Disease Prediction And Drug Recommendation Using Machine Learning

*Note: Sub-titles are not captured in Xplore and should not be used

1st Abu Said Akunji
Department of Computer Science
Kalinga Institute of Industrial Technology
Bhubaneswar, Odisha, India
tamimakunji@gmail.com

2nd Komalika Das
Department of Computer Science
Kalinga Institute of Industrial Technology
Bhubaneswar, Odisha, India
komalikadas0025@gmail.com

***Abstract*—The analysis and exploration of extensive datasets require the application of data mining techniques to identify meaningful patterns and trends. One critical domain where these techniques prove valuable is medical databases. Across the globe, many individuals face challenges related to health conditions and medical diagnoses. Although hospital information systems (HIS) generate vast amounts of data, extracting useful insights from diagnostic records remains a complex task. By simply inputting their symptoms, patients can swiftly receive insights into potential illnesses along with suitable medications, leveraging the methodologies proposed in this study. This research presents a system that recommends medications based on symptoms provided by users. To achieve disease prediction, four distinct machine learning algorithms—Decision Tree Classifier, Random Forest Classifier, Naïve Bayes Classifier, and Support Vector Machine (SVM)—are employed for symptom analysis. Furthermore, a Collaborative Filtering technique is implemented to suggest appropriate medications. A detailed discussion of each model and methodology is provided in this paper. The experimental results obtained from this study can serve as a foundation for future research and various applications in the medical field.**

***Index Terms*—Data mining, disease prediction, drug recommendation, Decision Tree, Random Forest, Naïve Bayes, Support Vector Machine (SVM), Collaborative Filtering**

I. Introduction

A recommendation system is generally defined as a system designed to predict how a user would rate a particular item and subsequently ranks these predictions accordingly. Prominent technology companies such as Google, Instagram, Spotify, Amazon, Reddit, and Netflix widely employ such systems. Based on a user's profile, a recommender model can evaluate whether a particular product or service would be preferred by that user. These models offer advantages to both service providers and users by reducing the effort and cost associated with discovering and selecting products in an online marketplace. The application of recommendation systems spans various domains, including medication recommendation platforms, product suggestions for e-commerce, playlist generation for media streaming services, and content recommendations for social networks. The core functionality of these systems is focused on numerically estimating a user's preference for items they have not yet interacted with, with the overarching goal of helping users identify relevant options.

One of the most common concerns individuals face when dealing with medical conditions is selecting a reliable healthcare provider. It is well established that a person's overall well-being significantly influences their quality of life. According to a 2013 study conducted by the Pew Internet and American Life Project, approximately 58.99% of Americans have searched for health-related information online, with 35.6% focusing on self-diagnosing medical conditions. As public interest in health and medical diagnoses continues to grow, concerns regarding medical errors also persist, leading to substantial fatalities.

Research indicates that medication-related errors account for over 200,000 deaths annually in China and more than 100,000 in the United States. Notably, approximately 42% of these errors are attributed to physicians prescribing medications based on their limited personal experience. Therefore, selecting a competent medical professional for diagnosis and treatment is a crucial decision for patients. The advancement of data mining and recommendation technologies has opened new possibilities for leveraging insights from medical diagnosis records, patient reviews, and medication ratings. These advancements aim to support healthcare providers in prescribing appropriate treatments while mitigating the risk of medication errors.

This study focuses on the development and implementation of a global disease prediction and drug recommendation framework that integrates multiple data mining methodologies. Various predictive algorithms and recommendation strategies are employed to aggregate and analyze data from diverse sources. The subsequent sections of this paper elaborate on data collection, preprocessing techniques, methodology, experimental results, conclusions, and potential directions for future research.

II. Related Work

In recent years, the use of intelligent systems for recommending treatments and diagnosing illnesses has grown significantly. These systems aim to deliver tailored medical

advice to healthcare providers and patients, enhancing clinical outcomes. A major focus has been the integration of artificial intelligence (AI) and machine learning (ML) to personalize recommendations. While AI/ML models can effectively identify suitable treatments and predict outcomes, challenges remain in evaluating their reliability and accuracy.

Several studies have explored different approaches. Gupta et al. employed machine learning models such as Decision Trees, Random Forests, and Naive Bayes to design a system for disease diagnosis and treatment recommendation, achieving up to 98% accuracy. Bao et al. developed a universal medicine recommender using SVM, ID3, and neural networks, with SVM showing the best performance. Zhang et al. proposed a hybrid model combining ANN and Case-Based Reasoning to aid GPs in prescribing.

Research by Kononenko et al. and Olsen et al. validates the role of ML in improving diagnostic precision and lowering healthcare costs, especially for heart-related conditions. Hussein et al. introduced a highly accurate clinical diagnosis system using random forest classifiers.

Rustam et al. presented a real-time diagnostic and preventive model that reached 99.9% accuracy. Bhat and Aishwarya created a hybrid recommender with a 75% success rate for newly launched drugs. Feldman et al. stressed the need for scalable AI systems in personalized care.

