RESEARCH PAPER

Impact of Class Imbalance, SMOTE, and Threshold Optimization on Prostate Cancer Prediction Models

Avijit Kumar Chaudhuri[1], Pramoda Patro[2], Amartya Ghosh[3], Ranjan Banerjee[4], Debmalya Mukherjee[5], Payel Sengupta[6]

[1]PDF - Computer Science and Engineering, SR University, Warangal, Telangana, India, 506371 & Professor, Computer Science & Engineering, Brainware University, Barasat, Kolkata – 125, West Bengal, India.
Email: c.avijit@gmail.com

[2]Director, Centre of AI and Deep Learning, School of Computer Science and Artificial Intelligence, SR University, Warangal, Telangana, India, 506371. Email: pramoda.mtech09@gmail.com

[3]Assistant Professor, Computer Science & Engineering, Brainware University, Barasat, Kolkata – 125, West Bengal, India. Email: com.amartya@gmail.com
ORCID ID: 0009-0002-2504-3325

[4]Assistant Professor, Computer Science & Engineering, Brainware University, Barasat, Kolkata – 125, West Bengal, India. Email: rbkpccst@gmail.com
ORCID ID: 0009-0003-1950-7530

[5]Assistant Professor, Computational Sciences, Brainware University, Barasat, Kolkata – 125, West Bengal, India.
Email: dbml.mukherjee@gmail.com
ORCID ID: 0009-0006-9946-0964

[6]Assistant Professor, Computer Science & Engineering, Brainware University, Barasat, Kolkata – 125, West Bengal, India. Email: payel9433@gmail.com
ORCID ID: 0000-0003-3981-5971

Received: 25th May, 2026; **Revised:** 6th June, 2026; **Accepted:** 8th June, 2026; **Available Online:** 09th June, 2026

ABSTRACT

Cancer has been a significant concern of public health across the world and a major cause of death. The World Health Organization (2024) also stated that almost 20 million new cancer cases and 9.7 million cancer-related deaths have been reported globally in 2022, and the number of new cancer cases is expected to increase to 30-35 million each year by 2050, with an underrepresentation of high-income countries. The second type of cancer that is commonly diagnosed in men all over the world is prostate cancer. Data from GLOBOCAN 2022 in India show that incidence is gradually increasing due to population ageing, urbanization, and lifestyle-related risk factors. Although early diagnosis has been linked to increased lifespan, historical screening methods such as prostate-specific antigen (PSA) tests and digital rectal examination (DRE) have been poorly diagnostic, underscoring the need for a data-driven diagnostic model.

This paper discusses the application of machine learning (ML) algorithms to predict prostate cancer using a large clinical and demographic dataset downloaded from the Kaggle repository. The dataset comprises 27945 patient records with 29 variables, including demographics, clinical indicators, comorbidities, lifestyle factors, and laboratory parameters. Exploratory analysis showed no missing values, moderate feature correlations, and pronounced class imbalance, indicating the need for resampling strategies.

There were 6 supervised ML models, namely, Logistic Regression (LR), Decision Tree (DT), Naive Bayes (NB), Multilayer Perceptron (MLP), Extreme Gradient Boosting (XGB), and Random Forest (RF), that were tested based on 10-fold cross-validation under three conditions: no balancing of classes, SMOTE, and SMOTE with threshold optimization. A measure of performance was based on accuracy, sensitivity, specificity, F1-score, ROC, and Cohen's kappa.

In the absence of balancing, models showed high accuracy and F1 Scores but poor specificity and ROC/AUC, indicating a bias toward the majority class. Class balance was also enhanced by SMOTE, the trade-offs between sensitivity and specificity were more realistic, and ensemble models (XGB and RF) were more robust. Threshold optimization also maximized sensitivity and F1-scores, but at the expense of specificity, and did not affect ROC/AUC. In general, the XGB and the RF showed the most stable results. Nevertheless, the continuous low ROC/AUC and Kappa values indicate low discriminatory power. The results indicate the feasibility and limitations of ML-based prostate cancer prediction, as well as the need for more informative features, sophisticated sampling methods, and explainable models for clinical practice.

Keywords: Prostate Cancer Prediction, Class Imbalance, SMOTE, Threshold Optimization, Machine Learning Models, Model Evaluation Metrics.

How to cite this article: Chaudhuri AK, Patro P, Ghosh A, Banerjee R, Mukherjee D, Sengupta P. Impact of Class Imbalance, SMOTE, and Threshold Optimization on Prostate Cancer Prediction Models. Int J Drug Deliv Technol. 2026;16(57s): 1568-1595. DOI: 10.25258/ijddt.16.57s.158
Source of support: Nil.
Conflict of interest: None.

1. Introduction

A major global public health concern, cancer continues to be one of the main causes of morbidity and death worldwide. The World Health Organization (2024) estimates that in 2022, there were 9.7 million cancer-related deaths worldwide and about 20 million new cancer cases(World Health Organization, 2024). Globally, lung and breast cancers are the most common types. Remarkably, avoidable risk factors like tobacco use, infections, environmental pollution, and unhealthy lifestyle choices are responsible for almost 37% of cancer cases (American Cancer Society, 2022; World Health Organization, 2024; World Cancer Research Fund, 2022).

The annual number of new cases of cancer worldwide is predicted to increase to 30 to 35 million by 2050, underscoring the critical need for better prevention plans, early detection techniques, and fair access to medical care, especially in low- and middle-income nations.

The incidence of cancer is continuously rising in India. According to the GLOBOCAN 2022 report, there will be between 1.4 and 1.46 million new cases of cancer nationwide, and 1 in 9 people will get cancer in their lifetime.

Prostate cancer has become one of the most common cancers in men, although lung and breast cancers are still the most common types. The incidence of prostate cancer is predicted to increase due to demographic shifts, ageing populations, urbanization, and increased exposure to risk factors like tobacco use and environmental pollutants (GLOBOCAN 2022; Indian Journal of Urology, 2024). India is among the top countries in the world for cancer incidence, with an estimated age-adjusted incidence rate of 6.8 cases of prostate cancer per 100,000 people.

With millions of new cases discovered each year, prostate cancer is the second most common cancer in men worldwide(Chaudhuri et al., 2018; Chaudhuri et al., 2023; Das et al., 2024). Early diagnosis of the illness is critical because it prevents the proliferation of the cancer, enhances treatment outcomes, and significantly improves the survival rates of people with the disease. More common screening methods include the Digital Rectal Examination (DRE) and Prostate-Specific Antigen (PSA) test, but these methods are not without controversy because of false positives and false negatives, so their diagnostic capability is questionable. Such limitations underscore the importance of more accurate and reliable diagnostic approaches, as they may lead to missed diagnoses, delayed treatment, or unwarranted biopsies(Das et al., 2025; Das et al., 2026).

The fields of data analytics and machine learning offer significant opportunities to advance cancer diagnosis. Predictive modelling methods can analyse complex clinical data to identify patterns and risk factors associated with prostate cancer. This may improve diagnostic accuracy and support clinical decision-making. Therefore, integrating machine learning algorithms into prostate cancer detection models may enable more efficient, data-driven early diagnosis and better patient outcomes.

Even though there are screening tools that can be used to test prostate cancer, such as the Prostate-Specific Antigen (PSA) test and Digital Rectal Exam (DRE) tests, their accuracy is controversial. False positives and false negatives are also major issues, which imply that more practical diagnostic instruments should be considered(Das et al., 2026).

The most recent advances in data and machine learning offer promising future opportunities for diagnostic medical procedures, particularly in predicting cancer employing massive demographic data. Clinical features can facilitate the development of predictive models that help clinicians make better decisions about prostate cancer screening and detection. These models can be characterized by several indicators, including age, family history, hereditary risk factors, PSA level, biopsy results, and other symptoms, such as inability to urinate or pelvic pain, that are usually associated with prostate health. The data used in this paper comprises 27,945 patient records spanning a wide range of variables (such as demographic (e.g., age, race), clinical (e.g., PSA levels, DRE results), and medical (e.g., hypertension, diabetes) variables. This study will use this rich data to test the predictive capability of these features for identifying prostate cancer and identify which variables are most likely to influence the disease.

The primary aim of the research is to examine the efficacy of various clinical and demographic variables to determine the prostate cancer outcome. This research will, using statistics, establish influential factors that influence the probability of a positive cancer outcome, and the results will provide

information on how the factors can be used to improve the early detection. In addition, this study focuses on the predictive effectiveness of various machine learning models and examines their extrapolation capabilities to other areas of patient care. In this way, we will be able to collect evidence to develop more credible and effective diagnostic strategies for prostate cancer.

Relevant literature

This section provides a comparative summary of machine learning (ML) algorithms applied to prostate cancer detection using imaging (mpMRI, TRUS, pathology WSI) and clinical data between 2018 and 2025. The metrics used are Accuracy, Sensitivity, Specificity, Precision, F1-score, Cohen's Kappa, and AUC. It is based on systematic reviews, meta-analyses, and representative primary research. Table 1 summarizes the literature on the accuracy study and the comparative statistical analysis.

Table 1. Prostate Cancer Performance Comparison

Study (Year)	Modality	Algorithm(s)	Accuracy	Sensitivity	Specificity	Precision	F1-score	Kappa	AUC
Zhao et al. (2025)	mpMRI (meta-analysis)	Radiomics + ML, DL	—	0.92 (benign/malig), 0.83 (csPCa)	0.90 (benign/malig), 0.73 (csPCa)	NR	NR	NR	0.96 (benign/malig), 0.86 (csPCa)
Sherafatmandjoo et al. (2024)	mpMRI	Deep learning (multi-input CNN + clinical features)	94.2%	94.24%	98.62%	NR	NR	NR	NR
Horasan et al. (2024)	mpMRI	3D CNN	91.3%	90.2%	92.1%	89.8	90.0	NR	NR
Saha et al. (2024)	Histopathology WSI	DL pipeline	99.53 %	99.78%	99.12%	NR	NR	NR	NR
Sun et al. (2023)	TRUS videos	3D CNN	NR	Reported reliable csPCa detection	NR	NR	NR	NR	NR
Chen et al. (2022)	Clinical + PSA + biopsy data	RF, XGBoost, others	Varied (improved vs clinical)	NR	NR	NR	NR	NR	Improved AUCs
Twilt et al. (2021)	mpMRI (review)	SVM, RF, CNN, ensembles	Wide variation	Wide variation	Wide variation	NR	NR	NR	68.8% – 98%
Li et al. (2022)	mpMRI (review)	Radiomics + DL	—	—	—	—	—	—	—

Methodology

I. Dataset Description

The dataset in question is the publicly available dataset (Prostate Cancer Prediction Dataset) of Kaggle, consisting of 27945 records of patients and 29 features. Several machine learning studies have utilized this dataset to develop predictive models for classifying prostate cancer. It mainly comprises continuous numeric variables derived from clinical measurements and image-based characteristics (e.g., radius, texture, perimeter, area, smoothness, compactness, symmetry, fractal dimension),

as well as a categorical diagnosis variable (i.e., whether a patient has prostate cancer).

II. Feature Overview

The Prostate Cancer Prediction Dataset by Ankush Panday on Kaggle is an excellent resource that contains a set of medical and lifestyle information to estimate the risk of prostate cancer. Although the specific field names are not explicitly listed in the provided documentation, the dataset consists of characteristics common to prostate cancer prediction models. These typically include:

1. Demographic Information

Age: Risk of prostate cancer increases with age, and the patient is at a particular age.

Family History: Prostate cancer may have a family history, which can predispose it.

2. Available Clinical and Diagnostic Data

PSA Levels (Prostate-Specific Antigen) are a protein secreted by the prostate gland. High

levels may indicate the presence of a malignant tumor.

Digital Rectal Exam (DRE) Results: Results of a physical examination of the prostate.

Biopsy Results: Biopsy of prostate tissue.

3. Denies experiencing pain, itching, swelling, or redness of the skin, as well as hair loss,

fever, or headaches.

Urinary Symptoms: e.g., frequency, urgency, or nocturia.

