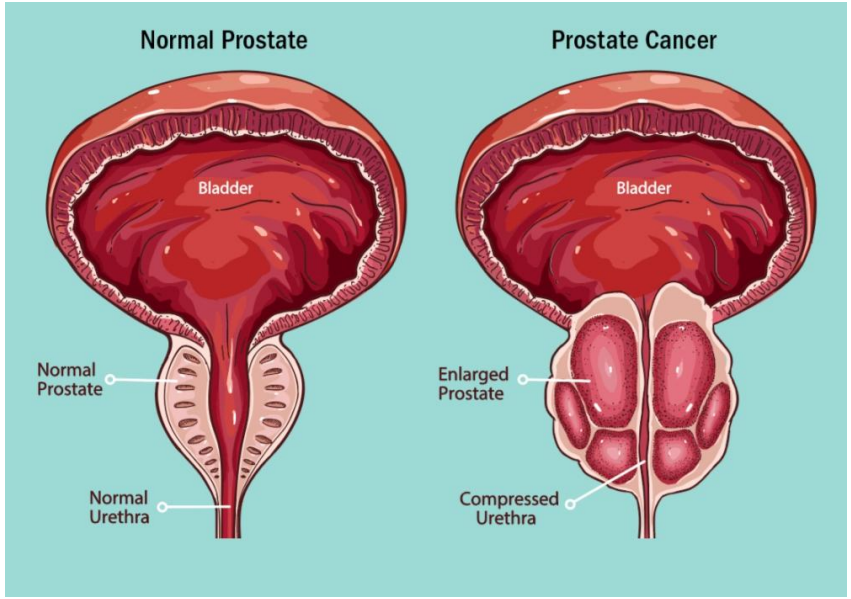
ML Project : Prostate Cancer Detection System

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

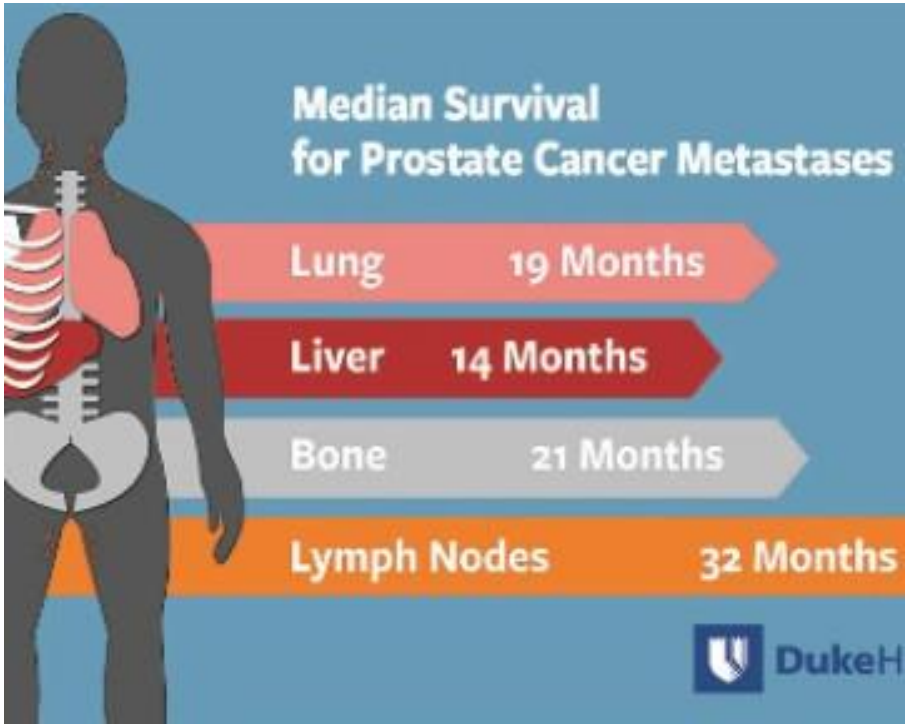
Motivation



Need for Early and Accurate Detection

- Small size
- Difficult to detect
- Generally grows slowly
- May cause Erectile Dysfunction

Motivation



Increasing Death Toll

- Poor Diagnostic System(Conventional systems)
- Approx. nine among 1 lakh men in India,Suffer
- 2020 Stats-
  - 1.4 million **new cases**
  - 375,000 **deaths**

