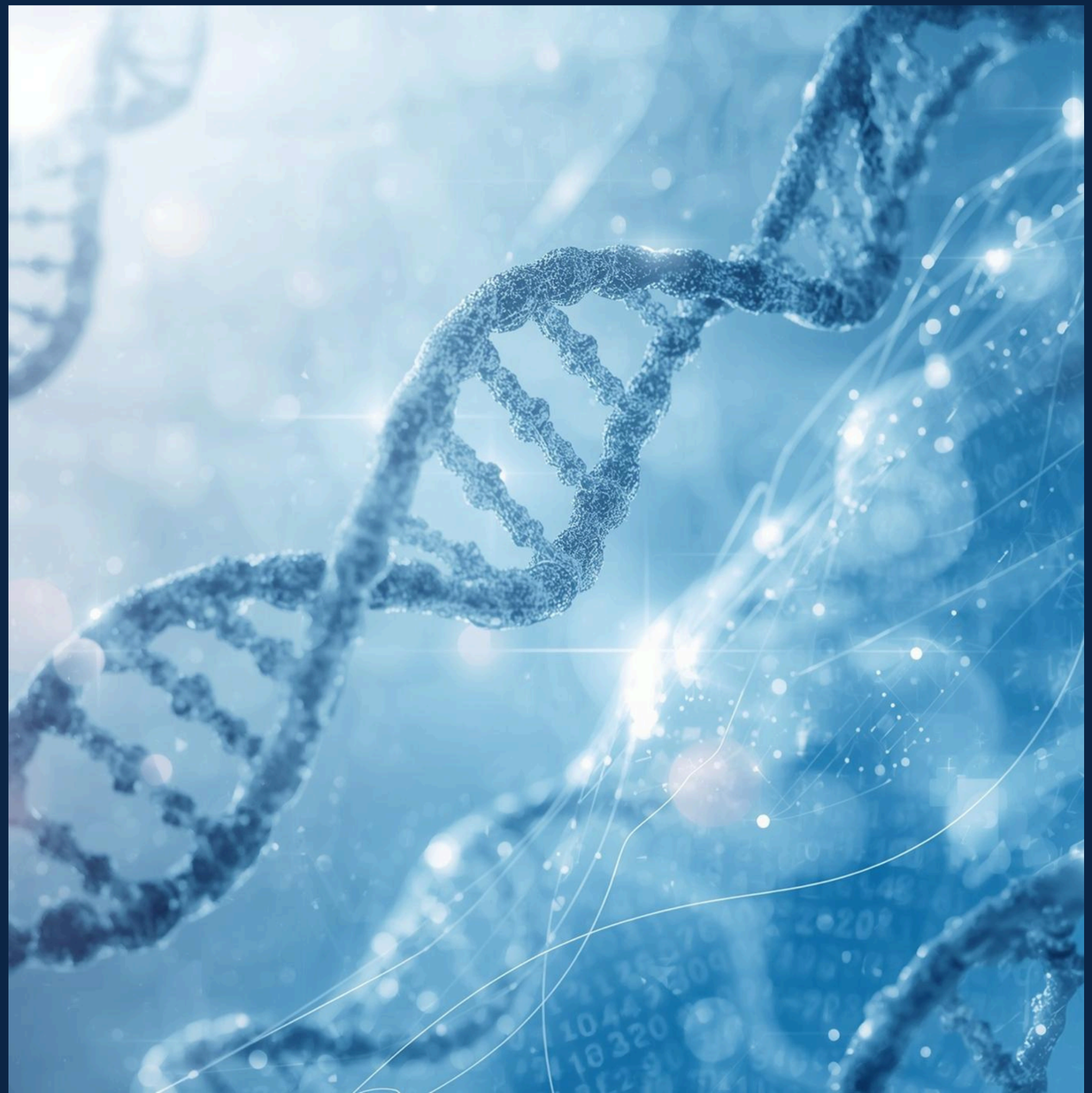
HEALTH INNOVATIONS

Breast Cancer Detection using Machine Learning



Presented by:-

AYUSH GANGWAR
ADITYA RAJ
PRTHAM KUMAR
SANJEEV MAURYA
JATIN SHARMA
ABHISEK YADAV

Project Motivation

🩺 **Breast cancer is life-threatening**

- Many women lose their lives due to late diagnosis. We wanted to help change that.

