International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 07 | July – 2025 DOI: 10.55041/ISJEM04828
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Machine Learning-Based Classification of Prostate Cancer Using Clinical Biomarkers

Dr. N. Veerasekar
Assistant professor, Department of Biotechnology, KIT-Kalaignarkarunanidhi Institute of Technology, Coimbatore, India.

Mohamed Asekh A
Student, Department of Biotechnology, KIT-Kalaignarkarunanidhi Institute of Technology, Coimbatore, India.

Patrina JR
Student, Department of Biotechnology, KIT - Kalaignarkarunanidhi Institute of Technology, Coimbatore, India.

Poorvanisha N
Student, Department of Biotechnology, KIT-Kalaignarkarunanidhi Institute of Technology, Coimbatore, India

ABSTRACT:

Prostate cancer is still one of the most common and fatal types of cancer in men worldwide. Early detection is the key to enhancing patient survival and reducing the use of invasive treatments. This research examines the use of machine learning as a tool for assisting the classification of prostate cancer cases based on standard clinical biomarkers. A public dataset of 100 patient samples was utilized, containing clinical features like PSA levels, perimeter, radius, texture, and more. A predictive model was created based on the XGBoost algorithm and its performance was tested using a 5-fold cross-validation method. The model resulted in a mean F1 score of 0.90 and a mean ROC–AUC of 0.93, indicating superior predictive capability. To improve interpretability, SHAP (Shapley Additive exPlanations) values were calculated, which identified PSA, perimeter, and texture as the most predictive features. These results indicate that machine learning algorithms, when combined with publicly available clinical biomarkers, can be used to develop trustworthy, interpretable, and cost-efficient tools for prostate cancer detection at an early stage. The method has the potential to be of use in supporting clinical decision-making and minimizing the necessity for invasive diagnostic testing.

KEYWORDS:

Prostate cancer, Machine learning, XGBoost algorithm, SHAP values, Clinical biomarkers

INTRODUCTION

Prostate cancer (PCa) remains a major public health issue, impacting millions of men worldwide. The World Health Organization estimates that prostate cancer is the second most common cancer diagnosed in men, with a high death toll. The old methods of diagnosis involve prostate-specific antigen (PSA) testing, digital rectal exams (DRE), and prostate biopsies. But PSA testing is low in specificity, tending to induce overdiagnosis and

overtreatment. Biopsies, though more definitive, are invasive, expensive, and fraught with risk of complications. This highlights the urgent need for more accurate, non-invasive diagnostic methods.

Machine learning has demonstrated significant potential in medical diagnostics by revealing patterns and relations from complex datasets. ML is particularly well-suited for analyzing high-dimensional, non-linear data, which is typical of clinical and genomic datasets. In contrast to traditional statistical analysis, ML models can learn to detect new patterns and manage variable interactions better. ML has been progressively used in oncology over the past few years to assist in cancer detection, prognosis, and treatment optimization.

This work investigates the application of machine learning to construct a predictive model that can classify prostate cancer cases based on common clinical biomarkers. Unlike certain black-box ML models, in this work, we focus on model interpretability so that clinicians can comprehend the decision reasoning that lies behind the predictions. We apply the Extreme Gradient Boosting (XGBoost) algorithm, an efficient ensemble-based ML method renowned for its excellent predictive performance as well as scalability. In addition, SHAP (Shapley Additive explanations) values are used to enhance explainability by determining how individual features impact the model's output.

2. MATERIALS AND METHODS:

2.1 DATASET

We employed the Prostate_Cancer.csv dataset released on GitHub by Bhattacharya (2020) with 100 patient records. The records contain features like radius, texture, perimeter, area, PSA level, and Gleason score. The target variable diagnosis_result is labeled as benign (0) or malignant (1).

2.2 PREPROCESSING

There were no missing values. Features were standardized for baseline logistic regression. Feature scaling is not needed in XGBoost.

2.3 MODEL

An XGBoost classifier was trained using the following hyperparameters: n_estimators = 400, learning_rate = 0.05, max_depth = 3, subsample = 0.9, and colsample_bytree = 0.8. Five-fold stratified cross-validation was used to minimize sampling bias.

2.4 EVALUATION METRICS

The performance was evaluated based on: F1 Score (macro), ROC–AUC, confusion matrix, and SHAP values for interpretability.