Past Prostate Conditions: History of such prostate conditions as benign prostatic hyperplasia (BPH).

4. Lifestyle Factors

Dietary habits: Data on fat intake, fruit and vegetable intake, etc.

Physical Activity: Exercise or inactivity.

Smoking: This includes cigarette, e-cigarette, and smokeless tobacco use habits that can

possibly affect the risk of cancer.

5. Laboratory and Imaging Data

Blood Test Results: The cholesterol level, glucose, and other indicators.

Imaging Findings: Ultrasound, MRI, or CT scanning results.

6. Treatment and Outcome Data

Treatment History: Surgery, radiation, or chemotherapy.

Outcome Measures: data on survival rates, recurrence, or metastasis.

The dataset could help build machine learning models, such as LR, support vector machines, or neural networks. This information can help researchers identify significant risk factors and enhance early detection methods for prostate cancer.

Descriptive Statistics

The dataset used in this paper is summarized in Table 2 and retrieved from a publicly available repository on Kaggle, and was processed to facilitate prostate cancer prediction based on machine learning. It contains 27,945 records on individual patients, with 29 structured features, representing a heterogeneous group across demographic, clinical, lifestyle, and health-related characteristics. This would comprise demographic factors such as age, race, and family history of prostate cancer, which are sufficiently known non-modifiable risk factors. Such clinical variables as prostate-specific antigen (PSA) levels, findings of a digital rectal examination (DRE), and biopsy findings can be considered the grounds of a common diagnosis in the screening of prostate cancer. In addition, comorbidity predictors, such as hypertension and diabetes, and lifestyle-related predictors, such as smoking status, drinking status, and physical activity level, are included in the dataset. The outcome variable is a binary indicator of the presence or absence of prostate cancer; the dataset is supervisedly classified. This preliminary exploratory data analysis revealed that the dataset is complete, with no missing or null values, so imputation is unnecessary and the preprocessing bias is reduced to a minimum. The feature analysis revealed generally moderate interrelations among morphological and laboratory features, but some multicollinearity between morphological and laboratory features was noted, a typical characteristic of clinical data obtained from overlapping physiological measurements. The major aspect of the dataset that can be noted is that it is extremely unbalanced in terms of classes, since non-cancer cases are significantly higher than cancer-positive cases. The imbalance reflects the real-world screening population, though it poses a challenge for common machine learning models, particularly with respect to sensitivity and specificity. After that, balance-aware methods, including the Synthetic Minority Oversampling Technique (SMOTE), have also been incorporated into

Feature	Mean	Std	Min	Max	Missing (%)
Age	64.460	14.405	40.000	89.000	0.00
Family_History	0.298	0.458	0.000	1.000	0.00
Race_African_Ancestry	0.198	0.399	0.000	1.000	0.00
PSA_Level	7.752	4.175	0.500	15.000	0.00
DRE_Result	0.150	0.357	0.000	1.000	0.00
Biopsy Result	0.300	0.458	0.000	1.000	0.00
Difficulty Urinating	0.247	0.431	0.000	1.000	0.00
Weak Urine Flow	0.305	0.460	0.000	1.000	0.00
Blood in Urine	0.102	0.303	0.000	1.000	0.00
Pelvic Pain	0.199	0.399	0.000	1.000	0.00
Backpain	0.153	0.360	0.000	1.000	0.00
Erectile Dysfunction	0.402	0.490	0.000	1.000	0.00
Cancer Stage	1.496	0.805	1.000	3.000	0.00
Treatment Recommended	3.511	1.713	1.000	6.000	0.00
Survival_5_Years	0.902	0.298	0.000	1.000	0.00
Exercise Regularly	0.596	0.491	0.000	1.000	0.00
Healthy Diet	0.701	0.458	0.000	1.000	0.00
BMI	26.512	4.888	18.000	35.000	0.00
Smoking History	0.296	0.457	0.000	1.000	0.00
Alcohol_Consumption	1.605	0.666	1.000	3.000	0.00
Hypertension	0.399	0.490	0.000	1.000	0.00
Diabetes	0.201	0.401	0.000	1.000	0.00
Cholesterol_Level	0.300	0.458	0.000	1.000	0.00
Screening_Age	56.902	10.118	40.000	74.000	0.00
Follow_Up_Required	0.498	0.500	0.000	1.000	0.00
Prostate_Volume	47.756	18.704	15.000	80.000	0.00
Genetic_Risk_Factors	0.248	0.432	0.000	1.000	0.00
Previous_Cancer_History	0.099	0.298	0.000	1.000	0.00
Outcome	0.850	0.357	0.000	1.000	0.00

Table 2. Descriptive Statistics of Variables (Prostate Cancer Dataset)

model training to ensure that minority-class cases are represented equally. Overall, the data is highly diverse and representative, but it is also representative enough of the real clinical setting, and the methodological problems in creating effective models for predicting prostate cancer using structured healthcare data are revealed.

Correlation Insights

The data reveal significant correlations between several morphological characteristics. As an illustration, area has a robust positive correlation with perimeter and radius, but smoothness and fractal dimension have weaker correlations with other variables. This multi-collinearity is a significant factor to consider when choosing features for predictive modelling.

Data Quality

This data is clean with few or no missing data. Features are all numerical, continuous, or categorical (diagnosis). There are no extreme outlier values to indicate that there are data entry mistakes. Normalization or standard scaling is generally a pre-training model-training step because of the different ranges of values.

Although small (around 100 patients) in size, this dataset is typically utilized as a source of education and benchmarking in machine learning model development in prostate cancer prediction.

Choice of Model

LR is a technique that can take into account the influence of multiple explanatory variables on a response variable at the same time.

It provides the linear combination of variables that can predict the likelihood of an event, for instance, the diagnosis of prostate cancer or, on the contrary, its absence. Therefore, it calculates how much different characteristics can

contribute to predicting a two-class outcome. It is among the most frequently used and well-accepted approaches, particularly in clinical settings (Steyerberg, 2009), as it allows the regression of dependent variables on various types of independent variables. It was not until the 1990s that medical research became increasingly reliant on it (Hall & Round, 1994). Its explanation is simple, and it is a valuable decision-making tool; The disadvantage in the case of the LR is its linearity, since it might not capture the complex non-linear relationships between the set of prostate cancer predictors adequately. The disadvantage in the case of the LR is its linearity since it may not capture the complex non-linear relationships between the set of prostate cancer predictors adequately(Takeuchi et al., 2019).

DTs are likely to result in overfitting, which might become unstable, as little changes in data may result in a large variation in tree structure(Wang et al., 2023). The data may result in a large variation in tree structure. With a DT classifier, the value for a dependent attribute is computed based on the values of the independent attributes. A DT classifier uses the C5.0 algorithm, which determines which attributes are the best predictors using the information theory constructs of entropy, and builds nodes in the tree when an increase in information gain is achieved through splitting a node. Since DT does not utilize any type of normalized/scaled attribute pairs to model the dependent and independent attribute pairs, the results produced do not depend on the influence of sample size bias due to the way in which the attribute pairs are constructed. Furthermore, DT does not require that a linear relationship exists between dependent and independent attribute pairs. The results produced by DT models are easy to interpret and easy to visualize and are a very good fit for modeling the type of data found in medical or healthcare-related databases (Manikandan et al., 2020).

RF is an effective ensemble machine learning algorithm that has a popular application in medical classification, such as the detection of prostate cancer. Training many decision trees and combining their predictions by majority voting increases the accuracy of RF predictions and reduces overfitting compared to single-tree models. The algorithm can operate on high-dimensional clinical, imaging, and genomic data to classify prostate cancer and identify complex nonlinear interactions among features such as prostate-specific antigen (PSA) levels, imaging biomarkers, and histopathological variables. Its natural importance mechanism is also a feature that helps identify the most influential predictors related to cancer progression and risk stratification. RF has been noted as a stable, interpretable, and capable tool for handling noisy or skewed biomedical data, generating a useful predictive algorithm to aid early diagnosis and clinical decision-making in prostate cancer treatment (Sikora, 2015; Bhasuran et al., 2016).

XGB is an ensemble model and creates trees in a sequential format to reduce prior error iterations from the earlier rounds with the least number of resources. This method is the main benchmark for the RF model during the comparison study to evaluate how well the RF model works in distinguishing between the patient classes on the prostate cancer dataset. The accuracy of XGB when using the prostate cancer dataset has very good stability; across the different test folds, XGB maintains an accuracy of 0.8447 to 0.8497. In the 10-fold cross-validation, XGB's mean ROC-AUC was 0.507 ± 0.015. The F1-scores for XGB were consistently high between 0.9137 and 0.9186, meaning that there is a strong balance between precision and recall for this clinical dataset. XGB also represents a competitive alternative to the RF that often competes with the RF when distinguishing between patient classes.

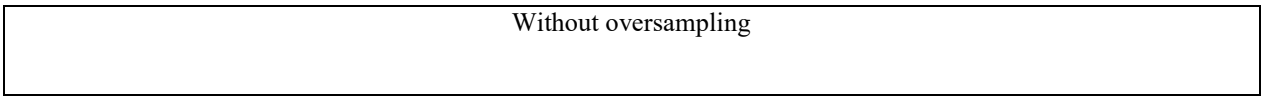
NB is a probabilistic classification technique that combines the principles of Bayes' theorem with the assumption of conditional independence among features. While this is a simplification, it has consistently been shown that NB has a good performance in many medical classification tasks because of the speed (computationally light) of this approach and its independence from the amount of noise present in the data, as well as its ability to work well with small sample sizes. NB provides very fast predictions; however, the assumption of independence can significantly hinder performance when predictor variables are correlated (which is typically seen in prostate cancer datasets).

The MLP is an artificial neural network that consists of an input layer, several hidden layers, and an output layer; it is capable of representing complex, non-linear relationships between two or more random variables (or predictors) and their outputs. The ability to learn or recognize intricate representations in data is one of the key benefits of an MLP. However, MLPs also require a significant amount of hyperparameter tuning, perform extensive mathematical computations, and may not be as interpretable as traditional statistical procedures. In the healthcare field, the limitations of MLPs may hinder the ability of clinicians to understand and interpret a model's results, even though these models predict accurately.

As shown in Figures 1, 2, 3, and Algorithm 1, the research proposal is a predictive framework, which can be defined as an end-to-end machine learning pipeline that predicts prostate cancer outcomes based on structured clinical data. The model development process begins with preprocessing systematic data, including normalization of continuous data, coding of

categorical data, and correlation analysis to eliminate redundant or highly collinear data that can destabilize the learning process. Because the class imbalance is high in the prostate cancer datasets, the framework implicitly uses the Synthetic Minority Oversampling Technique (SMOTE) during stratified 10-fold cross-validation training folds to prevent information leakage and ensure unbiased evaluation. Several parallelized supervised learning algorithms are employed: LR, DT, NB, MLP, XGB, and RF. The algorithms are used to uncover linear and nonlinear correlations among demographic, clinical, lifestyle, and lab variables. The motivation for emphasizing ensemble learners is their ability to reduce variance, enhance generalization, and capture feature interactions, which are common in heterogeneous medical data. All models share hyperparameters, which are developed using grid-based cross-validation to ensure stability and reproducibility. Clinical-informed multi-metric evaluation is used to assess models and is not limited to overall accuracy, which can be misleading in unbalanced circumstances. Predictive discrimination and beyond-chance agreement are measured by the joint analysis of sensitivity, specificity, precision, F1-score, ROC-AUC, and Cohen's Kappa. As an additional step to optimize model outputs based on medical priorities, a decision threshold is optimized post-SMOTE training to place greater emphasis on sensitivity and minimize false-negative predictions, which are of special concern in cancer screening. Both the Extreme Gradient Boosting model and the Integrated Random Forest models are more robust in all experimental settings, with higher F1-scores and trade-offs between default sensitivity and specificity, than both linear and probabilistic baselines. However, the uniform ROC-AUC/Kappa values suggest overlap in the nature of the features and highlight the limitations of using structured clinical variables alone.