❤️ **Personal connection to the cause**

- Many of us know someone affected by cancer. This project is our way of contributing.

⚖️ **Manual diagnosis has limitations**

- Human error, time pressure, and lack of resources can affect diagnosis quality

🧠 **AI can support doctors**

- Machine learning can assist doctors in making faster and more accurate decisions.

Objective

- To build a robust machine learning model for detecting breast cancer
- To classify tumors as benign (non-cancerous) or malignant (cancerous) based on diagnostic data
- To improve accuracy and reduce human error in diagnosis
- To support doctors with fast and reliable predictions

DATASET OVERVIEW

Source

The dataset is sourced from the **Wisconsin Breast Cancer Dataset**, widely used in cancer research.

Target

The target variable indicates whether tumors are **malignant (1)** or **benign (0)**, crucial for classification.

Records

There are a total of **569 records** in the dataset, providing a robust sample for analysis.

Challenge

A significant challenge is the **data imbalance**, with more benign cases than malignant ones affecting model training.

DATA CLEANING

Removed Irrelevant Columns:
Dropped id (non-predictive) and Unnamed: 32 (fully empty) to streamline the dataset.

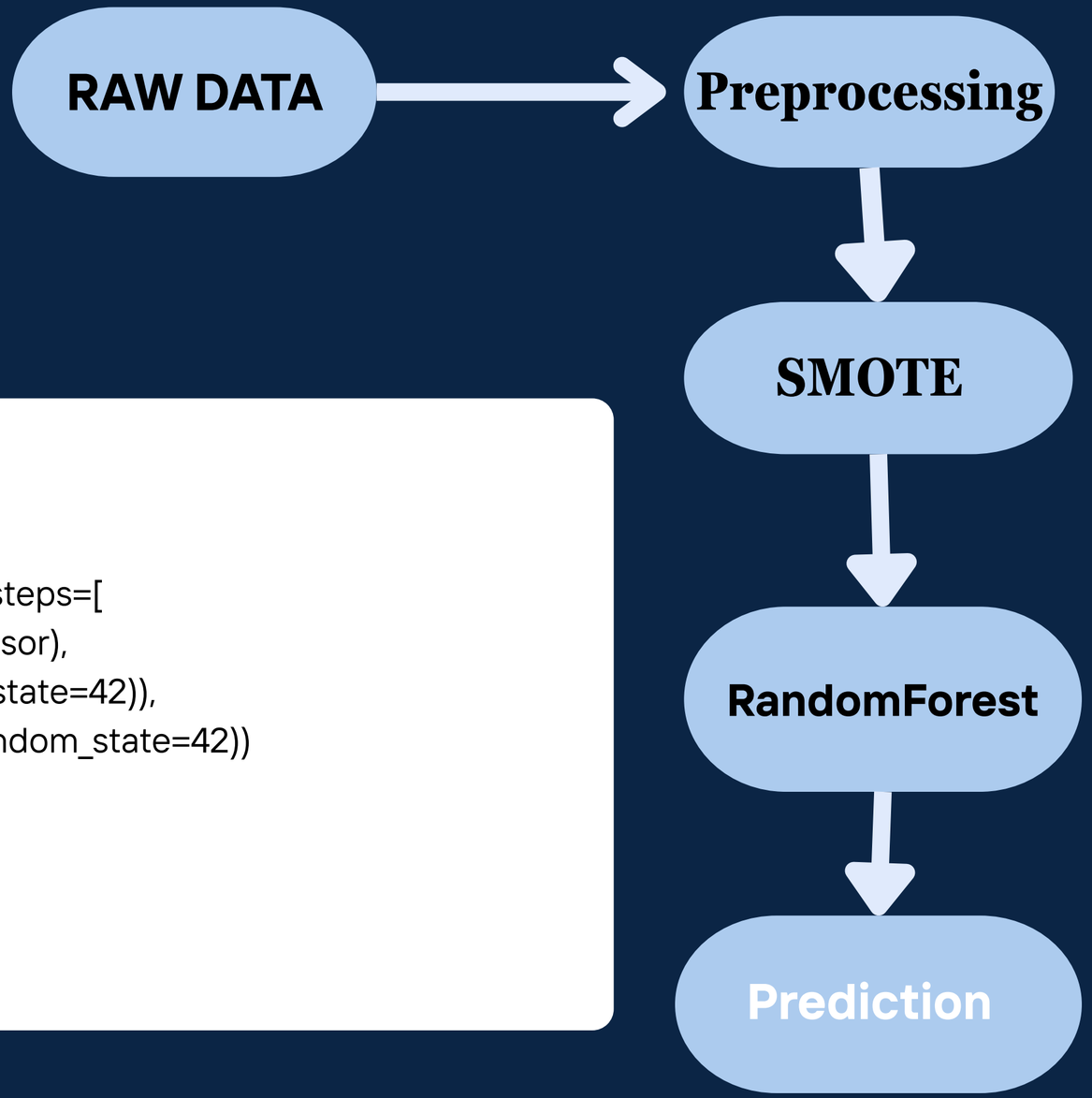
Checked for Missing Values:
No missing values found after cleaning — ensured complete and usable data.