Zhang et al. also built a real-time prediction system using the VFDT stream mining method, tackling issues like data complexity and size. Austin et al. compared flexible ML models like bagging and boosting for predicting heart failure with improved outcomes over traditional methods.

In the area of cardiac care, AbuKhousa et al. reviewed multiple predictive data mining models, highlighting issues like poor generalization due to limited data. Tran et al. provided an overview of recommender systems across various healthcare services, offering structured design insights.

For diabetic patients, Morales et al. developed a system using clustering and collaborative filtering, achieving moderate accuracy. Zhang et al. introduced iDoctor, a hybrid matrix factorization-based system using sentiment analysis and topic modeling to refine recommendations.

Lastly, Kuanr et al. applied multiple classifiers, including GBM and Decision Trees, to predict cervical cancer, with strong results from tree-based methods. Han et al. proposed a hybrid model for matching patients with doctors, outperforming traditional systems in accuracy.

III. Dataset and Pre-processing

The effectiveness of a machine learning-based disease diagnosis and recommendation system depends largely on the quality and structure of the dataset. In this work, a comprehensive dataset is utilized, consisting of multiple components that contribute to both disease prediction and personalized treatment recommendations. The dataset includes six key sections:

- Symptoms Dataset – A mapping of diseases to their associated symptoms.
- Disease Descriptions – Brief explanations of each disease.
- Medication Dataset – Recommended medicines for each diagnosed disease.
- Precaution Dataset – Suggested precautions for patients with specific conditions.
- Advise Plans – Advises of routines tailored for managing various diseases.
- Diet Plans – Diet ideas for managing various diseases.

A. Symptoms Dataset

The Symptoms Dataset serves as the core of the disease prediction system, as it links each disease with a list of symptoms that are indicative of its presence.

a) Handling Missing Symptoms: Some disease entries in the dataset may have missing or incomplete symptom data. To ensure consistency, missing symptoms are handled through the following approaches:

- Imputation by Mode: Missing symptom values for a given disease are imputed with the most frequently occurring symptom in the dataset. This ensures that the disease still has a complete set of symptoms for prediction, while maintaining logical consistency.
- Default Value Assignment: In cases where symptoms are completely missing, a default symptom value (such as 0 or a placeholder value) is assigned, indicating a lack of information. This helps the model recognize incomplete entries without discarding them entirely.

Disease	Symptom_1	Symptom_2	Symptom_3	Symptom_4
Fungal infection	itching	skin_rash	nodal_skin_eruptions	dischromic _patches
Fungal infection	skin_rash	nodal_skin_eruptions	dischromic _patches	NaN
Fungal infection	itching	nodal_skin_eruptions	dischromic _patches	NaN
Fungal infection	itching	skin_rash	dischromic _patches	NaN
Fungal infection	itching	skin_rash	nodal_skin_eruptions	NaN

Fig. 1. Before data processing Symptoms Dataset

- Removing Irrelevant or Redundant Symptoms: To reduce noise and improve model efficiency, any irrelevant or redundant symptoms are removed from the dataset. Symptoms that are too general (e.g., ”pain”, ”discomfort”) or those that appear frequently across many diseases without providing distinguishing features are excluded from the final dataset. This step ensures that the model focuses on the most relevant and discriminative symptoms for each disease.
- Normalization and Scaling: For machine learning algorithms that are sensitive to the scale of input features (such as distance-based models), normalization techniques are applied. In this case, each symptom value is scaled to a range, ensuring that no symptom disproportionately influences the model due to its scale.

B. Disease Description Dataset

The Disease Descriptions Dataset provides brief textual explanations of various diseases, offering essential background

disease	symptom_1	symptom_2	symptom_3	symptom_4
Fungal infection	itching	skin_rash	nodal_skin_eruptions	dischromic _patches
Fungal infection	skin_rash	nodal_skin_eruptions	dischromic _patches	Unknown
Fungal infection	itching	nodal_skin_eruptions	dischromic _patches	Unknown
Fungal infection	itching	skin_rash	dischromic _patches	Unknown
Fungal infection	itching	skin_rash	nodal_skin_eruptions	Unknown

Fig. 2. After data processing Symptoms Dataset

information for users. While this dataset is not directly used for disease prediction, it plays a crucial role in enhancing the interpretability of the system by giving users meaningful explanations alongside diagnostic results. To ensure consistency and usability, the following preprocessing steps are applied:

a) Text Cleaning and Standardization: Since the dataset contains free-text descriptions of diseases, text preprocessing techniques are applied to ensure uniform formatting and eliminate inconsistencies. The key steps include:

- Lowercasing: All text is converted to lowercase to maintain consistency.
- Removing Special Characters and Punctuation: Unnecessary symbols such as @, #, $, %, &, * are removed, as they do not contribute to meaningful descriptions.
- Expanding Contractions: Common contractions (e.g., ”it is ” $\rightarrow$ ”it is”, ”don’t” $\rightarrow$ ”do not”) are expanded to improve readability and standardization.

