American Journal of Mathematical and Computer Modelling
2026; 11(1): 30-38
http://www.sciencepublishinggroup.com/j/ajmcm
doi: 10.11648/j.ajmcm.20261101.13

Classification of Prostate Tumors for Effective Diagnosis and Treatment

Kofuna Alfayo Odongo *, Charity Wamwea, Susan Mwelu

Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

Email address:

kofuna71@gmail.com(Kofuna Alfayo Odongo), cwamwea@jkuat.ac.ke(Charity Wamwea), smwelu@jkuat.ac.ke(Susan Mwelu)
*Corresponding author

To cite this article:

Kofuna Alfayo Odongo, Charity Wamwea, Susan Mwelu. (2026). Classification of Prostate Tumors for Effective Diagnosis and Treatment. *American Journal of Mathematical and Computer Modelling, 11*(1), 30-38. https://doi.org/10.11648/j.ajmcm.20261101.13

Received: 26 October 2025; **Accepted:** 08 November 2025 ; **Published:** 4 February 2026

Abstract: Prostate tumors are tumors that grow in or around the prostate gland of the male reproductive system. these tumors are classified into two, that is: benign and malignant. Malignant tissues are cancerous and more dangerous as they can easily spread to other soft tissues that are easily attacked by cancer cells, such as the lungs and liver. In contrast, benign tumors are not as dangerous and can be surgically removed and never regrow. The cancerous tumors has a high mortality rate, according to GLOBOCAN 2020, there were estimated 19.3 million cases with 10 million registered deaths over the same year. Current diagnosis of prostate tumor still lacks maximum precision, with the PSA and invasive biopsy methods suffering from most false positives and negatives. The study suggested a ML algorithm blending technique from three base models, SVM, RF, and XGBoost. For maximum accuracy and overfitting, the study performed SMOTE for class balancing and correlation elimination techniques. The blended model outperformed the base model with an accuracy of 0.981 at 0.05 confidence level, followed by RF and XGboost at 0.971. In other performance metrics, blended model had the highest sensitivity and specificity of 0.979 and 0.981, respectively. XGBoost and RF shared almost the same sensitivity of 0.97 and a higher specificity of 0.98. SVM had an overall low performance and was not recommended for such a task as a standalone model. The study recommends the incorporation of a blended model over each performance, although Random Forest and XGBoost are still viable for application.

Keywords: SVM, SMOTE, RF, GLOBOCAN, PSA, RF, XGBOOST

1. Introduction

Prostate tumors are tumors that occur in the prostate glands of a male human being. These tumors are of two types, that is, malignant and benign. Malignant tumors are cancerous and more dangerous as they can easily spread to other soft tissues that are easily attacked by cancer cells, such as the liver and the lungs. [1] Benign tumors, on the other hand, are not as dangerous as malignant tumors and hence can be handled by early surgical procedures, after which they cannot regrow as malignant tumors. According to GLOBOCAN 2022, prostate cancer is the second most dangerous of all cancers in men in the event that a person test positive for malignant.There were 34 million recorded new cases with 17 million reported deaths by prostate cancer over the same year. [2] Diagnosis and treatment of these tumors are therefore a matter of health concern because of the nature and harm they can cause the body. The challenge, therefore, lies in correctly diagnosing these tumors as either malignant or benign, as they tend to look alike and possess similar signs and symptoms.Most cases are recorded at old age of a patient, with majority surpassing 50 years limit. Sometimes patients do recounts having related problems at early stages of lifetime moments, probably being misdiagnosed. This is where the effects of false negative, where patients are treated for benign instead of malignant, and false positives, where patients goes for unnecessary biopsies when in reality an invasive method and/or radiology, therapy sessions were not really needed. Current diagnoses are still not accurate in identifying the type of tumor a patient may be dealing with. Current Diagnosis of prostate tumors involves clinical assessment, laboratory testing, imaging, and histopathology. The digital rectal examination (DRE) is often the first step, allowing clinicians to check the prostate for irregularities, but

