ISSN No:-2456-2165

The Prostate Cancer Detection Using K-Nearest Neighbor (KNN)

First A. Sudipta Saha B. Kazi Muntashir Fahad C. Jannatuil Ferdouse Jannat D. Mahamoda Rupa E. Rukaiya Islam

Abstract:- The ongoing threat that cancer poses to the health and prosperity of the world highlights the critical need for early identification and efficient treatment. Machine learning and artificial intelligence have become effective tools for the early detection of diseases like cancer. The K-Nearest Neighbors (KNN) method stands out among them for its efficiency and simplicity. The goal of this study is to use the R implementation of the KNN algorithm to advance the identification of prostate cancer. A difficult diagnostic challenge is presented by the complicated and multifaceted illness of prostate cancer. This work seeks to develop precision medicine in oncology by improving the accuracy and reliability of prostate cancer detection using the capabilities of KNN. The study examines the prostate cancer detection landscape, presents the KNN algorithm's uses in medicine, and describes the study's goals. The KNN model is trained and tested using a dataset that has been pre-processed and used in this methodology. The findings have the potential to revolutionize the detection of prostate cancer by offering a data-driven strategy to supplement healthcare professionals' clinical judgement, thereby improving patient outcomes, and even saving lives.

Keyword:- *Prostate Cancer, Early Detection, Machine Learning, K-Nearest Neighbors (KNN), Precision Medicine, Diagnosis, Artificial Intelligence.*

I. INTRODUCTION

The health of people is still being relentlessly attacked by cancer, which poses a threat to our society's prosperity. Early cancer identification is essential for effective treatment and better patient outcomes. The use of artificial intelligence and machine learning have become potent instruments in this era of technological innovation to help with the early identification of numerous diseases, including cancer. The K-Nearest Neighbors (KNN) method distinguishes out among these techniques for its ease of use and potency. Prostate cancer is one of the most prevalent and potentially lethal cancers that affects men worldwide. Early identification is crucial to enhance the likelihood of effective therapy and improve the patients' overall prognosis. The convergence of medicine and machine learning has paved the path for novel methods to cancer diagnosis in recent years. Among these methods, the K-Nearest Neighbors (KNN) algorithm, a tried-and-true machine learning method, has shown promise in assisting in the early identification of prostate cancer. This research embarks on a journey to harness the power of the KNN algorithm, implemented in R, to further the field of prostate cancer detection. Prostate cancer is known for its complex and multifactorial nature, making it a challenging disease to diagnose accurately. We want to improve the accuracy and dependability of prostate cancer detection by using KNN's capabilities, ultimately advancing precision medicine in the field of oncology. We will delve into the current landscape of prostate cancer diagnosis, introduce the KNN algorithm and its potential applications in medical contexts, and outline the objectives and significance of this research endeavor. Furthermore, we will elucidate the structure of this research paper and provide insights into the methodology employed to achieve our goals. The results of this study have the potential to completely change how prostate cancer is diagnosed by providing a reliable, data-driven strategy that enhances healthcare practitioners' existing knowledge. This study aims to usher in an era where early diagnosis of prostate cancer becomes not only a possibility, but a practical reality, ultimately improving patient outcomes and saving lives.

