Prostate Cancer Detection Using Machine Learning

Naman Garg(2022602) Bhavy Chawla(2022594) Akshat Wadhera(2022057) Pratham Gautam(2022372)

***Abstract*— Prostate cancer (PCa) is a serious malignancy that kills many men due to a weak detection system. Images from cancer patients include important and intricate details that are difficult for conventional diagnostic methods to extract. This project aims to solve the problem by using Machine Learning for diagnosis. In 2020, there were an estimated 1.4 million new prostate cancer cases, causing 3,50,000 deaths. Nine among 1,00,00 men has prostrate cancer in India, which shows an increasing trend every year. A poor diagnostic system cause deaths of Lakhs of men every year. Given the rising incidence of prostate cancer and the high stakes of early diagnosis, this project aims to address a critical gap in current medical practice. By utilizing machine learning, we hope to provide a powerful tool that aids clinicians in detecting and diagnosing prostate cancer more effectively, ultimately contributing to better patient outcomes and saving lives. Hence, we decided to work on this project. The project which was divided into two basic tasks ie. 'Cancer Detection' and 'Cancer Risk Prediction' have both been completed for which we trained over a 100 models to achieve good accuracies(on classification tasks) and low mse's(on regression tasks).**

I. INTRODUCTION

Prostate cancer is a significant health issue affecting men worldwide, characterized by the uncontrolled growth of cells in the prostate gland, part of the male reproductive system. The disease is often identified through prostate-specific antigen (PSA) blood tests, which flag abnormal growth, prompting further diagnostic procedures such as biopsies and imaging to determine the extent of the cancer. Based on the Gleason score, PSA levels, and imaging results, prostate cancer cases are assigned stages from 1 to 5, with higher stages indicating more advanced and dangerous conditions. While many prostate tumors remain small and manageable with active surveillance, aggressive forms of the disease can spread to bones and lymph nodes, causing severe symptoms and drastically lowering survival rates.

Prostate cancer is the second leading cause of cancer and cancer-related deaths in men, with over 1.2 million new cases diagnosed annually and approximately 350,000 deaths worldwide. The risk of developing the disease increases with age and is exacerbated by genetic factors, such as BRCA2 gene mutations. Despite advancements in treatment—ranging from radical prostatectomy and radiation therapy to hormone therapy and chemotherapy—prostate cancer remains a critical challenge in men's health. Particularly in advanced cases, where the cancer becomes resistant to treatments, the prognosis worsens significantly.

Prostate cancer may cause no signs or symptoms in its early stages.

Prostate cancer that's more advanced may cause signs and symptoms such as:

- Trouble urinating.
- Decreased force in the stream of urine.
- Blood in the urine.
- Blood in the semen.
- Bone pain.
- Losing weight without trying.
- Erectile dysfunction.

Prostate cancer is generally slow-growing, which allows for a range of treatment options, including surveillance methods like active monitoring and watchful waiting. These approaches are often recommended for men with slow-growing cancers that aren't causing significant symptoms, as they allow patients to avoid or delay more invasive treatments. However, if the cancer shows signs of progression, more active treatments such as surgery, radiation, and systemic therapies are considered.

Surgical options like radical prostatectomy can remove the prostate gland and provide a cure for cancers confined to the prostate. Technological advancements, including robotic-assisted surgery, have made these procedures less invasive, leading to quicker recovery times and fewer complications. Radiation therapy, either through external beam radiation or internal techniques like brachytherapy, is another effective treatment, especially for localized cancers.

Systemic therapies come into play when cancer spreads beyond the prostate. These treatments, such as hormone therapy and chemotherapy, target cancer cells throughout the body. Newer approaches like immunotherapy and targeted therapies are being explored for advanced cases, offering hope even for those with aggressive forms of the disease. Alongside these treatments, focal therapies such as high-intensity focused ultrasound and cryotherapy are emerging, focusing on treating tumors with minimal impact on surrounding tissues.

Prostate cancer treatment often comes with side effects, including urinary incontinence, erectile dysfunction, and infertility, but advances in medical care are continually improving how these issues are managed. Patients are encouraged to communicate with their healthcare providers about side effects, as personalized management plans can significantly improve quality of life during and after treatment.