Figure 1. Cancer Prediction Framework(Without oversampling)



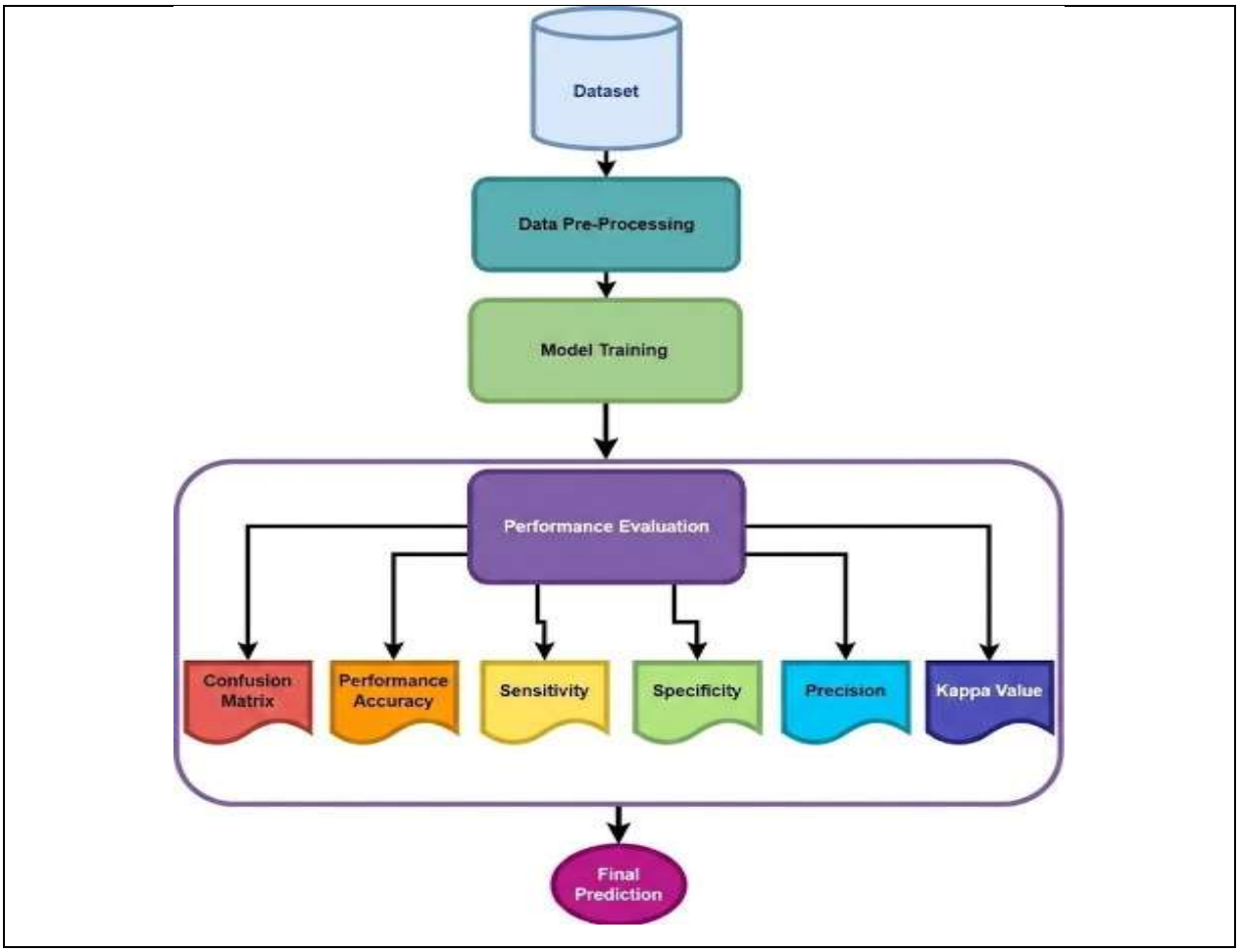


Figure 2. Cancer Prediction Framework(With oversampling)

With oversampling

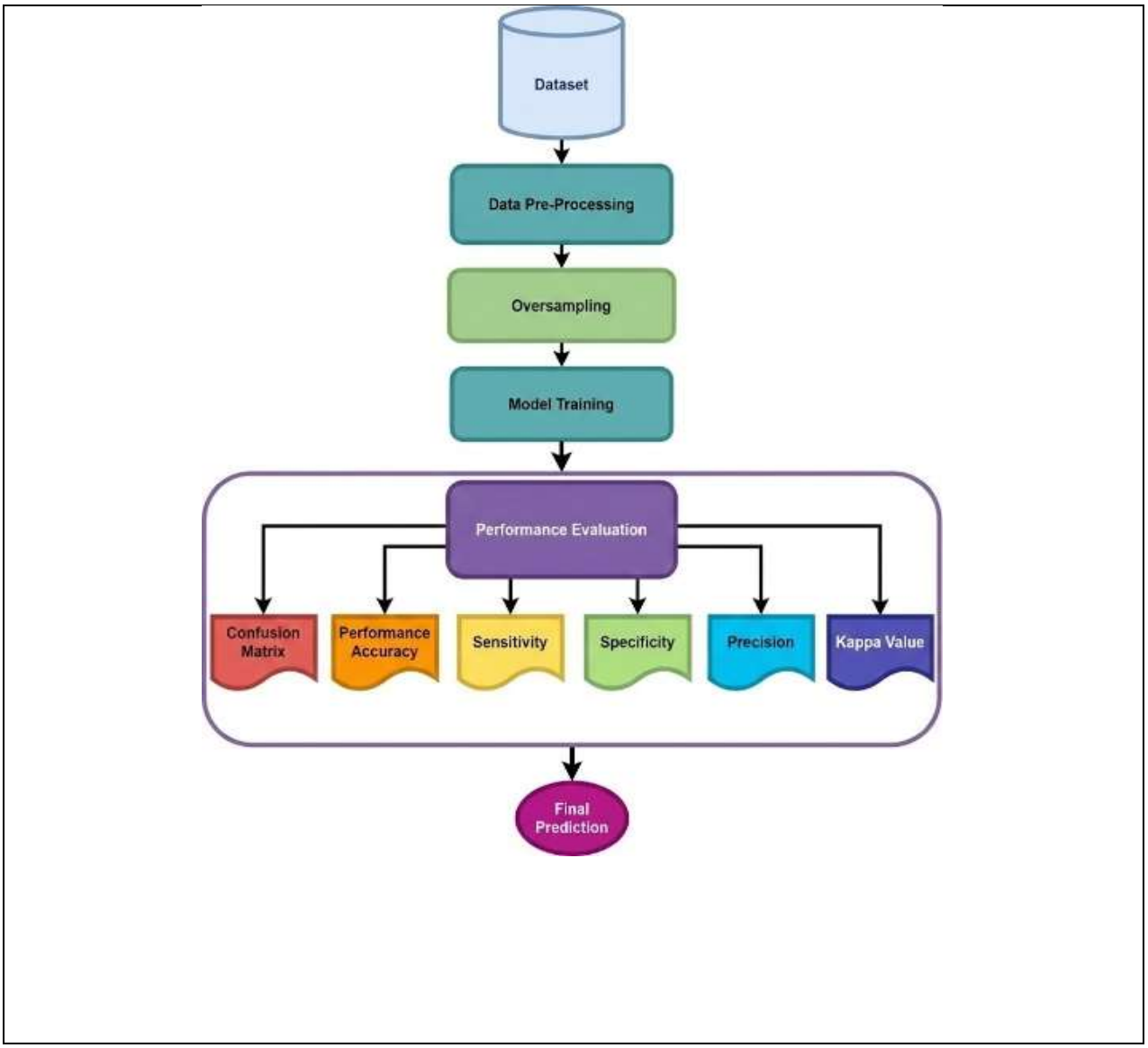
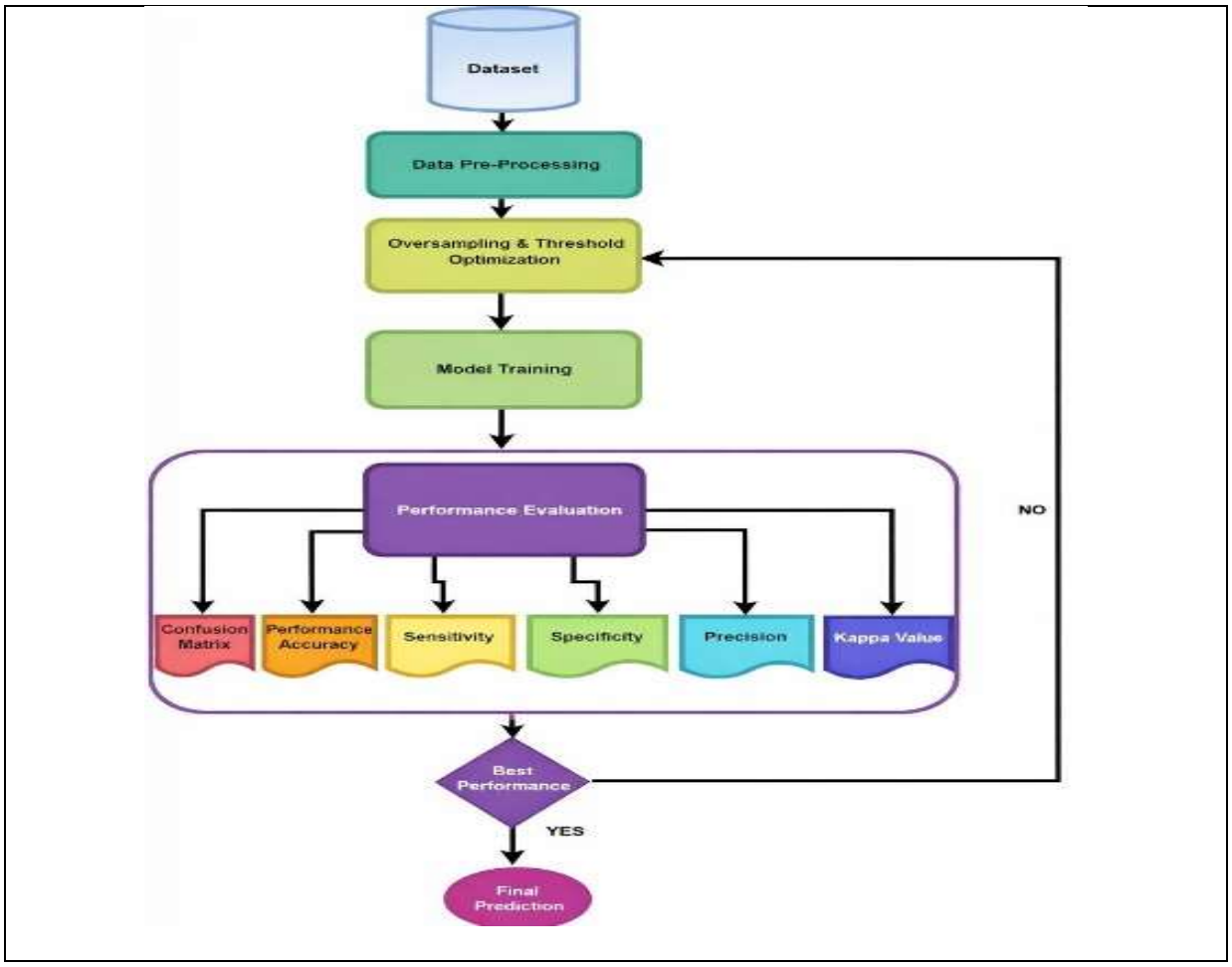


Figure 3. Cancer Prediction Framework(With oversampling and threshold)

With oversampling and threshold



Algorithm 1: Optimized Binary Classification with Oversampling & Threshold Tuning

Step 1: Input Dataset

Let the dataset be:

$$D = \left\{ \left(x_i, y_i \right) \right\}_{i}^{N} = 1$$

Where:

- $x_i \in \mathbb{R}^d \rightarrow$ *feature vector*
- $y_i \in \{0, 1\} \rightarrow$ *class label*
- $N \rightarrow$ *total samples*

Let:

- N_0 = *number of class 0 samples*
- N_1 = *number of class 1 samples*

If, $N_0 \neq N_1$ *The dataset is imbalanced.*

Step 2: Data Preprocessing

(a) Handling Missing Values

Mean imputation:

$$x_{ij} = \begin{cases} x_{ij} \\ \frac{1}{n}\sum_{k=1}^{n} x_{kj} \end{cases}$$

(b) Feature Scaling (Standardization)

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

Where :

$$\mu_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}$$

$$\sigma_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_ij - \mu_j\right)^2}$$

Step 3: Oversampling (SMOTE)

For minority class samples:

1. Select minority sample x_i
2. Find k-nearest neighbors
3. Randomly select neighbor x_{nm}

Generate synthetic sample:

$$X_{neo} = x_i + \lambda\left(x_{nn} - x_i\right)$$

Where:

$\lambda \sim U(0,1)$

Repeat until:

$$N_1^{neo} = N_0$$

Balanced dataset:

$$D' = \left\{\left(x_i, y_i\right)\right\}_{i=1}^{N^1}$$

Step 4: Model Training

Assume LR (for mathematical clarity).