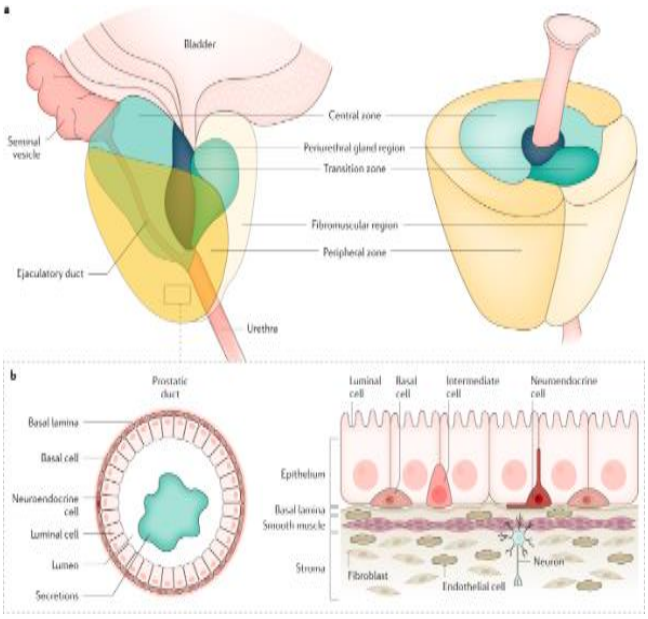

Fig 1: Prostate Cancer

ISSN No:-2456-2165

II. LITERATURE REVIEW

One of the most prevalent cancers that affect men globally is prostate cancer, with significant regional variations in incidence and fatality rates. This literature review seeks to present a summary of important prostate cancer-related topics, such as risk factors, diagnosis, treatment options, and new developments in research. The incidence rates of prostate cancer show significant regional variations. Asia and Africa have lower rates of prostate cancer than do North America, Western Europe, and Australia, according to [1]. Age, family history, ethnicity, and genetic variables are all risk factors for prostate cancer development [2]. African American males are at higher risk than Caucasian men, according to studies, underscoring the significance of genetic and racial variables [3]. Prostate cancer is diagnosed by a combination of clinical evaluation, digital rectal examination (DRE), prostate-specific antigen (PSA) testing, and prostate biopsy. The use of PSA as a screening tool is still debatable because to concerns regarding over diagnosis and overtreatment. [4]. According to the United States Preventive Services Task Force (USPSTF), shared decision-making between patients and medical practitioners is crucial when thinking about PSA screening. [5]. Prostate cancer treatment choices differ depending on the disease's stage, the patient's age, and general health. Common strategies include active surveillance, radical prostatectomy, radiation therapy, and androgen deprivation therapy (ADT). Hamady's work highlighted the significance of making personalized treatment decisions by showing that there was no discernible difference in survival between surgery and radiotherapy for localized prostate cancer. [6]. Novel therapeutic approaches have been created as a result of recent developments in the study of prostate cancer. Precision medicine, immunotherapy, and targeted medicines have showed promise in enhancing therapeutic outcomes [7]. Furthermore, genomic profiling has made it possible to identify particular genetic changes that may inform therapy choices [8]. An exciting area of ongoing research is the use of liquid biopsies to assess therapy response and identify recurrence [9]. Despite having a wide range of risk factors and treatment options, prostate cancer continues to be a major public health concern. In the developing fields of screening and diagnosis, coordinated decision-making is crucial. Recent scientific developments give patients with prostate cancer hope for better therapeutic strategies and outcomes.

Recent scientific developments give patients with prostate cancer hope for better therapeutic strategies and outcomes.

III. METHODOLOGY

For collecting the accurate resource, we have visit several website and gather some valuable research. We have found 396 articles from 4 different website. From Google scholar we got 251, Springer link 109, Science direct 34, IEEE Xplore 2 paper. Initially we have selected 396 in total and from there we have finally selected 43 paper.

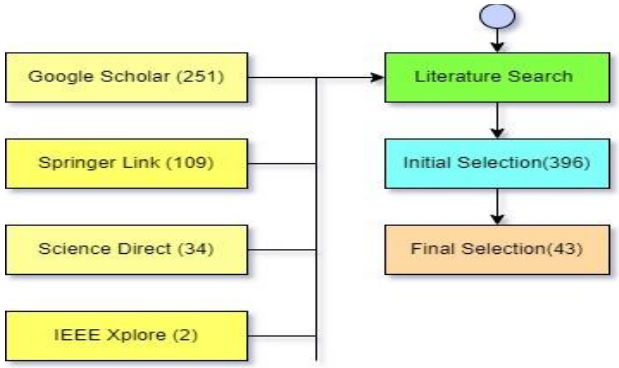

Fig 2: Resource Finding

- *Dataset*

The dataset has been collected from Kaggle and it is a numeric dataset. The dataset is in CSV format, most of the values are in numeric. The majority of the values in the dataset are in numeric CSV format. The data set has 100 observations and 10 variables, including 8 numeric variables, 1 categorical variable, and ID, as follows: Id, Radius, Texture, Perimeter, Area, Smoothness, Compactness, diagnosis result, Symmetry, and Fractal dimension. From all the attributes we try to find the diagnosis result. Here is the link of the dataset which we have used: https://www.kaggle.com/datasets/sajidsaifi/prostate-cancer.