Removed Duplicates:
Verified dataset integrity — no duplicate rows present

Encoded Target Labels:
Converted diagnosis from categorical to numeric:
M → 1 (Malignant), B → 0 (Benign)

Final Dataset Shape:
569 rows × 31 columns
(30 features + 1 target)

MODEL PIPELINING




```
pipeline = ImbPipeline(steps=[
('preproc', preprocessor),
('smote', SMOTE(random_state=42)),
('clf', RandomForestClassifier(random_state=42))
])
```

System Architecture: Three-Tier Design

Presentation Layer Front-End built with **HTML5, CSS3**, and JavaScript. Provides interactive forms and displays results.

Application Layer Flask-based server (app.py) handles Presentation Layer Front-End built with HTML5, CSS3, and JavaScript. Provides interactive forms and displays results. requests, processes data, and exposes REST API endpoints (

Data Layer Contains the **pre-trained Random Forest model (arya_best_cancer.joblib**) and uses **Pandas/NumPy** for data formatting.

0.9910

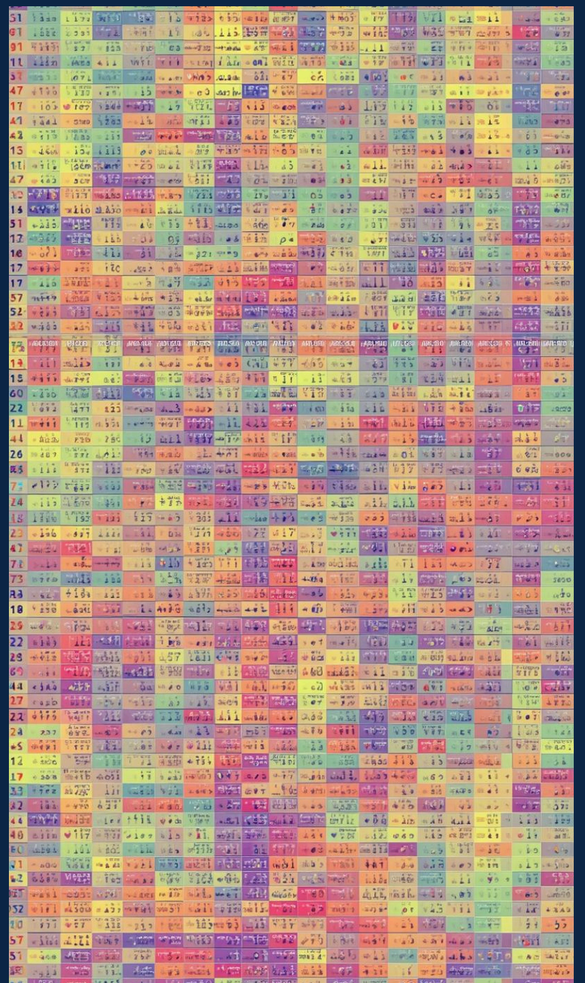
Optimal ROC AUC Score

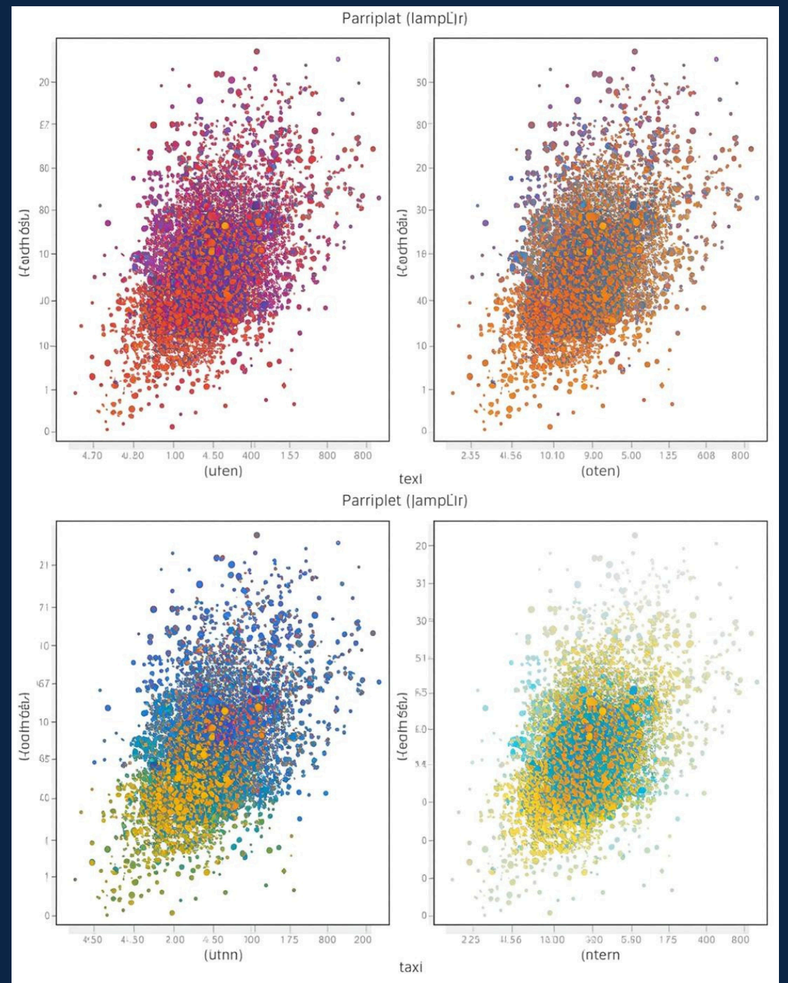
The achieved ROC AUC score demonstrates the model's excellent ability to distinguish between malignant and benign cases, enhancing diagnostic accuracy.