Disease	Description
Fungal infection	Fungal infection is a common skin condition ca...
Allergy	Allergy is an immune system reaction to a subs...
GERD	GERD (Gastroesophageal Reflux Disease) is a di...
Chronic cholestasis	Chronic cholestasis is a condition where bile ...
Drug Reaction	Drug Reaction occurs when the body reacts adve...
Peptic ulcer disease	Peptic ulcer disease involves sores that devel...
AIDS	AIDS (Acquired Immunodeficiency Syndrome) is a...
Diabetes	Diabetes is a chronic condition that affects h...

Fig. 3. Description Dataset

C. Medication Dataset

The Medication Dataset contains recommended medicines for each disease, serving as the foundation for the medicine recommendation system. Since the recommendation model relies on Collaborative Filtering, the dataset must be structured to support similarity-based recommendations.

D. Precaution Dataset

The Precaution Dataset provides preventive measures for each disease, helping users manage conditions effectively.

Disease	Medication
Fungal infection	['Antifungal Cream', 'Fluconazole', 'Terbinafi...
Allergy	['Antihistamines', 'Decongestants', 'Epinephri...
GERD	['Proton Pump Inhibitors (PPIs)', 'H2 Blockers...
Chronic cholestasis	['Ursodeoxycholic acid', 'Cholestyramine', 'Me...
Drug Reaction	['Antihistamines', 'Epinephrine', 'Corticoster...
Peptic ulcer disease	['Antibiotics', 'Proton Pump Inhibitors (PPIs)...
AIDS	['Antiretroviral drugs', 'Protease inhibitors'...
Diabetes	['Insulin', 'Metformin', 'Sulfonylureas', 'DPP...

Fig. 4. Medication Dataset

Since precautions are text-based and vary in format, preprocessing ensures consistency, clarity, and structured retrieval. Below are the key steps applied:

- Standardizing Precaution Text To ensure uniform formatting, the following text-cleaning steps are applied:
  1) Lowercasing: Converts all precaution text to lowercase for consistency.
  2) Removing Special Characters & Extra Spaces: Eliminates unnecessary punctuation, symbols, and excess spaces to maintain clarity.
  3) Expanding Abbreviations: Standardizes common medical abbreviations (e.g., ”BP” $\rightarrow$ ”blood pressure”, ”HR” $\rightarrow$ ”heart rate”).

Disease	Precaution_1	Precaution_2	Precaution_3	Precaution_4
Drug Reaction	stop irritation	consult nearest hospital	stop taking drug	follow up
Malaria	Consult nearest hospital	avoid oily food	avoid non veg food	keep mosquitos out
Allergy	apply calamine	cover area with bandage	NaN	use ice to compress itching
Hypothyroidism	reduce stress	exercise	eat healthy	get proper sleep
Psoriasis	wash hands with warm soapy water	stop bleeding using pressure	consult doctor	salt baths
GERD	avoid fatty spicy food	avoid lying down after eating	maintain healthy weight	exercise
Chronic cholestasis	cold baths	anti itch medicine	consult doctor	eat healthy
hepatitis A	Consult nearest hospital	wash hands through	avoid fatty spicy food	medication

Fig. 5. After processing Precautions Dataset

E. Advise Plan Dataset

The Advise Plan Dataset provides specific guidance for individuals diagnosed with various diseases. This dataset includes recommendations on what to do and what to avoid to manage symptoms effectively and prevent complications. Unlike the Precaution Dataset, which primarily focuses on general preventive measures, the Advise Plan Dataset offers more detailed, disease-specific guidance.

F. Diet Plan Dataset

The Diet Dataset plays a crucial role in disease management by providing tailored nutritional recommendations for various conditions. Proper dietary habits can significantly impact recovery, symptom control, and overall well-being.By providing disease-specific dietary recommendations, the system ensures that users receive nutrition guidance aligned with their medical

Unnamed: 0	disease	workout
0	Fungal infection	Avoid sugary foods
1	Fungal infection	Consume probiotics
2	Fungal infection	Increase intake of garlic
3	Fungal infection	Include yogurt in diet
4	Fungal infection	Limit processed foods
5	Fungal infection	Stay hydrated
6	Fungal infection	Consume green tea
7	Fungal infection	Eat foods rich in zinc

Fig. 6. Advise plan Dataset

condition.Each individual may require different nutritional needs based on their disease. The Diet Dataset allows for personalization, helping users adopt a diet suited to their health condition.

The Diet Dataset provides nutritional recommendations tailored to different diseases. It specifies the types of essential foods that can help manage symptoms and promote recovery. This dataset serves as a guideline for patients to follow a diet that complements their medical treatment.

