

Prostate Cancer Prediction

Intelligent Systems

Prof. Abdel-Rahman Header

Monday, 19 May 2022

عبدالله احمد مهدي
عبدالرحمن جمال احمد
عدنان مصطفى فريد
عمر اشرف عمر
علي احمد عبدالحميد

1 Contents

2	Introduction	3
2.1	Project problem	3
2.2	problem definition and importance.....	3
2.3	Problem analysis	4
2.3.1	Tests for prostate cancer	4
2.3.2	Risk Factors	4
2.3.3	Symptoms	4
2.3.4	Complications.....	5
2.4	Available algorithms.....	5
2.4.1	The selected algorithm	5
3	Methodology:.....	5
3.1	First step.....	5
3.2	Second step.....	6
3.3	Third step	6
3.4	Fourth step.....	6
3.5	Time Complexity and Computational cost.....	6
4	Dataset	6
4.1	Data sample	6
4.2	Data Description	7
5	Experimental simulation:	7
5.1	Tools and technologies used:.....	7
5.2	Structure and Workflow.....	8
5.3	Primary function	8
5.4	The Machine learning model	8
5.4.1	Importing the necessary package	8
5.4.2	Importing the dataset	9
5.4.3	Divide the dataset into train and test	9
5.4.4	Importing the random forest algorithm	9
5.4.5	Training the model.....	9
5.4.6	Getting the accuracy of the model	9
5.4.7	Saving the trained model	10
6	Running and installation:	10
6.1	First step.....	10
6.2	Second step.....	10

6.3	Third step	10
6.4	Fourth step.....	10
6.5	Fifth step	11
7	Results and technical discussion:.....	11
7.1	Main program results and output.....	11
7.2	Test/evaluation experimental Procedure	11
7.3	Quality of the test	11
7.4	analysis of the accuracy of the model	11
7.5	Optimize the quality.....	11
8	Conclusions:	12
9	References:	12

2 Introduction

The prostate is both an adornment organ of the male regenerative framework and a muscle-driven mechanical switch between urination and ejaculation. It is found as it were in a few warm-blooded animals. It varies between species anatomically, chemically, and physiologically. Anatomically, the prostate is found underneath the bladder, with the urethra passing through it. It is depicted in net life structures as comprising of projections and in microanatomy by zone. It is encompassed by a flexible, fibromuscular capsule and contains glandular tissue as well as connective tissue. (Fayed, 2021)

2.1 Project problem

Project problem is all about Prostate cancer, Prostate cancer (PCA) is the second most common cancer and the fifth leading cause of cancer-attributed death in men worldwide with an estimated incidence of 1 276 106 and 358 989 deaths in 2018. In the UK, around 47 200 new cases of PCA were reported in 2015, accounting for 26% of all new cancer cases in males. PCA deaths in the UK of were around 11 600 in 2016. The global projections of PCA incidence and mortality for 2030 are 1.7 and 0.5 million, respectively. The highest incidence of PCA is seen in western societies. The significant increase of PCA incidence and diagnosis over the last three decades can be attributed mainly to the widespread implementation of the prostate-specific antigen (PSA serum test after it had been introduced in the late 1980s, Prostate cancer is a disease common to elderly men, with more than 75% of cancers being diagnosed in men over the age of 65. In recent years, however, the incidence has increased in younger age groups, In the earlier stages of the disease, prostate cancer rarely causes any specific symptoms (Sheard, 2022).

2.2 problem definition and importance

Prostate cancer is cancer that occurs in the prostate. The prostate is a small walnut-shaped gland in males that produces the seminal fluid that nourishes and transports sperm, Prostate cancer is one of the most common types of cancer. Many prostate cancers grow slowly and are confined to the prostate gland, where they may not cause serious harm. However, while some types of prostate cancer grow slowly and may need minimal or even no treatment, other types are aggressive and can spread quickly. Prostate cancer is one of the leading causes of morbidity and mortality in the world. According to World Health Organization reports, an estimated 513,000 cases and 255,000 deaths were attributed to prostate cancer in 1999.1 Each year in the United States, approximately 220,000 new cases of prostate cancer will be diagnosed, and 30,000 men will die of prostate cancer. Refinements in early detection and treatment of prostate cancer can lead to theoretic cure of a potentially disabling and deadly disease.