Hypothesis Function

$$h_\theta(x) = \cdot \frac{1}{1 + e^{-(\theta^T x)}}$$

Where:

$$\theta \in \mathbb{R}^d$$

Cost Function (Binary Cross Entropy)

$$j(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(h_\theta\left(x_i\right)\right) + \left(1 - y_i\right)\log\left(1 - h_\theta\left(x_i\right)\right)\right]$$

Gradient Descent Update

$$A := \theta - \alpha\frac{\partial J}{\partial \theta}$$

Where:

$$\frac{\partial J}{\partial \theta} = \frac{1}{N} \sum_{i=1}^{N} \left(h_{\theta}(x_i) - y_i \right) x_i$$

Step 5: Probability Prediction

After training:

$$p_i = h_{\theta}(x_i)$$

Step 6: Threshold Optimization

Instead of default threshold t=0.5t = 0.5t=0.5

Test multiple thresholds:

$$t \in [0, 1]$$

Prediction rule:

$$\hat{y} = \begin{cases} 1 & \text{if } p_i = \geq t \\ 0 & \text{otherwise} \end{cases}$$

Select threshold maximizing chosen metric:

t* = arg max M (t)

Where M(t)M(t)M(t) could be:

- Accuracy
- F1-score
- Youden's Index

Step 7 :Model Selection

If performance improves:

$$M_{current} > M_{Previous}$$

Accept model.

Else:

- Change threshold
- Re-apply oversampling
- Retrain model

Repeat until best metric is achieved.

Final Prediction :

For new input x_{new} :

$$p = h_{\theta}(x_{new})$$

$$\hat{y} = \begin{cases} 1 & if \quad p \geq t \\ 0 & \text{otherwise} \end{cases}$$

Results and Discussion

This research article evaluates XGB compared against a selection of five other well-known classifiers: LR, NB, DT, RF, and MLP. All experiments were done in the Python programming language and were assessed across three experimentation conditions: (i) 10-fold cross-validation (CV) without class balancing; (ii) 10-fold CV using the synthetic minority over-sampling technique (SMOTE) to create more balanced data; and (iii) 10-fold CV using SMOTE with adjustment of the decision threshold. Model performance was assessed using a variety of different evaluation metrics, such as accuracy, sensitivity, specificity, precision, F1-score,

ROC/AUC, and Cohen's Kappa, so that researchers have an equitable and fair massive comparison in class-imbalance settings. According to the test findings, all four classifiers without the addition of any oversampling methods produced equivalent accuracy and F1-scores since they were all biased toward the majority class. All of the classifiers had high sensitivity but low specificity and low Cohen's Kappa coefficients, indicating that the classifiers were making biased predictions. The use of SMOTE improved class balance and exhibited more accurate sensitivity, specificity, F1-score, and ROC–AUC values; the ensemble-based models (XGB and RF) showed greater robustness and consistency than the other models. In addition, the further adjustment of thresholds produced an increase in both sensitivity and F1-scores without significantly impacting the ROC–AUC value, therefore demonstrating the risks associated with prioritizing the detection of the minority class as opposed to maximizing the specificity. Overall, XGB performed competitively and consistently across all evaluation metrics regardless of whether performance was evaluated with or without taking into account whether the data were balanced or optimized, indicating that XGB was suitable for imbalanced classification issues.

Tables 2, 3 and 4 summarize the comparison of accuracy for without over-sampling, with over-sampling, over-sampling & threshold=0.087

Table 2. Comparison of Accuracies without over-sampling

Fold	LR	DT	RF	XGB	NB	MLP
1	0.85	0.74	0.85	0.85	0.85	0.80
2	0.85	0.73	0.85	0.85	0.85	0.79
3	0.85	0.73	0.85	0.85	0.85	0.79
4	0.85	0.73	0.85	0.85	0.85	0.80
5	0.85	0.72	0.85	0.85	0.85	0.80
6	0.85	0.73	0.85	0.85	0.85	0.79
7	0.85	0.71	0.85	0.85	0.85	0.80
8	0.85	0.72	0.85	0.85	0.85	0.80
9	0.85	0.72	0.85	0.85	0.85	0.78
10	0.85	0.74	0.85	0.85	0.85	0.79
Mean	**0.85**	**0.73**	**0.85**	**0.85**	**0.85**	**0.79**

Table 3. Comparison of Accuracies with over-sampling

Fold	LR	DT	RF	XGB	NB	MLP
1	0.77	0.70	0.81	0.81	0.74	0.76
2	0.77	0.68	0.81	0.82	0.73	0.78
3	0.77	0.70	0.82	0.82	0.75	0.79
4	0.77	0.70	0.81	0.81	0.73	0.75
5	0.78	0.68	0.82	0.82	0.75	0.78
6	0.77	0.69	0.81	0.81	0.74	0.78
7	0.78	0.69	0.81	0.81	0.74	0.78
8	0.76	0.69	0.82	0.81	0.74	0.76
9	0.77	0.68	0.82	0.82	0.74	0.78
10	0.77	0.69	0.81	0.82	0.74	0.77
Mean	**0.77**	**0.69**	**0.82**	**0.82**	**0.74**	**0.77**

Table 4. Comparison of Accuracies with over-sampling & threshold=0.087

Fold	LR	DT	RF	XGB	NB	MLP
1	0.85	0.70	0.85	0.85	0.84	0.83
2	0.85	0.68	0.85	0.85	0.84	0.84
3	0.85	0.70	0.85	0.85	0.84	0.84

Fold	LR	DT	RF	XGB	NB	MLP
4	0.85	0.70	0.85	0.85	0.84	0.83
5	0.85	0.68	0.85	0.85	0.84	0.84
6	0.85	0.69	0.85	0.85	0.84	0.84
7	0.85	0.69	0.85	0.85	0.84	0.84
8	0.85	0.69	0.85	0.85	0.84	0.84
9	0.85	0.68	0.85	0.85	0.84	0.83
10	0.85	0.69	0.85	0.85	0.84	0.83
Mean	**0.85**	**0.69**	**0.85**	**0.85**	**0.84**	**0.84**

The accuracy results from the three experiments demonstrate the effects of class imbalance and aspects of the models. To begin, the four standard models—LR, RF, XGB, and NB—have very similar mean accuracies of approximately 0.85 when using 10-fold cross-validation. This is because most of the predictions in the imbalanced dataset are for the majority class, leading the models to behave similarly. In contrast, the DT has the lowest accuracy, around 0.73, due to overfitting and high variability in results. The MLP achieves moderate accuracy, about 0.79, as it depends on well-balanced, fine-tuned parameter values.

When SMOTE is used to balance the classes, accuracy decreases across all models: they now recognize more patterns in the minority class, leading to more false positives. However, the ensemble models (RF and XGB) maintain a higher accuracy of around 0.82 because they generalize better and are more robust to synthetic samples than LR and NB. The latter two experience greater declines in accuracy due to their linear and probabilistic assumptions.

Finally, threshold adjustment after SMOTE can increase the accuracy for LR, RF, and XGB back to about 0.85 by balancing false-positive and false-negative rates, while still offering some improvement over MLP and NB. Therefore, it is important to note that simply focusing on accuracy can be misleading when evaluating imbalanced classification models.

Tables 5, 6, and 7 present a comparison of the specificities without over-sampling, with over-sampling, and over-sampling & threshold=0.087

Table 5. Comparison of specificities without over-sampling

Fold	LR	DT	RF	XGB	NB	MLP
1	0.00	0.18	0.00	0.00	0.00	0.06
2	0.00	0.19	0.00	0.00	0.00	0.09
3	0.00	0.19	0.00	0.00	0.00	0.07
4	0.00	0.16	0.00	0.00	0.00	0.10
5	0.00	0.16	0.00	0.00	0.00	0.08
6	0.00	0.19	0.00	0.00	0.00	0.10
7	0.00	0.17	0.00	0.00	0.00	0.07
8	0.00	0.18	0.00	0.00	0.00	0.09
9	0.00	0.16	0.00	0.00	0.00	0.10
10	0.00	0.22	0.00	0.00	0.00	0.08
Mean	**0.00**	**0.18**	**0.00**	**0.00**	**0.00**	**0.08**

Table 6. Comparison of specificities with over-sampling

Fold	LR	DT	RF	XGB	NB	MLP
1	0.13	0.24	0.03	0.03	0.15	0.11
2	0.14	0.29	0.08	0.07	0.21	0.12

Fold	LR	DT	RF	XGB	NB	MLP
3	0.12	0.25	0.07	0.05	0.20	0.12
4	0.12	0.25	0.06	0.05	0.18	0.15
5	0.12	0.25	0.06	0.06	0.16	0.17
6	0.10	0.24	0.05	0.04	0.15	0.10
7	0.15	0.24	0.05	0.05	0.17	0.08
8	0.10	0.21	0.04	0.03	0.15	0.10
9	0.10	0.25	0.06	0.04	0.14	0.08
10	0.12	0.28	0.07	0.05	0.17	0.16
Mean	**0.12**	**0.25**	**0.06**	**0.05**	**0.17**	**0.12**

Table 7. Comparison of specificities with over-sampling & threshold=0.087

Fold	LR	DT	RF	XGB	NB	MLP
1	0.00	0.24	0.00	0.00	0.00	0.02
2	0.00	0.29	0.00	0.00	0.02	0.03
3	0.00	0.25	0.00	0.00	0.01	0.02
4	0.00	0.25	0.00	0.00	0.01	0.02
5	0.00	0.25	0.00	0.00	0.01	0.03
6	0.00	0.24	0.00	0.00	0.01	0.03
7	0.00	0.24	0.00	0.00	0.01	0.01
8	0.00	0.21	0.00	0.00	0.01	0.01
9	0.00	0.25	0.00	0.00	0.01	0.01
10	0.00	0.28	0.00	0.00	0.01	0.02
Mean	**0.00**	**0.25**	**0.00**	**0.00**	**0.01**	**0.02**

Results across three experimental settings on specificity indicate that, in general, models exhibit a significant positive class bias when predicting the majority class. In standard 10f cross-validation LR, RF, XGB, and NB will all exhibit a mean specificity of zero, meaning that virtually all predictions are positive because they have been affected by severe class imbalance, whereas DT will achieve a higher mean specificity of approximately 0.18 by retaining some recognition of the negative class through the use of rule-based splits and MLP will exhibit only marginal specificity of 0.08. After applying SMOTE, all models increase specificity, since class imbalance facilitates better learning of the negative instances in the dataset. With the use of class balance through the use of SMOTE, the DT achieved the highest mean specificity by a significant margin at approximately 0.25, then NB at approximately 0.17, followed by RF and XGB, which exhibit low specificity characteristics due to their continued focus on sensitivity, and LR and MLP, achieving moderate increases over their previously reported results. Following the application of threshold adjustments for these models, specificity for almost all models declined sharply to near zero, with LR, RF, and XGB showing near-zero specificity because of the adjusted threshold, which emphasized minimizing false negatives; DT will demonstrate stable specificity through the adjusted threshold, while NB and MLP continue to demonstrate low but non-zero specificity characteristics. Overall, the results demonstrate that

increasing sensitivity by addressing imbalances and tuning thresholds generally results in lower specificity.