its sensitivity is low, and results depend on the examiner's experience. The prostate-specific antigen (PSA) blood test detects elevated protein levels associated with malignancy, yet it lacks specificity, as levels can also rise in benign prostatic hyperplasia or inflammation. Trans-rectal ultrasound (TRUS) provides imaging for volume estimation and biopsy guidance, but cannot reliably differentiate benign from malignant tissue. The prostate biopsy, particularly MRI-targeted biopsy, remains the gold standard for confirming diagnosis, although it is invasive and subject to sampling errors. [2] Multiparametric MRI (mpMRI) enhances detection by combining anatomical and functional imaging sequences, improving accuracy in identifying clinically significant tumors. However, mpMRI is costly, expertise-dependent, and less accessible in low-resource settings. [11] Advanced imaging, such as PSMA-PET/CT, offers superior sensitivity for staging and recurrence detection but is similarly expensive and not widely available. This study aimed to address the limitations of current prostate tumor diagnosis methods by developing a blended machine learning approach that improved diagnostic accuracy and reliability. Current techniques often suffer from low specificity, invasive-ness, and sampling errors, leading to misclassifications or overdiagnosis. To overcome these challenges, we employed an ensemble of three algorithms, Random Forest, Support Vector Machine, and Extreme Gradient Boosting, combined through a blending technique. Each base model captured unique patterns within the dataset. RF handled nonlinear interactions, SVM optimized classification boundaries, and XGBoost enhanced predictive performance through gradient boosting. Their blended outputs were integrated into a logistic regression meta-classifier, which learned the optimal combination of predictions to produce a more accurate final classification of prostate tumors. The study addressed class imbalance, a common issue where one class is more dominant over the other. We applied the Synthetic Minority Oversampling Technique to generate synthetic samples of the minority class to ensure a balanced learning and prevent model bias toward the dominant class. The proposed framework aims to improve and enhance sensitivity, specificity, and overall diagnostic precision in prostate tumor detection. This would help clinicians and experts solve or easily and correctly diagnose new patients, an improvement to currently used procedures.

2. Methodology

2.1. Class Balancing using SMOTE

When modeling and doing analysis using machine learning algorithms, an imbalanced dataset may increase the accuracy of the most dominant class, which may cause false predictions as the model may tend to classify more of the dominant class. Synthetic Minority Oversampling Technique (SMOTE) creates synthetic data points between existing minority class samples. [4]

Letting x_i and x_j be two samples from the minority class. A new synthetic data point x_n is generated as,

$$x_n = x_i + \lambda(x_j - x_i) \tag{1}$$

Where,
x_i and x_j are two randomly selected minority class samples such that $x_i \neq x_j$.
$\lambda \in [0, 1]$ is a random number ensuring variability in the synthetic sample.

Stepwise SMOTE Algorithm

Select a minority class instance x_i.
Find its k nearest neighbors within the minority class.
Randomly select one neighbor x_j.
Generate a new sample using Equation (1).
Repeat the process until the desired balance is achieved.

This study employed SMOTE for balancing tumor classes, pervious studies and ensembles mostly focused on learning data with significantly high class different, approaches that the majority call more attention for example in random forest an unbalanced data causes a higher out-of-bag errors, less support vectors in SVM etc. This approach gave all classes equally chances, Which enables improved and more accurate analysis.

2.2. Normalization (Feature Scaling)

Normalization ensures all features contribute equally to the model. The two common methods are the min-max scaling and standardization

2.2.1. Min-Max Scaling

Min-Max normalization transform a feature X to a fixed range of [0, 1]

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{2}$$

where,
-X is the original feature value,
-$X_{\min}$ and $X_{\max}$ are the minimum and maximum values,
-X_{scaled} is the normalized value.

2.2.2. Standardization (Z-score Normalization)

Standardization is a statistical technique used in data preprocessing to make different variables more comparable. It's like translating all these different data "languages" into one universal dialect. It is achieved by the following mathematical procedure.

$$Z = \frac{X - \mu}{\sigma} \tag{3}$$

where,
- Z is the standardized value (z-score).
- X is the original value.
- μ is the mean of the dataset.
- σ is the standard deviation of the dataset.

2.3. Feature Elimination (Correlation)

The study focuses on seven predictor variables, which are numeric, and one response variable, which is ranked by class. We aim to focus only on features that efficiently contribute to the classification and prediction of prostate tumors for cancer classification. The study suggests Pearson's correlation technique to eliminate one predictor variable that is highly correlated with the others, to remove redundancy.