However, controversy exists because no conclusive direct evidence demonstrates that early detection and treatment improve length and quality of life (j.Godec, 2016).

2.3 Problem analysis

Cancer starts when cells in the body begin to grow out of control. Cells in nearly any part of the body can become cancer cells, and can then spread to other areas of the body, figure 1 shows the difference between normal prostate and cancerous prostate, The accumulating abnormal cells form a tumour that can grow to invade nearby tissue. In time, some abnormal cells can break away and spread (metastasize) to other parts of the body. (wiki, 2022)

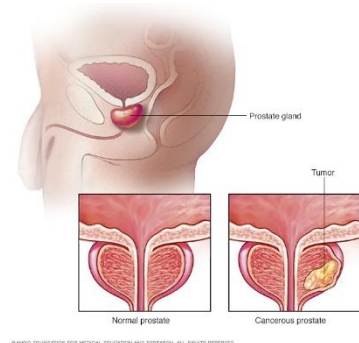


Figure 1. prostate

2.3.1 Tests for prostate cancer (Clinic, 2022)

There's no single test for prostate cancer, All the tests used to help diagnose the condition have benefits and risks that your doctor should discuss with you, the most commonly used tests for prostate cancer are:

- blood tests
- physical examination of your prostate
- MRI scan
- biopsy

2.3.2 Risk Factors (Clinic, 2022)

Factors that can increase your risk of prostate cancer include:

- Older age. Your risk of prostate cancer increases as you age. It's most common after age 50.
- Race. For reasons not yet determined, Black people have a greater risk of prostate cancer than do people of other races. In Black people, prostate cancer is also more likely to be aggressive or advanced.
- Family history. If a blood relative, such as a parent, sibling or child, has been diagnosed with prostate cancer, your risk may be increased. Also, if you have a family history of genes that increase the risk of breast cancer (BRCA1 or BRCA2) or a very strong family history of breast cancer, your risk of prostate cancer may be higher.
- Obesity. People who are obese may have a higher risk of prostate cancer compared with people considered to have a healthy weight, though studies have had mixed results. In obese people, the cancer is more likely to be more aggressive and more likely to return after initial treatment.

2.3.3 Symptoms (Clinic, 2022)

Prostate cancer may cause no signs of symptoms in its early stages but prostate cancer that's more advanced may cause signs and symptoms (Clinic, 2022) such as:

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

2.3.4 Complications (Clinic, 2022)

Complications of prostate cancer and its treatments (Clinic, 2022) include:

- Cancer that spreads (metastasizes). Prostate cancer can spread to nearby organs, such as your bladder, or travel through your bloodstream or lymphatic system to your bones or other organs. Prostate cancer that spreads to the bones can cause pain and broken bones. Once prostate cancer has spread to other areas of the body, it may still respond to treatment and may be controlled, but it's unlikely to be cured.
- Incontinence. Both prostate cancer and its treatment can cause urinary incontinence. Treatment for incontinence depends on the type you have; how severe it is and the likelihood it will improve over time. Treatment options may include medications, catheters and surgery.
- Erectile dysfunction. Erectile dysfunction can result from prostate cancer or its treatment, including surgery, radiation or hormone treatments. Medications, vacuum devices that assist in achieving erection and surgery are available to treat erectile dysfunction.

2.4 Available algorithms

Our problem is supervised machine learning problem because there are two cases, first case is malignant and the second case is benign, there are too many machine learning algorithms that can help in solving our supervised problem like Random Forest classifier, Neighbors Classifier, Decision Tree. Each algorithm has its configurations and properties

2.4.1 The selected algorithm (Yiu, 2019)

The selected algorithm is random forest classifier, it is popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

3 Methodology:



Figure 2. process flowchart

Figure 2 shows the methodology of the system involves extracting some features from the medical attributes of the prostate, nearly 10 properties depending on our data, and then introducing these features to intelligent learning model that uses Random Forest classifier, then the smart learning model determine whether tumour benign or malignant.

