# Classification in Medicine

Common algorithms for diagnosis and prediction tasks

### Logistic Regression

Linear model for binary/multi-class classification with probability outputs

- ✓ Highly interpretable
- ✓ Fast training
- ✗ Assumes linearity
- ✗ Limited complexity

### Random Forests

Ensemble of decision trees for robust, non-linear classification

- ✓ Handles non-linearity
- ✓ Feature importance
- ✗ Less interpretable
- ✗ Can overfit

### Support Vector Machines

Maximum margin classifier with kernel tricks for non-linear boundaries

- ✓ Effective in high-dim
- ✓ Versatile kernels
- ✗ Slow on large data
- ✗ Hard to interpret

### Neural Networks

Deep learning models for complex pattern recognition

- ✓ Highest performance
- ✓ Automatic features
- ✗ Black box
- ✗ Needs large data

## 1. Logistic Regression

**Overview:** Logistic regression is a fundamental statistical method that models the probability of a binary outcome using a logistic (sigmoid) function. Despite its name, it's a classification algorithm that outputs probabilities between 0 and 1.

### How It Works:

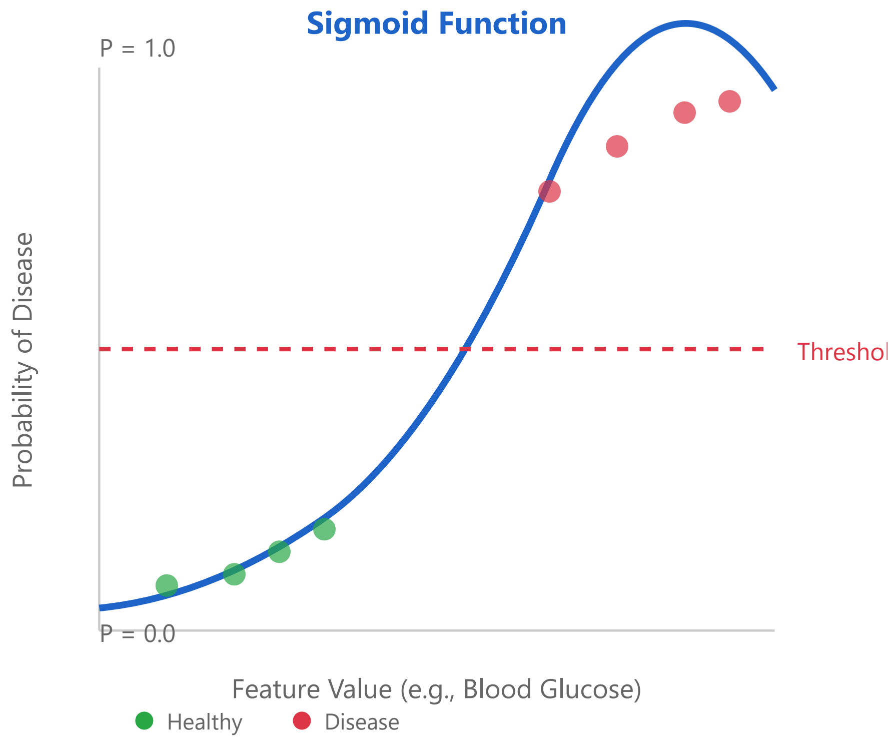
- Uses a linear combination of input features
- Applies sigmoid function: $P(y=1) = 1 / (1 + e^{(-z)})$
- Classifies based on probability threshold (typically 0.5)
- Optimized using maximum likelihood estimation

#### Medical Application Example:

**Disease Risk Prediction:** Predicting diabetes risk based on age, BMI, blood glucose, and family history. Each coefficient shows how much each factor increases or decreases the log-odds of diabetes. *Example:* If age coefficient = 0.05, each year increases diabetes odds by $e^{0.05} \approx 1.05$ times.

