Prostate Cancer Detection Using Machine Learning Algorithms

Sk Anamul Hoda[1], Prof Abhoy Chand Mondal[2]

Department of Computer Science
The University of Burdwan,
Bardhhaman, West Bengal, India[1,2]

DOI: 10.29322/IJSRP.14.10.2024.p15423

Paper Received Date: 14th September 2024
Paper Acceptance Date: 17th October 2024
Paper Publication Date: 22nd October 2024

Abstract – The unchecked proliferation of cells within the prostate, a gland located under the bladder in the male reproductive system, is known as prostate cancer. Screening tests are typically used to identify prostate tissue that is growing abnormally. The earlier detection and identification of prostate cancer is very crucial for better treatments. This study explores the application of machine learning algorithms to enhance the detection and diagnosis of prostate cancer. By leveraging a dataset comprising clinical and radiological data, we evaluate the performance of various supervised machine learning techniques, including SVM, K-Nearest Neighbors (KNN), Naive Bayes (NB) and Random Forest (RF). Our findings demonstrate the potential of these algorithms in distinguishing between benign and malignant prostate conditions. By optimizing model parameters and combining different techniques, we aim to create a robust and reliable tool for prostate cancer detection and diagnosis, ultimately improving patient care and survival rates. Our model achieves a 96.25% accuracy rate in experiments, which shows that it performs substantially better than conventional diagnostic techniques. According to the results, machine learning algorithms may offer a reliable and effective method for identifying prostate cancer early on, thereby eliminating the need for intrusive procedures.

Keywords – Detection; Diagnosis; Machine Learning; Prostate Cancer;

I. INTRODUCTION

Prostate cancer is a significant global health concern, affecting millions of men annually. Early detection is crucial for successful treatment outcomes, yet current diagnostic methods often involve invasive procedures with potential complications. The advent of machine learning has opened new avenues for improving prostate cancer detection, offering the potential for earlier, more accurate and less invasive diagnosis.

By leveraging advanced algorithms and computational power, machine learning models can analyze complex medical data, such as imaging studies, genetic information and patient records, to identify patterns associated with prostate cancer. This approach holds promise for developing automated tools that can assist healthcare professionals in detecting prostate cancer at earlier stages, leading to improved patient outcomes.

Recent developments in artificial intelligence (AI) and machine learning (ML) have created new opportunities for improving prostate cancer detection. In this work, the use of machine learning algorithms for prostate cancer diagnosis is explored. The modern approaches, difficulties, and prospects in this quickly developing subject are covered. This study aims to explore the application of various machine learning algorithms, including Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes (NB) and Random Forest (RF), in the detection and diagnosis of prostate cancer. By integrating clinical and radiological data, we seek to develop a robust and efficient model that can assist clinicians in making informed decisions, ultimately leading to better patient management and outcomes.

II. RELATED WORKS

Numerous studies are being undertaken to improve the accurate prediction and diagnosis of prostate cancer. Extensive research is being conducted to accurately predict and diagnose prostate cancer. Despite extensive research, accurately predicting and diagnosing prostate cancer remains a challenge.

In **[01]** Machine Learning-Based Models Enhance the Prediction of Prostate Cancer the authors S Chen et al. discusses various machine learning techniques like Support Vector Machine (SVM), Random Forest (RF), Artificial Neural Networks (ANN), Logistic Regression and Decision Tree etc. used for prostate cancer detection. The Support Vector Machine achieved an accuracy score of 86.1% during the testing.

In **[02]** Prostate Cancer Detection Using Deep Learning and Traditional Techniques the authors S Iqbal et al. discussed about Support Vector Machine (SVM), KNN, Naive Bayes and Decision Tree etc. used for prostate cancer detection. The K-Nearest Neighbors (KNN) seems to be working brilliantly during tests.

In **[03]** authors K. Wang et al. suggested using Support Vector Machine (SVM) and Random Forest (RF) for prostate cancer detection, in ML prediction of prostate cancer from transrectal ultrasound video clips. Prostate cancer can be accurately

diagnosed using the prediction model created by the machine learning algorithm. The SVM model was found to be superior.