Disease	Diet
Fungal infection	['Antifungal Diet', 'Probiotics', 'Garlic', 'C...
Allergy	['Elimination Diet', 'Omega-3-rich foods', 'Vi...
GERD	['Low-Acid Diet', 'Fiber-rich foods', 'Ginger'...
Chronic cholestasis	['Low-Fat Diet', 'High-Fiber Diet', 'Lean prot...
Drug Reaction	['Antihistamine Diet', 'Omega-3-rich foods', '...
Peptic ulcer disease	['Low-Acid Diet', 'Fiber-rich foods', 'Ginger'...
AIDS	['Balanced Diet', 'Protein-rich foods', 'Fruit...
Diabetes	['Low-Glycemic Diet', 'Fiber-rich foods', 'Lea...

Fig. 7. Diet Dataset

IV. Methodology

The methodology of this system focuses on two key components: **Disease Prediction and Medicine Recommendation**. The approach leverages machine learning models to diagnose diseases based on symptoms and uses collaborative filtering techniques to suggest appropriate medications. Additionally, the system is deployed using Flask to provide real-time accessibility.

To construct an effective Disease Prediction and Drug Recommendation system, it is essential to address two fundamental subcategories:

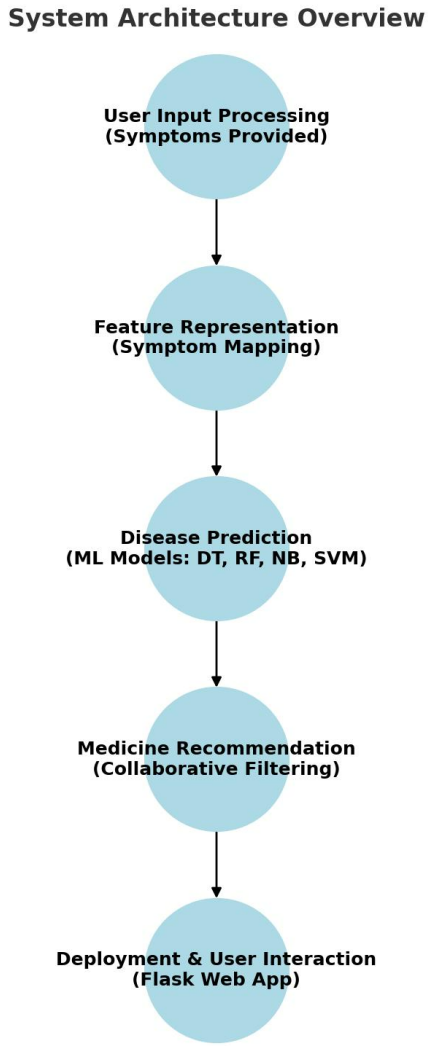

Fig. 8. Design implementation pipeline and dataflow of the Prediction & Recommendation System

A. *Disease Prediction*

A probabilistic model that predicts diseases based on observed symptoms serves as the foundation for the disease prediction system. For this purpose, we utilize the Disease-Symptom Knowledge Database, which comprises 132 symptoms associated with over 40 distinct diseases. In this study, various methodologies have been explored to enhance the accuracy of disease prediction. A key challenge arises from the dataset's single-feature nature, where symptoms are the sole input. Training a classifier with only one type of information often reduces its robustness, ultimately impacting prediction accuracy when incorporating additional variables.

To address this limitation, we aim to generate multiple data points for each disease by considering all relevant symptoms, their respective importance, and their frequency of occurrence. We propose four distinct strategies, each designed to optimize different aspects of disease prediction. These approaches are

evaluated using four classification models, and their accuracy and predictive performance are compared to identify the most effective technique. The following sections provide a detailed overview of the implemented methodologies:

1) APPROACH 1: DECISION TREE (DT): The Decision Tree (DT) is a widely used classification algorithm that operates through a hierarchical structure, enabling step-by-step decision-making based on input symptoms. It works by recursively splitting the dataset into smaller subsets based on symptom-based conditions until a final classification (disease) is reached.

In this approach, each node in the tree represents a symptom, and each branch represents a decision based on that symptom's presence or absence. The tree continues to split until it reaches a terminal node, which corresponds to a specific disease. For instance, if a patient reports fever and cough, the tree follows a path that considers these symptoms, leading to a probable classification such as flu or pneumonia. If additional symptoms like body aches or fatigue are present, the tree refines the diagnosis further.

This approach is particularly useful in disease classification because many medical diagnoses rely on step-by-step symptom assessment. The model effectively translates this process into an automated system that can provide quick and transparent predictions, aiding in preliminary diagnosis and decision support.

2) APPROACH 2: RANDOM FOREST (RF): The Random Forest (RF) approach is an advanced ensemble learning technique that improves the accuracy and reliability of disease classification. Unlike a single decision tree, which can sometimes overfit the training data, Random Forest constructs multiple decision trees and combines their outputs to generate a final prediction. This ensemble strategy enhances the model's ability to generalize and handle complex symptom-disease relationships.