Correlation (Pearson's r)

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \cdot \sum(y_i - \bar{y})^2}} \tag{4}$$

Variable Elimination

From each highly correlated pair, surpassing a threshold of 0.9, one variable is removed. The choice of which variable to keep depends on domain knowledge, interpretability, predictive usefulness, and relationship with other variables. The aim is to reduce redundancy among predictors, avoid multicollinearity, and ensure that the model remains interpretable and stable.

2.4. Non Parametric Models

A general statistical non-parametric model takes the form of;

$$Y = M(X_i) + \epsilon_i \; for \; i = 1, 2, 3, 4, ...n \tag{5}$$

Where,

i) Y is the response variable representing diagnosis class, which are Benign and Malignant for our case.

ii) $X_i = x_1, x_2, x_3, ...x_7$ are the predictor variables, including Area, Fractional dimension, etc, which are all in continuous form.

iii) ϵ_i is the error term and has the following assumptions.

1.$E[\epsilon] = 0$, i.e., the mean of the error should be zero.

2.$Var[\epsilon] = \sigma_\epsilon^2$,i.e., variance of error function is σ^2.

3. There is zero interaction between the error functions and the predictor variables.

iv) $M(X)$ is the mean function, which can also be expressed as $E(Y/X)$. Estimation of the mean function can use different models such as random forest, gradient boosting,support vectors etc.

2.4.1. Support Vector Machine

The Support Vector Machine is a supervised learning algorithm that classifies binary data by constructing an optimal separating hyperplane in a transformed feature space. [6] For a binary classification problem with training data

$$\{(x_i, y_i)\}_{i=1}^n, \quad x_i \in \mathbb{R}^m, \; y_i \in \{-1, +1\}, \tag{6}$$

SVM aims to solve the following optimization problem:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^n \xi_i, \tag{7}$$

subject to

$$y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \tag{8}$$

where C controls the penalty for misclassification and overfitting, and $\phi(x_i)$ represents a nonlinear mapping that transforms the data into a higher-dimensional feature space.

To efficiently perform this transformation without explicitly computing $\phi(x)$, the Radial Basis Function (RBF) kernel is employed,

$$K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2), \tag{9}$$

where $\gamma > 0$ determines the influence of individual training samples. A smaller γ results in smoother decision boundaries, whereas a larger γ allows more complex separations. The RBF kernel enables nonlinear decision boundaries by implicitly mapping data into an infinite-dimensional space.

The resulting SVM decision function for a new observation x is expressed as,

$$f(x) = \text{sign}\left(\sum_{i=1}^n \alpha_i y_i K(x_i, x) + b\right), \tag{10}$$

where α_i are the Lagrange multipliers obtained during model training.

The posterior class probabilities, which are essential for our next step in blending, are estimated through logistic calibration using *Platte scaling* [8]:

$$P(y = 1|x) = \frac{1}{1 + \exp(Af(x) + B)}, \tag{11}$$

where A and B are calibration parameters fitted from validation data. These probability estimates, $P(y = 1|x)$, serve as soft outputs of the SVM and are subsequently used as input features for the blended meta-classifier in the blending framework, thereby improving the overall diagnostic accuracy of the new model.

2.4.2. Random Forest

The Random Forest algorithm is an ensemble-based machine learning method that constructs multiple decision trees and aggregates their predictions to improve classification accuracy and reduce overfitting. [3] It operates on the principles of bootstrap aggregation (bagging) and random feature selection, ensuring that each tree in the forest is trained on a random subset of both data samples and features.

Given,

$$D = \{(x_i, y_i)\}_{i=1}^n, \quad x_i \in \mathbb{R}^m, \; y_i \in \{0, 1\}, \tag{12}$$

Random Forest generates T decision trees $h_t(x)$, each trained on a bootstrap sample drawn with replacement from D. At every internal node, the algorithm randomly selects a subset of features and determines the best splitting point that minimizes impurity. For a given feature A with possible split point s, the data is partitioned into left (D_L) and right (D_R) subsets as:

$$D_L = \{(x_i, y_i)|A(x_i) \leq s\}, \quad D_R = \{(x_i, y_i)|A(x_i) > s\}. \tag{13}$$

The impurity at each node can be measured using either the Gini index or entropy,

Gini Impurity

$$G(D) = 1 - \sum_{k=1}^{K} p_k^2, \tag{14}$$

where p_k is the proportion of samples belonging to class k in node D.