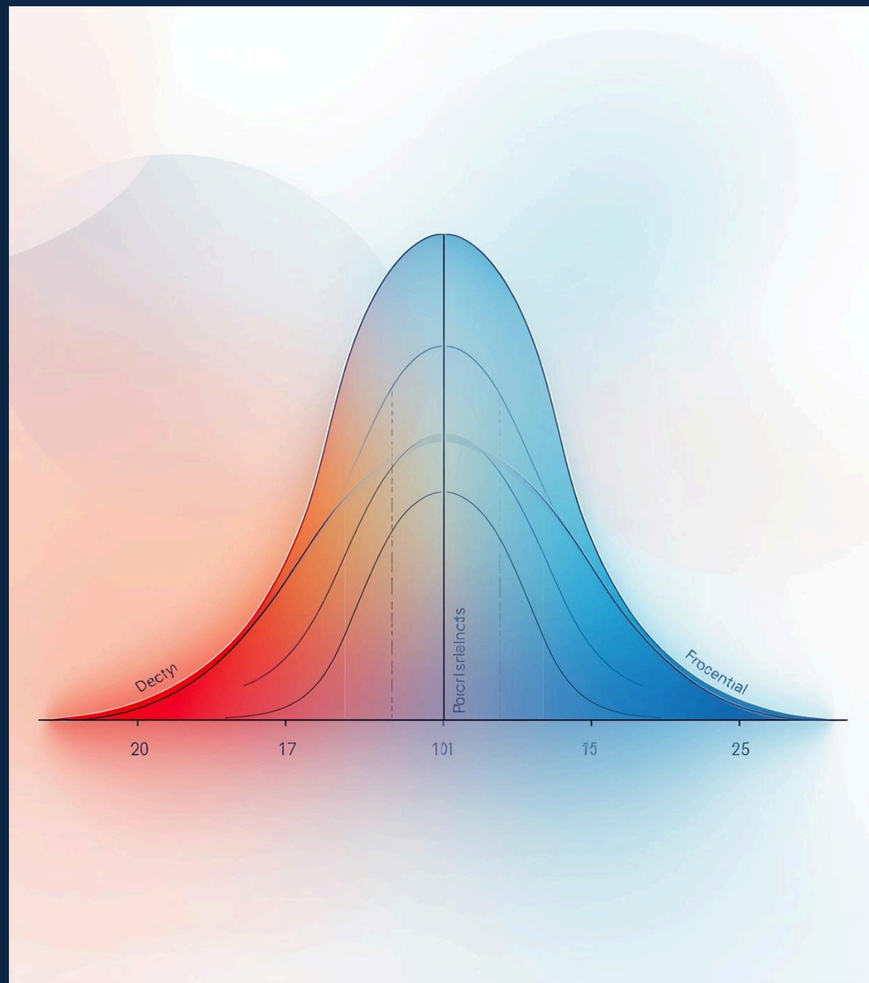
Data Analysis Insights

Key Feature Identification

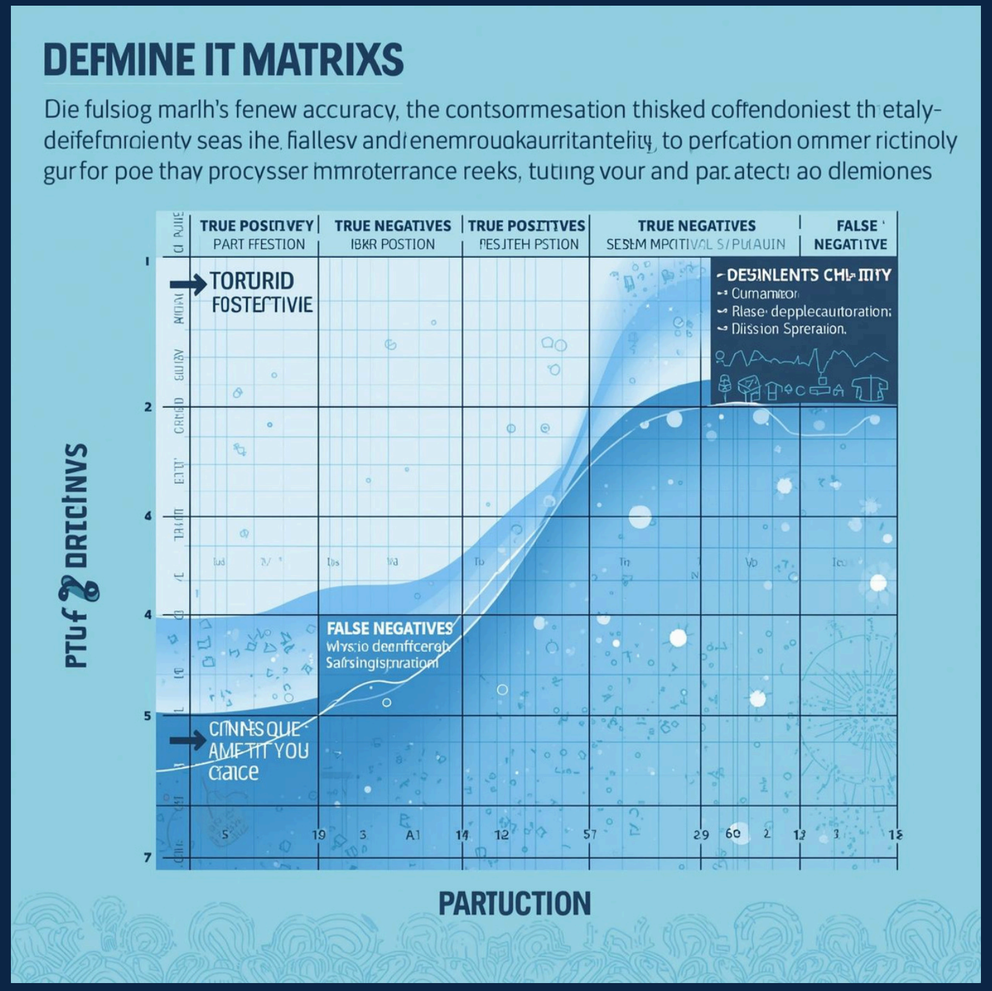
This section highlights **top predictive features** essential for improving breast cancer detection accuracy using machine learning.