Random Forest operates by creating a collection of decision trees, each trained on a randomly selected subset of the dataset. During prediction, each tree votes on the most likely disease based on the given symptoms. The final classification is determined by majority voting, meaning the disease with the most votes from different trees is selected as the predicted outcome. For example, if a patient reports symptoms such as cough, fever, and fatigue, some trees might classify the condition as influenza, while others may suggest bronchitis. The disease with the highest number of votes is chosen as the final diagnosis, reducing the likelihood of errors caused by a single misclassified decision tree.

In the context of disease prediction, Random Forest offers a balanced trade-off between accuracy and generalization. Since medical symptoms often have overlapping patterns across multiple diseases, using multiple decision trees helps improve classification confidence. The ability of Random Forest to process complex, multi-symptom data makes it a strong approach for reliable and practical disease diagnosis in real-world applications.

3) APPROACH 3: NAIVE BAYES (NB): The Naïve Bayes (NB) algorithm is a probabilistic classification approach based on Bayes' theorem, which calculates the likelihood of a disease occurring given a set of symptoms. It assumes that all symptoms contribute independently to the final classification, making it computationally efficient and well-suited for medical diagnosis.

Naïve Bayes calculates the probability of each possible disease based on the symptoms provided by the user. It determines the likelihood of a disease occurring by multiplying the probability of each symptom appearing in cases of that disease. The disease with the highest probability is then selected as the predicted diagnosis.

For example, if a patient reports fever, cough, and body aches, the model evaluates the probability of each disease based on past occurrences in the dataset. If flu has a high probability of causing these symptoms, it is chosen as the most likely diagnosis. Mathematically, the probability of a disease D given a set of symptoms S_1, S_2, …, S_n is calculated as:

$$P(\frac{D}{S_1, S_2, \ldots, S_n}) = \frac{P(D) \cdot P(\frac{S_1}{D}) \cdot P(\frac{S_2}{D}) \cdots P(\frac{S_n}{D})}{P(S_1, S_2, \ldots, S_n)} \quad (1)$$

Naïve Bayes is particularly useful when quick predictions are needed based on a structured dataset. Its probabilistic nature makes it highly interpretable and valuable for disease classification tasks, especially in scenarios where multiple possible diagnoses need to be considered. By using probability-based reasoning, the model offers a reliable, lightweight, and effective method for disease diagnosis in a medical recommendation system.

4) APPROACH 4: SUPPORT VECTOR MACHINE (SVM): The Support Vector Machine (SVM) is a powerful classification algorithm that works by finding the optimal boundary (hyperplane) that separates different disease categories based on input symptoms. It is particularly useful when dealing with complex datasets where symptoms do not have a clear linear relationship with diseases.

SVM maps the symptom data into a higher-dimensional space and then finds the best possible hyperplane that separates different disease classes. If symptoms from different diseases are difficult to distinguish, SVM uses a kernel trick to transform the data into a form where clear separation is possible. For example, if a patient reports headache, nausea, and dizziness, the model analyzes these symptoms and positions them in a multi-dimensional space. The algorithm then determines which disease category the symptoms fall into by identifying the closest decision boundary.

Mathematically, SVM aims to maximize the margin between disease categories while minimizing classification errors. The decision boundary is defined as:

$$w \cdot x + b = 0 \quad (2)$$

SVM is a strong choice for medical diagnosis because it can handle both linear and non-linear relationships between

symptoms and diseases. Its ability to create a clear separation between different disease categories makes it highly accurate for classification tasks. By integrating SVM into the disease prediction module, the system can provide precise and reliable diagnoses, even in cases where symptoms overlap between multiple diseases.

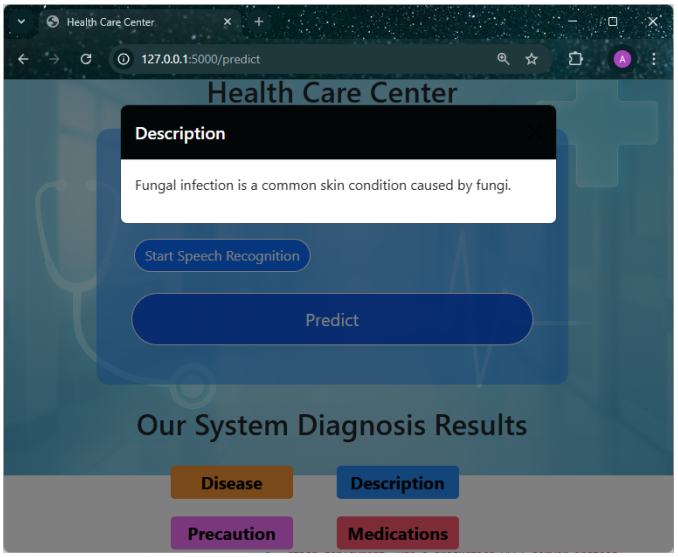

Fig. 9. The Disease Prediction and Description Interface

here the user input was a textual script : "I have cough and itching."