In this paper, we aim to explore the use of machine learning techniques in the early detection and diagnosis of prostate cancer as well as the risk level prediction,

which can lead to improved patient outcomes. With advancements in medical imaging and data analysis, machine learning models have the potential to enhance the accuracy and efficiency of cancer diagnosis, providing a promising avenue for reducing the mortality rate associated with this widespread disease.

II. LITERATURE SURVEY

In recent years, machine learning (ML) has become increasingly prominent in the diagnosis of prostate cancer (PC). Several studies have leveraged various ML algorithms, such as Support Vector Machine (SVM), Least Squares Support Vector Machine (LS-SVM), Artificial Neural Networks (ANN), and Random Forest (RF), to enhance the prediction and classification of PC, particularly in distinguishing significant PC cases from benign and insignificant cases. These models, applied to large datasets with clinical features such as PSA levels, age, prostate volume, and digital rectal examination (DRE) results, have shown high accuracy and diagnostic power in reducing unnecessary biopsies.

SVM, first introduced by Cortes and Vapnik (1995), has been widely used due to its ability to project data into a higher-dimensional feature space and find an optimal hyperplane for classification. The kernel trick enhances its efficiency, allowing it to deal with non-linear data. In comparison, LS-SVM, proposed by Suykens and Vandewalle (1999), simplifies the process by solving a set of linear equations instead of a quadratic programming problem, making it computationally more efficient while maintaining comparable generalization performance. Empirical results from various studies have shown that both SVM and LS-SVM are effective in binary classification tasks, such as distinguishing between significant and insignificant PC.

ANN, inspired by the biological neural network, has shown impressive results in PC diagnosis. The network, composed of layers of nodes (neurons), has demonstrated the ability to model complex non-linear relationships between input variables. Studies have highlighted that ANN can achieve higher sensitivity and accuracy compared to other methods. In the context of prostate cancer, ANN achieved the highest accuracy of 95.27

RF, an ensemble learning method, has also been widely studied due to its robustness and superior performance in multi-class classification problems. The RF algorithm builds multiple decision trees, where each tree independently predicts the output class. The final classification is determined by majority voting. Studies have shown that RF achieved the highest accuracy (79.41

Overall, the literature demonstrates that ML techniques, especially ANN and RF, have significantly improved the diagnostic accuracy of prostate cancer models. ANN performed best in detecting significant PC, while RF was the most effective in multi-class classification tasks. These results suggest that ML models hold great potential for enhancing PC diagnosis, but further validation and refinement of these models, especially in distinguishing significant from insignificant PC cases, are necessary. Future research should focus on testing these models on diverse populations and incorporating additional clinical data to improve predictive power.

A cohort of 427 consecutive patients with a PI-RADS score of 3 or higher who underwent biopsy was included in this study. Out of these patients, 175 had clinically significant prostate cancer (PCa), while 252 did not. The dataset consisted of 5,832 2D DWI slices containing prostate glands. For classification purposes, patients with a Gleason score of 7 or higher (GG 2) were considered to have clinically significant PCa, whereas those with a Gleason score of 6 or lower (GG = 1) or no cancer (GG = 0) were classified as without clinically significant PCa.

The DWI data was acquired from January 2014 to July 2017 using a Philips Achieva 3T MR imaging scanner. The imaging was performed using four b-values (0, 100, 400, 1000 s/mm²) and other settings such as TR (5000–7000 ms), TE (61 ms), slice thickness (3 mm), and FOV (240 mm × 240 mm). The images were resized to 144 × 144 pixels and center-cropped to 66 × 66 pixels, ensuring the prostate was covered. CNNs were designed to handle DWI data with six channels (ADC, b0, b100, b400, b1000, and b1600).

The DWI dataset was divided into three sets: training (271 patients, 3,692 slices), validation (48 patients, 654 slices), and test (108 patients, 1,486 slices) with a 64%, 11%, and 25% ratio, respectively. This separation ensured a balanced distribution of PCa and non-PCa patients. The dataset was normalized using a z-score normalization technique.

The proposed pipeline involved three stages: five individually trained CNNs for slice-level classification, extraction of first-order statistical features from CNN outputs, and classification into PCa and non-PCa using a Random Forest classifier. The CNNs utilized a ResNet architecture, chosen for its effectiveness in computer vision tasks. The 41-layer deep ResNet model underwent hyperparameter fine-tuning to achieve optimal results for slice-level classification. The model used Stochastic Gradient Descent for optimization with a learning rate of 0.001, reduced by a factor of 10 when improvement plateaued, and was trained with a batch size of 8. Binary cross-entropy was employed as the loss function due to dataset imbalance.