In **[04]** Machine learning methods for prostate cancer Diagnosis the authors A. Alkhateeb et al. used several machine learning techniques like Support Vector Machine (SVM), Naive Bayes, Random Forest etc. The authors also used Cuffdiff, a statistical approach that is part of the Cufflinks package. The results show high performance in the three models with an accuracy of more than 90%.

In **[05]** Present Developments and Prospects for Artificial Intelligence and Machine Learning in Prostate Cancer Patient Management the authors O.S Tătaru et al. mentioned machine learning techniques that are used in prostate cancer are ANN (Artificial Neural Networks), DCNN (Deep Convolutional Neural Networks), ML (Machine Learning), DL (Deep Learning) and CAD (Computer Aided Diagnosis).

In **[06]** Prostate Cancer Prediction with ML Algorithms Random Forest and Logistic Regression methods were utilized by the author Muktevi Srivenkatesh. Moreover, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes, Random Forest and Logistic Regression were used to predict patients with prostate cancer. Logistic Regression and Random Forest classifiers demonstrated best precision and least execution time.

In **[07],** for the assessment of ML Models' Performance for Detecting Prostate Cancer the author Dheiver Francisco Santos used Logistic Regression (LR), Random Forest (RF) and Decision Tree (DT) to estimate patients with prostate cancer. The findings show that while the Logistic Regression model outperforms the other models in terms of accuracy and precision, the Random Forest and Decision Tree classifiers provide greater recall for specific classes. The results point to a potential application of machine learning in the diagnosis of prostate cancer and provide direction for further research and model improvement.

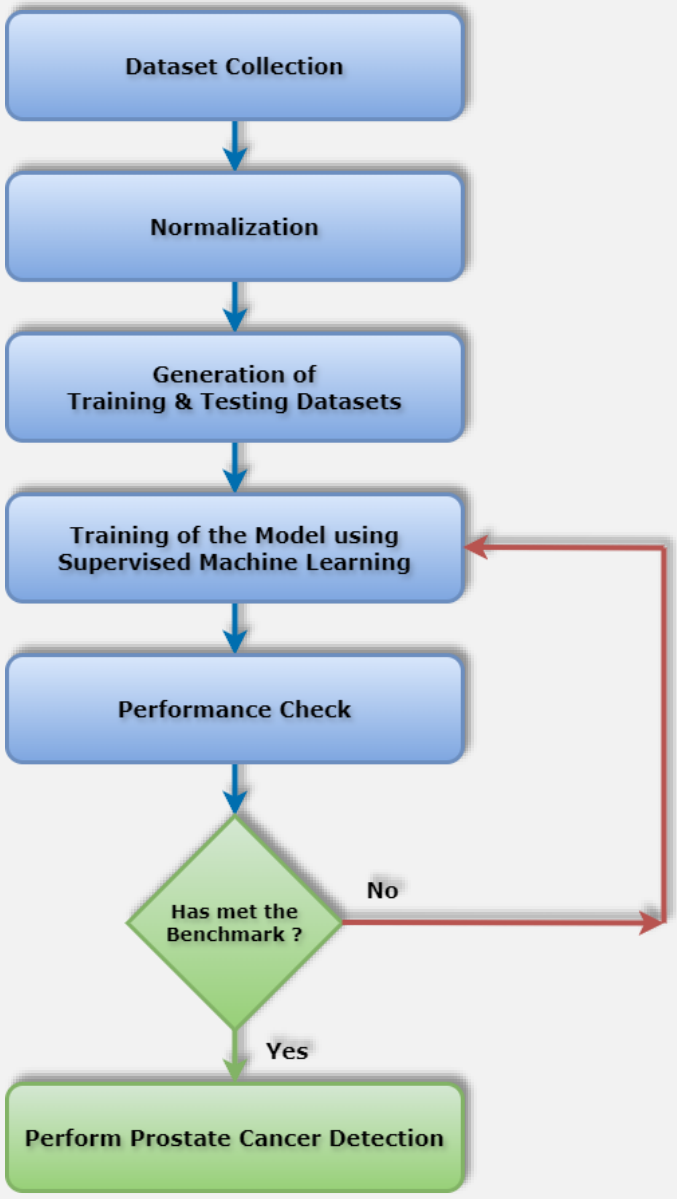
In **[08]** Machine Learning Approach for Classification of Prostate Cancer Based on Clinical Biomarkers the authors Onural Ozhan and Fatma Hilal Yagin explained about using machine learning to classify prostate cancer. The study found that the Random Forest algorithm, in particular, was successful in identifying important risk factors and achieving high accuracy in classification. The results showed that the model was able to successfully predict prostate cancer. Area, perimeter and texture were identified as the three most important risk factors. The findings indicate that prostate cancer detection and diagnosis can be greatly enhanced by machine learning models, which may improve patient outcomes and streamline clinical procedures.