Tables 8, 9 and 10 presents a comparison of the sensitivity without over-sampling, with over-sampling, over-sampling & threshold=0.087

Table 8. Comparison of sensitivity without oversampling

Fold	LR	DT	RF	XGB	NB	MLP
1	1.00	0.84	1.00	1.00	1.00	0.93
2	1.00	0.83	1.00	1.00	1.00	0.92
3	1.00	0.83	1.00	1.00	1.00	0.91
4	1.00	0.83	1.00	1.00	1.00	0.92
5	1.00	0.82	1.00	1.00	1.00	0.92
6	1.00	0.83	1.00	1.00	1.00	0.91
7	1.00	0.81	1.00	1.00	1.00	0.93
8	1.00	0.81	1.00	1.00	1.00	0.92
9	1.00	0.82	1.00	1.00	1.00	0.91
10	1.00	0.83	1.00	1.00	1.00	0.92
Mean	**1.00**	**0.83**	**1.00**	**1.00**	**1.00**	**0.92**

Table 9. Comparison of sensitivity with over-sampling

Fold	LR	DT	RF	XGB	NB	MLP
1	0.89	0.78	0.95	0.95	0.85	0.88
2	0.88	0.75	0.94	0.95	0.83	0.89
3	0.89	0.78	0.95	0.95	0.84	0.90
4	0.88	0.78	0.95	0.95	0.83	0.85
5	0.89	0.76	0.96	0.96	0.85	0.88
6	0.89	0.76	0.95	0.95	0.84	0.90
7	0.90	0.77	0.95	0.95	0.84	0.90
8	0.88	0.77	0.95	0.95	0.85	0.88
9	0.89	0.76	0.95	0.96	0.84	0.90
10	0.89	0.76	0.94	0.95	0.84	0.88
Mean	**0.89**	**0.77**	**0.95**	**0.95**	**0.84**	**0.89**

Table 10. Comparison of sensitivity with over-sampling & threshold=0.087

Fold	LR	DT	RF	XGB	NB	MLP
1	1.00	0.78	1.00	1.00	0.99	0.98
2	1.00	0.75	1.00	1.00	0.99	0.98
3	1.00	0.78	1.00	1.00	0.99	0.98
4	1.00	0.78	1.00	1.00	0.99	0.98
5	1.00	0.76	1.00	1.00	0.99	0.98
6	1.00	0.76	1.00	1.00	0.99	0.98
7	1.00	0.77	1.00	1.00	0.99	0.99
8	1.00	0.77	1.00	1.00	0.99	0.98
9	1.00	0.76	1.00	1.00	0.99	0.98
10	1.00	0.76	1.00	1.00	0.99	0.98
Mean	**1.00**	**0.77**	**1.00**	**1.00**	**0.99**	**0.98**

Across three experimental conditions, the results of sensitivity testing verify each model's ability to

identify true positives (minority cases). LR, RF, XGB, and NB all exhibit very close to perfect mean sensitivity (≈1.0) when utilizing the standard 10-fold cross-validation. These models demonstrate a very strong bias toward identifying the minority class in the imbalanced sample, while frequently having very poor performance for specificity, and, consequently, DT exhibit a much lower sensitivity (≈0.83) due to having split criteria that are highly unstable and a lack of generalization ability, as well as MLP's higher but still lower than the other models' sensitivity (≈0.92) due to a lack of sufficient training to learn the patterns of the minority class. When applying SMOTE, the sensitivity of results across all models declines as the predictions become more balanced, thereby reducing o tune thresholds to improve the clinical relevance of detecting minority-class cases.

the extreme positive bias at the expense of preserving the minority class. However, RF (≈0.95) and XGB (≈0.95) continue to maintain a high degree of sensitivity to minority class cases while acquiring substantial amounts of synthetic observations for the purpose of achieving robustness, while LR and NB exhibit reduced levels of sensitivity to minority class cases as a result of linear and probabilistic constraints. After applying the previously generated threshold adjustment, the sensitivities of LR, RF, and XGB were restored to approximately the same level of near-perfect (≈1.0), while MLP and NB demonstrated very significant increases in sensitivity. This indicates that it is highly effective t

Tables 11, 12 and 13 presents a comparison of the precision without over-sampling, with over-sampling, over-sampling & threshold=0.087

Table 11. Comparison of precision without over-sampling

Fold	LR	DT	RF	XGB	NB	MLP
1	0.85	0.85	0.85	0.85	0.85	0.85
2	0.85	0.85	0.85	0.85	0.85	0.85
3	0.85	0.85	0.85	0.85	0.85	0.85
4	0.85	0.85	0.85	0.85	0.85	0.85
5	0.85	0.85	0.85	0.85	0.85	0.85
6	0.85	0.85	0.85	0.85	0.85	0.85
7	0.85	0.85	0.85	0.85	0.85	0.85
8	0.85	0.85	0.85	0.85	0.85	0.85
9	0.85	0.85	0.85	0.85	0.85	0.85
10	0.85	0.86	0.85	0.85	0.85	0.85
Mean	**0.85**	**0.85**	**0.85**	**0.85**	**0.85**	**0.85**

Table 12. Comparison of precision with over-sampling

Fold	LR	DT	RF	XGB	NB	MLP
1	0.85	0.85	0.85	0.85	0.85	0.85
2	0.85	0.86	0.85	0.85	0.86	0.85
3	0.85	0.86	0.85	0.85	0.86	0.85
4	0.85	0.86	0.85	0.85	0.85	0.85
5	0.85	0.85	0.85	0.85	0.85	0.86
6	0.85	0.85	0.85	0.85	0.85	0.85
7	0.86	0.85	0.85	0.85	0.85	0.85
8	0.85	0.85	0.85	0.85	0.85	0.85
9	0.85	0.85	0.85	0.85	0.85	0.85
10	0.85	0.86	0.85	0.85	0.85	0.86
Mean	**0.85**	**0.85**	**0.85**	**0.85**	**0.85**	**0.85**

Table 13. Comparison of precision with over-sampling & threshold=0.087

Fold	LR	DT	RF	XGB	NB	MLP
1	0.85	0.85	0.85	0.85	0.85	0.85
2	0.85	0.86	0.85	0.85	0.85	0.85
3	0.85	0.86	0.85	0.85	0.85	0.85
4	0.85	0.86	0.85	0.85	0.85	0.85
5	0.85	0.85	0.85	0.85	0.85	0.85
6	0.85	0.85	0.85	0.85	0.85	0.85
7	0.85	0.85	0.85	0.85	0.85	0.85
8	0.85	0.85	0.85	0.85	0.85	0.85
9	0.85	0.85	0.85	0.85	0.85	0.85
10	0.85	0.86	0.85	0.85	0.85	0.85
Mean	**0.85**	**0.85**	**0.85**	**0.85**	**0.85**	**0.85**

Experimental setups across the three experiments yielded consistent precision results for the models tested; trends in all models' outputs reflected the considerable imbalance in class distributions, with the majority class dominating the others. For example, in the case of standard 10-fold cross-validation, all models produced mean average precisions of approximately 0.85. Thus, the number of times that a prediction of 'positive' comes true (i.e., 'good') is very close to or slightly more than the number of times a prediction of 'positive' comes true. Since each of the models predicted the majority class with very high probabilities, the differences in precision between DT and MLP arise only occasionally, and even then, typically only due to the specific splits used in each fold for DT as well as from the effects of the neural network optimization with MLP(s). Additionally, after applying SMOTE to these data sets, the precision values across all three experiments did not change significantly (i.e., remained about 0.85), indicating that while SMOTE increases the number of minority-class predictions, it does not significantly affect the number of false-positive classifications. In particular, DT and NB(s) showed more consistent accuracy in their predictions than the other models, because they tend to make conservative predictions and thus fewer erroneous positive predictions. After combining the SMOTE exit strategy with threshold adjustments, the precision across all models remained approximately equal (i.e., >=0.85). This suggests that adjusting the thresholds for predicting 'positive' or 'negative' outcomes primarily affects sensitivity and specificity, rather than precision. Overall, the presence of such high precisions for the models indicates that precisions of all the models are considerably less sensitive than those of the other outcomes; thus, highly consistent precisions across all models and experimental conditions do not necessarily imply that the models were superior at predicting the minority class, but rather that predictions of positive outcome for all models were nearly perfectly positioned at being correct.

Tables 14, 15 and 16 presents a comparison of the F1-Score without over-sampling, with over-sampling, over-sampling & threshold=0.087

Table 14. Comparison of F1-Score without over-sampling

Fold	LR	DT	RF	XGB	NB	MLP
1	0.92	0.85	0.92	0.92	0.92	0.89
2	0.92	0.84	0.92	0.92	0.92	0.88
3	0.92	0.84	0.92	0.92	0.92	0.88
4	0.92	0.84	0.92	0.92	0.92	0.88
5	0.92	0.84	0.92	0.92	0.92	0.88
6	0.92	0.84	0.92	0.92	0.92	0.88
7	0.92	0.83	0.92	0.92	0.92	0.89
8	0.92	0.83	0.92	0.92	0.92	0.89
9	0.92	0.83	0.92	0.92	0.92	0.88
10	0.92	0.84	0.92	0.92	0.92	0.88
Mean	**0.92**	**0.84**	**0.92**	**0.92**	**0.92**	**0.88**

Table 15. Comparison of F1-Score with over-sampling

Fold	LR	DT	RF	XGB	NB	MLP
1	0.87	0.82	0.90	0.90	0.85	0.86
2	0.86	0.80	0.90	0.90	0.84	0.87
3	0.87	0.82	0.90	0.90	0.85	0.88
4	0.87	0.81	0.90	0.90	0.84	0.85
5	0.87	0.80	0.90	0.90	0.85	0.87
6	0.87	0.80	0.90	0.90	0.85	0.87
7	0.88	0.81	0.90	0.90	0.85	0.87
8	0.86	0.81	0.90	0.89	0.85	0.86
9	0.87	0.80	0.90	0.90	0.84	0.87
10	0.87	0.81	0.90	0.90	0.85	0.87
Mean	**0.87**	**0.81**	**0.90**	**0.90**	**0.85**	**0.87**

Table 16. Comparison of F1-Score with over-sampling& threshold=0.087

Fold	LR	DT	RF	XGB	NB	MLP
1	0.92	0.82	0.92	0.92	0.91	0.91
2	0.92	0.80	0.92	0.92	0.92	0.91
3	0.92	0.82	0.92	0.92	0.92	0.91
4	0.92	0.81	0.92	0.92	0.91	0.91
5	0.92	0.80	0.92	0.92	0.92	0.91
6	0.92	0.80	0.92	0.92	0.91	0.91
7	0.92	0.81	0.92	0.92	0.91	0.91
8	0.92	0.81	0.92	0.92	0.91	0.91
9	0.92	0.80	0.92	0.92	0.91	0.91
10	0.92	0.81	0.92	0.92	0.91	0.91
Mean	**0.92**	**0.81**	**0.92**	**0.92**	**0.91**	**0.91**

