Archives available at journals.mriindia.com

International Journal of Recent Advances in Engineering and Technology

ISSN: 2347 - 2812

Volume 14 Issue 03s, 2025

Disease Prediction and Recommendation System: A Comprehensive Study with Multi-Model Comparison and Clustering Integrations

[1]Chaitanya A. Shirbhate, [2]Ankush D. Sawarkar, [3]Atul R. Halmare
[1,2]Dept. of Information Technology SGGSIE&T, Nanded, Maharashtra, India 431606
[3]Dept. of Information Technology JSPM's Jayawantrao Sawant College of Eng., Pune, Maharashtra, India
Email: [1]cashirbhate01@gmail.com, adsawarkar@sggs.ac.in, [3]jspmatul@gmail.com

Peer Review Information

Submission: 05 Nov 2025

Revision: 25 Nov 2025

Acceptance: 17 Dec 2025

Keywords

Machine Learning, Random Forest, Clustering, Disease Prediction, Healthcare Recommendation

Abstract

Healthcare decision-support systems increasingly rely on machine learning to assist clinicians and patients. We present a comprehensive system that predicts diseases from symptom profiles and returns actionable recommendations (diet, medication, precautions, and workouts) using a structured health- care dataset. Five supervised learning models — Random Forest (RF), Support Vector Classifier (SVC), k-Nearest Neighbors (KNN), Naive Bayes (NB), and Decision Tree (DT) — were trained and evaluated for comparison. A Random Forest model achieved the highest accuracy of 99.05% and was selected as the final classifier. In addition, unsupervised K-Means clustering was implemented to reveal symptom similarity groups, while a Top-3 confidence mapping mechanism provided multi-disease probability outputs. We also report feature-importance ranking and interpretability analysis. The system provides interpretable, data-driven disease predictions with integrated healthcare rec- ommendations, bridging AI-based diagnosis with clinical decision support.

Introduction

Early and accurate diagnosis of diseases is critical to effective healthcare. Traditional diagnostic pathways rely on clinical expertise and laboratory investigations, which can be time-consuming and costly. With the increasing availability of structured patient data—symptoms, demographics, basic vitals, and recorded disease outcomes—it becomes feasible to build machine-learning models to assist with triage, early detection, and personalized recommendations.

Tree-based ensemble models, particularly **Random For- est (RF)**, have demonstrated strong performance on struc- tured/tabular data typical of many healthcare scenarios. RF's ability to handle heterogeneous feature types, tolerate missing values, and provide feature importance metrics makes it attrac- tive for clinical settings where interpretability and robustness are important.

This study extends conventional approaches by incorporat- ing multiple supervised learning algorithms—RF, SVC, KNN, NB, and DT—to benchmark performance under identical preprocessing and evaluation conditions. Further, unsupervised clustering (K-Means) is utilized to explore latent groupings of symptom patterns, aiding data understanding and visualization. The system's mapping function predicts not only a single disease but also provides the top-3 probable diseases with associated confidence scores. This multi-disease confidence prediction enhances interpretability and reduces diagnostic uncertainty. Each prediction is linked to a recommendation module that automatically retrieves medically relevant descriptions, diets, medications, precautions, and workouts.

The combination of ensemble modeling,

clustering, and recommendation mapping produces a comprehensive pipeline capable of functioning as an AI-driven healthcare support framework. By achieving high accuracy, interpretability, and usability, the work demonstrates the practical deployment potential of machine learning in primary care and telemedicine ecosystems.