3.1 First step

First step is entering the medical attributes of the prostate, attributes include Radius, symmetry, Texture, Perimeter, Area, smoothness, Compactness, Fracture dimension.

3.2 Second step

Second step is extracting the features that have a good correlation with the remaining variables and eliminate the features that have a low correlation like Fracture dimension, Texture.

3.3 Third step

Third step is passing the data to the trained model.

3.4 Fourth step

Fourth step is getting the prediction results from the trained model.

3.5 Time Complexity and Computational cost (Yiu, 2019)

The computational complexity at test time for a Random Forest of size T and maximum depth D (excluding the root) is $O(T \cdot D)$. However, the computational cost can be lower if trees are not balanced. It must be noted that unification costs of ensemble methods, left out in the theoretical cost computation as they are often negligible, in the regime where Random Forest operates become quite significant and dominate the total cost. Another important cost to be considered in our method is memory space, exponential in the depth of the tree: $O(2^D)$.

4 Dataset

This is the dataset of 100 patients to implement the machine learning algorithm and thereby interpreting results, The data set consists of 100 observations and 10 variables (out of which 8 numeric variables and one categorical variable and is ID) which are as follows:

- Id
- Radius
- Texture
- Perimeter
- Area
- Smoothness
- Compactness
- Diagnosis Result
- Symmetry
- Fractal dimension

4.1 Data sample

	id	diagnosis_result	radius	texture	perimeter	area	smoothness	compactness	symmetry	fractal_dimension
0	1	M	23	12	151	954	0.143	0.278	0.242	0.079
1	2	B	9	13	133	1326	0.143	0.079	0.181	0.057
2	3	M	21	27	130	1203	0.125	0.160	0.207	0.060
3	4	M	14	16	78	386	0.070	0.284	0.260	0.097
4	5	M	9	19	135	1297	0.141	0.133	0.181	0.059

Figure 3. Data Sample

Figure 3 shows Data sample contains all the medical attributes of the prostate including diagnosis result, radius, texture, perimeter, area, smoothness, compactness, symmetry, fractal dimension, also note that in the diagnosis results column M indicates to malignant and B indicates to benign.

4.2 Data Description

```
data.describe()
```

	diagnosis_result	radius	area	smoothness	compactness	symmetry
count	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000
mean	0.380000	16.850000	702.880000	0.102730	0.126700	0.193170
std	0.487832	4.879094	319.710895	0.014642	0.061144	0.030785
min	0.000000	9.000000	202.000000	0.070000	0.038000	0.135000
25%	0.000000	12.000000	476.750000	0.093500	0.080500	0.172000
50%	0.000000	17.000000	644.000000	0.102000	0.118500	0.190000
75%	1.000000	21.000000	917.000000	0.112000	0.157000	0.209000
max	1.000000	25.000000	1878.000000	0.143000	0.345000	0.304000

Figure 4. Data Description

Figure 4 shows Data description contains information about the data like the mean, count, min, max, std,

5 Experimental simulation:

The experimental simulation of our project will describe the programming languages and the environments that we used in implementing the prostate machine learning model and the random forest algorithm that we selected in the model, also will describe the test cases that we used and the results of that test cases.