Literature Review

Diagnosis of prostate cancer in a Chinese population by using machine learning methods

- **Used Methods-** Support Vector Machine(Svm),Least Square SVM,Random Forests(RF),Artificial Neural Networks(ANN)
- **Input-**Cohort of **1625** Chinese men with prostate biopsies
- **ANN- Highest Accuracy**(95.27%) ,**AUC Value**(0.9755)
- **RF-**Highest Performance in Benign,Significant and Insignificant Cases|| **Accuracy** (97.41%), **F1 Score**(0.8290)

Literature Review

Prostate Cancer Detection using Deep Convolutional Neural Networks

- **Used Methods-** Deep Convolutional Neural Networks(CNN)
- **Input-** DWI images of 427 patients(175 patients with **significant prostate cancer**) || **Testing set** (108 patients) **Training** (319)

Slice Level (Cancerous Region)	**Patient Level** (Benign or malignant)
AUC Value(0.87)	**AUC Value**(0.84)
Confidence Interval 95%(0.84-0.90)	Confidence Interval 95%(0.76-0.91)

Dataset description

Overview:

The *Prostate MRI and Ultrasound With Pathology and Coordinates of Tracked Biopsy* dataset is a comprehensive imaging collection from The Cancer Imaging Archive (TCIA), consisting of **1,151 subjects** who underwent biopsies due to prostate cancer suspicion. The dataset integrates **MRI** and **Ultrasound** imaging data, along with biopsy results, offering a rich source of information for advanced research on prostate cancer detection and treatment.

```
prostarte-mri-us-biopsy/
 Biopsy Overlays(3D-Slicer)/
    Biopsy Overlays(3D-Slicer)/
        Prostate-MRI-US-Biopsy-{patient_id}/
            Data/
              Bx-{S.No.}-Benign.fcsv
 STLs/
    STLs/
        Prostate-MRI-US-Biopsy-{patient_id}
 prostate-mri-us-biopsy/
    Prostate-MRI-US-Biopsy/
        Prostate-MRI-US-Biopsy-{patient_id}
 TCIA Biopsy Data_2020-07-14.xlsx
 Target Data_2019-12-05.xlsx
 metadata.csv
```

Dataset description

Key Features:

- **Data Types**: Includes ultrasound (US) and MRI imaging data, biopsy pathology, and spatial coordinates of biopsy cores.
- **Imaging Modalities**:
  - **MRI**: Multi-parametric MRI sequences, such as **T2-weighted**, **diffusion-weighted**, and **perfusion-weighted**.
  - **Ultrasound**: 3D transrectal ultrasound scans that are fused with preoperative MRI for targeted biopsy.
- **Biopsy Data**: Systematic biopsies using a 12-core template and targeted biopsies based on MRI fusion with real-time tracking.
- **Size**: Dataset size is approximately **80GB**, comprising **102,397 DICOM images**.
- Consists of mainly three metadata sheets namely Target Data_2019-12-05.xlsx,metadata.csv, TCIA Biopsy Data_2020-07-14.xlsx.

Technology:

- **MRI Scanners**: MRI scans were conducted on **Siemens Trio, Verio, Skyra 3 Tesla scanners**.
- **Ultrasound Systems**: Ultrasound was performed using **Hitachi Hi-Vision 5500** and **Noblus C41V probes**.
- **Biopsy Core Tracking**: The Artemis biopsy system was used to track biopsy core locations with mechanical arm kinematics, recording exact positions of both systematic and targeted biopsies.

Dataset description

Applications:

- **Prostate Cancer Research**: Enables development of AI and machine learning models for prostate cancer detection and prognosis.
- **Clinical Tools**: Provides resources to enhance diagnostic accuracy with MRI-guided biopsies.
- **3D Visualization**: Offers **STL files** and biopsy overlays for 3D visualization of prostate anatomy and biopsy cores.