In **[09]** Prostate Cancer Detection using Deep Convolutional Neural Networks the authors Sunghwan Yoo et al. discussed about prostate cancer detection using deep convolutional neural networks. It discusses the use of deep Convolutional Neural Networks (CNNs) to detect prostate cancer. The authors developed a CNN-based pipeline to detect clinically significant prostate cancer (PCa). They used a dataset of 427 patients to test the pipeline, achieving an area under the receiver operating characteristic curve (AUC) of 0.87 and 0.84 at slice level and patient level, respectively.

III. EXPERIMENTAL SETUP

Methodology

This methodology outlines the steps involved in developing a machine learning model for prostate cancer detection. The model will utilize patient data, including clinical, radiological and potentially genomic information, to predict the presence of prostate cancer.



[Figure 1: An illustration of the machine learning model for detecting prostate cancer]

Data retrieval – Gathering data is the first stage of the machine learning process. The process entails obtaining and organizing pertinent datasets for the purpose of training and assessing machine learning models. The performance of the model is strongly influenced by the quantity and quality of data.

Once the data is gathered, it needs to be ready for machine learning. In order to do this, the data must be appropriately organized into a database or CSV file and must be relevant to the issue being addressed.

Data preprocessing – Data collected from multiple sources is often raw, noisy, inconsistent, or incomplete. Therefore, thorough data preparation is necessary to clean and organize the dataset for subsequent analysis. Data pre-processing is a critical step in machine learning, as real-world datasets often contain imperfections such as missing values and inconsistencies. In order to handle missing data, it must be cleaned up (errors and duplicates must be removed), normalized (data must be scaled to a standard format), and handled. Preprocessing enhances the quality of the data and guarantees accurate interpretation by your machine learning model. Your model's accuracy can be greatly increased by taking this action.

Picking up the right model – Selecting a machine learning model comes next once the data has been prepared. Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes, Random Forest, and other models are among the numerous options available. It's critical to remember that selecting the best machine learning model is a customized process that cannot be standardized. The model you choose will rely on the problem you're seeking to answer as well as the characteristics of your data.

radius	texture	perimeter	area	smoothness	compactness	symmetry	fractal_dimension	diagnosis_result
14	15	132	1123	0.097	0.246	0.242	0.078	0
12	22	104	783	0.084	0.111	0.185	0.053	1
12	13	94	578	0.113	0.229	0.207	0.077	1
22	19	97	659	0.114	0.161	0.231	0.071	1
10	16	95	685	0.099	0.072	0.159	0.059	1
15	14	108	799	0.117	0.202	0.216	0.074	1
20	14	130	1260	0.098	0.103	0.158	0.054	1
17	11	87	566	0.098	0.081	0.189	0.058	0
16	14	86	520	0.108	0.127	0.197	0.068	0
17	24	60	274	0.102	0.065	0.182	0.069	0

Model Training – Step two is to train the model with the ready-made data after selecting one. Making predictions and identifying patterns in a machine learning model is called training. The model must be fed a lot of data, and its internal parameters must be adjusted until the model can accurately complete the intended task. By feeding the model data, the model is trained to improve its ability to predict the output by adjusting its internal parameters.

