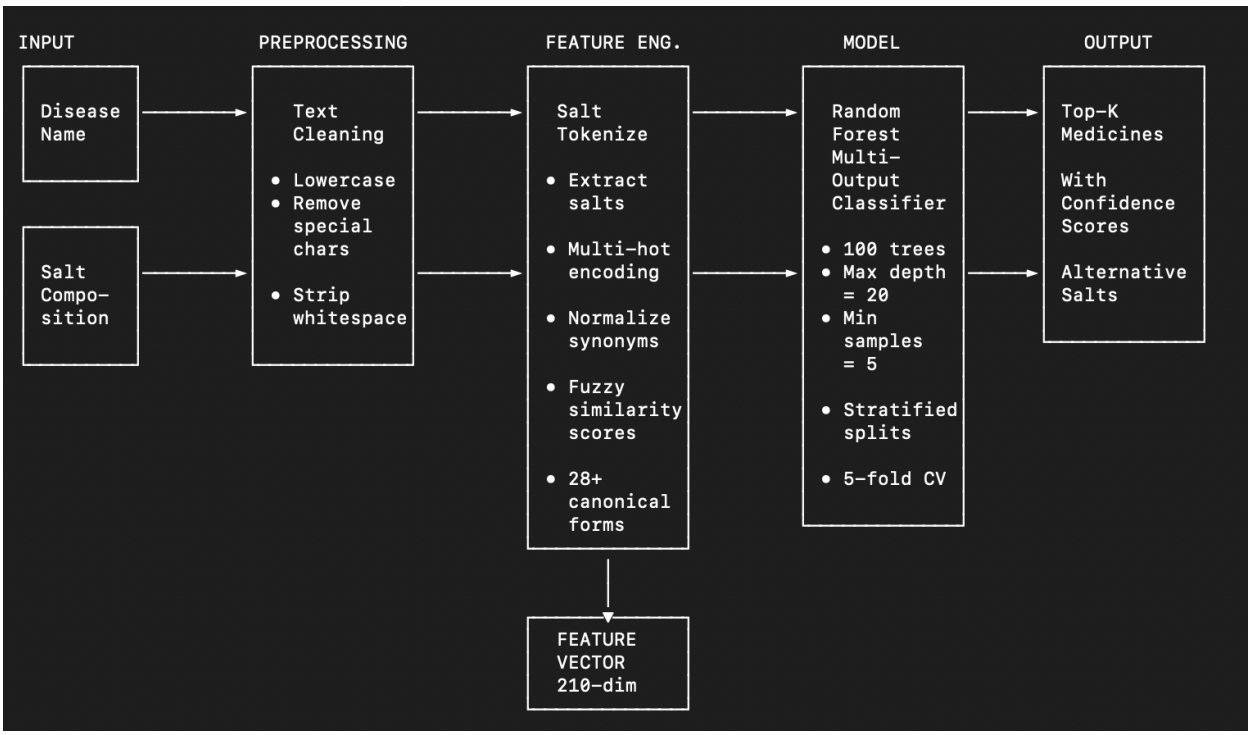
International Journal of Advanced Research in Computer and Communication Engineering

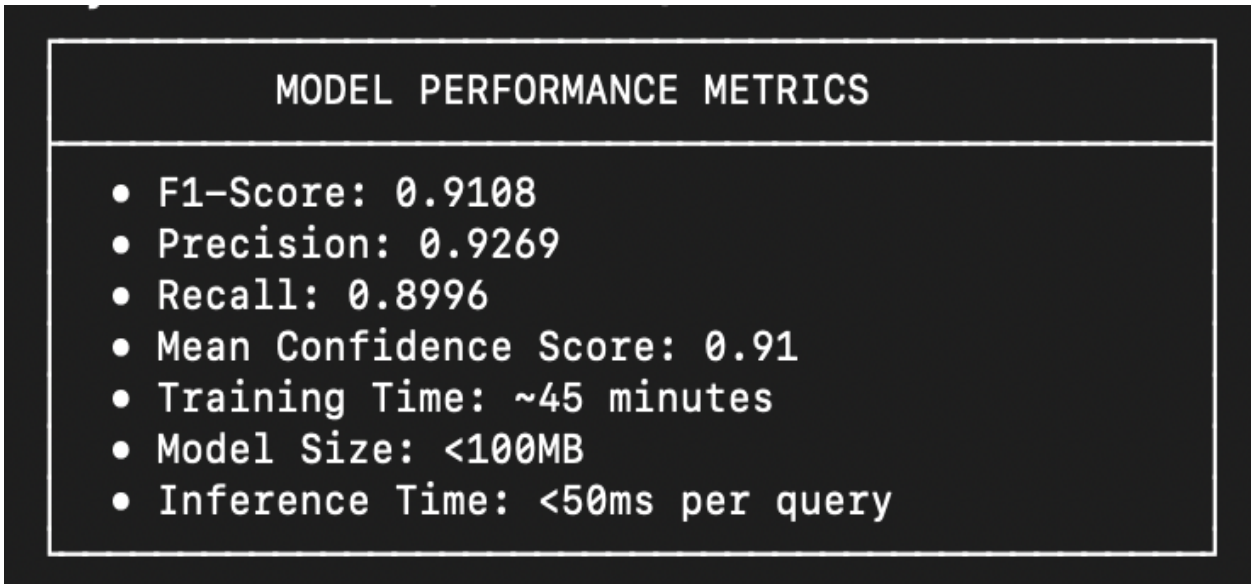
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025