Access and Citation:

- **DOI**: 10.7937/TCIA.2020.A61IOC1A.
- **License**: The dataset is available under the **CC BY 4.0 license** and can be accessed from The Cancer Imaging Archive.

metadata.csv

	Series UID	Collection	3rd Party Analysis	Data Description URI	Subject ID	Study UID
0	1.3.6.1.4.1.14519.5.2.1.140367896789002601449 3...	Prostate-MRI-US-Biopsy	NaN	https://doi.org/10.7937/TCIA.2020.A61IOC1A	Prostate-MRI-US-Biopsy-0001	1.3.6.1.4.1.14519.5.2.1.168053951941444894929 2...
1	1.3.6.1.4.1.14519.5.2.1.266717969984343981963 0...	Prostate-MRI-US-Biopsy	NaN	https://doi.org/10.7937/TCIA.2020.A61IOC1A	Prostate-MRI-US-Biopsy-0001	1.3.6.1.4.1.14519.5.2.1.855483049219656583677 2...
2	1.3.6.1.4.1.14519.5.2.1.120228593041312099989 2...	Prostate-MRI-US-Biopsy	NaN	https://doi.org/10.7937/TCIA.2020.A61IOC1A	Prostate-MRI-US-Biopsy-0001	1.3.6.1.4.1.14519.5.2.1.301993288926669228449 8...
3	1.3.6.1.4.1.14519.5.2.1.186749128823666050588 7...	Prostate-MRI-US-Biopsy	NaN	https://doi.org/10.7937/TCIA.2020.A61IOC1A	Prostate-MRI-US-Biopsy-0002	...
4	1.3.6.1.4.1.14519.5.2.1.200776032537717955457 1...	Prostate-MRI-US-Biopsy	NaN	https://doi.org/10.7937/TCIA.2020.A61IOC1A	Prostate-MRI-US-Biopsy-0002	1 ./Prostate-MRI-US-Biopsy/Prostate-MRI-US-Biops... 2023-09-12T21:29:45.762
						2 ./Prostate-MRI-US-Biopsy/Prostate-MRI-US-Biops... 2023-09-12T21:29:48.683
						3 ./Prostate-MRI-US-Biopsy/Prostate-MRI-US-Biops... 2023-09-12T21:29:52.657
						4 ./Prostate-MRI-US-Biopsy/Prostate-MRI-US-Biops... 2023-09-12T21:29:56.385

TCIA Biopsy_data.xlsx

	PSA (ng/mL)	Primary Gleason	Secondary Gleason	Cancer Length (mm)	% Cancer in Core	Core Fragment #1 Tissue Length (mm)	Core Fragment #2 Tissue Length (mm)	Core Fragment #3 Tissue Length (mm)	Bx Tip X (MRI Coord)	Bx Tip Y (MRI Coord)	...	Bx Tip Y (US Coord)
0	10.9	NaN	NaN	NaN	NaN	10.0	3.0	NaN	-8.915	34.791	...	12.970
1	10.9	3.0	4.0	4.0	50.0	14.0	NaN	NaN	-5.644	23.161	...	6.628
2	10.9	3.0	3.0	2.0	10.0	10.0	2.0	1.0	-9.642	22.070	...	10.436
3	10.9	NaN	NaN	NaN	NaN	12.0	4.0	NaN				
4	10.9	NaN	NaN	NaN	NaN	14.0	NaN	NaN				
...												
3	Prostate-MRI-US-Biopsy-0001											

TCIA_Target_data.xlsx

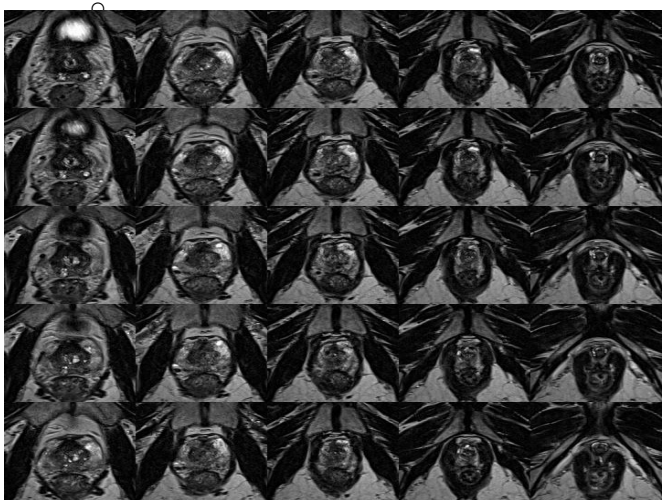
	UCLA Score (Similar to PIRADS v2)	ROI Volume (cc)	Target No.
0	3	0.834323	1
1	3	0.834323	1
2	1	0.364729	1
3	3	0.364729	1
4	3	0.884436	1

	seriesInstanceUID_US \
0	1.3.6.1.4.1.14519.5.2.1.1403678967890026014493...
1	1.3.6.1.4.1.14519.5.2.1.1202285930413120999892...
2	1.3.6.1.4.1.14519.5.2.1.9782151183163602689533...
3	1.3.6.1.4.1.14519.5.2.1.2007760325377179554571...
4	1.3.6.1.4.1.14519.5.2.1.9358385420720256123337...