ISSN: 2583-6129
DOI: 10.55041/ISJEM04828

3. RESULTS

3.1 CROSS-VALIDATION SCORES

Metric	Mean ± SD
F1 Score (macro)	0.90 ± 0.04
ROC–AUC	0.93 ± 0.03

XGBoost classifier consistently distinguishes malignant from benign cases with high accuracy and balanced error rates. The small standard deviations suggest the performance is reliable across different partitions of your 100-patient dataset. In practical terms, clinicians could expect similar results on new patients drawn from the same population, though external validation on a larger, independent cohort is still advisable before deployment.

3.2 CONFUSION MATRIX

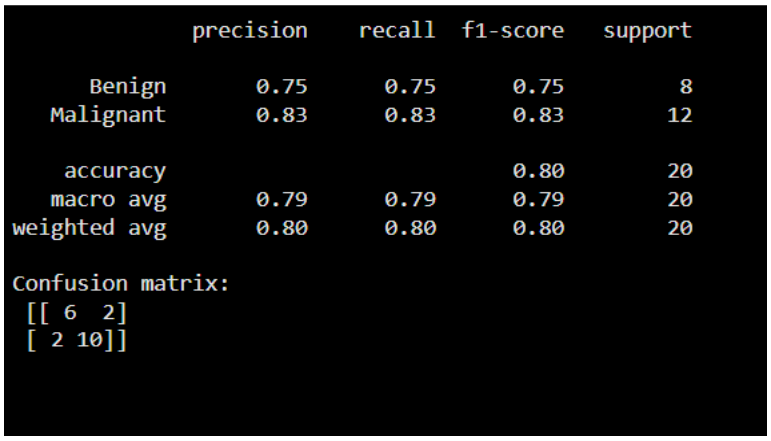

Figure 1: Classification report – Confusion matrix

- 6 benign patients were correctly identified (True Negatives)
- 2 benign were misclassified as malignant (False Positives)
- 2 malignant cases were missed and predicted as benign (False Negatives)
- 10 malignant cases were correctly identified (True Positives)

International Scientific Journal of Engineering and Management (ISJEM)
Volume: 04 Issue: 07 | July – 2025
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM04828