A stacked generalization method was used to improve patient-level performance by combining predictions from the five CNNs. The patient-level AUC significantly improved (AUC: 0.84, CI: 0.76–0.91) compared to a single CNN (AUC: 0.71, CI: 0.61–0.81). First-order statistical features such as mean, standard deviation, skewness, and kurtosis were extracted from slice-level probabilities, and important features were selected using a decision-tree-based feature selector. A Random Forest classifier trained on these features yielded robust patient-level classification results.

The model's performance was evaluated using ROC curves and AUC scores, with the slice-level classifica-

tion achieving an AUC of 0.87 (CI: 0.84–0.90) and the patient-level classification achieving an AUC of 0.84 (CI: 0.76–0.91). The computational time for training all five CNNs was approximately 6 hours, with additional time for training the Random Forest classifier and testing patient-level classification.

This method demonstrates the efficacy of using CNNs combined with first-order statistical feature extraction and Random Forest classifiers for detecting clinically significant PCa from DWI MRI scans, providing a promising tool for PCa detection and diagnosis.

III. DATASET

The dataset is structured in a well-organized folder system to facilitate efficient access and usability for research purposes, particularly in the study of prostate cancer using MRI, ultrasound, and biopsy data. The primary data folder contains various subfolders, each storing different types of information essential for analysis, such as medical images, 3D models, biopsy overlays, and annotations.

A. Dataset Structure and Organization

The dataset is meticulously organized to ensure efficient access and usability for research, particularly in the study of prostate cancer using MRI, ultrasound, and biopsy data. The primary data folder consists of several subfolders, each dedicated to storing different types of information essential for analysis, such as medical images, 3D models, biopsy results, and annotations.The Directory Structure is given below:

```
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
```

B. Biospsy Overlays(3D)

This sub-directory includes directories for each patient, and within each, there are files like Bx-S.No.-Benign.fcsv, which represent biopsy points and classifications (e.g., benign or cancerous tissue). These .fcsv files contain coordinates that indicate where biopsy samples were taken relative to the prostate's anatomy in the 3D space additionally also indicating the location of presence of cancerous tissue.

C. STLs

In this subfolder, STL files were used to store 3D surface models that are crucial for visualizing and analyzing anatomical structures. These files are a part of the 3D reconstruction process, where the data from ultrasound was converted into 3D models. The .stl (stereolithography) format was used in 3D printing and medical imaging.All in all this sub-directory contains .stl ultrasound files arranged into sub-folders named patient-id wise.

D. prostate-mri-us-biopsy

This contains mri scans stored in the form of.dcm format.These DICOM files are the standard format for medical imaging data and include not only the images but also metadata such as patient information, scan parameters, and anatomical details. Each patient in the dataset underwent multiple MRI scans, captured in various planes and orientations (such as axial, sagittal, and coronal views), providing a comprehensive view of the prostate anatomy.These dicom(.dcm) files were arranged as per patient-id into subdirectories.

E. TCIA Biopsy Data_2020-07-14.xlsx

This file contains detailed clinical and biopsy-related information for each patient in the dataset. The key data points provided include:

- **PSA (ng/mL):** Prostate-Specific Antigen level, a biomarker used to screen for prostate cancer.
- **Primary Gleason:** The Gleason score for the most common cancerous tissue observed in the biopsy.
- **Secondary Gleason:** The Gleason score for the second most common cancerous tissue type.
- **Cancer Length (mm):** The length of cancerous tissue measured in millimeters.
- **% Cancer in Core:** The percentage of the biopsy core that is affected by cancer.
- **Core Fragment Tissue Lengths (mm):**
  - **Core Fragment #1 Tissue Length**
  - **Core Fragment #2 Tissue Length**
  - **Core Fragment #3 Tissue Length**
- **Bx Coordinates:** Biopsy needle tip and base coordinates in both MRI and ultrasound spaces:
  - **MRI Coordinates:** Bx Tip X, Y, Z and Bx Base X, Y, Z.
  - **US Coordinates:** Bx Tip X, Y, Z and Bx Base X, Y, Z.
- **Prostate Volume (CC):** The volume of the prostate gland measured in cubic centimeters.
- **Core Label:** The label assigned to each biopsy core sample.
- **Series Instance UID (US and MRI):** Unique identifiers for the ultrasound and MRI imaging series.
- **Patient Number:** A unique identifier for each patient in the dataset.