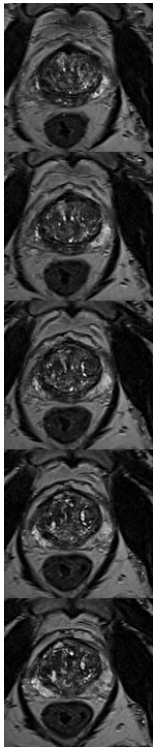
	seriesInstanceUID_MR \
0	1.3.6.1.4.1.14519.5.2.1.2667179699843439819630...
1	1.3.6.1.4.1.14519.5.2.1.2667179699843439819630...
2	1.3.6.1.4.1.14519.5.2.1.1867491288236660505887...
3	1.3.6.1.4.1.14519.5.2.1.1867491288236660505887...
4	1.3.6.1.4.1.14519.5.2.1.1345819869189093607538...

	Patient ID
0	Prostate-MRI-US-Biopsy-0001
1	Prostate-MRI-US-Biopsy-0001
2	Prostate-MRI-US-Biopsy-0002
3	Prostate-MRI-US-Biopsy-0002
4	Prostate-MRI-US-Biopsy-0003

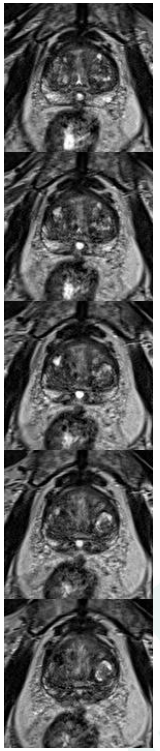
Methodology



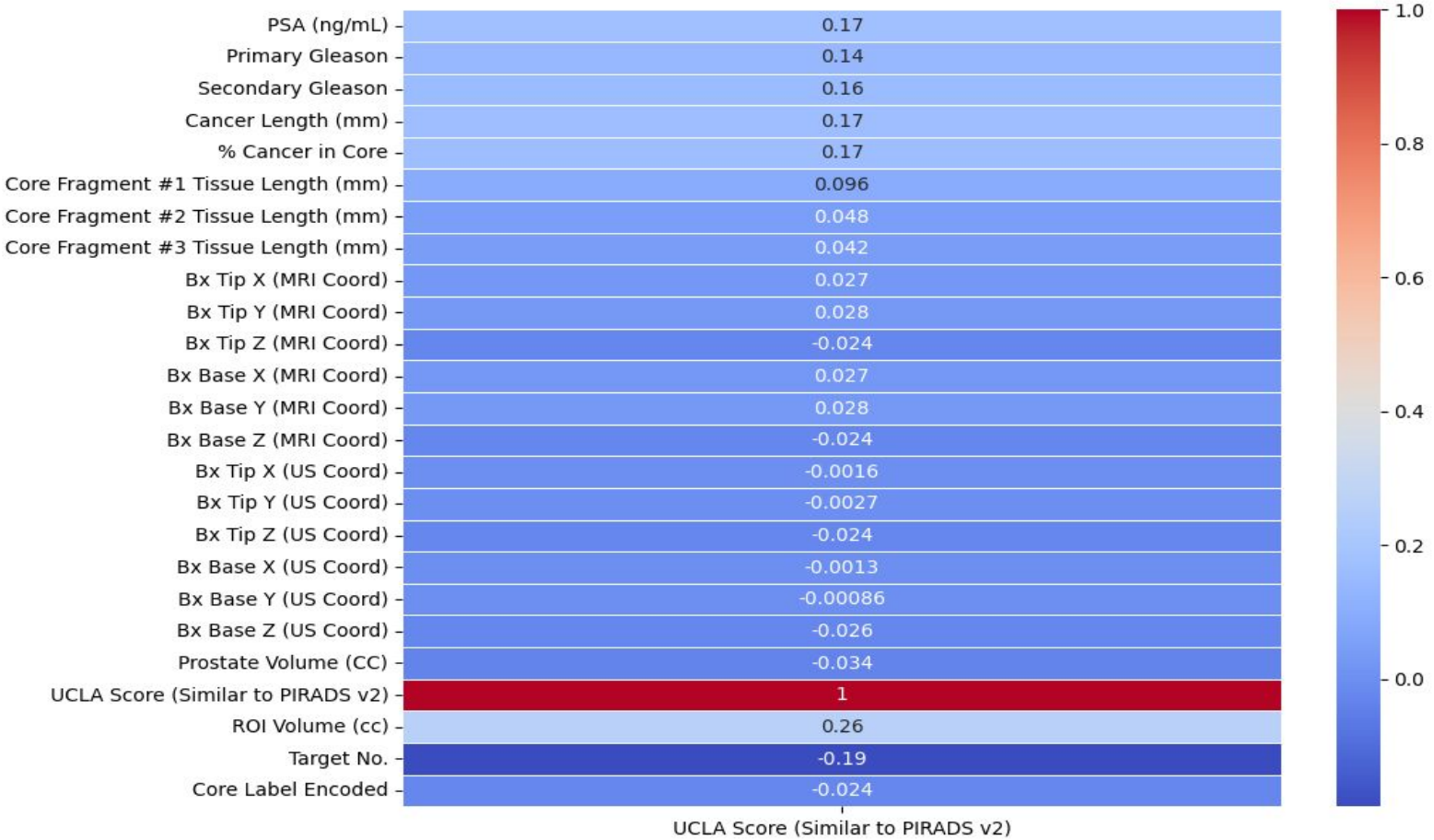
Collage of inputs given to model for image classification