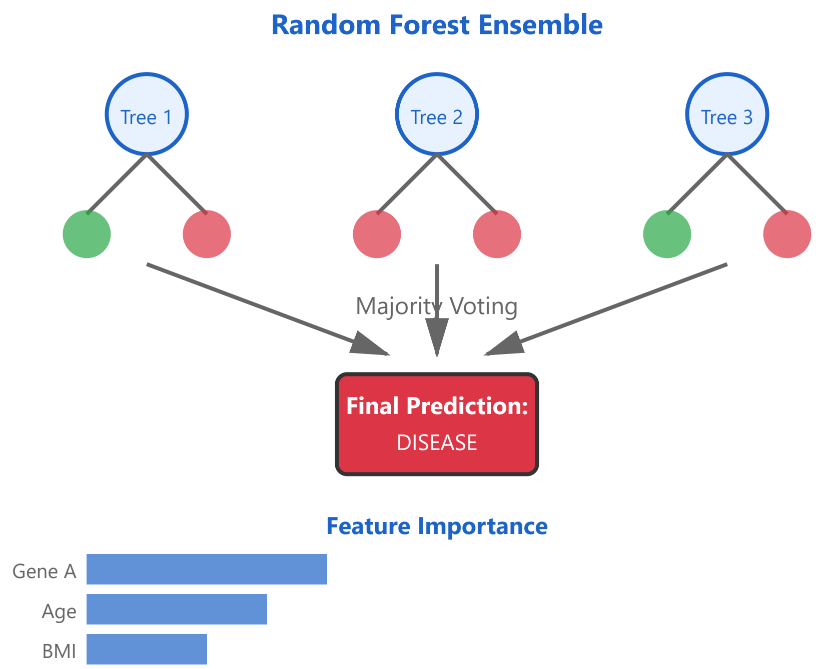
#### Key Strengths in Medicine:

- Provides interpretable odds ratios for clinical decision-making
- Fast to train and deploy in clinical systems
- Well-established statistical framework with confidence intervals
- Works well with limited data



## 2. Random Forests

**Overview:** Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the class that is the mode of the classes from individual trees. It combines the predictions of many trees to reduce overfitting and improve accuracy.




### How It Works:

- Creates multiple decision trees using bootstrap sampling
- Each tree uses a random subset of features at each split
- Aggregates predictions through majority voting (classification)
- Provides feature importance rankings

#### Medical Application Example:

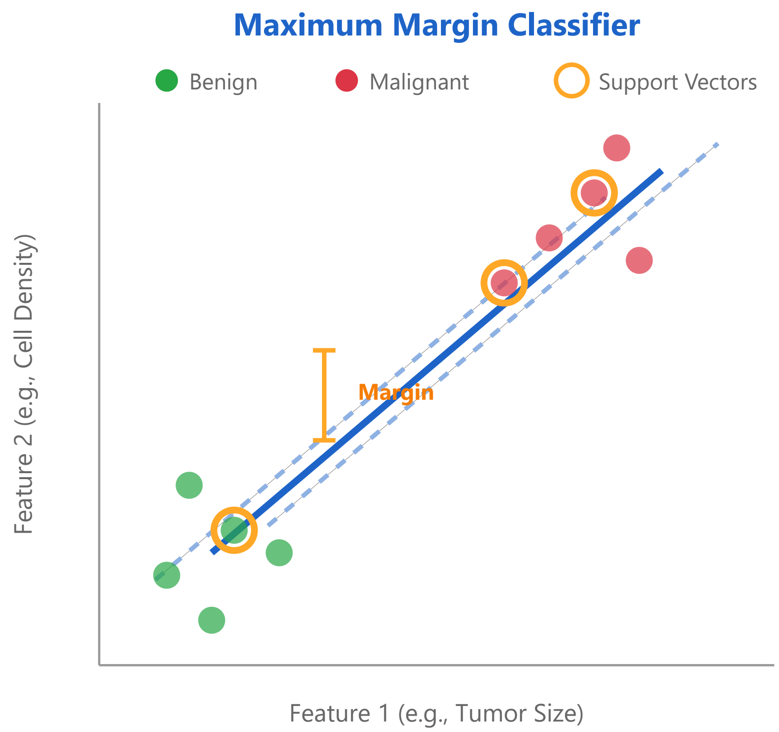
**Cancer Diagnosis:** Classifying tumor types based on gene expression data, imaging features, and patient demographics. Random forests can handle hundreds of genes simultaneously and identify which genes are most important for classification. *Example:* Using 500 trees to analyze 1,000 genomic features, identifying the top 20 genes that best distinguish between benign and malignant tumors.

**Key Strengths in Medicine:**

- Handles complex, non-linear relationships in clinical data
- Robust to outliers and missing values
- Provides feature importance for biomarker discovery
- Minimal hyperparameter tuning required

## 3. Support Vector Machines (SVM)

**Overview:** Support Vector Machine is a powerful classification algorithm that finds the optimal hyperplane (decision boundary) that maximizes the margin between different classes. It can handle both linear and non-linear classification through kernel functions.