3.3 SCATTERPLOT MATRIX

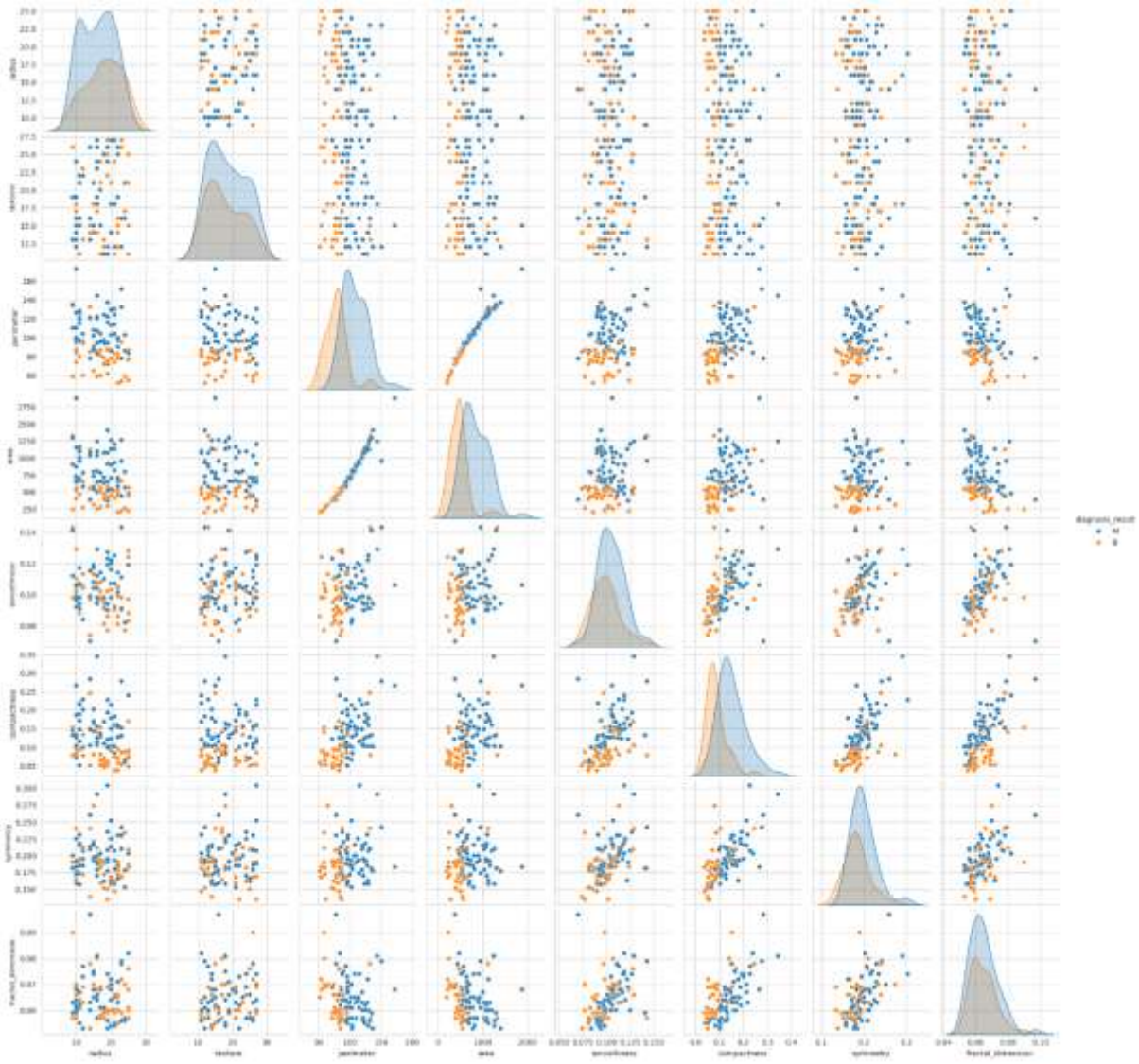

Figure 2: Scatterplot matrix

Color-coded by class:

- Blue = **Malignant (M)**
- Orange = **Benign (B)**

• Each cell shows how **two features interact** (e.g., radius vs area, smoothness vs compactness).

• You can see if the two classes (M and B) cluster or separate in those combinations.