F. Target Data_2019-12-05.xlsx

This file contains the target label data, including UCLA scores and other key labels, which will be used for prostate cancer risk prediction. The UCLA score serves as a key target variable in our model, helping to assess the risk of prostate cancer based on biopsy and imaging data(part II).

G. metadata.csv

The `metadata.csv` file contains details such as timestamps, information about the devices used for imaging, study dates, and the number of images available for each patient. It provides essential context regarding the technical aspects of the dataset.

Each of the aforementioned files—`metadata.csv`, `Target Data_2019-12-05.xlsx`, and `TCIA Biopsy Data_2020-07-14.xlsx`—contains patient and series identifiers, which are crucial for linking the clinical data with the corresponding images.

IV. Methodology and Model Details

This study proposed a structured approach for processing and analyzing medical imaging data to detect and classify prostate cancer. The methodology was divided into two parts: **Part I** focused on cancer detection (binary classification), and **Part II** addressed cancer risk prediction (multi-class classification). Below, we detail the methods employed in each part.

A. Part I: Cancer Detection (Binary Classification)

The first step involved converting DICOM images, commonly used in medical imaging, into JPG format. This conversion enabled easier manipulation and the application of standard image processing techniques. Once converted, the images underwent several preprocessing steps, including:

- **Resizing**: Standardizing image resolution.
- **Cropping**: Ensuring the region of interest (ROI) was properly focused.
- **Patient-wise collages**: Generating collages to consolidate relevant image information.

After preprocessing, the images were categorized into two distinct folders or linked to a metadata sheet that associated binary labels: **positive** for images with cancerous regions and **negative** for those without. This classification was based on the cancer presence in the core samples (*Percent Cancer in Core*). Samples with a cancer percentage greater than 0 were labeled as 1, and those with 0 percent were labeled as 0. These binary labels served as the output variables for classification.

In addition to the preprocessing above, we extracted both pixel data (around 49152 pixels per image) and specific image features to use them as the two sub-parts in this task as inputs for models other than CNN(ie. two subparts per model). Feature extraction focused on capturing statistical and texture-based properties of the images(Canny Edges , Gradient Histogram , LBP ,Entropy):

- **Multilayer Perceptron (MLP)**: Used for its capability to generalize well on pixel data. Activation functions such as *tanh*, *ReLU*, and *logistic* were employed to learn complex nonlinear patterns.
- **Other Models**: Included for robustness in classification tasks:
  - Decision Trees (DT)
  - Naive Bayes (NB)
  - Random Forests (RF)
  - Logistic Regression (Log R)

By integrating visual information from the images and quantitative data from metadata, this approach provided a cohesive and effective system for cancer detection.

B. Part II: Cancer Risk Prediction

The second part of the study focused on predicting the risk level of prostate cancer using the UCLA score. This involved utilizing the biopsy sheet (**TCIA Biopsy Data_2020-07-14.xlsx**) containing critical metadata, such as:

- Percentage of cancer in the core biopsy samples
- PSA levels
- Primary and secondary Gleason indices
- Biopsy overlay coordinates
- Cancerous tissue measurements etc.

The target sheet (**Target Data_2019-12-05.xlsx**) served as the output variable, indicating the risk level.

1) Model Training and Optimization: Over 70 models were trained for this task, incorporating both classifiers and regressors to improve classification accuracy and minimize mean squared error (MSE) for regressors. Models used included:

- Logistic Regression
- Support Vector Machines (SVMs)
- Naive Bayes
- Decision Trees
- Random Forests
- K-Means Clustering
- Gradient Boosting (including AdaBoost and XGBoost)
- Voting Classifiers

2) Dimensionality Reduction and Boosting: To enhance model performance, **Principal Component Analysis (PCA)** was applied, with the number of components varied from 2 to 15. Boosting techniques were extensively employed to refine model accuracy for classifiers and reduce error metrics for regressors. These steps ensured the models could effectively handle the complexity and diversity of the metadata.