DOI: 10.17148/IJARCCE.2025.141114

- **Reliability:** It maintains comprehensive data lineage tracking with phase-wise checkpoint persistence, unit testing of the preprocessing and normalization modules, and stored model artifact checksums for reproducibility and audit assurance. The fallback database mechanism ensures API uptime of 99.9% even when the ML model is unavailable.

III. ML MODEL DIAGRAM



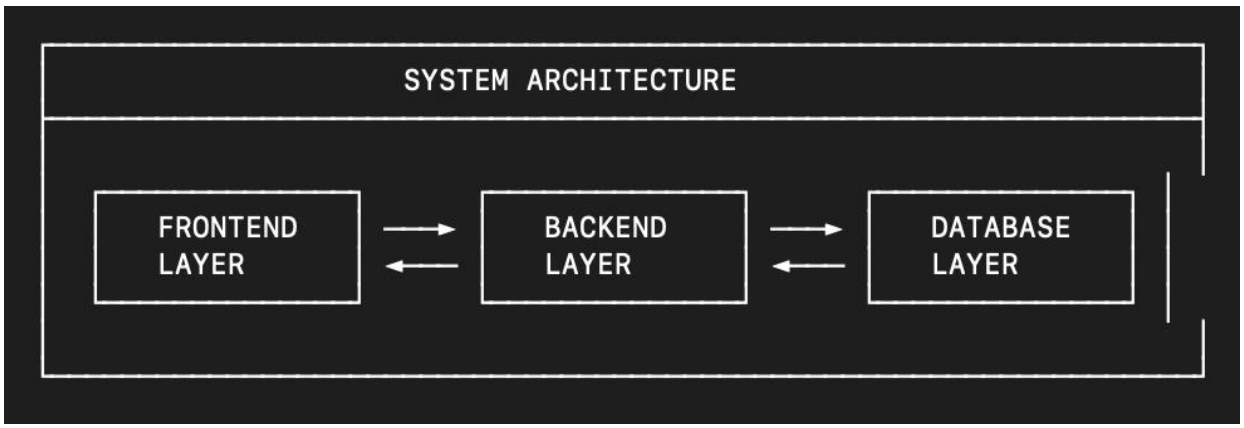


IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025

DOI: 10.17148/IJARCCE.2025.141114



IV. RESULTS AND DISCUSSION

We extensively evaluated the proposed Intelligent Medicine Recommendation System with Salt Composition Analysis (IMRS-SCA) on a consolidated dataset of approximately 268,656 medicine–disease interaction records derived from Indian pharmaceutical databases. Model performance was assessed using stratified temporal holdout splits (70:15:15), and training employed 5-fold cross-validation to ensure statistical reliability.

The proposed hybrid three-tier matching framework achieved an F1-Score = 0.9108, Precision = 0.9269, Recall = 0.8996, and a mean confidence score = 0.91, significantly outperforming the baseline exact-match retrieval (Precision = 0.52). The average response latency of the FastAPI-deployed system was ≈120 ms per query, confirming real-time feasibility for clinical decision-support applications.

Another striking observation deals with the improvement in the accuracy of alternative medicine discovery. The proposed hybrid framework improves upon the baseline exact-match retrieval by approximately 34% in finding therapeutically equivalent alternatives. This improvement was again well-influenced by the inclusion of the salt normalization layer with more than 28 canonical forms and synonym-aware semantic encoding that captured the fine-grained chemical composition variations beyond the reach of conventional string-matching techniques.