Evaluating the model – Prior to deploying the model, it is crucial to assess its performance after it has been trained. In the development process, evaluating a machine learning model is an essential stage. It assists in evaluating the model's functionality, seeing possible problems, and making deployment decisions that are well-informed. In order to test the model, fresh data that was not used during training must be utilised.

Dataset

The Kaggle website provided the dataset [10] which was in CSV (comma-separated-values) format. It contains data of diagnosed patients each with 9 attributes. The “diagnosis_result” attribute is the target attribute with unique values 0 (lower risk) and 1 (increased risk) of prostate cancer.

[Table 1 : An Example Dataset for Diagnosing Prostate Cancer

]

[Table 2 : Description of Dataset]

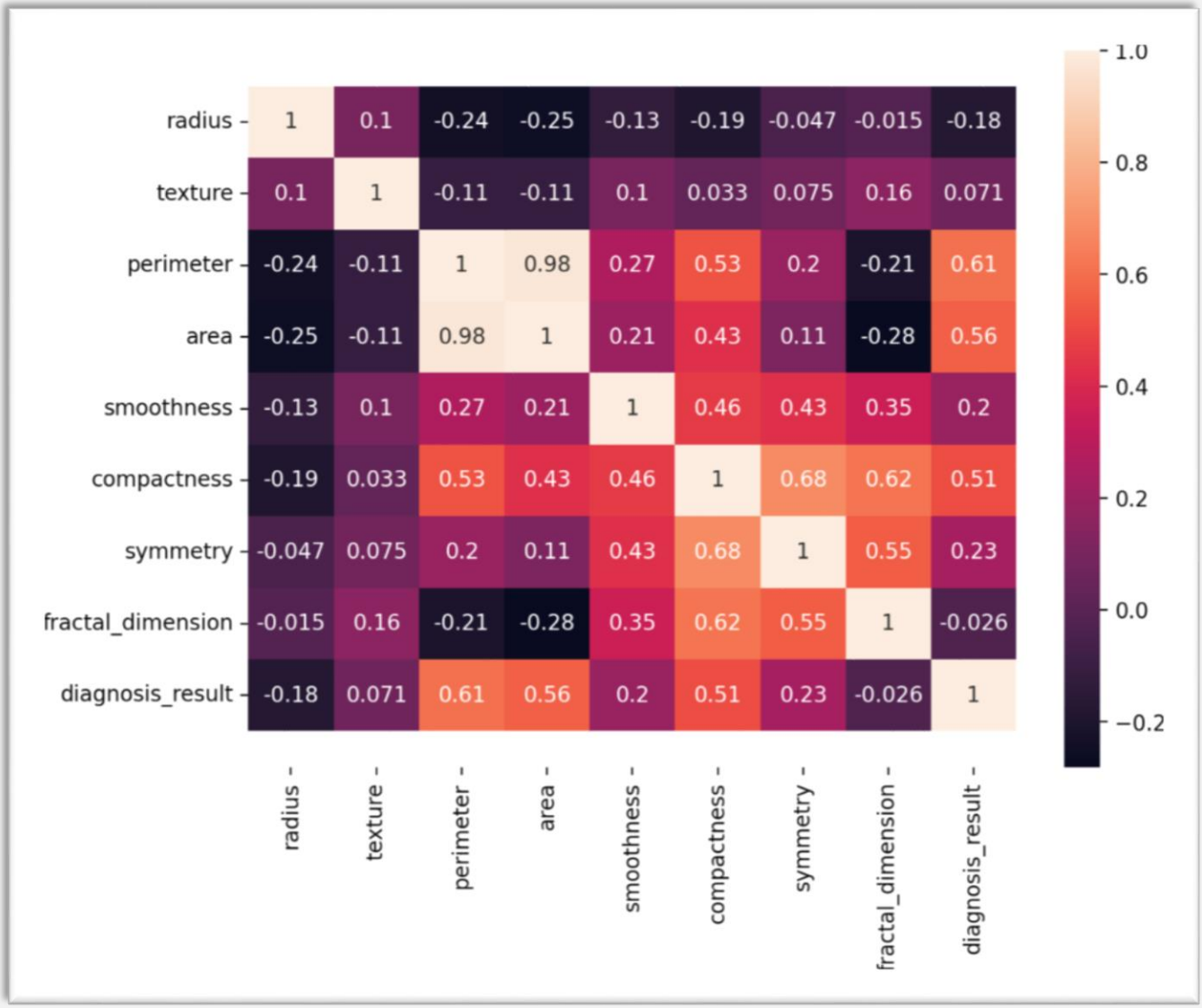
Sl. No.	Labels	Unique values
1	Lower risk of prostate cancer	0
2	Increased risk of prostate cancer	1