B. *Medicine Recommendation*

After diagnosing a disease, the next step is to recommend suitable medications for treatment. Traditional approaches rely on manually curated mappings between diseases and medicines, which can be rigid and may not adapt well to new medical data. To enhance the recommendation process, this system employs Item-Based Collaborative Filtering (IBCF), a technique that identifies relationships between diseases and their commonly prescribed medicines based on historical data.

Since the dataset contains disease names, symptoms, and corresponding medicine names, the recommendation system works by identifying patterns in which medicines are prescribed together for specific diseases. Instead of analyzing patient behavior, IBCF determines how often specific medicines appear for the same disease and makes recommendations based on these associations.

For example, if the dataset shows that Disease A is commonly treated with Medicine X and Medicine Y, the model learns that these two medicines are frequently prescribed together. When a user is diagnosed with Disease A, the system will recommend both Medicine X and Medicine Y, even if only one was initially considered.

The similarity between medicines is determined by how often they co-occur within the same disease records. The system can use measures like cosine similarity or Jaccard similarity to quantify how related two medicines are based on their presence in the dataset. Since the dataset contains disease names, symptoms, and corresponding medicines, Item-Based Collaborative Filtering is well-suited for this system. It identifies patterns in how medicines are prescribed for different diseases and makes recommendations based on past data rather than predefined rules. This allows for a dynamic, scalable, and data-driven approach to medicine recommendation, ensuring that users receive relevant treatment suggestions based on historical prescription trends. It also recommends the precaution, diet, advises for the disease predicted. according to the disease the data are referred and recommended with the medicine recommendation.

Algorithm: Drug Recommendation System using Collaborative Filtering:

Input:

- x: Disease

Output:

- Recommended drugs
- Disease and corresponding side effects
- Medication
- Precautions
- Advises to avoid or recover from the disease
- Diet Suggestions

Steps:

- Extract dataset rows corresponding to the given disease.
- Identify medicines associated with the extracted disease from the dataset.
- Calculate similarity scores between medicines based on their co-occurrence in different diseases.
- Rank medicines based on their frequency of association with the diagnosed disease.
- Filter out duplicate entries and retain the most relevant medicines.
- Encode ratings from the side effects dataset.
- Generate a list of recommended drugs.
- Retrieve additional disease-related information, including:
  a. Description of the disease to provide insights into its causes and symptoms.
  b. Medication details explaining how each prescribed medicine helps.
  c. Precautionary measures to prevent worsening of the condition.
  d. Recovery advice with guidelines on how to improve health and manage the disease.
  e. Diet suggestions to support faster recovery and maintain overall well-being.
- Display the complete recommendation, including the medicines, disease-related details, and recovery plan.

There are some other Modals that are recommended by clicking the button of that particular feature. Such as Description, Precaution, Workouts and Diets

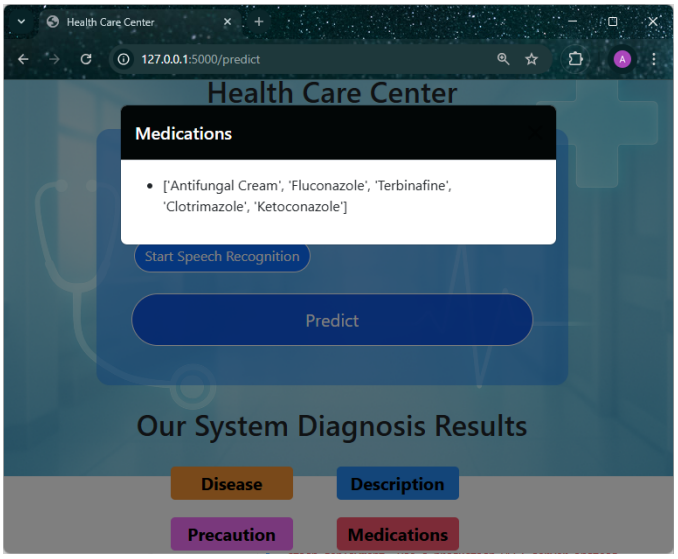

Fig. 10. The Medication Recommendation Interface

V. Experimental Results

A. Disease Prediction:

a) APPROACH 1: Decision Tree: The Decision Tree model was evaluated for disease prediction using a dataset containing symptoms and corresponding diseases. The data was split into training and testing sets to ensure balanced learning. The model effectively mapped symptoms to diseases, achieving high accuracy while maintaining fast computation. Performance was measured using accuracy, precision, recall, and F1-score. While the Decision Tree provided clear and interpretable decision paths, it showed signs of overfitting, performing slightly better on training data than on unseen test cases. A sample test case with symptoms like fever, sore throat, and fatigue successfully predicted influenza, demonstrating the model's effectiveness. However, its accuracy can be improved by applying pruning techniques to reduce overfitting and enhance generalization.