V. Results and Analysis

A. Dataset Overview - Results of EDA

The dataset used for analysis is the *Prostate MRI and Ultrasound with Pathology and Coordinates of Tracked Biopsy*. It includes imaging data from 1,151 subjects, providing MRI and ultrasound scans, biopsy coordinates, and pathology results. The data consists of 3D multiparametric

MRI and ultrasound images, STL files for biopsy core locations, and biopsy overlays.

EDA suggests that each patient on an average underwent *22 MRI scans,US and biopsys combined.*The dataset additionally is divided into *2779 series* for over *1 lakh images* including *MRI and UltraSound Scans.*A total of *44 different protocols* were used to record MRI scans of which *t2spcrstaxial oblProstate* is the most used.The dataset also has over *24000 MRI and US scans.*

Finally the heatmap suggests high correlation between UCLA score(output label) with the input features PSA,gleason scores,Cancer Length and percentage cancer in core.

B. Model Identification

Part I - Cancer Detection(Binary Classification):

- **Multilayer Perceptron (MLP)**: Due to their ability to generalize well for pixel data, MLP classifiers have been employed with activation functions such as *tanh*, *ReLU*, and *logistic*. These activation functions help the model learn complex nonlinear patterns in the data.
- **Convolutional Neural Networks (CNNs)**: Due to their vast use in image classification, we identified CNNs as appropriate for achieving high accuracy in classifying patient-wise JPG image collages as cancerous or non-cancerous. CNNs effectively learn spatial hierarchies of features, making them suitable for this task.
- **Other Models**: The following models have also been included due to their classification capabilities:
  - **Decision Trees (DT)**
  - **Naive Bayes (NB)**
  - **Random Forests (RF)**
  - **Logistic Regression (Log R)**

Part II - Cancer risk multiclass-classification(UCLA scores): As we are supposed to classify metadata such as PSA scores,Gleason Indices,Tumor Measurements into mainly 5 levels of risks identified some commonly used models and their ensembles for this task(we used both classifiers and regressors):

- Logistic Regression
- Support Vector Machines
- Naive Bayes
- Decision Trees
- Random Forest
- K-means Clustering
- Gradient Boosting

C. Model Training

- Part I: Cancer Detection:
  For the task of binary classification, we trained multiple models to identify cancerous cases based on imaging data and derived features. A combination of pixel data (from MRI and ultrasound scans) and extracted metadata features was used. We achieved the final accuracies as:
  - Decision Trees-65.862
  - GNB on Pixel Data-65.86
  - Decision Trees after feature extraction - 71.85
  - GNB after feature extraction-75.44
  - CNN on image data-76.047
  - BNB on Pixel Data- 79.041
  - Random Forest after feature extraction -82.0359
  - BNB after feature extraction- 84.431
  - Random Forest- 84.431
  - logistic regression - 85.0299
  - MLP - 85.0299
- Part II: Cancer Risk Prediction: For UCLA score classification, models were trained using metadata features (PSA levels, Gleason scores, tumor measurements ,etc.) with PCA dimensionality reduction. Logistic Regression with 7 PCA components achieved the highest accuracy (60.71) for classifiers. An MSE of 0.4265 using support vector regressor using 15 PCA components was acheived. We also did hyperparameterization but it did not yield satisfactory results however the best accuracy we were able to acheive using this were GBR and Random Forest each with an accuracy of about 58.33 for the classification task and lowest MSE for regression task was with Support Vector Regression (0.4344).

VI. CONCLUSIONS

In this study, we addressed the critical challenge of prostate cancer detection by proposing an automated system utilizing machine learning techniques for accurate and reliable diagnosis. Catering to this deadline we focused on the preprocessing of prostate MRI and biopsy images, along with the integration of metadata to enhance cancer detection , and also model training.

Subsequently, we applied various machine learning models, including Decision Trees, Random Forest, and Support Vector Machines (SVM), K-means Clustering, and other tuning techniques to predict cancer risk in terms of the UCLA Prostate Cancer Index and to classify images using CNNs,MLP and other classification models.

In the end we were able to achieve considerably good accuracy scores and low MSE scores for both the parts I and II thus being able to effectively detect and risk classify US images for effective aiding of prostrate cancer diagnosis.