- *Pre-processing*

We have loaded the dataset from the local storage and view the columns names and the values. When we checked the data type, we discovered that the diagnosis result was of the character data type, but the integer data types for Id, Radius, Texture, Perimeter, and Area were Smoothness, Compactness, Symmetry, and Fractal Dimension. Firstly, we gather knowledge about the data type. For better efficient value we summarize the data and found mean, median and mode. The attribute id is an identical number and has no impact on the other data of the dataset so we decided to remove this attribute. Converting the diagnosis-result column to numeric. Visualizing the values using histogram. Missing value shows us that in which rows the value is missing. Normalizing the data is a important part. Using min-max normalization the whole data are normalized. But the chosen data set has no missing values. Finding the co-relation between the attribute and have the graphical value.

ISSN No:-2456-2165

➢ *Classification*

Here in the data set the classification has been done depend on the values. The data has been classified for training and test data. By depending on the training data, the testing data has been checked. We have divided the dataset in 65% and 35% where training is 65% and testing is 35%. Then depending on the training and test data the value of k has been found.

➢ *Model Selection*

For this project we are using a k-Nearest Neighbors (KNN) as our algorithm. We have found a dataset in CSV format and for CSV format the Using KNN is better to apply. It is one of the supervised learning techniques. There are several types of algorithms but K-Nearest Neighbor is one of the easy machine learning algorithms. If new data appeared then it can be classified into category by using KNN. KNN don't make any assumption with undelay data, also for our convenience we are using KNN. One of the most versatile algorithms which can be used for any CSV dataset. It helps to easily classify any data from a dataset. For better accuracy we have use the co relation wise feature. The prediction has been made from the training data to the testing data. As the model performance using metrics like accuracy, precision, recall, and presents a confusion matrix, helping assess the model in predicting prostate cancer diagnosis results.

```
> summary(mydata)
       id          diagnosis_result       radius          texture         perimeter
 Min.   :  1.00   Length:100         Min.   : 9.00   Min.   :11.00   Min.   : 52.00
 1st Qu.: 25.75   Class :character   1st Qu.:12.00   1st Qu.:14.00   1st Qu.: 82.50
 Median : 50.50   Mode  :character   Median :17.00   Median :17.50   Median : 94.00
 Mean   : 50.50                      Mean   :16.85   Mean   :18.23   Mean   : 96.78
 3rd Qu.: 75.25                      3rd Qu.:21.00   3rd Qu.:22.25   3rd Qu.:114.25
 Max.   :100.00                      Max.   :25.00   Max.   :27.00   Max.   :172.00
      area         smoothness       compactness        symmetry       fractal_dimension
 Min.   : 202.0   Min.   :0.0700   Min.   :0.0380   Min.   :0.1350   Min.   :0.05300
 1st Qu.: 476.8   1st Qu.:0.0935   1st Qu.:0.0805   1st Qu.:0.1720   1st Qu.:0.05900
 Median : 644.0   Median :0.1020   Median :0.1185   Median :0.1900   Median :0.06300
 Mean   : 702.9   Mean   :0.1027   Mean   :0.1267   Mean   :0.1932   Mean   :0.06469
 3rd Qu.: 917.0   3rd Qu.:0.1120   3rd Qu.:0.1570   3rd Qu.:0.2090   3rd Qu.:0.06900
 Max.   :1878.0   Max.   :0.1430   Max.   :0.3450   Max.   :0.3040   Max.   :0.09700
```