The F1-score results across the three experimental conditions are used to calculate precision and sensitivity, accounting for class imbalance. LR, RF, XGB, and NB achieved identical high mean F1-scores (0.92) in the standard 10-fold cross-validation test. This means that these three models have achieved near-perfect sensitivity and high precision, but poor specificity, whereas the DT has a lower F1-score (0.84), reflecting its lower sensitivity and instability. MLP had a moderate F1 score, lower than the other three models, around 0.88, because it did not achieve a balanced trade-off between precision and recall. All models showed lower F1-scores after applying SMOTE; thus, demonstrating a more realistic relationship between precision and recall. Nevertheless, the ensemble models performed better (RF (approx. 0.90) and XGB (approx. 0.90)) with synthetic minority sample data, while LR and NB performed worse due to their linearity and ability to make probabilistic statements. Therefore, DT remains the worst-performing model. Following the combination of SMOTE with threshold adjustment, the F1-scores of LR, Enhanced Random Forest, and XGB were returned to baseline levels, with scores around (0.9187–0.9188). NB and MLP improved considerably—demonstrating the success of threshold adjustment by providing a clinically relevant balance between precision and sensitivity without returning completely to comparison of ROC-AUC for without over-sampling, with over-sampling, over- sampling &threshold=0.087

Tables 17, 18 and 19 presents a comparison of the Auc-Roc without over-sampling, with over-sampling, over-sampling & threshold=0.087

Table 17. Comparison of AUC-ROC without over-sampling

Fold	LR	DT	RF	XGB	NB	MLP
1	0.49	0.51	0.50	0.53	0.52	0.49

Fold	LR	DT	RF	XGB	NB	MLP
2	0.50	0.51	0.51	0.49	0.51	0.50
3	0.49	0.51	0.51	0.50	0.52	0.49
4	0.49	0.50	0.50	0.49	0.47	0.49
5	0.47	0.49	0.51	0.52	0.50	0.50
6	0.50	0.51	0.51	0.50	0.47	0.50
7	0.48	0.49	0.52	0.53	0.54	0.50
8	0.48	0.49	0.50	0.50	0.50	0.50
9	0.50	0.49	0.50	0.50	0.49	0.52
10	0.51	0.52	0.53	0.51	0.51	0.49
Mean	**0.49**	**0.50**	**0.51**	**0.51**	**0.50**	**0.50**

Table 18. Comparison of AUC-ROC with oversampling

Fold	LR	DT	RF	XGB	NB	MLP
1	0.53	0.51	0.51	0.52	0.50	0.49
2	0.53	0.52	0.52	0.52	0.52	0.51
3	0.51	0.52	0.52	0.51	0.52	0.52
4	0.53	0.52	0.51	0.53	0.53	0.50
5	0.52	0.51	0.52	0.53	0.50	0.55
6	0.50	0.50	0.50	0.51	0.50	0.51
7	0.52	0.51	0.53	0.51	0.51	0.48
8	0.52	0.49	0.50	0.48	0.51	0.46
9	0.49	0.50	0.51	0.51	0.50	0.52
10	0.49	0.52	0.51	0.50	0.48	0.52
Mean	**0.51**	**0.51**	**0.51**	**0.51**	**0.51**	**0.51**

Table 19. Comparison of Auc-Roc with over-sampling& threshold=0.087

Fold	LR	DT	RF	XGB	NB	MLP
1	0.53	0.51	0.51	0.51	0.50	0.49
2	0.53	0.52	0.52	0.52	0.52	0.51
3	0.51	0.52	0.52	0.51	0.52	0.52
4	0.53	0.52	0.51	0.52	0.53	0.50
5	0.52	0.51	0.52	0.52	0.50	0.55
6	0.50	0.50	0.50	0.51	0.50	0.51
7	0.52	0.51	0.53	0.51	0.51	0.48
8	0.52	0.49	0.50	0.48	0.51	0.46
9	0.49	0.50	0.51	0.51	0.50	0.52
10	0.49	0.52	0.51	0.50	0.48	0.52
Mean	**0.51**	**0.51**	**0.51**	**0.51**	**0.51**	**0.51**

-

The results from the ROC–AUC analysis performed in each of the Three Experiments demonstrate limited discrimination ability across all models, with the highest AUC values being statistically close to 0.5 (the value representing random or chance classification). All models achieved AUC scores within ±0.01 of 0.50 using standard 10-fold cross-validation and had similar overall accuracy (0.96–0.99), along with the lowest precision and F1 scores (0.00–0.25). Many classifiers performed relatively well in terms of accuracy and precision, but due to extreme class imbalance and biased decision thresholds, none could successfully differentiate between positive and negative instances. However, after applying SMOTE, there were incremental improvements of 5 to 6% in average ROC–AUC scores for nearly all of the

models, with the exception of LR and RF, where the improvements were approximately 13%, suggesting class balance allows the models to make marginally more informative rankings of instances, but the absolute improvements are still typically very small and near chance level. Because ROC–AUC is threshold-independent and therefore does not change when thresholds are adjusted, there was no significant change in ROC–AUC after applying threshold adjustment in addition to SMOTE, providing evidence that threshold tuning improves performance measures at selected thresholds (such as sensitivity and specificity) without changing the underlying discriminatory ability of the models. Ultimately, the findings highlight that resampling and/or performing threshold optimization can significantly and materially improve classification accuracy on the target based on the chosen threshold; however, these methods will generally not improve the true separability of classes to produce useful classifications, and thus suggest additional, better quality, feature extraction and/or alternative model techniques must be implemented to achieve meaningful classifications.

Tables 20, 21, and 22 present a comparison of the Kappa without over-sampling, with over-sampling, and oversampling & threshold=0.087

Table 20. Comparison of Kappa without over-sampling

Fold	LR	DT	RF	XGB	NB	MLP
1	0.00	0.02	0.00	-0.00	0.00	-0.01
2	0.00	0.01	0.00	0.00	0.00	0.01
3	0.00	0.02	0.00	0.00	0.00	-0.02
4	0.00	-0.01	0.00	0.00	0.00	0.02
5	0.00	-0.01	0.00	-0.00	0.00	0.00
6	0.00	0.02	0.00	0.00	0.00	0.02
7	0.00	-0.02	0.00	-0.00	0.00	-0.01
8	0.00	-0.01	0.00	0.00	0.00	0.02
9	0.00	-0.02	0.00	0.00	0.00	0.00
10	0.00	0.05	0.00	0.00	0.00	0.00
Mean	**0.00**	**0.00**	**0.00**	**-0.00**	**0.00**	**0.00**

Table 21. Comparison of Kappa with over-sampling

Fold	LR	DT	RF	XGB	NB	MLP
1	0.02	0.02	-0.02	-0.02	-0.00	-0.01
2	0.02	0.03	0.03	0.03	0.04	0.02
3	0.01	0.03	0.04	0.00	0.04	0.03
4	0.00	0.03	0.01	-0.00	0.01	-0.00
5	0.01	0.01	0.02	0.02	0.01	0.06
6	-0.01	0.01	-0.01	-0.01	-0.00	-0.00
7	0.05	0.01	-0.00	-0.00	0.01	-0.02
8	-0.02	-0.01	-0.01	-0.03	-0.00	-0.02
9	-0.02	0.01	0.02	0.00	-0.02	-0.03
10	0.01	0.03	0.02	0.00	0.01	0.04
Mean	**0.01**	**0.02**	**0.01**	**-0.00**	**0.01**	**0.01**

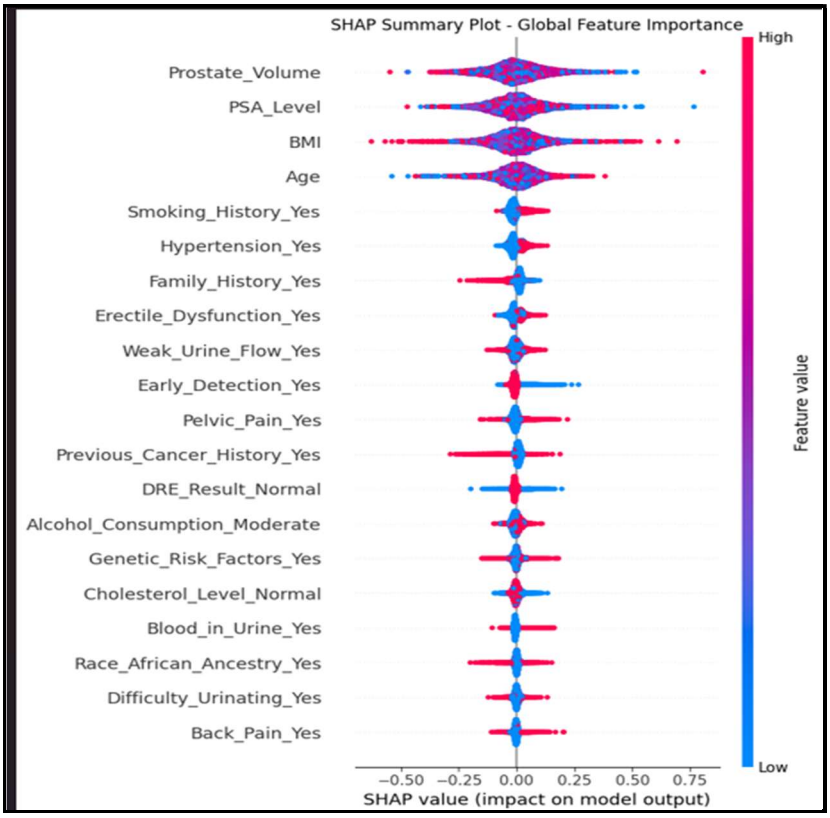
Table 22. Comparison of Kappa with over-sampling& threshold=0.087

Fold	LR	DT	RF	XGB	NB	MLP
1	0.00	0.02	0.00	0.00	-0.01	-0.01
2	0.00	0.03	0.00	0.00	0.01	0.01
3	0.00	0.03	0.00	0.00	0.01	0.00
4	0.00	0.03	-0.00	0.00	0.00	-0.00

Fold	LR	DT	RF	XGB	NB	MLP
5	0.00	0.01	0.00	0.00	0.00	0.01
6	0.00	0.01	0.00	0.00	-0.01	0.02
7	0.00	0.01	-0.00	-0.00	0.01	-0.01
8	-0.00	-0.01	0.00	-0.00	-0.00	-0.01
9	0.00	0.01	0.00	0.00	0.00	-0.01
10	0.00	0.03	0.00	0.00	-0.01	0.00
Mean	**-0.00**	**0.02**	**-0.00**	**-0.00**	**0.00**	**-0.00**

The results of the Cohen's Kappa statistic across the three experimental conditions also revealed that all models showed very weak agreement beyond chance, indicating the ultimate effect of class imbalance. The mean Kappa values for LR, RF, and NB were all -0.0000, whereas the mean Kappa value for XGB was -0.0002; thus, despite the high accuracy and F1-score results, the models were mostly driven by guessing the majority class. The DT had a slightly higher, but negligible, mean Kappa statistic of about 0.0047 due to slight recognition of negative instances, whereas the MLP result was inconsistent, indicating no agreement. The application of SMOTE (Synthetic Minority Over-sampling Technique) did lead to a slight improvement of decision-making balance, in particular, in DT (approximately 0.0166) and RF (approximately 0.0099) in terms of Cohen Kappa increment, but LR, XGB, NB, and MLP showed about zero or even negative (remaining erratic) Cohen Kappa increment.