Negative Label sample



Positive Label sample


PSA (ng/mL)
Primary Gleason
Secondary Gleason
Cancer Length (mm)
% Cancer in Core
Core Fragment #1 Tissue Length (mm)
Core Fragment #2 Tissue Length (mm)
Core Fragment #3 Tissue Length (mm)
Bx Tip X (MRI Coord)
Bx Tip Y (MRI Coord)
Bx Tip Z (MRI Coord)
Bx Base X (MRI Coord)
Bx Base Y (MRI Coord)
Bx Base Z (MRI Coord)
Bx Tip X (US Coord)
Bx Tip Y (US Coord)
Bx Tip Z (US Coord)
Bx Base X (US Coord)
Bx Base Y (US Coord)
Bx Base Z (US Coord)
Prostate Volume (CC)
UCLA Score (Similar to PIRADS v2)
ROI Volume (cc)
Target No.
Core Label Encoded
0.17
0.14
0.16
0.17
0.17
0.096
0.048
0.042
0.027
0.028
-0.024
0.027
0.028
-0.024
-0.0016
-0.0027
-0.024
-0.0013
-0.00086
-0.026
-0.034
1
0.26
-0.19
-0.024
UCLA Score (Similar to PIRADS v2)
1.0
0.8
0.6
0.4
0.2
0.0

Methodology

The methodology was divided into two main parts→:

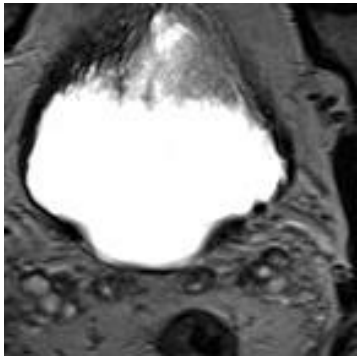
- Part I: Binary classification for cancer detection.
- Part II: Multi-class classification for cancer risk prediction.

Analyzing and Classifying:

1. **Data Preparation:**
    - Conversion of DICOM images to JPG for easier processing.
    - Preprocessing:
        - Resizing for standardized resolution.
        - Cropping for Region of Interest (ROI).
        - Patient-wise collages for consolidated information.
    - Binary labels:
        - Positive (cancer present, % > 0).
        - Negative (cancer absent, % = 0).
2. **Features Used in Detection:**
    - Pixel data (~49152 per image).
    - Statistical and texture-based features (Canny Edges, Gradient Histogram, LBP, Entropy).