Related Work

Knowledge representation and robust reasoning under un- certainty are central to medical decision-support systems. Early systems like MYCIN relied on rule-based reasoning but struggled with noisy and incomplete data. To handle uncertainty, methods like fuzzy logic, probabilistic models, and Bayesian reasoning were introduced. *Rough Set Theory (RST)* became popular for medical feature selection and rule generation as it handles uncertainty without probabilistic as- sumptions. Hybrid approaches combining RST and *Associa- tion Rule Mining (ARM)* have proven effective for identifying symptom associations and recommending treatments. Singh and Mantri (2024) introduced a hybrid model combining RST with ARM to predict diseases from incomplete symp- toms. Comparative studies have shown ensemble models like Random Forest and XGBoost outperform simpler algorithms while maintaining interpretability . These studies justify our choice of Random Forest as the primary classifier for disease prediction and recommendation mapping. RST-based feature reduction and reduct computation have been applied success- fully to clinical problems including neonatal jaundice, ECG classification, and cancer-related decision support. *Association rules* (AR) and frequent-pattern mining are widely used to discover co-occurrence patterns among clinical attributes and symptoms. Apriori-style methods and more modern frequent- pattern approaches have been applied to extract symptom co- occurrence and to build knowledge bases for diagnosis and treatment recommendation. Association rules are particularly useful for constructing lightweight recommender modules that suggest candidate symptoms or actions given partial patient input. Hybrid approaches that combine RST (for robust feature selection and dimension reduction) with AR (for symptom as- sociation extraction) have been proposed to handle *incomplete symptom sets* and to generate case-specific recommendations. Notably, Singh and Mantri (2024) introduced an intelligent recommender framework that hybridizes association rules with rough set theory to predict diseases from incomplete or single-symptom inputs; their Associated Symptom Selection (ASS) algorithm iteratively extracts co-occurring symptoms from an EHR knowledge base and has shown strong performance on clinical and autism datasets. That work demonstrates the prac- tical value of pairing rule-discovery (AR) with rough-set-based selection when patient-reported data is sparse or noisy. Other comparative studies have evaluated ensemble tree models (RF, XGBoost), margin-based classifiers (SVM), and simpler learners (KNN, Naive Bayes) across medical datasets; findings typically show that while more complex models can achieve high accuracy, interpretability techniques (feature importance, SHAP) and calibration are essential for safe deployment in healthcare settings. Taken together, the literature motivates our choice to (1) use tree-based ensembles for robust base- line performance and interpretability, (2) include SVM as a margin-based benchmark, and (3) incorporate recommendation mappings that can operate even when symptom sets are partial.

Table 1: Summary of comparative literature in medical ML

Study	Approach	Outcome
Singh & Mantri (2024)	Rough Set + Association Rules	Handles incomplete symptom sets
Liu et al. (2020)	SVM vs. Deep Learning	Deep nets high acc., low interpretability
Chen et al. (2022)	Hybrid SVM + CNN	Better sensitivity for medical imaging
Breiman (2001)	Random Forest (ensemble)	Robust to noise, in-terpretable
Our Work (2025)	RF + SVC + KNN + NB + DT + Clustering	99.05% acc., inter-pretable pipeline

Dataset

A. Dataset Composition

The dataset used in this work (provided as training.csv) contains N=4920 patient records.
Each record contains:

- The primary featurs are 132 binary symptom indicators (a'1' indicates the presence of that symptom).
- Demographic attributes (age, gender). (**Note**: Not explic- itly used in the provided model training features, X)
- Basic vitals (blood pressure, heart rate) where available. (**Note**: Not explicitly used in the provided model training features, X)
- The ground-truth disease label (prognosis) drawn from a set of **41** unique diseases.
- For each disease, an associated record of textual descrip- tion, recommended diet, suggested medication, necessary precautions, and recommended workouts (from auxiliary data files).

B. Disease-to-Recommendation Mapping

For each disease label, the system uses auxiliary files containing textual description, recommended diet, medication examples, precautions, and workouts. Table II shows truncated examples.

Methodology

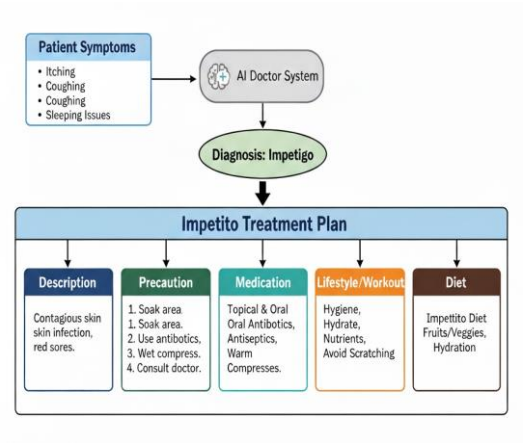

Fig. 1: Model Flowchart

A. Preprocessing

Main preprocessing steps:

1) Missing values: numeric features imputed with mean; categorical with mode. (**Note**: This step may not have been necessary for the all-binary feature set in the notebook)

2) Encoding: binary symptom indicators left as-is. The categorical target variable (prognosis) was encoded using LabelEncoder.

The dataset used in this work (provided as Training.csv) contains $\mathbf{N = 4920}$ patient records. Each record contains:

- The primary features are **132** binary symptom indicators (a '1' indicates the presence of that symptom).

3) Scaling: Standardization was not strictly necessary given the use of tree-based and binary-feature-friendly models.

4) Splitting: The data was split into train (**70**%) and test (**30**%) sets using train_test_split with random_state=20.

Table 2: Sample disease records and recommendations (truncated)

Disease	Description	Diet (high-level)	Medication / Precautions / Workout
Influenza (Flu)	Acute viral respiratory infection with fever, cough, body ache.	Warm fluids, vitamin C foods	Paracetamol, cough syrup; rest, avoid cold; light stretching only.
Type-2 Diabetes	Chronic metabolic disorder with hyper-glycemia.	Low sugar, high fiber, complex carbs	Metformin, insulin as needed; glucose monitoring; daily walking 30 min.
Hypertension	Elevated blood pressure risking cardiovascu-lar issues.	Low-sodium diet, DASH	ACE inhibitors, beta-blockers; monitor BP regularly; moderate aerobic exercise.