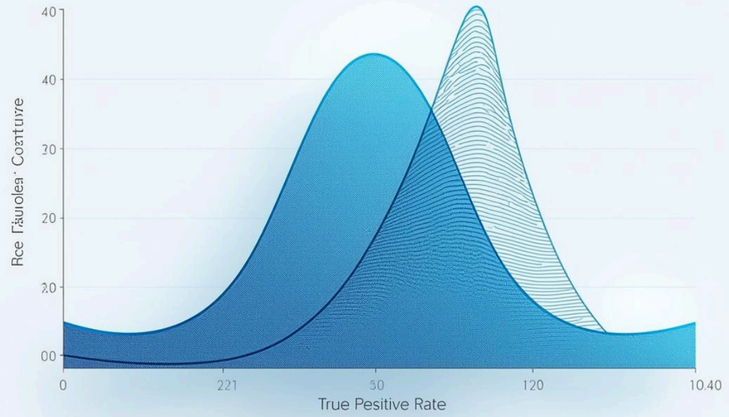






Final Results





Requirements for Prediction

✅ Input Required

30 Diagnostic Features from patient data
(e.g., radius_mean, texture_mean, concavity_worst, area_mean, etc.)
All features must be : Numerical
Cleaned (no missing values)
Scaled (StandardScaler applied)

🧠 Preprocessing Step

- Missing value imputation (Median)
- Feature scaling (StandardScaler)
- Categorical encoding (if any)

🔄 Format

```
X_new = pd.DataFrame([patient_features])
prediction = model.predict(X_new)
```

🎯 Output

Prediction Label:
0 → Benign
1 → Malignant
Probability Score:
e.g., 0.87 → 87% chance of malignancy

FUTURE SCOPE

- Successfully built a breast cancer classification model using the WDBC dataset
- Achieved 97% accuracy and ROC AUC of 0.9964 — indicating excellent predictive performance
- Handled class imbalance using SMOTE for fair and balanced learning
- Identified top predictive features through correlation analysis and EDA
- Model is scalable, interpretable, and ready for deployment in real-world scenarios
- Demonstrates how machine learning can assist in early cancer detection and patient care

Conclusion and Impact

- Successfully built a breast cancer classification model using the WDBC dataset
- Achieved 97% accuracy and ROC AUC of 0.9964 — indicating excellent predictive performance
- Handled class imbalance using SMOTE for fair and balanced learning
- Identified top predictive features through correlation analysis and EDA
- Model is scalable, interpretable, and ready for deployment in real-world scenarios
- Demonstrates how machine learning can assist in early cancer detection and patient care

Useful Hyperlink