Fig 3 Summary

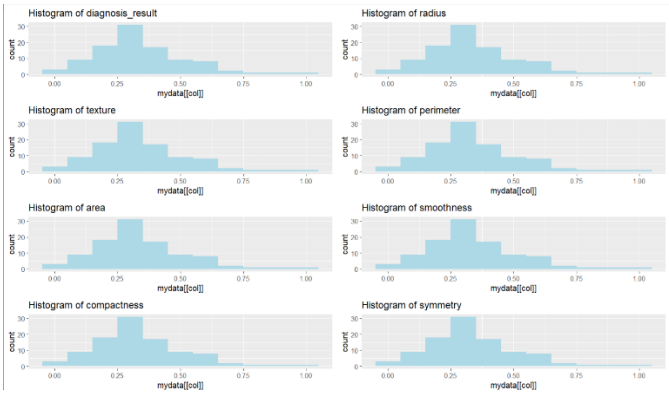

Fig 4 Histogram

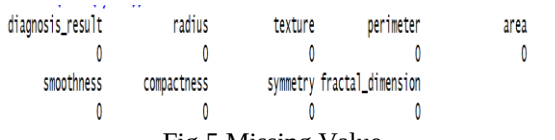

Fig 5 Missing Value

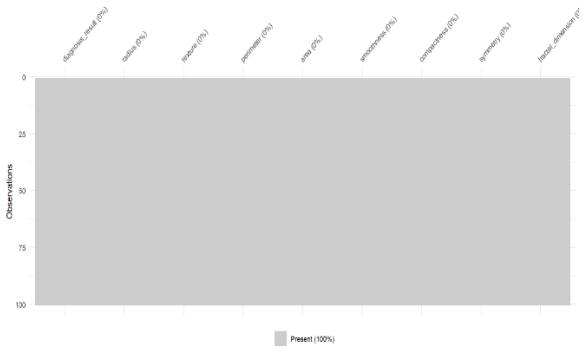

Fig 6 Missing Value Visualization (Observation vs Present)

Table 1 Normalization

```
> normalize(mydata)
   diagnosis_result      radius     texture  perimeter      area   smoothness  compactness
1      0.0010447496 0.012227084 0.006369671 0.08038608 0.5079773 5.591167e-05 1.277981e-04
2      0.0005122574 0.004772195 0.006902163 0.07080122 0.7060643 5.591167e-05 2.183218e-05
3      0.0010447496 0.011162100 0.014357053 0.06920374 0.6405678 4.632682e-05 6.496404e-05
4      0.0010447496 0.007434655 0.008499640 0.04151415 0.2055217 1.703975e-05 1.309931e-04
5      0.0010447496 0.004772195 0.010097116 0.07186620 0.6906221 5.484669e-05 5.058675e-05
6      0.0005122574 0.013292069 0.013292069 0.04417661 0.2539785 4.792429e-05 7.028896e-05
7      0.0010447496 0.008499640 0.013824561 0.06387882 0.5537716 3.035205e-05 3.780694e-05
8      0.0010447496 0.007967147 0.009564624 0.04790406 0.3077602 4.313186e-05 6.762650e-05
9      0.0010447496 0.010097116 0.012759577 0.04683907 0.2768757 4.739180e-05 8.253628e-05
10     0.0010447496 0.013292069 0.005837179 0.04470910 0.2534460 4.313186e-05 1.075634e-04
11     0.0010447496 0.012759577 0.011162100 0.05482646 0.4249085 2.342965e-05 1.544227e-05
12     0.0010447496 0.009032132 0.007967147 0.05535895 0.4158561 3.141704e-05 4.845678e-05
13     0.0005122574 0.007434655 0.007967147 0.07026873 0.5979684 3.141704e-05 1.107584e-04
14     0.0010447496 0.006369671 0.011694592 0.05535895 0.4169211 2.449464e-05 3.301451e-05
15     0.0010447496 0.006369671 0.006902163 0.05003403 0.3077602 3.993691e-05 1.017060e-04
16     0.0010447496 0.011694592 0.010097116 0.05163150 0.3508921 4.046940e-05 6.496404e-05
17     0.0010447496 0.005304687 0.008499640 0.05056652 0.3647369 3.248202e-05 1.810473e-05
18     0.0010447496 0.007967147 0.007434655 0.05748892 0.4254410 4.206688e-05 8.732871e-05
19     0.0010447496 0.010629608 0.007434655 0.06920374 0.6709199 3.194953e-05 3.461199e-05
20     0.0005122574 0.009032132 0.005837179 0.04630658 0.3013703 3.194953e-05 2.289716e-05
21     0.0005122574 0.008499640 0.007434655 0.04577409 0.2768757 3.727445e-05 4.739180e-05
22     0.0005122574 0.009032132 0.012759577 0.03192929 0.1458826 3.407950e-05 1.437729e-05
23     0.0010447496 0.010629608 0.014357053 0.05482646 0.3748542 3.674196e-05 9.371862e-05
24     0.0010447496 0.010097116 0.006369671 0.07293119 0.7475987 2.981956e-05 3.407950e-05
25     0.0010447496 0.004772195 0.006902163 0.05855390 0.4818851 3.940442e-05 5.750915e-05
26     0.0010447496 0.010097116 0.014357053 0.06174885 0.4861451 4.313186e-05 1.011735e-04
27     0.0010447496 0.005304687 0.012759577 0.05163150 0.3434372 3.567697e-05 7.934133e-05
28     0.0010447496 0.008499640 0.012759577 0.06494381 0.5825262 2.981956e-05 3.674196e-05
29     0.0010447496 0.007967147 0.007967147 0.05429396 0.3897640 3.727445e-05 7.028896e-05
```