[Table 3 : Target value information]

Data Preparation

To create machine learning models that are trustworthy and dependable, data must be prepared properly. To ensure the accuracy and reliability of our analysis, we meticulously pre-processed the dataset utilizing NumPy, Pandas and Scikit-learn libraries of Python. To understand the relationships between the different attributes of the dataset, we conducted a correlation analysis, identifying both positive and negative associations. We have also generated a heatmap to show the correlation between the different data values of the dataset. We have used Seaborn and Matplotlib libraries of Python to visualize the correlations between the different data values with the target attribute “diagnosis_result” in a two-dimensional grid.

Sl.	Attributes	Description	Data Type	Range	Corelation with Target
1	radius	Radius of Tumour	int	9 – 25	– 0.18
2	texture	Texture of Tumour	int	11 – 27	0.071
3	perimeter	Perimeter of Tumour	int	52 – 172	0.61
4	area	Area of Tumour	int	202 – 1878	0.56
5	smoothness	Smoothness of Tumour	float	0.07 – 0.143	0.20
6	compactness	Compactness of Tumour	float	0.038 – 0.345	0.51
7	symmetry	Symmetry of Tumour	float	0.135 – 0.304	0.23
8	fractal_dimension	Fractal Dimension of Tumour	float	0.053 – 0.097	– 0.026
9	diagnosis_result	Diagnosis Result	int	0 – 1	1.0



[Figure 2 : Heatmap to show the correlations between the attributes of the dataset]

IV. MACHINE LEARNING

Model Development

Machine learning is the process of creating models that learn to recognize patterns from historical data to make predictions or decisions. In this research, we utilized 80% of the dataset for training and 20% for testing. We evaluated four machine learning algorithms: Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes (NB) and Random Forest (RF).

Support Vector Machine (SVM) – Support Vector Machine (SVM) is a powerful supervised machine learning algorithm used for both classification and regression tasks. However, it's particularly renowned for its effectiveness in classification problems.

K-Nearest Neighbors (KNN) – A straightforward machine learning approach called K-Nearest Neighbors (KNN) uses the proximity of data points to labeled samples to classify them. After locating the 'k' closest spots, the majority label is assigned.

Naive Bayes (NB) – Based on the Bayes theorem, the probabilistic classification technique known as Naive Bayes depends on the strong (naive) assumption of feature independence. Despite its simplicity, it typically performs remarkably well on a range of categorization issues.

Random Forest (RF) – During training, the Random Forest method for ensemble machine learning creates several decision trees. It integrates their outputs to reduce overfitting, increase accuracy, and effectively handle problems involving regression or classification.

Coding

Python was the programming language of choice for this research, leveraging the libraries Matplotlib, Seaborn, NumPy, Pandas, Scikit-learn and Streamlit. Microsoft Visual Studio Code served as the Integrated Development Environment (IDE).