Entropy

$$H(D) = -\sum_{k=1}^{K} p_k \log_2(p_k), \tag{15}$$

which quantifies the uncertainty of class distribution in node D.

The optimal split is selected by minimizing the weighted impurity after the split:

$$\text{Split Gain} = I(D) - \left(\frac{|D_L|}{|D|} I(D_L) + \frac{|D_R|}{|D|} I(D_R) \right), \tag{16}$$

where $I(D)$ represents either $G(D)$ or $H(D)$.

Once all trees are trained, the final classification for an unseen instance x is obtained through majority voting:

$$\hat{y} = \text{majority\_vote}\{h_t(x)\}_{t=1}^{T}. \tag{17}$$

For blending, the class probability of the Random Forest model is calculated as the proportion of trees predicting the positive class,

$$P(y = 1|x) = \frac{1}{T} \sum_{t=1}^{T} I(h_t(x) = 1), \tag{18}$$

where $I(\cdot)$ is an indicator function equal to 1 if tree t classifies x as malignant, and 0 as benign.

2.4.3. Extreme Gradient Boosting

Extreme Gradient Boosting is a powerful and efficient implementation of gradient boosting decision trees that optimizes predictive accuracy through sequential model improvement, i.e, the next tree corrects the errors made from the previous trees. [5, 12] Unlike Random Forest, which builds trees independently, XGBoost constructs trees sequentially, where each new tree is trained to correct the residual errors of the previous ensemble. This process minimizes a regularized objective function that balances model fit and complexity.

Given,

$$D = \{(x_i, y_i)\}_{i=1}^{n}, \quad x_i \in \mathbb{R}^m, \; y_i \in \{0, 1\}, \tag{19}$$

the model predicts $\hat{y}_i$ as the sum of outputs from K trees,

$$\hat{y}_i = \sum_{t=1}^{K} f_t(x_i), \quad f_t \in \mathcal{F}, \tag{20}$$

where $\mathcal{F}$ is the space of regression trees. The objective function at iteration t is given by;

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t), \tag{21}$$

where $l(\cdot)$ is a differentiable loss function (e.g., logistic loss) and $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \|w\|^2$ is the regularization term controlling tree complexity. The model uses first, and second-order gradients, Gradient and Hessian respectively, to optimize the objective, improving both convergence and stability. For classification, XGBoost applies a logistic transformation to convert the additive tree outputs into probabilities.

$$P(y = 1|x) = \frac{1}{1 + \exp(-\hat{y}_i)}. \tag{22}$$

These probabilistic outputs are kept in a matrix form for blending, as they provide calibrated confidence estimates that can be combined with predictions from other base models in the meta-classifier layer. The inclusion of regularization, shrinkage, and column sub sampling makes XGBoost highly resistant to overfitting and effective for complex, high-dimensional medical datasets such as prostate cancer features.

2.4.4. Blending Methodology

Blending is an ensemble learning technique that enhances predictive accuracy by combining the strengths of multiple base models. [9, 10] In this study, three classifiers—Random Forest, Support Vector Machine, and Extreme Gradient Boosting- were used as base learners. Each model was first trained independently on the training dataset D_{train} and subsequently used to generate class probability predictions on a separate validation dataset D_{val}. These probability outputs formed the input features for a higher-level model known as the meta-classifier.

Let the dataset be represented as $D = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^m$ denotes the input features and $y_i \in \{0, 1\}$ represents the class labels (benign or malignant). After training, each base model produces a probability estimate for the positive class.

Letting,
$P_{RF}(y = 1|x_i) = K$,
$P_{SVM}(y = 1|x_i) = M$,
$P_{XGB}(y = 1|x_i) = N$.