5.1 Tools and technologies used:

We used so many tools to implement and visualize our intelligent model like:

- Python programming language
- Html
- CSS
- Fast API framework
- Uvicorn asynchronous server gateway
- NumPy library
- Pandas' library
- Matplotlib library
- Sklearn library
- Pickle library

5.2 Structure and Workflow

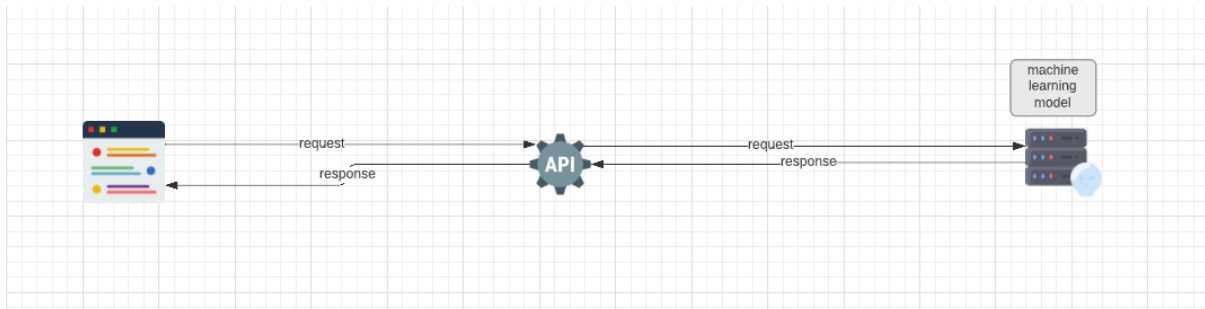


Figure 5.workflow

Figure 5 shows all the structure and components of the system, there are two main components in the system:

- Web Server API
- The Machine learning model

A user sends a request to the API interface and the API processes the request and pass it to the machine learning model then the model make predictions based on the given request data and response back to the API with the result and finally the API passes the result to the user.

5.3 Primary function

```

@app.post("/prostates", status_code=status.HTTP_201_CREATED)
def create_heart(request: Request, Radius: str = Form(...), Texture: str = Form(...), Perimeter: str = Form(...))
    prediction = classifier.predict([[Radius, Area,
                                     Smoothness, Compactness, Symmetry]])
    if prediction[0] == 0:
        return templates.TemplateResponse("result.html", {"request": request, "result": "Malignant Tumor"})
    else:
        return templates.TemplateResponse("result.html", {"request": request, "result": "Benign Tumor"})
  
```

Figure 6.API primary function

Figure 6 shows the API the primary function that receives the requests and pass it to the intelligent model.

5.4 The Machine learning model

For the machine learning model, it's divided into seven parts:

5.4.1 Importing the necessary package

```

import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report
import os
from sklearn.metrics import accuracy_score
import pickle
  
```

Figure 7.importing packages

Figure 7 shows all the imported packages including NumPy, pickle, ...

5.4.2 Importing the dataset

```
data = pd.read_csv('./Prostate_Cancer.csv')
```

```
data.head()
```

	id	diagnosis_result	radius	texture	perimeter	area	smoothness	compactness	symmetry	fractal_dimension
0	1	M	23	12	151	954	0.143	0.278	0.242	0.079
1	2	B	9	13	133	1326	0.143	0.079	0.181	0.057
2	3	M	21	27	130	1203	0.125	0.160	0.207	0.060
3	4	M	14	16	78	386	0.070	0.284	0.260	0.097
4	5	M	9	19	135	1297	0.141	0.133	0.181	0.059

Figure 8.importing dataset

Figure 8 shows the first 10 records of the data including all the medical attributes

5.4.3 Divide the dataset into train and test

```
X = data.drop(['diagnosis_result'], axis=1) # Features
y = data['diagnosis_result'] # Labels
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state=10)
```

Figure 9.dividing the dataset

Figure 9 shows dividing the data into features and labels also showing the function that is responsible for splitting the dataset into training data and testing data .

5.4.4 Importing the random forest algorithm

```
from sklearn.ensemble import RandomForestClassifier
```

```
forest = RandomForestClassifier(n_estimators = 50)
```

Figure 10.importing the random forest

Figure 10 shows how to import the random forest algorithm from the Sklearn package and how to set the configuration of the algorithm.

5.4.5 Training the model

```
forest.fit(X_train,y_train)
```

```
pred_forest = forest.predict(X_test)
```

Figure 11.training the model

Figure 11 shows how to train the model on the given data set, Also Getting the accuracy of the model.