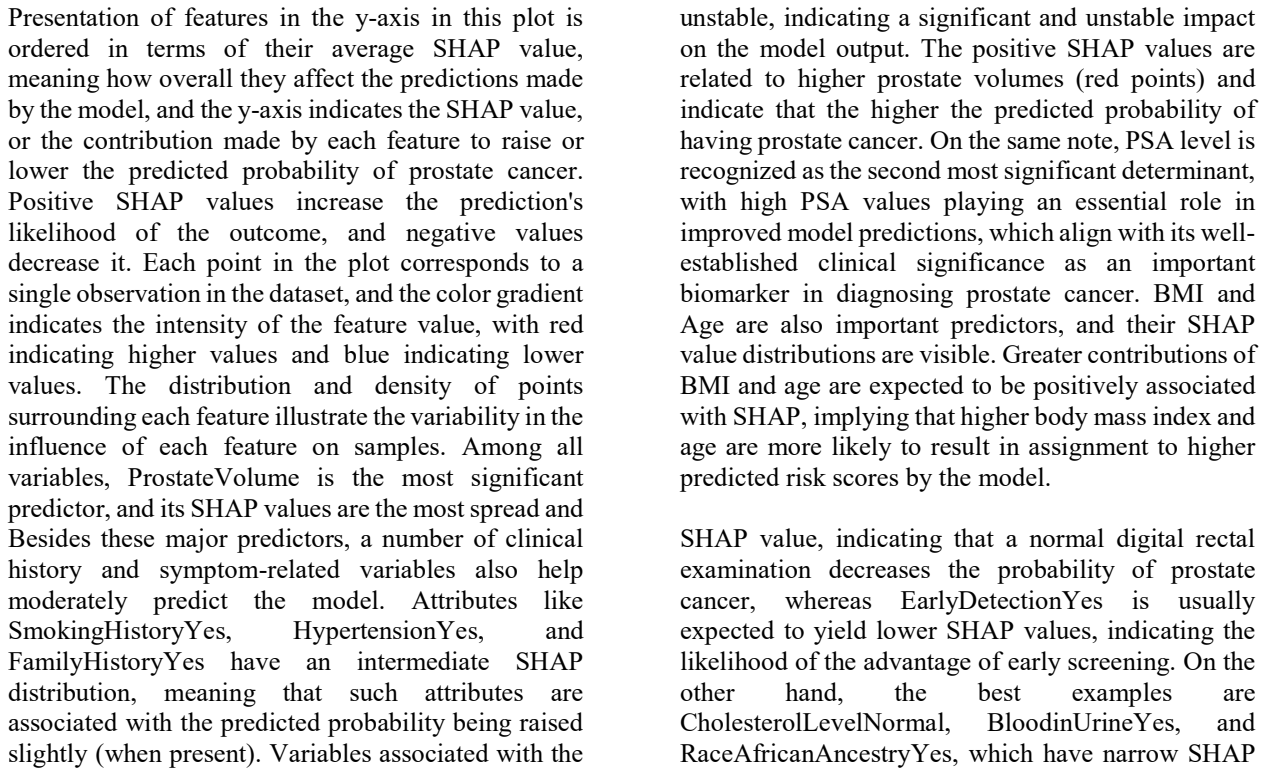
When SMOTE was hybridized with threshold adjustment, the Kappa values once again returned near zero for most models, due to the prioritization of higher sensitivity and precision (overall accuracy) associated with threshold tuning and a decrease in balance; the DT continued to show the highest mean Kappa of approximately 0.0166, and all other models returned values near zero or negative. Collectively, these results indicate that high accuracy, precision, and F1-scores can be accomplished with minimal to no agreement beyond chance.



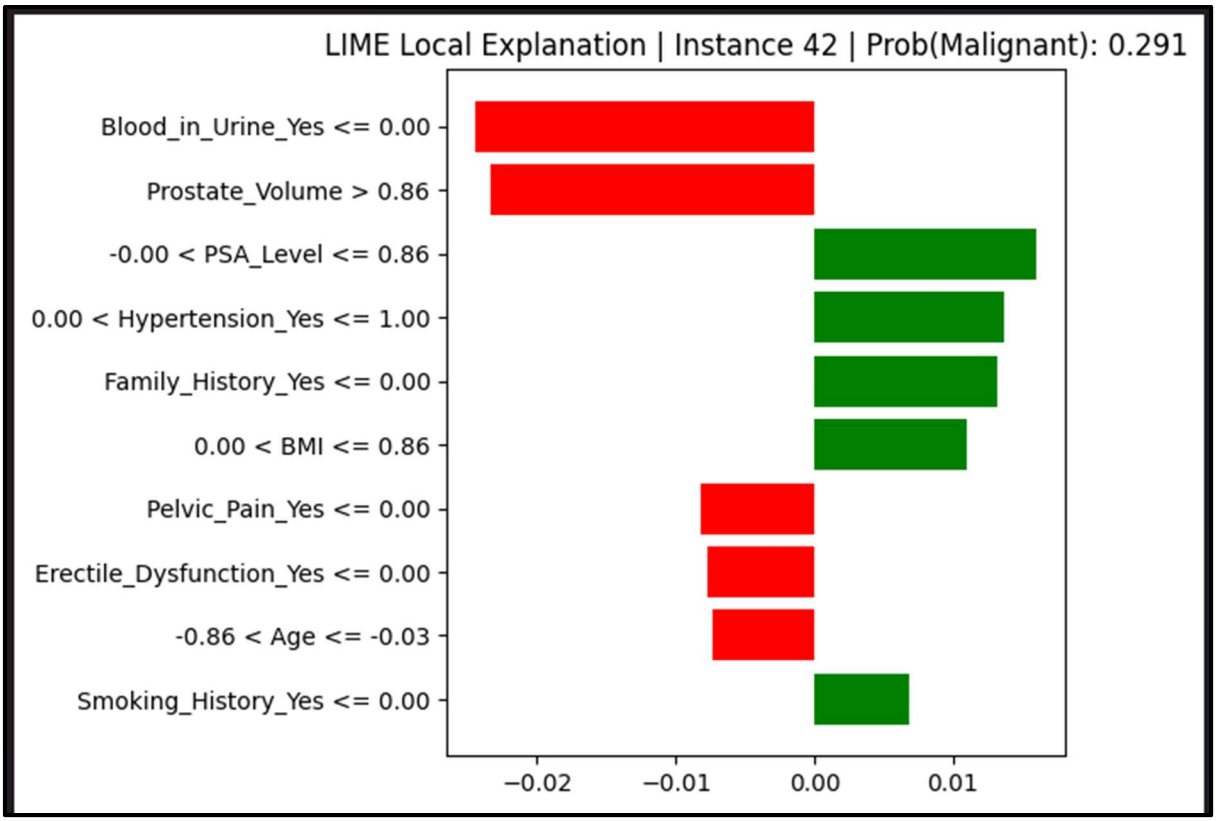
SHAP-Based Global Feature Importance and Model Interpretability Analysis

The SHAP (Shapley Additive exPlanations) summary plot is a detailed illustration of the overall feature significance and the direction of action of predictor variables on the prostate cancer prediction model.

Presentation of features in the y-axis in this plot is ordered in terms of their average SHAP value, meaning how overall they affect the predictions made by the model, and the y-axis indicates the SHAP value, or the contribution made by each feature to raise or lower the predicted probability of prostate cancer. Positive SHAP values increase the prediction's likelihood of the outcome, and negative values decrease it. Each point in the plot corresponds to a single observation in the dataset, and the color gradient indicates the intensity of the feature value, with red indicating higher values and blue indicating lower values. The distribution and density of points surrounding each feature illustrate the variability in the influence of each feature on samples. Among all variables, ProstateVolume is the most significant predictor, and its SHAP values are the most spread and Besides these major predictors, a number of clinical history and symptom-related variables also help moderately predict the model. Attributes like SmokingHistoryYes, HypertensionYes, and FamilyHistoryYes have an intermediate SHAP distribution, meaning that such attributes are associated with the predicted probability being raised slightly (when present). Variables associated with the symptoms, such as Erectile Dysfunction, Weak Urine Flow, Pelvic Pain, Difficulty Urinating, and Back Pain, exhibit smaller but significant positive contributions when the condition is reported, indicating that urinary and pelvic symptoms are supportive in risk estimation. There are also varying levels of medical and lifestyle influences, such as PreviousCancerHistoryYes, AlcoholConsumptionModerate, and GeneticRiskFactorsYes, which have the greatest impact among other lifestyle-related variables. Diagnostic indicators provide additional information: DREResultNormal is more likely to yield a negative SHAP value, indicating that a normal digital rectal examination decreases the probability of prostate cancer, whereas EarlyDetectionYes is usually expected to yield lower SHAP values, indicating the likelihood of the advantage of early screening. On the other hand, the best examples are CholesterolLevelNormal, BloodinUrineYes, and RaceAfricanAncestryYes, which have narrow SHAP distributions centered around zero and have little overall effect on the model's predictions. On the whole, the SHAP analysis has shown that the predictive model is mainly propelled by the key clinical biomarkers, specifically, the prostate volume and the PSA level, followed by the demographic pressure of age and BMI, with other contextual factors being the medical history, genetic predisposition, and symptomatic indicators. Such a hierarchical structure of importance makes the model fit within the known clinical context and increases interpretability by showing the contribution of each feature to the prediction results across the entire dataset.



LIME-Based Local Interpretability Analysis



The local interpretability analysis based on LIME is performed by analyzing the results of CIF scenarios.

The LIME (Local interpretable model-agnostic explanations) visualization shows how each feature contributes to the prediction of a particular instance (Instance 42) in the prostate cancer prediction model. Unlike global interpretability models like SHAP, LIME aims to interpret the model's behavior locally, providing an understanding of the impact of various variables on the prediction of a single observation. The predicted likelihood of the instance being malignant in this case is therefore 0.291, which represents a fairly low likelihood of prostate cancer that is predicted. The horizontal bar chart illustrates the most powerful features used to make this prediction. The characteristics exemplified by features in the green color are positively associated with the malignant class, and those exemplified by features in the red color are negatively associated with the malignant class, thereby increasing the prediction probability and moving it towards the benign, respectively. The size of the bar indicates the feature's impact on the final decision. Among the features with positive contributions, the strongest positive effects are observed for PSA_Level, Hypertension_Yes, and Family_History_Yes, indicating that moderate PSA levels, hypertension, and certain family history conditions contribute slightly to the model of malignancy. Moreover, there is a positive contribution to the prediction with regard to the variables of BMI and the variable of Smoking_History_Yes, implying that these two variables have a small impact on predicting the risk in this particular case of the patient.

On the other hand, a few variables negatively affect the prediction and, consequently, reduce the likelihood of malignancy. The most significant negative predictors are Blood in Urine Yes=0 (no blood in urine) and Prostate Volume=0.86, which significantly push the prediction value towards the benign outcome. Other variables, like Pelvic_Pain Yes = 0 or less, Erectile dysfunction Yes = 0 or less, and Age in the lower normalized range, each have negative contributions, indicating that the absence of this symptom, or, rather, younger age, would lower the predicted risk of cancer in this case. These findings underline the importance of LIME in providing a clear description of the model's decision by approximating a complex prediction model locally into a more interpretable one. All in all, the analysis shows that although some clinical indicators exert a slight positive pressure on the presence of malignancy, the absence of critical symptoms and certain patient characteristics has a more significant negative impact, leading to a final forecast in favor of the benign type. This local interpretability makes the predictive model

more transparent, enabling clinicians to better understand how specific patient characteristics affect its diagnostic recommendations.

