Machine Learning Approach

1. The Type of Machine Learning

Supervised Learning (multi-label classification)

AI is given symptoms (input features) and must predict possible diseases (labels).

Training data: historical cases where patients reported symptoms and doctors confirmed conditions.

That means we already know the correct outputs during training → this is the definition of supervised learning.

Relevance:

Fits perfectly since we want the chatbot to "learn from examples" and generalize to new users.

Each patient may have multiple possible conditions at once → multi-label supervised learning is ideal.

2.Machine learning Algorithms

A) Logistic Regression

Function: Models the probability that a condition is present given the symptoms.

Strengths:

Probabilistic output: Can say "flu = 72%, migraine = 15%", which is perfect for suggesting "top 3 conditions" in the chatbot.

Interpretability: We can inspect coefficients to see which symptoms contributed most to predicting flu, which is important for medical explainability.

Scales well: Works efficiently with large numbers of features (like hundreds of symptoms).

Relevance:

Patients want probabilities, not just yes/no answers. Logistic Regression provides this in a medically meaningful way.

Easy to communicate back to users: "flu is more likely because you reported fever + cough + fatigue."

B) Random Forest

Function: Uses many decision trees to capture complex, non-linear relationships between symptoms and diseases.

Strengths:

Handles non-linearities & interactions: Example: “fever + cough + chest pain” → pneumonia (but fever alone does not). Logistic Regression might miss such non-linear combinations.

Robust to noise: If a patient forgets to mention one symptom, Random Forest still works well.

Feature importance: Can rank which symptoms are most informative overall (e.g., “shortness of breath” is highly predictive for heart conditions).

Relevance:

Medical symptoms rarely act alone; it’s often combinations that matter. Random Forest captures this better.

Works well on tabular clinical data (symptom flags, demographics, duration).