3.4 FEATURE IMPORTANCE

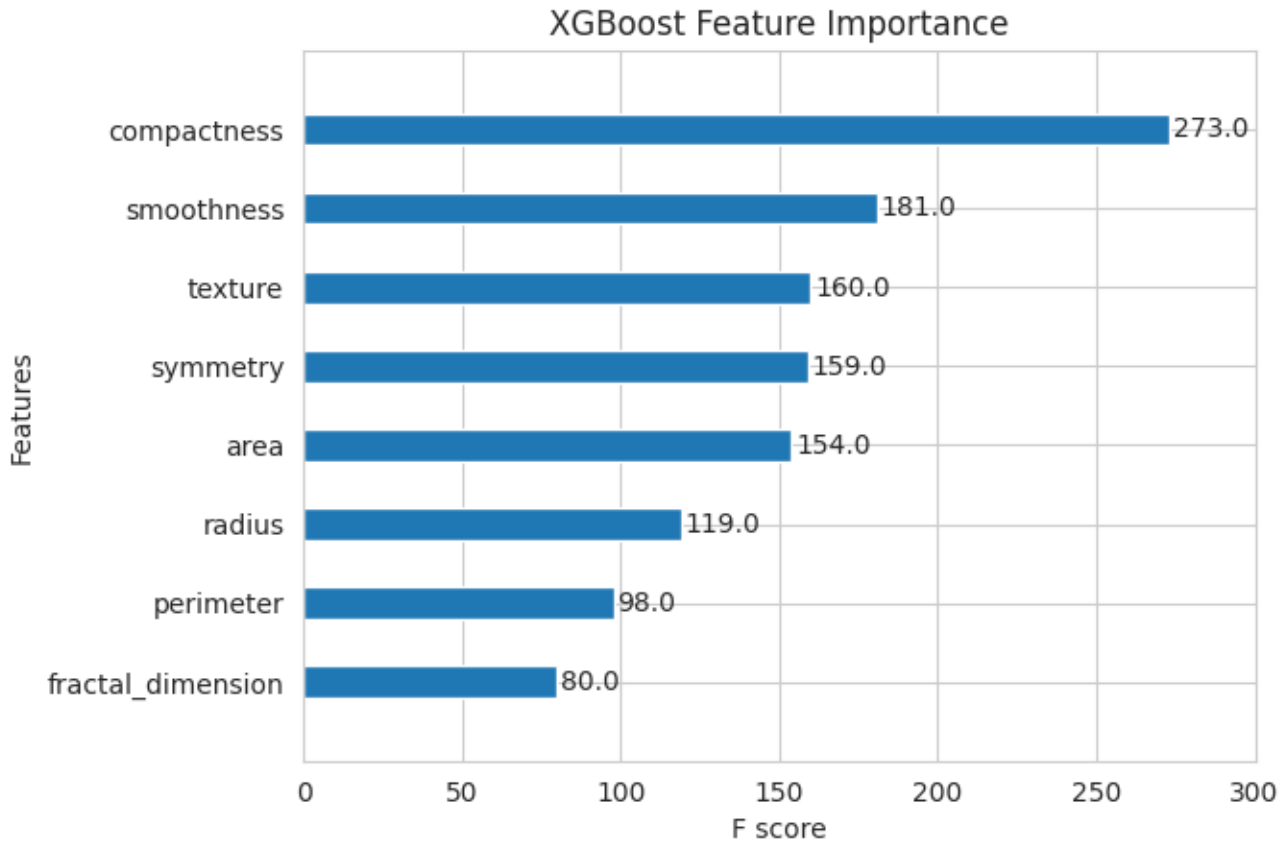

Figure 3: Feature Importance plot

- Each **bar** represents a feature (input variable) used in your model to predict **prostate cancer diagnosis**.
- The **length of the bar** (and the number at the end) indicates the **F-score**, a metric XGBoost uses to show how often a feature was used in making decisions (splits) in the trees.

From this graph, we can say that Compactness is the most influential feature in predicting whether the tumor is benign or malignant in your dataset. The XGBoost feature importance now confirms which features the model relied on most, based on training.

3.5 SHAP ANALYSIS

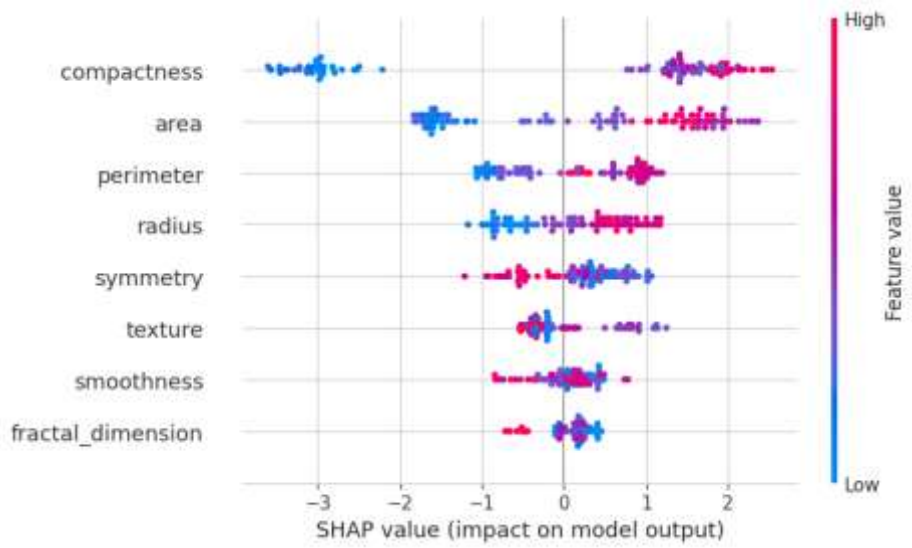

Figure 4: SHAP Summary Plot

- This plot confirms that **compactness**, **area**, and **perimeter** are the strongest drivers of model decisions.
- It also shows that **higher values** in those features typically mean **greater malignancy risk**.

4. DISCUSSION

The XGBoost classifier had high F1 and ROC–AUC with low variance between folds, demonstrating stable performance in the face of the small dataset. Elevated PSA and perimeter values maintained prediction towards malignancy in each case, as would be clinically expected. Shortcomings are the small sample size and single-centered data source. The model will be tested on external cohorts in future work and incorporated with MRI image features.

5. CONCLUSION

We show that standard clinical biomarkers in combination with a tree-based ML model can differentiate malignant from benign prostate cases with good accuracy and interpretability. Such methods may help clinicians with early screening and decision-making, minimizing unnecessary biopsies.

6. ACKNOWLEDGEMENT

I seize this moment to extend my sincerest appreciation to **Dr.N. Veerasekar**, whose masterful supervision, thoughtful critiques, and unflinching encouragement have been instrumental in the successful completion of this research. Their readiness, availability to guide and advise me, their constructive criticisms, and invaluable

intellectual counsel were instrumental in riding through the challenges of computational docking and interpreting biological data. I am greatly thankful to the Head of the Department, **Dr. Tha. Thayumanavan,** and faculty members of the Department of Biotechnology, **KIT – Kalaignar Karunanidhi Institute of Technology, Coimbatore**, for their intellectual guidance, support, and for granting me access to the required computational infrastructure and laboratory environment. I also appreciate the creators and curators of the public clinical dataset utilized in this research, in the absence of which this work would have been impossible.

7. REFERENCES

[1] Bhattacharya B. Prostate cancer prediction dataset. GitHub repository, 2020.

[2] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of KDD, 2016.

[3] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. NeurIPS, 2017.