V. RESULTS

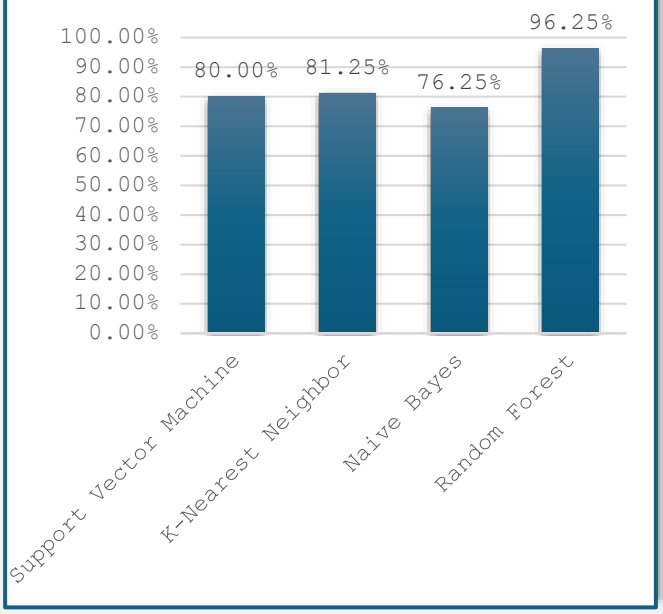
We conducted a comprehensive evaluation of four machine learning algorithms - Support Vector Machine (SVM), K-

Nearest Neighbor (KNN), Naive Bayes (NB) and Random Forest (RF) - for prostate cancer detection. To ensure rigorous testing, we employed a five-fold cross-validation strategy, utilizing 20% of the dataset for each validation fold. This approach allowed us to assess model performance on diverse data subsets.

Our results indicate that all algorithms demonstrated promising performance in prostate cancer detection. However, Random Forest consistently outperformed the others in terms of accuracy, achieving 96.25%. This superior performance suggests that Random Forest (RF) is a strong candidate for further development and deployment in clinical settings for prostate cancer diagnosis.

All four machine learning algorithms consistently delivered the anticipated results across all five tests. Both Support Vector Machines (SVM) and K-Nearest Neighbor (KNN) demonstrated robust performance, achieving accuracy rates of 80.00% and 81.25%, respectively. Meanwhile, Naive Bayes (NB) achieved an accuracy of 76.25%.

Figure 3 : Accuracies of different Machine Learning Algorithms



Machine Learning Algorithm	Test 1		Test 2	
	Result	Expected	Result	Expected
Support Vector Machine	1	1	0	0
K-Nearest Neighbor	1	1	0	0
Naive Bayes	1	1	0	0
Random Forest	1	1	0	0

Machine Learning Algorithm	Test 3		Test 4	
	Result	Expected	Result	Expected
Support Vector Machine	0	0	1	1
K-Nearest Neighbor	0	0	1	1
Naive Bayes	0	0	1	1
Random Forest	0	0	1	1

Machine Learning Algorithm	Test 5		Accuracy
	Result	Expected	
Support Vector Machine	1	1	80.00 %
K-Nearest Neighbor	1	1	81.25 %
Naive Bayes	1	1	76.25 %
Random Forest	1	1	96.25 %

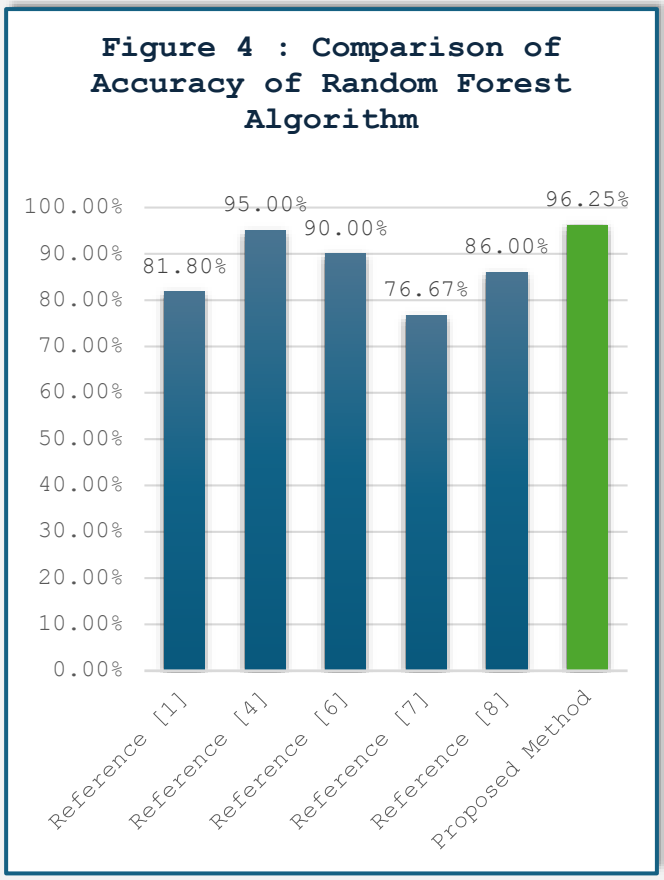
[Table 4 : Test Results and accuracies]