5.4.6 Getting the accuracy of the model

```
score=accuracy_score(y_test,pred_forest)
```

Figure 12.getting the accuracy

Figure 12 shows the function that is responsible for getting the accuracy of the model and in our model, it was 90 % accuracy.

5.4.7 Saving the trained model

```
pickle_out = open("classifier.pkl", "wb")
pickle.dump(forest, pickle_out)
pickle_out.close()
```

Python

Figure 13. saving the model

Figure 13 shows pickle function that is responsible for saving our trained model for future use.

6 Running and installation:

This section will explain all the steps to get the model up and running.

6.1 First step

First step is navigating to the project directory.

6.2 Second step

Second step is running the API with python command.

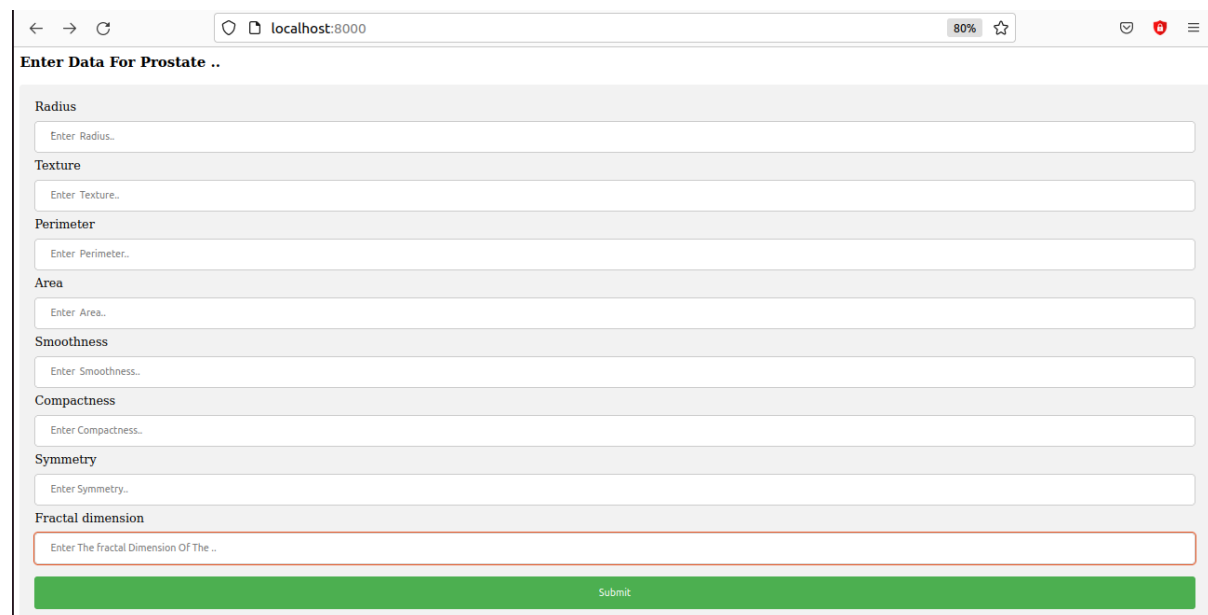
\$ Python main.py

6.3 Third step

Third step is accessing the service page by writing the IP address (local host) in the browser followed to the port number 8000.

6.4 Fourth step

Fourth step is entering all the required data fields.



The screenshot shows a web browser window with the address bar set to `localhost:8000`. The page title is "Enter Data For Prostate ..". The form contains the following fields:

- Radius:
- Texture:
- Perimeter:
- Area:
- Smoothness:
- Compactness:
- Symmetry:
- Fractal dimension:

A green "Submit" button is located at the bottom of the form.

Figure 14. input fields

Figure 14 shows all the required prostate medical attributes.

6.5 Fifth step

fifth step is clicking on the submit button to send the request to the server and then get the results back from the server.