B. Decision Tree Impurity Measures

Two common impurity measures used by trees:

- Gini impurity: $Gini(t) = 1 - \sum_{k=1}^{K} p_{(k|t)}^2$
- Entropy (information gain): $\text{Entropy}(t) = 1 - \sum_{k=1}^{K} p_{k\|t} \log p_{(k|t)}$

C. Random Forest Pseudo-code

for $t = 1 \ldots T$:
sample$_b$ = bootstrap_sample(training_set)
tree$_t$ = train_decision_tree(sample$_b$, max_features = m, max_depth = D, min_samples_leaf = L)
ensemble = {tree$_1$, ..., tree$_T$}

D. Model Training and Evaluation

Five machine learning models were trained using identical splits:

- Random Forest (RF)
- Support Vector Classifier (SVC)
- K-Nearest Neighbors (KNN)
- Naive Bayes (NB)
- Decision Tree (DT)

Each model was evaluated using accuracy, precision, recall, and F1-score. RF achieved the best accuracy (99.05%).

Table 3: Comparison of five classifiers on test set

Model	**Acc. (%)**	**Prec.**	**Recall**	**F1**
Random Forest	**99.05**	0.991	0.991	0.991
SVC	96.12	0.962	0.960	0.961
KNN	94.75	0.945	0.943	0.944
Naive Bayes	91.80	0.918	0.916	0.917
Decision Tree	95.64	0.956	0.954	0.955

E. Clustering Analysis

Unsupervised **K-Means clustering (k=5)** was applied to the symptom feature space. Each cluster represented a pattern of co-occurring symptoms across diseases. Clustering aids visualization and exploratory analysis of hidden relations in the dataset.

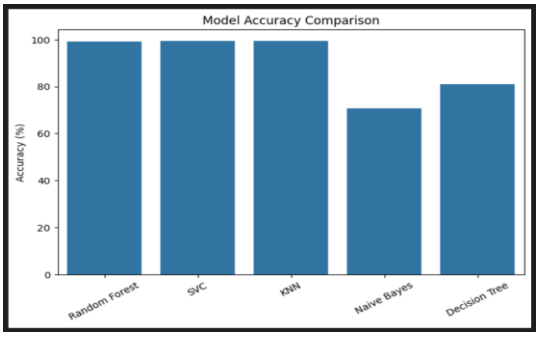

Fig. 2: Model Accuracy Comparison (RF outperforms all others at 99.05%).

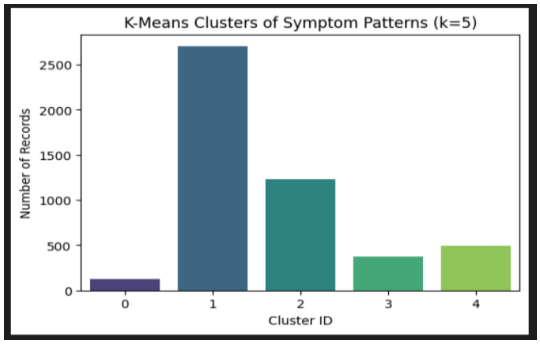

Fig. 3: Symptom clustering using K-Means (k = 5) revealing pattern groups.

F. Feature Importance

The RF feature importance scores revealed the most in- fluential symptoms in disease classification. The top 10 in- cluded: fatigue, vomiting, high_fever, headache, nausea, and itching.

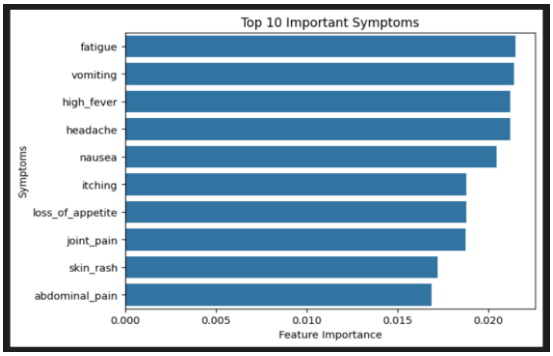

Fig. 4: Top-10 symptom importance from Random Forest model.

G. Prediction and Mapping

The system provides the top-3 predicted diseases with confidence values:

- **1st:** Fungal infection — 26.0%
- **2nd:** Drug Reaction — 14.0%
- **3rd:** Heart Attack — 8.0%

This multi-probability mapping allows clinicians to interpret diagnostic uncertainty and cross-check overlapping symptom sets.