## How It Works:

- Finds the hyperplane with maximum margin between classes
- Uses support vectors (critical points closest to decision boundary)
- Applies kernel trick for non-linear transformations
- Optimizes using quadratic programming

### Medical Application Example:

**Medical Image Classification:** Distinguishing between different types of brain lesions in MRI scans using texture features, intensity patterns, and spatial characteristics. SVMs excel with high-dimensional imaging data.
*Example:* Using an RBF kernel to classify brain tumors into gliomas, meningiomas, or metastases based on 200+ imaging features extracted from MRI scans.

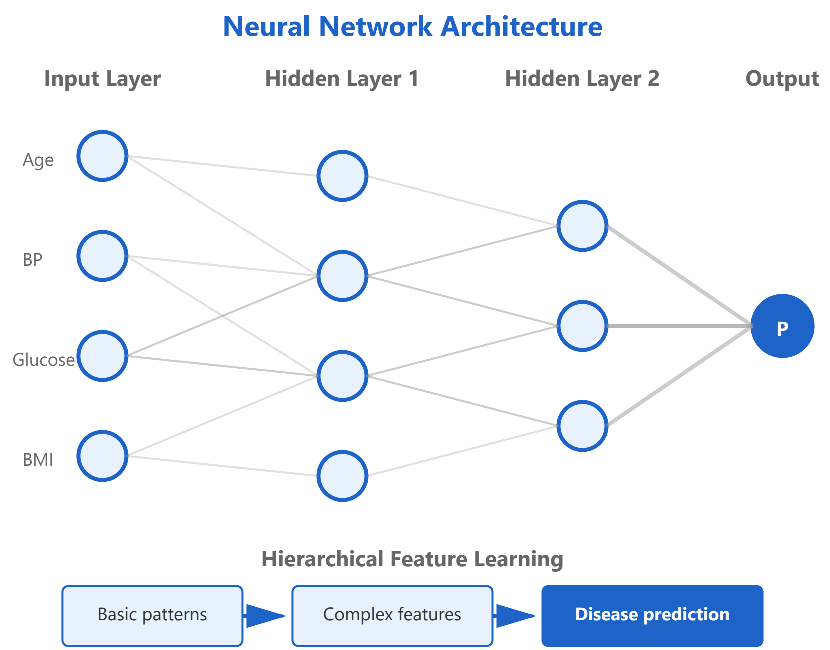
### Key Strengths in Medicine:

- Effective with high-dimensional medical data (genomics, imaging)
- Memory efficient (only uses support vectors)
- Flexible with different kernel functions for complex patterns

- Good performance with limited training samples

## 4. Neural Networks (Deep Learning)

**Overview:** Neural networks are computational models inspired by biological neural systems, consisting of interconnected layers of nodes (neurons) that can learn complex patterns through backpropagation. Deep learning uses multiple hidden layers to automatically extract hierarchical features.




### How It Works:

- Input layer receives raw data (images, sequences, tabular data)
- Hidden layers extract progressively abstract features
- Weights adjusted through backpropagation and gradient descent
- Activation functions introduce non-linearity (ReLU, sigmoid, tanh)

#### Medical Application Example:

**Chest X-ray Diagnosis:** Convolutional Neural Networks (CNNs) analyze chest X-rays to detect pneumonia, tuberculosis, lung cancer, and other conditions. The network automatically learns to identify relevant patterns like consolidations, nodules, and infiltrates.

*Example:* A CNN with 50 layers trained on 100,000 chest X-rays achieving 95% accuracy in detecting pneumonia, comparable to expert radiologists.

**Key Strengths in Medicine:**

- State-of-the-art performance on medical imaging tasks
- Automatic feature extraction eliminates manual engineering
- Handles multiple modalities (images, text, time-series)
- Transfer learning enables use of pre-trained models

## Algorithm Selection Guide

| Criterion | Logistic Regression | Random Forests | SVM | Neural Networks |
|---|---|---|---|---|
| **Data Size** | Small-Medium | Medium-Large | Small-Medium | Large-Very Large |
| **Interpretability** | ⭐⭐⭐⭐⭐ | ⭐⭐⭐ | ⭐⭐ | ⭐ |
| **Training Speed** | Very Fast | Fast | Slow | Very Slow |
| **Prediction Speed** | Very Fast | Fast | Medium | Fast |

| Criterion | Logistic Regression | Random Forests | SVM | Neural Networks |
|---|---|---|---|---|
| **Best Use Case** | Risk scoring, odds ratios needed | Tabular data, feature importance | High-dimensional data, limited samples | Images, complex patterns, large datasets |
| **Typical Accuracy** | Good | Very Good | Very Good | Excellent |