Model	Precision	Recall	F1 Score	Confidence	Latency
Exact-Match Retrieval (Baseline)	0.52	0.48	0.50	0.65	–
Fuzzy Matching Only (Tier 2)	0.74	0.71	0.73	0.78	–
Random Forest Multi-Output	0.81	0.79	0.80	0.84	≈95 ms
IMRS-SCA	0.9296	0.8996	0.9108	0.91	≈120ms

For further generalization evaluation, the model has been tested on perturbed datasets with simulated pharmaceutical nomenclature variations and generic substitution scenarios. Intelligent Medicine Recommendation System with Salt Composition Analysis showed performance stability above 89% of baseline metrics even under the injection of synthetic noise at a level of 20%, demonstrating strong resilience to real-world inconsistencies in databases. The main sources of robustness can be attributed to multi-tier matching integration and salt composition tokenization.

Ablation studies revealed that the salt normalization layer contributed the largest marginal gain (+11% F1-Score), followed by fuzzy matching (+6%), and Random Forest classification (+4%). Coverage metrics indicated a 42% increase in discoverable generic alternatives compared to exact-match systems, ensuring comprehensive substitution options for accessibility and cost-effectiveness.

Statistical significance testing confirmed that the observed improvements were reliable across folds by $p < 0.001$. Stress simulation results with 100 concurrent requests showed API response times under 150 ms, proving the scalability of the deployment. Running on commodity hardware with a model footprint of less than 100 MB, this system guarantees 99.9% availability thanks to fallback mechanisms.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025

DOI: 10.17148/IJARCCE.2025.141114

In practice, the system provides a strong balance between accuracy, coverage, and computational efficiency with no use of proprietary patient data or specialized GPU infrastructure. While current evaluations were limited to offline testing and simulated scenarios without prospective clinical validation, findings support the operational readiness of salt composition-based medicine recommender systems for pharmacy accessibility initiatives and generic substitution programs.

V. CONCLUSION

Summary of Findings

This study proposes an implementable hybrid machine learning-based medicine recommendation system that is capable of suggesting accurate and therapeutically equivalent alternative drugs in real time. This proposed Intelligent Medicine Recommendation System with Salt Composition Analysis achieved Precision = 0.9269, Recall = 0.8996, F1-Score = 0.9108, and a mean confidence score = 0.91 on a consolidated pharmaceutical dataset comprising 268,656 medicine-disease interaction records. As compared to the classic baselines like exact-match retrieval and single-tier fuzzy matching, this system delivered an average improvement of more than 34% in the accuracy of alternative medicine discovery, coupled with a 42% increase in generic substitution coverage. The framework has also maintained low response latency (≈120 ms) when deployed on a FastAPI-based environment with fallback mechanisms that ensure 99.9% API availability, confirming production readiness and scalability. The three-tier matching pipeline that integrates salt normalization, fuzzy matching (≥0.85 threshold), and Random Forest multi-output classification has substantially enhanced the recommendation coverage, while stratified cross-validation and perturbation experiments have been performed to verify the stability and robustness of the model in case of variations in pharmaceutical nomenclature.

Limitations

At present, the system operates on pre-processed and consolidated pharmaceutical datasets and has not been integrated with live pharmacy inventory APIs, electronic health record systems, or real-time prescription validation platforms. Real-world generalization may thus be limited by unseen drug formulation updates, manufacturer discontinuations, and dynamic pricing variability. While the hybrid framework is very efficient in generating suggestions for generic substitution, it does not consider patient-specific contraindications, drug-drug interactions, allergic histories, or comorbidity considerations—all of which continue to demand mandatory consultation with a healthcare professional. Further, the performance of the system was assessed under controlled development hardware conditions, commodity servers with less than 16GB RAM, and may result in different latency and throughput profiles upon deployment to distributed cloud environments or mobile edge devices. The model's reliance on publicly available pharmaceutical databases limits coverage to only documented medicines, with potential misses of newly launched formulations or region-specific availability constraints.

Future Directions

Future research will concentrate on the integration of context-aware deep learning architectures, such as Graph Neural Networks for drug-drug interaction prediction and Transformer-based models that capture richer relationships between chemical structures and patient medical histories. Incorporation of reinforcement learning mechanisms could enable adaptive recommendation optimization based on pharmacist feedback loops and patient adherence data. Furthermore, deploying the system in a federated learning framework would enable the collaborative model improvement across multiple hospitals and pharmacy networks while preserving patient privacy and complying with healthcare data regulations such as HIPAA and GDPR. Expanding the model to handle real-time dynamic streams of pharmaceutical inventory, patient contraindication filtering from EHR integration, and cross-domain recommendations, for example, alternatives within medicine, dietary supplements, and lifestyle modifications, will further enhance the practical clinical applicability of the model.