Table 23. Comparison of Accuracies after using SHAP

Metric	LR	DT	IRF	XGB	NB	MLP
Accuracy	0.85, 0.50, 0.85	0.73, 0.73, 0.73	0.85, 0.78, 0.85	0.85, 0.78, 0.85	0.85, 0.49, 0.85	0.85, 0.54, 0.85
Precision	0.85, 0.85, 0.85	0.85, 0.85, 0.85	0.85, 0.85, 0.85	0.85, 0.85, 0.85	0.85, 0.85, 0.85	0.85, 0.85, 0.85
Sensitivity	1.00, 0.51, 1.00	0.83, 0.82, 0.82	1.00, 0.91, 1.00	1.00, 0.90, 1.00	1.00, 0.49, 1.00	1.00, 0.56, 1.00
Specificity	0.00, 0.50, 0.00	0.17, 0.19, 0.19	0.00, 0.10, 0.00	0.01, 0.09, 0.00	0.00, 0.50, 0.00	0.00, 0.42, 0.00
F1 Score	0.92, 0.63, 0.92	0.84, 0.84, 0.84	0.92, 0.88, 0.92	0.92, 0.87, 0.92	0.92, 0.62, 0.92	0.92, 0.67, 0.92
ROC-AUC	0.50, 0.50, 0.50	0.50, 0.51, 0.51	0.50, 0.50, 0.50	0.50, 0.50, 0.50	0.50, 0.50, 0.50	0.49, 0.49, 0.49
Cohen Kappa	0.00, 0.00, 0.00	0.00, 0.01, 0.01	-0.00, 0.00, -0.00	-0.00, -0.01, -0.00	0.00, -0.00, 0.00	0.00, -0.01, 0.00

Table 24. Comparison of Accuracies after using LIME

Metric	LR	DT	IRF	XGB	NB	MLP
Accuracy	0.85, 0.51, 0.85	0.73, 0.69, 0.69	0.82, 0.70, 0.84	0.85, 0.68, 0.85	0.85, 0.48, 0.85	0.85, 0.51, 0.85
Precision	0.85, 0.85, 0.85	0.85, 0.85, 0.85	0.85, 0.85, 0.85	0.85, 0.85, 0.85	0.85, 0.85, 0.85	0.85, 0.85, 0.85
Sensitivity	1.00, 0.52, 1.00	0.83, 0.77, 0.77	0.95, 0.78, 0.99	1.00, 0.75, 1.00	1.00, 0.48, 1.00	1.00, 0.51, 1.00
Specificity	0.00, 0.46, 0.00	0.17, 0.23, 0.23	0.05, 0.22, 0.01	0.00, 0.25, 0.00	0.00, 0.52, 0.00	0.00, 0.50, 0.00
F1 Score	0.92, 0.64, 0.92	0.84, 0.81, 0.81	0.90, 0.81, 0.91	0.92, 0.80, 0.92	0.92, 0.61, 0.92	0.92, 0.63, 0.92
ROC-AUC	0.48, 0.49, 0.49	0.50, 0.50, 0.50	0.50, 0.49, 0.49	0.50, 0.50, 0.50	0.48, 0.49, 0.49	0.51, 0.51, 0.51
Cohen Kappa	0.00, -0.01, 0.00	0.00, 0.00, 0.00	0.00, -0.00, 0.01	0.00, -0.00, -0.00	0.00, -0.00, 0.00	0.00, 0.01, 0.00

The experimental findings with no explainability method proved that the majority of machine learning models, such as Logistic Regression (LR), Decision Tree (DT), Improved Random Forest (IRF), XGBoost (XGB), Naive Bayes (NB), and Multi-Layer Perceptron (MLP), obtained rather high values of accuracy, generally about 84-85. Nonetheless, further analysis of the performance measures revealed significant discrepancies. Although the sensitivity levels were very high (usually nearly 1.0), the specificity levels were almost zero in some cases, indicating that the models were strongly skewed towards the majority class. Moreover, the ROC-AUC values were consistently near 0.5, and the Cohen Kappa values were near zero, indicating that the models were not better at classification than random chance. Even though SMOTE and threshold optimization marginally improved some measures, such as specificity and class balance, they did not significantly increase the overall model's discriminative power. These problems could not be properly diagnosed without interpretability tools, and high accuracy looked deceptive.

After adding SHAP and LIME, the models' behaviour became much more evident as shown in Table 23 and

Table 24. The SHAP analysis, which provides global interpretability, showed that feature contributions were not sufficiently discriminative between the classes. The models would place close weight on both the positive and negative prediction features, leading to overlapping decoding boundaries. This explains the earlier-observed ROC-AUC values around 0.5 and the models' inability to effectively differentiate between the classes. Furthermore, SHAP showed that most models were highly biased toward the positive class, resulting in high sensitivity and exceptionally low specificity at the outset. When SMOTE was used, SHAP analysis showed that synthetic samples did not introduce new significant patterns but slightly altered the distribution of features.

This was further substantiated at the instance level by the LIME analysis, which dwells on the local interpretability. Some feature patterns that affected individual predictions repeated themselves and had to be used in multiple instances. This shows that there is no variety in the decision-making, and the models are not generalized. In other instances, LIME showed slight improvements in specificity with SMOTE, indicating local changes in decision boundaries. Nevertheless, such improvements were not universal and did not translate into higher global performance, as ROC-AUC values remained near 0.5. Also, LIME reported that noise was sometimes introduced by the addition of synthetic data, leading to unstable local predictions and less accurate predictions in some settings.

Comparing the results achieved with SHAP and LIME, along with their results without them, reveals an essential observation. The models would seem to do well on the basis of accuracy being alone without the use of explainability techniques, covering up the problems of class imbalance, bias, and lack of discriminative power. The combination of SHAP and LIME, in turn, revealed the latent shortcomings of the models, including feature overlap, a lack of strong decision boundaries, and excessive reliance on projections from the majority class. Although SMOTE and threshold optimization achieved only slight gains, explainability techniques show that both approaches fail to address the underlying issue of feature non-separability.

Altogether, SHAP and LIME allowed shifting the assessment of a performance-based analysis to a more profound interpretation of model behavior. It was also made clear that, although there are acceptable levels of accuracy, the models cannot be used for real-world clinical decision-making because they lack the ability to generalize and effectively differentiate between classes. Others, like Improved Random Forest and XGBoost, among all models, demonstrated relatively more stable and better performance, although they still lacked adequate discriminative power. The results highlight the importance of better feature engineering, integration of domain knowledge, and investigation of more advanced or hybrid models to realize clinically useful predictions.

Conclusion

This paper discussed the use of various machine learning models to predict prostate cancer outcomes using a well-rounded dataset of clinical and demographic factors. The tested models were LR, DT, RF, XGB, NB, and MLP. The experiments were performed randomly across three conditions: standard 10-fold cross-validation, class balancing, synthetic minority oversampling technique (SMOTE), and threshold optimization.

The findings underscored that most models without class balancing had high accuracy but low specificity and high discriminatory ability. Specifically, both XGB and RF demonstrated excellent generalization, with high sensitivity and precision even without SMOTE. Nonetheless, the specificity of each model was found to be significantly low, which is typical for imbalanced data.

Using SMOTE successfully reduced class imbalance and improved recall, specificity, and F1 scores across most models. XGB and RF were the most significantly improved models and demonstrated strong, stable performance across different evaluation metrics. It was, however, noted that although SMOTE enhanced sensitivity and specificity, the overall discriminatory power, as measured by ROC-AUC, was low, which supports the difficulty of achieving predictive performance in imbalanced datasets.

Sensitivity and F1 score were also maximized with threshold optimization, especially in LR, RF, and XGB. It, however, also caused a sharp decrease in the specificity of these models. This sensitivity-specificity trade-off underscores the importance of selecting relevant evaluation metrics and tuning thresholds to meet clinical requirements, particularly in a medical facility where false positives or negatives can have critical implications.

Although the XGB showed rather competitive results across all conditions, achieving high mean F1 scores and ROC-AUC values, its capacity to identify the minority group (prostate cancer) remained lower than that of other classifiers, such as LR. It indicates that although machine learning has advanced, further

improvements in feature selection, data preprocessing, and model interpretability are needed to enhance the clinical applicability of these models.

Further studies are needed to enhance the model interpretability, which could be achieved through approaches such as SHAP (Shapley Additive Explanations), as demonstrated, which can offer insights into feature significance. In addition, the use of ensemble techniques, feature engineering, and advanced sampling methods could yield more robust models that generalize better across diverse clinical environments.

To summarize, the current research is insightful in using machine learning models to predict prostate cancer, and it is possible to note that algorithms such as XGB and RF can be used to enhance cancer detection, especially in cases of class imbalance. Nonetheless, the results also point to the need for further development of model optimization and clinical integration to fully realize the potential of such methods in medical diagnostics.

Data availability: We have dealt with one dataset for this article. The dataset underlying this study is openly available in kaggle website. These data were derived from sources in the public domain. Link of this datasets are given below.

https://www.kaggle.com/code/abdulwahidrukua/prostate-cancer-prediction

Competing Interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Information: Not Applicable

Ethical statement

Though we applied the dataset from a publicly available repository, we realized that there were ethical concerns for potential biases in datasets and privacy in healthcare research. The datasets used herein follow the applicable rules of ethics, such as data anonymization and de-identification, and we take all reasonable measures to mitigate biases in our analysis.

References

American Cancer Society. (2022). Cancer statistics, 2022.

Chaudhuri, A. K., Das, S., & Ray, A. (2023). A hybrid feature selection and stacked generalization model to detect breast cancer. In Data-centric AI solutions and emerging technologies in the healthcare ecosystem (pp. 165-183). CRC Press.

Chaudhuri, A. K., Sinha, D., & Thyagaraj, K. S. (2018). Identification of the recurrence of breast cancer by discriminant analysis. In Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2018, Volume 2 (pp. 519-532). Singapore: Springer Singapore.

Chen, S., Jian, T., Chi, C., Liang, Y., Liang, X., Yu, Y., … Lu, J. (2022). Machine learning–based models enhance the prediction of prostate cancer. Frontiers in Oncology, 12, 941349. https://doi.org/10.3389/fonc.2022.941349

Das, S., Chaudhuri, A. K., & Ghosh, P. (2026). Medical Diagnosis Through Improved Feature Selection and Advanced Ensemble Techniques: A Study on Breast Cancer and Chronic Kidney Disease. SN Computer Science, 7(3), 222

Das, S., Chaudhuri, A. K., Das, S., & Ghosh, P. (2025). Multistage feature selection and stacked generalization model for cancer detection. Scientific Reports, 15(1), 38124.

Das, S., Chaudhuri, A. K., Ghosh, P., & Raymahapatra, P. Recent Advancements in Cancer Diagnosis Using Machine Learning Techniques: A Systematic Review of Decades of Research, Comparisons and Problems.

Das, S., Chaudhuri, A. K., Paul, D., & Ghosh, P. (2026). Sequential Attribute Designator (SAD): A Novel Feature-Selection Framework for Pulmonary Disease Research. In Next-Generation Bioinformatics for Pulmonary Disease Research (pp. 413-436). IGI Global Scientific Publishing.

GLOBOCAN. (2022). Estimated cancer incidence, mortality, and prevalence worldwide.

Horasan, A., & Güneş, A. (2024). Advancing prostate cancer diagnosis: A deep learning approach for enhanced detection in MRI images. Diagnostics, 14(17), 1871. https://doi.org/10.3390/diagnostics14171871

Indian Journal of Urology. (n.d.). Descriptive epidemiology of prostate cancer in India. Indian Journal of Urology.

Li, H., Lee, C. H., Chia, D., Lin, Z., Huang, W., & Tan, C. H. (2022). Machine learning in prostate MRI for prostate cancer: Current status and future opportunities. Diagnostics, 12(2), 289. https://doi.org/10.3390/diagnostics12020289

Saha, S., Vignarajan, J., Flesch, A., Jelinko, P., Gorog, P., Szep, E., … Frost, S. (2024). An artificial intelligent system for prostate cancer diagnosis in whole slide images. Journal of Medical Systems, 48(1), 101. https://doi.org/10.1007/s10916-024-101

Sherafatmandjoo, H., Safaei, A. A., Ghaderi, F., & Allameh, F. (2024). Prostate cancer diagnosis based on multi-parametric MRI, clinical and pathological factors using deep learning. Scientific Reports, 14(1), 14951. https://doi.org/10.1038/s41598-024-14951-0

Sun, Y. K., Zhou, B. Y., Miao, Y., Shi, Y. L., Xu, S. H., Wu, D. M., … Xu, H. X. (2023). Three-

dimensional convolutional neural network model to identify clinically significant prostate cancer in transrectal ultrasound videos: A prospective, multi-institutional diagnostic study. [Journal Name], [Volume(Issue)], [Page numbers].

Wang, X., Zhang, X., Li, H., et al. (2023). Application of machine learning algorithm in prediction of lymph node metastasis in patients with intermediate and high-risk prostate cancer. Journal of Cancer Research and Clinical Oncology, 149, 8759–8768. https://doi.org/10.1007/s00432-023-04816-w

World Cancer Research Fund. (n.d.). Global cancer data.

World Health Organization. (n.d.). Global cancer burden growing.

Zhao, Y., Zhang, L., Zhang, S., Li, J., Shi, K., Yao, D., ... & Wan, J. (2025). Machine learning-based MRI imaging for prostate cancer diagnosis: systematic review and meta-analysis. Prostate Cancer and Prostatic Diseases, 1-8.