Methodology

Part I - Cancer Detection (Binary Classification)



Preprocessing and Labeling:

- Images categorized with binary labels based on cancer presence in core samples.
- Combined visual and metadata inputs for detection.

Models Used:

1. **Multilayer Perceptron (MLP):**
   - Activation functions: tanh, ReLU, logistic.
   - Effective in capturing nonlinear patterns.
2. **Other Models for Robust Classification:**
   - Decision Trees (DT).
   - Naive Bayes (NB).
   - Random Forests (RF).
   - Logistic Regression (Log R).

Methodology

Part II - Cancer Risk Prediction

Objective:

- Predict risk level using UCLA score and biopsy metadata.

Metadata Features Used:

- Percentage of cancer in core biopsy samples.
- PSA levels.
- Gleason indices (primary/secondary).
- Biopsy overlay coordinates and tissue measurements.

Model Training and Optimization:

- **Models Trained:**
  - Logistic Regression, SVM, Naive Bayes, Decision Trees, Random Forests.
  - Advanced techniques: Gradient Boosting (AdaBoost, XGBoost), Voting Classifiers.
- **Optimization Steps:**
  - PCA (2 to 15 components).
  - Boosting techniques for higher accuracy and lower MSE.
- Total: Over 70 models tested for performance refinement.

Results

1. Binary Classification (Cancer Detection)

- **Models Used**:
  - Multilayer Perceptron (MLP), CNN, Decision Trees (DT), Naive Bayes (NB), Random Forest (RF), Logistic Regression.
- **Input Data**: MRI/Ultrasound pixel data + extracted metadata (PSA, Gleason scores, etc.).

2. Multiclass UCLA Risk Prediction

- **Models Used**: Logistic Regression, SVM, Gradient Boosting, Random Forest.
- **PCA Applied**: Reduced dimensions (2–15 components).

Task	Best Model	Metric
Classification (5 Levels)	Logistic Regression (PCA-7)	Accuracy: 60.71%
Regression (UCLA Risk)	SVR (PCA-15)	MSE: **0.4265**
Classification (RF/GBR)	Random Forest, GBR	Accuracy: 58.33%

Model	Accuracy
Decision Trees	65.86
Naive Bayes	75.44%
CNN	76.04%
Bernoulli Naive Bayes (BNB)	84.43%
Random Forest (with features)	84.43%
Logistic Regression/MLP	85.03%

Models we tried with execution results(part 1)

logistic regression - 85.0299%

RF- 84.431%

DT-65.862%

RF after feature extraction 82.0359%

DT feature extraction - 71.85

GNB Pixel Data-65.86

GNB FE-75.44

BNB Pixel Data- 79.041

BNB FE- 84.431

MLP Pixel Data - 85.029

Models we tried with execution results(part 2)

Logistic Regression: 0.5417

Decision Tree: 0.5060

Random Forest: 0.6012

Support Vector Classifier: 0.4881

K-Nearest Neighbors: 0.5000

Gradient Boosting Classifier: 0.5476

Naive Bayes: 0.3036

Number of PCA components: 2

Logistic Regression: 0.4107

Decision Tree: 0.3929

Random Forest: 0.4405

Support Vector Classifier: 0.4762

K-Nearest Neighbors: 0.4345

Gradient Boosting Classifier: 0.4405

cont

Naive Bayes: 0.4762

Number of PCA components: 3

Logistic Regression: 0.5774

Decision Tree: 0.4881

Random Forest: 0.5298

Support Vector Classifier: 0.5536

K-Nearest Neighbors: 0.5000

Gradient Boosting Classifier: 0.5119

Naive Bayes: 0.5179

---

cont

Number of PCA components: 4

Logistic Regression: 0.5714

Decision Tree: 0.4107

Random Forest: 0.5179

Support Vector Classifier: 0.5357

K-Nearest Neighbors: 0.4643

Gradient Boosting Classifier: 0.5060

regressors

Training with 2 PCA components...