b) APPROACH 2: Random Forest: The Random Forest model was tested for disease prediction by training on a dataset containing symptoms and corresponding diseases. By combining multiple decision trees, it improved accuracy and reduced overfitting compared to a single Decision Tree. The dataset was divided into training and testing sets, and performance was evaluated using accuracy, precision, recall, and F1-score. The model demonstrated strong generalization, handling new symptom patterns more effectively. In a test case with symptoms like high fever, body aches, and chills, the model accurately predicted influenza. While Random Forest provided reliable predictions, it required more computational resources than simpler models. Overall, it proved to be a robust approach for disease classification.

c) APPROACH 3: Naive Bayes: The Naïve Bayes model was applied to disease prediction, leveraging its probabilistic approach to classify diseases based on symptoms. The dataset was divided into training and testing sets to assess its performance using accuracy, precision, recall, and F1-score. Despite its assumption of feature independence, the model performed well on structured symptom data, offering fast and efficient predictions. When tested with symptoms like nausea, stomach pain, and vomiting, it correctly identified food poisoning as the most likely condition.

d) APPROACH 4: SVM: The Support Vector Machine (SVM) model was used for disease prediction, leveraging its ability to classify diseases by finding optimal decision boundaries between symptom patterns. The dataset was split into training and testing sets, and performance was evaluated using accuracy, precision, recall, and F1-score. SVM effectively handled high-dimensional data and showed strong generalization, making accurate predictions even for complex symptom combinations. In a test case with symptoms like persistent cough, chest pain, and fatigue, the model correctly predicted pneumonia. While SVM demonstrated high accuracy, it required more computational power and longer training time compared to simpler models. Despite this, it proved to be a reliable choice for disease classification, particularly in handling diverse symptom data.

B. Comparison of Performance Values:

Table 1 shows the accuracy values of each approach as described above.

Classifier Model	Accuracy (%)
Decision Tree	97.46
Random Forest	99.19
Naive Bayes	100.00
SVM	100.00

TABLE I
ACCURACY COMPARISON OF CLASSIFIER MODELS

C. Drug Recommendation:

In this project, medicine recommendations were generated based on the predicted disease using Item-Based Collaborative Filtering. The final recommendation was determined by selecting medicines associated with the identified disease from the dataset.

To ensure relevant suggestions, multiple factors were considered before recommending a medication. Since different medicines can be prescribed for the same condition, the system relied on dataset-based filtering rather than subjective reviews, ensuring that only appropriate and frequently used medications were suggested. Additionally, along with the recommended medicines, precautionary measures, recovery advice, dietary recommendations, and a brief disease description were provided to enhance treatment guidance.

a) APPROACH : Item Based Collaborative Filtering: The medicine recommendation system was tested using **Item-Based Collaborative Filtering**, where symptoms were taken as input, and the trained models predicted the most likely disease. Based on the predicted disease, the system retrieved relevant medicines from the dataset along with essential health-

related information, including disease description, precautions, recovery advice, and dietary recommendations.

Performance evaluation involved verifying whether the recommended medicines aligned with standard treatments for the predicted disease. A test case where symptoms such as fever, cough, and fatigue were input resulted in the system predicting influenza and suggesting appropriate medications along with supportive recovery guidelines.

The system effectively provided accurate medicine recommendations and additional health guidance. However, its performance depended on the accuracy of disease prediction. Future improvements could involve refining similarity calculations and integrating patient feedback to enhance personalization. Overall, the approach proved to be a reliable method for recommending medicines and health management strategies based on predicted diseases.

VI. Discussion

All the functions implemented in this study have been consolidated into a single file, incorporating the most accurate techniques. The program first predicts diseases based on input symptoms, which then serve as input for the drug recommendation system. The system subsequently provides the recommended medication along with a list of its potential side effects as output.

VII. Conclusion and Future Scope

Drug recommendation systems are widely used in modern online services, and as the demand for such systems continues to rise, automation has become essential. To address this need, we developed a medication recommendation system. The key findings from our project are as follows:

- Successfully designed a drug recommendation framework that suggests medications along with potential adverse effects based on user-provided symptoms.
- Implemented three distinct models: one for sentiment analysis, another for disease prediction, and a third for drug recommendation.
- Evaluated multiple strategies for each of these models.
- Achieved high accuracy across all three models, enhancing the reliability of the overall recommendation system.
- A crucial future direction for this work is to further improve the accuracy of disease prediction and drug recommendation.

Acknowledgment

We express our sincere gratitude to our mentor and institution for their invaluable guidance and support throughout this research. Their insights and encouragement were instrumental in shaping our work. We also extend our appreciation to the contributors of publicly available datasets, which played a crucial role in our study. Lastly, we acknowledge the efforts of our peers for their constructive feedback, helping us refine our approach to disease prediction and drug recommendation using machine learning.