These probabilities are then combined to form a new meta-dataset, where each observation is represented by the three predicted probabilities corresponding to the same instance in the validation set. The meta-classifier employed in this study is a logistic regression model, which learns to optimally combine the predictions from the base models to produce a final blended output. The logistic regression model is defined as;

$$z_i = w_0 + w_1 K + w_2 M + w_3 N, \tag{23}$$

where w_0 is the intercept, and w_1, w_2, w_3 represent the coefficients that determine the contribution of each base model.

The resulting linear combination z_i is then passed through a sigmoid activation function to generate the final predicted probability for the malignant class:

$$P(y = 1|x_i) = \frac{1}{1 + e^{-z_i}}. \tag{24}$$

The final class prediction is made based on a decision threshold τ (typically $\tau = 0.5$):

$$\hat{y}_i = \begin{cases} 1, & \text{if } P(y = 1|x_i) \geq 0.5, \\ 0, & \text{otherwise.} \end{cases} \tag{25}$$

Blending framework leverages the complementary strengths of the three base models. The Random Forest provides robustness against overfitting, the SVM ensures optimal decision boundaries, and XGBoost captures complex nonlinear relationships. By integrating their probabilistic predictions through logistic regression, the blended model reduces both bias and variance, leading to improved classification performance and more reliable prostate tumor diagnosis.

2.5. Performance Metrics

The performance of the blended prostate tumor classification model was evaluated using several standard metrics derived from the confusion matrix, including accuracy, precision, recall (sensitivity), specificity, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC). These measures provide a comprehensive view of the model's diagnostic performance.

Accuracy

Accuracy measures the overall proportion of correctly classified instances and is expressed as;

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. While accuracy provides an overall indication of model performance, it may not fully reflect effectiveness when the dataset is imbalanced.

Precision

Precision quantifies the proportion of positive predictions that are truly positive and is calculated as;

$$\text{Precision} = \frac{TP}{TP + FP}$$

It indicates the reliability of positive predictions and helps reduce false alarms that could lead to unnecessary medical procedures.

Recall (Sensitivity)

Sensitivity, or recall, measures the model's ability to correctly identify actual positive cases.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

A high sensitivity indicates that the model effectively detects true cancer cases, minimizing the likelihood of missed diagnoses.

Specificity

Specificity evaluates the model's capacity to correctly identify the negative cases and is defined as;

$$\text{Specificity} = \frac{TN}{TN + FP}$$

It complements sensitivity by assessing how well the model avoids false positives, thereby preventing overdiagnosis.

F1-Score

The F1-score represents the harmonic mean of precision and sensitivity, balancing both metrics in a single measure:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$$

It is particularly valuable when dealing with imbalanced class distributions, offering a balanced performance measure.

2.6. Data Description

The study used a secondary dataset from online data source for academic purposes, a source also trusted by clinical officers all over the world. Ethical consideration was not necessary for this study as no patients' personal information was needed or used. NCD dataset for prostate cancer tumors for December 2024 from KAGGLE was used in general methodology development, with original data having a sample size of 600, malignant cases were 381 and 219 benign cases. data consisted of 9 variables, eight predictors which included fractional dimension, radius symmetry e.t.c, and diagnosis result as our response variable.

3. Results

3.1. Data Balancing

The original data had nine variables of study, one response, and eight predictor variables, with a sample size of 600 responses for prostate tumors. Malignant cases were dominant over benign cases. We applied a Synthetic Oversampling Technique to the dataset, enabling us to achieve a 50-50 chance on both classes, as shown in Figure 1 below.

Figure 1. *Balanced class distributions of Benign and Malignancy after performing SMOTE.*

3.2. Variable Reduction

There are so many steps in performing feature selection or variable reduction. These include the supervised principal component analysis, Lasso, and Ridge, etc, but the study opted for a correlation elimination technique because of data structure and feature importance. This was seen to be more appropriate because of the type of dataset we were dealing with. Figure 2 is a heatmap of correlations between the predictor variable. There was redundancy between two of our variables, that is, area and perimeter. We therefore had two redundant variables, that is, area and perimeter.