Linear Regression - PCA 2 components:

MSE: 0.6492, MAE: 0.6459, R2: 0.0306

Decision Tree Regression - PCA 2 components:

cont

MSE: 1.2321, MAE: 0.7083, R2: -0.8397

Random Forest Regression - PCA 2 components:

MSE: 0.7204, MAE: 0.6554, R2: -0.0756

Support Vector Regression - PCA 2 components:

MSE: 0.5620, MAE: 0.6001, R2: 0.1609

K-Nearest Neighbors Regression - PCA 2 components:

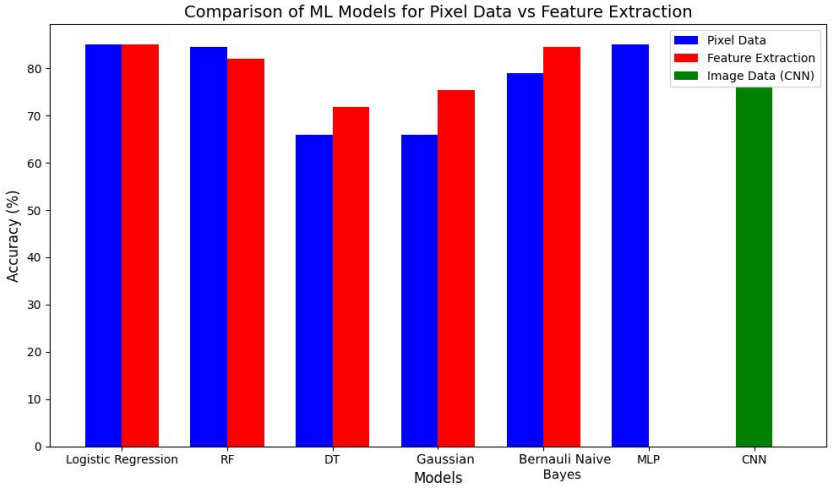
MSE: 0.7764, MAE: 0.6821, R2: -0.1593

Gradient Boosting Regression - PCA 2 components:

MSE: 0.6369, MAE: 0.6379, R2: 0.0491

We trained more but couldn't add due to space constraints

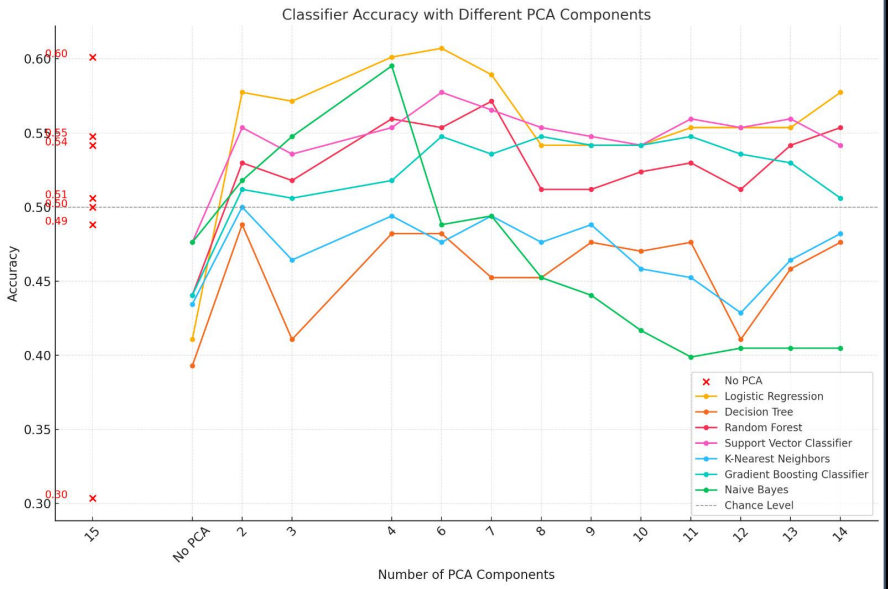
Results



Feature Extraction For Models

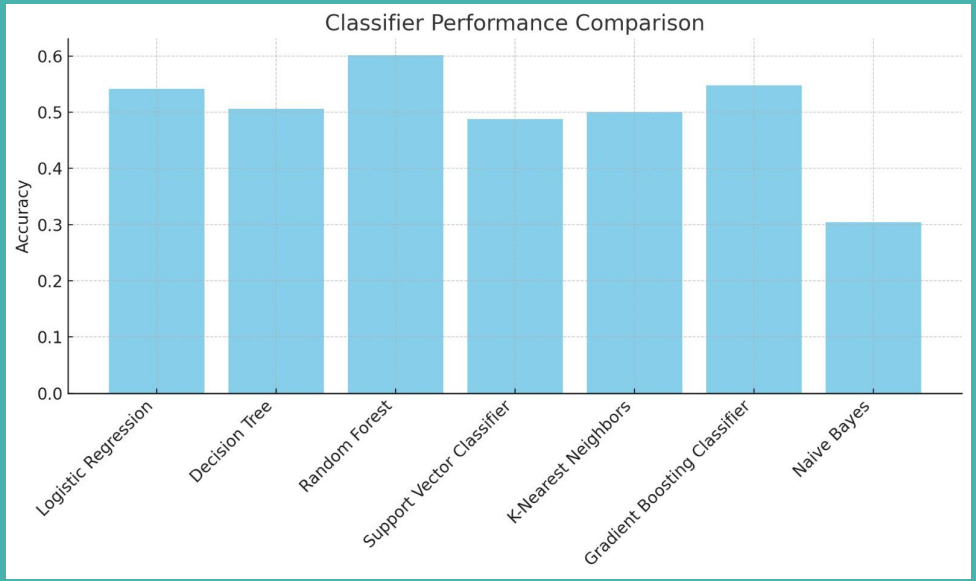
Error Comparison

Results(Classifier)



Accuracy Comparison

Error Comparison



Analysis

Exploratory Data Analysis (EDA)

- Dataset: Prostate MRI and Ultrasound Biopsy Data (1,151 patients, 24,000+ scans).
  - Avg. **22 procedures per patient** (MRI/US/biopsy combined).
  - Over **100,000 total images**.
- MRI Protocols: 44 different methods used; *t2spcrstaxial oblProstate* most frequent.
- Correlation Insights:
  - UCLA scores highly correlated with **PSA levels**, **Gleason scores**, **Cancer length**, and **% cancer in core**.

Model Observations

- **Cancer Detection**:
  - Best performing models: **MLP and Logistic Regression (85.03%)**.
  - **Random Forest (84.43%)** effective with feature extraction.
  - Boosted Naive Bayes also showed strong performance.
- **UCLA Risk Prediction**:
  - Logistic Regression (PCA-7) achieved **highest accuracy** (60.71%).
  - Support Vector Regression (PCA-15) achieved **lowest MSE (0.4265)**.
  - Hyperparameter tuning showed limited impact on improving accuracy.

Conclusion

- Addressed the critical challenge of **prostate cancer detection** by proposing an automated system using **machine learning techniques** for reliable and accurate diagnosis.
- Focused on preprocessing prostate MRI/biopsy images and integrating metadata to enhance detection and risk prediction.

Key Achievements:

- Implemented a variety of models: **Decision Trees**, **Random Forest**, **SVM**, **CNNs**, and **MLP**, alongside tuning techniques.
- Successfully predicted cancer risk using the **UCLA Prostate Cancer Index** and classified images effectively.
- **Results**:
  - High accuracy scores for cancer detection (**85.03% with MLP/Logistic Regression**).
  - Low MSE for UCLA risk prediction (**0.4265 with Support Vector Regression**).
  - Moderate Accuracy for part 2 of about **60.71 percent** upon boiling down to 7 components

Impact:

- Demonstrated that combining image data and metadata enhances model performance.
- Effective detection and risk classification achieved for MRI/US images.

Future Work:

- Further improve UCLA risk classification through:
  - Advanced feature extraction (e.g., tumor shape and texture).
  - Deep learning models tailored for risk prediction tasks.
- Address noise and variability in imaging data through enhanced preprocessing techniques.

Timeline

Project planning and lit review

Understanding the **current state** of diagnostics and machine learning models associated with them.

Data Preparation

Data preprocessing and **collection** is complete (converting dicom to jpg ,assigning labels etc)

Identifying models

Use of **CNN** (unsupervised learning). Use of supervised learning techniques like **SVM,Decision Trees ,Random forest** using score techniques like **UCLA** score.

Model Development

Implemented the suggested **models** and obtained satisfactory results.

Testing Phase

K-Cross Validation was used on training/testing data formulation to avoid **overfitting**.

Evaluate and optimise

Use of various performance measures like the **confusion matrix ,receiver operating curve ,Hyper-parameter Tuning,etc.**

Thank You