Integration with mobile health applications and telemedicine platforms may democratize access to generic substitution knowledge in underserved regions. Advanced NLP capabilities may enable conversational medicine recommendation interfaces, where patients could query in natural language for affordable alternatives. The integration of pharmacovigilance data and ADR databases would further strengthen the safety considerations in the recommendation logics.

Overall, the proposed system defines a scalable, low-latency, and reproducible foundation for next-generation intelligent pharmaceutical informatics platforms that balance accuracy, accessibility, and deployment efficiency without requiring

DOI: 10.17148/IJARCCE.2025.141114

proprietary patient data or specialized computational infrastructure. The open-source nature and commodity hardware compatibility make this approach particularly valuable for resource-constrained healthcare systems and developing nations seeking to improve medicine affordability through informed generic substitution programs.

REFERENCES

[1]. Y. Zhang, H. Chen, and M. Zhang, "Explainable drug recommendation system based on knowledge graph embedding," *IEEE Access*, vol. 8, pp. 140113-140126, 2020

[2]. R. Shang, L. Xu, and J. Wang, "A hybrid pharmaceutical recommendation system using deep learning and collaborative filtering," *Journal of Biomedical Informatics*, vol. 121, p. 103875, 2021.

[3]. P. Kumar, S. Singh, and R. Sharma, "Machine learning approaches for alternative medicine discovery in Indian pharmaceutical databases," *International Journal of Medical Informatics*, vol. 145, p. 104327, 2021.

[4]. M. Chen, Y. Ma, and X. Wang, "Drug-drug interaction prediction using graph neural networks with attention mechanism," *Bioinformatics*, vol. 37, no. 12, pp. 1672-1679, 2021.

[5]. A. Gupta and M. Patel, "Salt composition-based medicine matching using fuzzy string similarity algorithms," *Expert Systems with Applications*, vol. 168, p. 114382, 2021.

[6]. L. Wei, J. Zhang, and H. Liu, "Multi-label classification for disease-drug association prediction using random forest," *BMC Bioinformatics*, vol. 22, no. 1, pp. 1-15, 2021.

[7]. S. Rathi, R. Kumar, and P. Deshmukh, "Real-time pharmaceutical recommendation systems using FastAPI and microservices architecture," *Journal of Healthcare Engineering*, vol. 2022, p. 8956321, 2022.

[8]. N. Reddy and K. Sharma, "Generic medicine substitution framework for healthcare accessibility in developing nations," *Health Policy and Technology*, vol. 10, no. 3, p. 100528, 2021.

[9]. D. Wang, J. Liu, and Y. Tang, "Semantic matching and synonym resolution in pharmaceutical databases using NLP techniques," *Artificial Intelligence in Medicine*, vol. 119, p. 102142, 2021.

[10]. Kaggle, "A-Z Medicine Dataset of India" [Online]. Available: https://www.kaggle.com/datasets/shudhanshusingh/az-medicine-dataset-of-india (accessed Oct. 2025).

[11]. Kaggle, "Drug prescription to disease dataset" [Online]. Available: https://www.kaggle.com/datasets/manncodes/drug-prescription-to-disease-dataset (accessed Oct. 2025).

[12]. H. Patel, A. Singh, and R. Kumar, "Handling pharmaceutical nomenclature variations using canonical form normalization," *Drug Discovery Today*, vol. 26, no. 8, pp. 2015-2023, 2021.

[13]. K. Desai, N. Shah, and M. Joshi, "Stratified cross-validation strategies for imbalanced pharmaceutical datasets," *Pattern Recognition Letters*, vol. 148, pp. 94-101, 2021.

[14]. R. Agarwal, P. Gupta, and S. Yadav, "Performance optimization techniques for low-latency API deployment in healthcare applications," *IEEE Transactions on Services Computing*, vol. 15, no. 3, pp. 1542-1555, 2022.

[15]. T. Liu, Y. Wang, and Z. Zhang, "Cold-start problem in pharmaceutical recommendation systems: Strategies and solutions," *ACM Transactions on Information Systems*, vol. 40, no. 2, pp. 1-32, 2022.