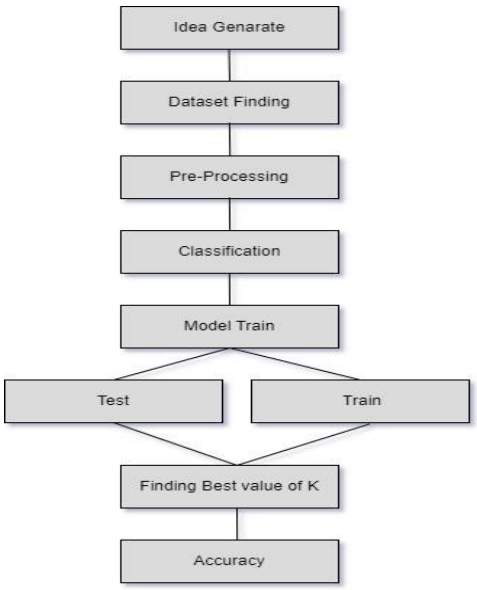

Fig 7 Implementation Block Diagram

ISSN No:-2456-2165

IV. RESULT & ANALYSIS:

In the system the accuracy has been 91.43%. Our dataset has been collected from online website. It is a trained and tested depending on the dataset. In the field of medical science, the accuracy rate is pretty much acceptable. The value of k=6 here in our model. The whole model run according to the data input.

```
> cat("Best k:", best_k, "\n")
Best k: 6
> cat("Best Accuracy:", round(best_accuracy * 100, 2), "%\n")
Best Accuracy: 91.43 %
```

Fig 8 Accuracy of the implementation.

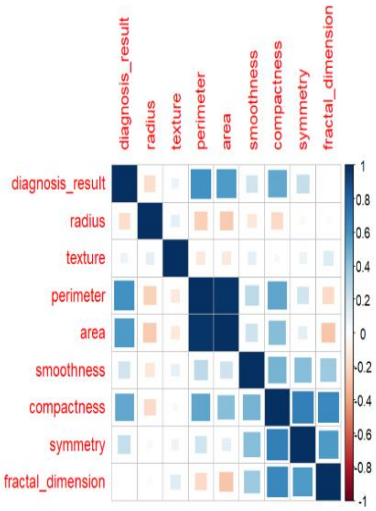

Fig 9 Confusion Matrix with labels

All the attributes are in the confusion matrix and by this we got the idea of which attribute related to others. Model accuracy curve visualize the whole model accuracy in a graphical way. This model will help to detect the prostate cancer in the early stage by diagnosis result of patients and we hope that it will cure in the early stage.

```
> print("Recall (Sensitivity):")
[1] "Recall (Sensitivity):"
> print(recall)
        1         2
0.9090909 0.9583333
>
> print("Precision:")
[1] "Precision:"
> print(precision)
        1         2
0.9090909 0.9583333
```

Fig 10 The accuracy rate at the end.