7 Results and technical discussion:

This section will explain all the results getting back from the model.

7.1 Main program results and output

The output of the model will be either malignant (0) or benign (1).

7.2 Test/evaluation experimental Procedure

our data split into two parts train data that represents 80% of the data and the test data that represents 20% .

7.3 Quality of the test

The results of our test cases are getting 90% accuracy, The accuracy of the model is good but, in the future, we maybe add more datasets and use optimizations techniques to improve the accuracy.

7.4 analysis of the accuracy of the model

The Random Forest is a powerful tool for classification problems, but as with many machine learning algorithms, it can take a little effort to understand exactly what is being predicted and what it means in context. Luckily, Scikit-Learn makes it pretty easy to run a Random Forest and interpret the results. In this post I'll walk through the process of training a straightforward Random Forest model and evaluating its performance using confusion matrices and classification reports. I'll even show you how to make a color-coded confusion matrix using Seaborn and Matplotlib, also normally accuracy is not the metric we use to judge the performance of a classification model for reasons such as possible imbalances in data leading to high accuracy due to imbalanced predictions to one class. However, for simplicity reasons I included it above. I also included the F1 score, which measures the harmonic mean between precision and recall. The F1 score metric is able to penalize large differences between precision. Generally speaking, we would prefer to determine a classification's performance by its precision, recall, or F1 score. (Kreiger, 2020)

7.5 Optimize the quality (Kreiger, 2020)

There are so many optimization techniques that can be used to optimize the quality like:

- Machine learning model results will change if the training dataset is changed.
- Stochastic machine learning algorithms use randomness during learning, ensuring a different model is trained each run.
- Differences in the development environment, such as software versions and CPU type, can cause rounding error differences in predictions and model evaluations.
- Specify the maximum depth of the trees. By default, trees are expanded until all leaves are either pure or contain less than the minimum samples for the split. This can still cause the trees to overfit or underfit. Play with the hyperparameter to find an optimal number for max depth.
- Increase or decrease the number of estimators. How does changing the number of trees affect performance? More trees usually mean higher accuracy at the cost of slower learning. If you wish to speed up your random forest, lower the number of estimators. If you want to increase the accuracy of your model, increase the number of trees.

- Specify the maximum number of features to be included at each node split. This depends very heavily on your dataset. If your independent variables are highly correlated, you'll want to decrease the maximum number of features. If your input attributes are not correlated and your model is suffering from low accuracy, increase the number of features to be included.

8 Conclusions:

In conclusion we would like to say that AI and Machine Learning is paving the way for the future and one of their largest and most useful applications is in the medical field. Research on curing and treating cancer has been a prominent focus for years. With the recent advancements of AI and Machine Learning, it will soon be possible to use them to make more insightful observations and create better treatment plans.

Also, there are a lot of things that we have not identified which ones we will use in the project as hyperparameters values, optimization algorithm, and the model that we will use it because all these things are depend on try and enhance during implementation phase but, we will roughly choose all these things at the end of project.

9 References

- Clinic, M. (2022, 5 25). Retrieved from <https://www.mayoclinic.org/diseases-conditions/prostate-cancer/symptoms-causes/syc-20353087>
- Fayed, L. (2021, Dec 12).
- j.Godec, M. a. (2016). *Prostate Cancer book*.
- Kreiger, J. (2020, Jan 13). Retrieved from <https://medium.com/analytics-vidhya/evaluating-a-random-forest-model-9d165595ad56>
- Sheard, R. (2022). *understanding prostate cancer book*.
- Veldhuis, W. (2022). Retrieved from <https://www.quantib.com/the-ultimate-guide-to-ai-in-prostate-cancer>
- wiki. (2022, May 26). Retrieved from https://en.wikipedia.org/wiki/Prostate_cancer
- Winslow, T. (2021, 5 5). Retrieved from <https://www.cancer.gov/about-cancer/understanding/what-is-cancer>
- Yiu, T. (2019, Jun 12). Retrieved from <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>