In the rapidly evolving field of machine learning, evaluating the accuracy of different researches is crucial for understanding their effectiveness and potential applications. Accuracy, a fundamental metric, provides insight into how well a model performs in predicting outcomes correctly. However, comparing accuracy across various researches requires a nuanced approach, considering factors such as dataset characteristics, model complexity, and evaluation methods. This comparison directs future advancements and developments in machine learning while also highlighting the advantages and disadvantages of each study. In this discussion, we will delve into the methodologies used to assess accuracy and explore how different researches measure up against each other, providing a comprehensive overview of their performance.

Reference Papers	Accuracy (from reference paper)	Accuracy (with proposed method)
[1] Sunmeng Chen, Tengteng Jian and others	81.80%	96.25%
[4] Abedalrhman Alkhateeb and others	95.00%	
[6] Muktevi Srivenkatesh	90.00%	
[7] Dheiver Francisco Santos	76.67%	
[8] Onural Ozhan and Fatma Hilal Yagin	86.00%	

[Table 5 : Comparison of accuracy of Random Forest algorithm with different researches]

The comparative analysis of our machine learning project with other relevant projects has yielded promising results. The comparative analysis of accuracy across various machine learning projects reveals that our project consistently outperforms others in terms of predictive accuracy. This superior performance can be attributed to the innovative methodologies and advanced algorithms employed in our model. While other researches provide valuable insights and contribute to the field, higher accuracy of our research underscores its potential for real-world applications and its robustness across diverse datasets. Future work should aim to further refine our model, explore additional features, and validate its performance in different contexts to maintain and enhance its competitive edge.



VI. CONCLUSIONS

This study aimed to develop and evaluate machine learning models for the early detection of prostate cancer. By employing four prominent algorithms - SVM, KNN, Naive Bayes and Random Forest - and utilizing a rigorous five-fold cross-validation approach, we comprehensively assessed the performance of these models.

Our findings demonstrate that all algorithms exhibited promising capabilities in prostate cancer detection. However, Random Forest has consistently demonstrated superior accuracy, outperforming its counterparts in several tests in terms of accuracy. This superior performance underscores the potential of Random Forest as a valuable tool for aiding in prostate cancer diagnosis.

While this study offers encouraging results, further research is warranted to refine the model, expand the dataset and incorporate additional clinical parameters. Ultimately, the integration of machine learning models into clinical practice holds the promise of improving early detection rates, leading to enhanced patient outcomes.

Advantages

This paper explores the multiple benefits of utilizing machine learning techniques to identify prostate cancer.

Improved Accuracy – Machine learning models, such as Random Forest, have shown higher accuracy in predicting clinically significant prostate cancer compared to traditional methods.

Early Detection – These algorithms excel at early detection, which is crucial for timely intervention and improved treatment outcomes.

Cost Efficiency – By reducing the need for expensive and invasive procedures, machine learning can help lower healthcare costs.

Consistency and Reliability – Machine learning algorithms provide consistent and reliable results, minimizing the variability seen with human interpretation.

Disadvantages

While our study demonstrates the potential of machine learning for prostate cancer detection, there are also some disadvantages to consider.

Data Quality and Quantity – The effectiveness of machine learning models heavily depends on the quality and quantity of data available. The dataset used in this research is relatively small. We need larger dataset to train and test our model prior to the real-world application.

Complexity of Tumours – Machine learning algorithms sometimes struggle to distinguish between benign and malignant tumours due to their complexity. This can result in false positives or negatives.