References

[1] F. O. Isinkaye, Y. O. Folajimi, and B. A. Ojokoh, "Fundamentals, methodologies, and assessment criteria of recommendation systems," Egyptian Informatics Journal, vol. 16, no. 3, pp. 261–273, Nov. 2015.

[2] M. A. N. Banu and B. Gomathy, "Utilizing data mining techniques for disease prediction systems," International Journal of Technical Research and Applications, vol. 1, no. 5, pp. 41–45, 2013.

[3] H. Wang, Q. Gu, J. Wei, Z. Cao, and Q. Liu, "Enhancing drug repositioning through mining drug-disease interactions: The intersection of recommendation systems and genome-wide association studies," Clinical Pharmacology and Therapeutics, vol. 97, no. 5, pp. 451–454, May 2015.

[4] S. A. Alsaif, M. S. Hidri, I. Ferjani, H. A. Eleraky, and A. Hidri, "A bi-directional NLP-powered recommendation system for job seekers and recruiters," Big Data and Cognitive Computing, vol. 6, no. 4, p. 147, Dec. 2022.

[5] J. P. Gupta, A. Singh, and R. K. Kumar, "Machine learning-driven disease prediction and drug recommendation framework," International Journal of Advanced Research in Engineering and Technology (IJARET), vol. 12, no. 3, pp. 673–683, 2021.

[6] Y. Bao and X. Jiang, "Framework for an intelligent medicine recommender system," Proceedings of the IEEE 11th Conference on Industrial Electronics and Applications (ICIEA), Jun. 2016, pp. 1383–1388.

[7] Q. Zhang, G. Zhang, J. Lu, and D. Wu, "Hybrid recommendation system framework for personalized clinical prescriptions," Proceedings of the 10th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), Nov. 2015, pp. 189–195.

[8] U. Bhimavarapu, N. Chintalapudi, and G. Battineni, "Ensuring fairness and safety in emergency drug recommendations using a stacked artificial neural network approach," Algorithms, vol. 15, no. 6, p. 186, May 2022.

[9] J. Chen, K. Li, H. Rong, K. Bilal, N. Yang, and K. Li, "Integrating big data mining and cloud computing for disease diagnosis and treatment recommendations," Information Sciences, vol. 435, pp. 124–149, Apr. 2018.

[10] I. Kononenko, I. Bratko, and M. Kukar, "Leveraging machine learning for medical diagnosis applications," Machine Learning and Data Mining: Methods and Applications, vol. 389, p. 408, Jun. 1997.

[11] C. R. Olsen, R. J. Mentz, K. J. Anstrom, D. Page, and P. A. Patel, "Applying machine learning techniques for heart failure diagnosis, classification, and prognosis," American Heart Journal, vol. 229, pp. 1–17, Nov. 2020.

[12] A. S. Hussein, W. M. Omar, X. Li, and M. Ati, "Chronic disease diagnosis prediction and recommendation system for efficient healthcare solutions," Proceedings of the IEEE-EMBS Conference on Biomedical Engineering and Science, Dec. 2012, pp. 209–214.

[13] F. Rustam, Z. Imtiaz, A. Mehmood, V. Rupapara, G. S. Choi, S. Din, and I. Ashraf, "A machine learning-based system for automated disease diagnosis and precautionary recommendations," Multimedia Tools and Applications, vol. 81, no. 22, pp. 31929–31952, Sep. 2022.

[14] S. Bhat and K. Aishwarya, "Developing an item-based hybrid recommender system for newly introduced pharmaceutical drugs," Proceedings of the International Conference on Advanced Computing, Communication, and Informatics (ICACCI), Aug. 2013, pp. 2107–2111.

[15] K. Feldman, D. Davis, and N. V. Chawla, "Scaling personalized healthcare through disease prediction algorithms: A case study," Journal of Biomedical Informatics, vol. 57, pp. 377–385, Oct. 2015.

[16] Y. Zhang, S. Fong, J. Fiaidhi, and S. Mohammed, "A real-time clinical decision support system leveraging data stream mining," Journal of Biomedical Biotechnology, vol. 2012, pp. 1–8, May 2012.

[17] P. C. Austin, J. V. Tu, J. E. Ho, D. Levy, and D. S. Lee, "Employing machine learning techniques for disease classification and prediction: A heart failure case study," Journal of Clinical Epidemiology, vol. 66, no. 4, pp. 398–407, Apr. 2013.

[18] E. AbuKhousa and P. Campbell, "Clinical decision support via predictive data mining: A review of heart disease prediction systems," Proceedings of the International Conference on Innovations in Information Technology (IIT), Mar. 2012, pp. 267–272.

[19] L. F. G. Morales, P. Valdiviezo-Diaz, R. Reátegui, and L. Barba-Guaman, "Developing and evaluating a collaborative filtering and clustering-based drug recommendation system for diabetes," Journal of Medical Internet Research, vol. 24, no. 7, Jul. 2022, Art. no. e37233.