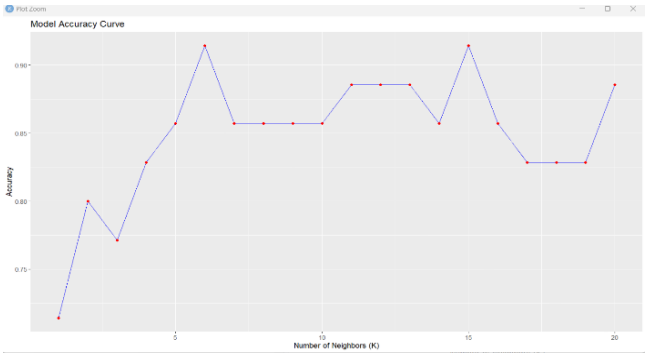

Fig 4: Model Accuracy Curve

V. CONCLUSION

In this paper our concept is to check the diagonesis_result and observe the prostate cancer rate for a patient. Prostate cancer in the new life common disease for nowadays. To identify the cancer in a short time is the core of this research. Most of the result conducted according to the features and we also try this to complete this research. The work has been done in a short dataset if the dataset range become high then we could get a good accuracy in future we want to work on it probably.

REFERENCES

[1]. Iqbal, Saqib, et al. "Prostate cancer detection using deep learning and traditional techniques." *IEEE Access* 9 (2021): 27085-27100.

[2]. Lorenz, A., et al. "Comparison of different neuro-fuzzy classification systems for the detection of prostate cancer in ultrasonic images." *1997 IEEE ultrasonics symposium proceedings. an international symposium (Cat. No. 97CH36118)*. Vol. 2. IEEE, 1997.

[3]. Deev, V., Solovieva, S., Andreev, E., Protoshchak, V., Karpushchenko, E., Sleptsov, A., Kartsova, L., Bessonova, E., Legin, A. and Kirsanov, D., 2020. Prostate cancer screening using chemometric processing of GC–MS profiles obtained in the headspace above urine samples. *Journal of Chromatography B*, *1155*, p.122298.

[4]. Llobet R, Pérez-Cortés JC, Toselli AH, Juan A. Computer-aided detection of prostate cancer. International Journal of Medical Informatics. 2007 Jul 1;76(7):547-56.

[5]. Wen, H., Li, S., Li, W., Li, J., & Yin, C. (2018, December). Comparision of four machine learning techniques for the prediction of prostate cancer survivability. In *2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)* (pp. 112-116). IEEE.

[6]. Llobet, Rafael, et al. "Computer-aided prostate cancer detection in ultrasonographic images." *Pattern Recognition and Image Analysis: First Iberian Conference, IbPRIA 2003, Puerto de Andratx, Mallorca, Spain, JUne 4-6, 2003. Proceedings 1*. Springer Berlin Heidelberg, 2003.
[7]. Boutilier, Justin J., et al. "Models for predicting objective function weights in prostate cancer IMRT." *Medical physics* 42.4 (2015): 1586-1595.
[8]. Hun, Chang Cui, et al. "Comparison Between K-Nearest Neighbor (KNN) and Decision Tree (DT) Classifier for Glandular Components." *Proceedings of the 11th International Conference on Robotics, Vision, Signal Processing and Power Applications: Enhancing Research and Innovation through the Fourth Industrial Revolution*. Singapore: Springer Singapore, 2022.
[9]. Madabhushi, Anant, et al. "Comparing ensembles of learners: Detecting prostate cancer from high resolution mri." *Computer Vision Approaches to Medical Image Analysis: Second International ECCV Workshop, CVAMIA 2006 Graz, Austria, May 12, 2006 Revised Papers 2*. Springer Berlin Heidelberg, 2006.