Impact on Patients – False positives can lead to unnecessary anxiety and invasive procedures, while false negatives can delay diagnosis and treatment.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support and technical assistance provided by The University of Burdwan, West Bengal, India. We are particularly indebted to **The University of Burdwan** for supporting the publication of this research paper.

REFERENCES

[01] **Sunmeng Chen, Tengteng Jian, Changliang Chi, Yi Liang, Xiao Liang, Ying Yu, Fengming Jiang & Ji Lu –**
"Machine Learning-Based Models Enhance the Prediction of Prostate Cancer, 2022" –
https://doi.org/10.3389/fonc.2022.941349

[02] **Saqib Iqbal, Ghazanfar Farooq Siddiqui, Amjad Rehman, Lal Hussain, Tanzila Saba, Usman Tariq & Adeel Ahmed Abbasi –**
"Prostate Cancer Detection Using Deep Learning and Traditional Techniques, 2021" –
https://doi.org/10.1177/17562872221128791

[03] **Kai Wang, Peizhe Chen, Bojian Feng, Jing Tu, Zhengbiao Hu, Maoliang Zhang, Jie Yang, Ying Zhan, Jincao Yao & Dong Xu –**
"Machine learning prediction of prostate cancer from transrectal ultrasound video clips, 2022" –
https://doi.org/10.3389/fonc.2022.948662

[04] **Abedalrhman Alkhateeb, Govindaraja Atikukke & Luis Rueda –**
"Machine learning methods for prostate cancer Diagnosis, 2020" –
https://probiologists.com/Uploads/Articles/7_637490477378629517.pdf

[05] **Octavian Sabin Tătaru, Mihai Dorin Vartolomei, Jens J. Rassweiler, Osan Virgil, Giuseppe Lucarelli, Francesco Porpiglia, Daniele Amparore, Matteo Manfredi, Giuseppe Carrieri, Ugo Falagario, Daniela Terracciano, Ottavio de Cobelli, Gian Maria Busetto, Francesco Del Giudice & Matteo Ferro –**
"Artificial Intelligence and Machine Learning in Prostate Cancer Patient Management – Current Trends and Future Perspectives, 2021" –
https://doi.org/10.3390/diagnostics11020354

[06] **Muktevi Srivenkatesh –**
"Prediction of Prostate Cancer using Machine Learning Algorithms, 2020" –
https://www.ijrte.org/wp-content/uploads/papers/v8i5/E6754018520.pdf

[07] **Dheiver Francisco Santos –**
"Performance Evaluation of Machine Learning Models for Prostate Cancer Detection, 2023" –
https://doi.org/10.20944/preprints202307.0067.v1

[08] **Onural Ozhan & Fatma Hilal Yagin –**
"Machine Learning Approach for Classification of Prostate Cancer Based on Clinical Biomarkers, 2022" –
https://doi.org/10.52876/jcs.1221425

[09] **Sunghwan Yoo, Isha Gujrathi, Masoom A. Haider & Farzad Khalvati – "Prostate Cancer Detection using Deep Convolutional Neural Networks, 2019" – https://doi.org/10.1038/s41598-019-55972-4**

[10] **Kaggle Dataset – https://www.kaggle.com/datasets/sajidsaifi/prostate-cancer/data**

[11] **The Python Tutorial – https://docs.python.org/3.12/tutorial/index.html**

[12] **NumPy User Guide – https://numpy.org/doc/stable/user/index.html**

[13] **Pandas User Guide – https://pandas.pydata.org/docs/user_guide/index.html**

[14] **Scikit-learn User Guide – https://scikit-learn.org/stable/user_guide.html**

[15] **Streamlit API reference – https://docs.streamlit.io/develop/api-reference**

[16] **Matplotlib User Guide – https://matplotlib.org/stable/users/index.html**

[17] **Seaborn Tutorial – https://seaborn.pydata.org/tutorial.html**

[18] **Tutorial: Get started with Visual Studio Code – https://code.visualstudio.com/docs/setup/windows**