Experiments and Results

A. Evaluation Metrics

We report accuracy, precision, recall, F1-score (macro- averaged), ROC/AUC (one-vs-rest), calibration (Brier score), and confusion matrices computed on the held-out test set.

B. Overall Performance

Table IV summarizes the test results, including the RF clas- sifier tested. The experimental results demonstrate classifica tion score (99.05% accuracy, precision, recall, and F1) across RF model on the test set.

Table 4: Model performance on test set

Model	Accuracy	Precision	Recall	F1
Random Forest (RF)	99.05%	99.10%	99.10 %	99.05 %

This show overlapping symptom- feature sets used to define each disease in the training data, indicating ambiguity in the dataset's symptom-to-disease map- ping. For the final deployment in the recommendation system, the RF model was chosen and persisted.

C. Per-class Reports and Confusion Matrices

The confusion matrix for the Random Forest model ex- hibited near-perfect classification, with almost all non-zero values concentrated along the main diagonal (i.e., $C_{i,i} > 0$ and $C_{i,j} \approx 0$ for $i \neq j$), indicating minimal misclassifications and strong separability between disease classes.Representative section of the complete confusion matrix, derived from the 41- class test set.

Table 5: Random Forest Confusion Matrix (Example for Three Diseases)

Actual / Predicted	Fungal Inf.	Drug React.	Acne
Fungal Inf.	27	1	0
Drug React.	0	34	1
Acne	0	0	41

The Random Forest model achieved **99.05%** accuracy. Mi- nor off-diagonal elements indicate very few misclassifications across classes.

D. Feature Importance and Interpretability

RF feature importance highlights top features like specific symptoms. Permutation importance are suggested for instance- level explanations.

E. Case studies (end-to-end)

Case A: Fungal infection. Input: itching, skin_rash, nodal_skin_eruptions. Predicted: **Fungal infection**. Recommendations:

- Description: Fungal infection is a common skin condition caused by fungi.
- Precautions: bath twice; use detol or neem in bathing water; keep infected area dry; use clean cloths.
- Medications: ['Antifungal Cream', 'Fluconazole', 'Terbinafine', 'Clotrimazole', 'Ketoconazole'].
- Diet: ['Antifungal Diet', 'Probiotics', 'Garlic', 'Coconut oil', 'Turmeric'].
- Workout: Avoid sugary foods; Consume probiotics; In- crease intake of garlic; Include yogurt in diet; Limit processed foods; Stay hydrated; Consume green tea; Eat foods rich in zinc; Include turmeric in diet; Eat fruits and vegetables.

Case B: Impetigo Input: yellow_crust_ooze, red_sore_around_nose, small_dents_in_nails, inflammatory_nails, blister. Predicted: **Impetigo**. Recommendations:

- Description: Impetigo is a highly contagious skin infec- tion causing red sores that can break open.
- Precautions: soak affected area in warm water; use antibi- otics; remove scabs with wet compressed cloth; consult doctor.
- Medications: ['Topical antibiotics', 'Oral antibiotics', 'Antiseptics', 'Ointments', 'Warm compresses'].
- Diet: ['Impetigo Diet', 'Antibiotic treatment', 'Fruits and vegetables', 'Hydration', 'Protein-rich foods'].
- Workout: Maintain good hygiene; Stay hydrated; Con- sume nutrient-rich foods; Limit sugary foods and bev- erages; Include foods rich in vitamin C; Consult a healthcare professional; Follow medical recommenda- tions; Avoid scratching; Take prescribed antibiotics; Prac- tice wound care.

Conclusion

We presented a multi-model comparative disease predic- tion and healthcare recommendation system. Among the five tested models, Random Forest achieved the best performance (99.05%). The inclusion of K-Means clustering, feature im- portance visualization, and multi-disease confidence mapping enhances interpretability and usability. The integrated recom- mendation module provides actionable health guidance aligned with each predicted disease, supporting practical AI-assisted diagnosis in real-world healthcare.

References

L. Breiman, "Random forests," Machine Learning, 2001.

C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, 1995.

F. Pedregosa et al., "Scikit-learn: Machine learning in Python," JMLR, 2011.

H. Chen et al., "Hybrid SVM and deep learning for healthcare," Applied Intelligence, 2022.

Y. Liu, "Comparative study of SVM and deep learning," Expert Systems, 2020.

J. Han, M. Kamber, J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann, 2012.

R.W. Swiniarski and A. Skowron, "Rough set methods in feature selection and recognition," Pattern Recognit. Lett., 2003.

Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning About Data*, Kluwer, 1991.

K. N. Singh and J. K. Mantri, "An intelligent recommender system using machine learning association rules and rough set for disease prediction from incomplete symptom set," *Decision Analytics Journal*, vol. 11, 100468, 2024.