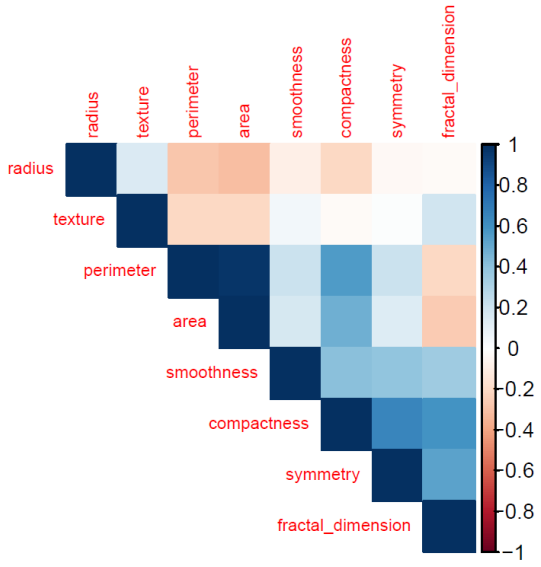

Figure 2. *Correlation heatmap showing rate of correlation between predictor variables in prostate tumor classification.*

A redundant variable is one that does not add new information because it is highly correlated (or strongly dependent) on another variable already in the dataset, as seen with area and perimeter in our dataset. It had a positive correlation of more than 0.98. Through the correlation elimination method, we opted to keep perimeter and eliminate area because of consistency with other variables, because area could still be represented with radius, one of our variables in the dataset, since area is just a constant multiplied by a radius squared.

3.3. Fitted Models

3.3.1. Support Vector Machine

The SVM model was trained under the C-classification approach with a radial basis function (RBF) kernel. Our cost parameter was set to 1, which balanced the margin width and classification errors, avoiding too much strictness. We used 146 support vectors because the margin of separation between malignant and benign cases was so complex and not easily separable with a few points.
The model achieved an accuracy of 88%, at a confidence level of 95% with an interval ranging between 82.2% to 92.%. The model exhibited a sensitivity of 90% and a specificity of 83.5%, showing the rate at which the positive and the negative cases were correctly classified by the model. The model does not look perfect but still could be viable for some none clinical applications or further tuning or blending for improved performance.

3.3.2. Random Forest

The Random Forest model was utilized on a dataset to categorize observations according to multiple attributes, employing 500 decision trees. The model showed that out-of-bag error flattened immediately after 200 trees. The model stayed with the 500 decision trees, with 3 variables selected at each split, also a standard one in random forests, good at handling overfitting. The Out-of-Bag (OOB) error in our Random Forest output was 2.25%. The out-of-bag (OOB) error rate, a dependable performance metric for Random Forests, signifies that the model is well-calibrated, avoiding both underfitting and overfitting. This makes OOB a reliable and unbiased measure without needing an explicit validation set.
The model achieved an overall accuracy of 97.1% at a 95% confidence level with an interval of 95.1% to 99.7%, showing a strong reliability on its prediction. With a sensitivity and specificity of 98% and 96..7% respectively, this is an indication of good classification for both classes with minimal errors.

3.3.3. Extreme Gradient Boosting

The training process of our XGBoost model ran for almost 200 boosting rounds (iterations). From our evaluation log, the model's training AUC started high, 0.977 at iteration 1, and improved rapidly, reaching 1.000 by iteration 199, meaning it achieved a perfect discrimination in the training set. Also, the validation training AUC started at 0.934 and plateaued at 0.998 by iteration 200 as shown in Table 1, indicating a well generalization on the unseen data.
The model achieved an overall accuracy of 97.1% at a confidence level of 95% with an interval of 93.5% to 99.1%. The predictions were highly reliable and significant.

Table 1. *Training and Validation log loss across all Iterations.*

Iterations	Training log	Val log
1	0.9769304	0.9340000
2	0.9932345	0.9612667
—	—	—
199	1.0000000	0.9977333
200	1.0000000	0.9977333

3.3.4. Blended Logistic Classifier

After successfully fitting and evaluating the performance of our three base models, we fitted a logistic meta classifier from the probability predictions from validation sets to train into a final model. The model achieved an accuracy of 98% at a 95% confidence level with an interval of 95% to 99%. With a sensitivity and specificity of 97.9% and 98.1% respectively, the model recorded a reliable and stronger performance than

individual machine learning models.

4. Performance Evaluation and Discussion

Random forest exhibited a recommendable accuracy of 0.97 and the highest sensitivity. this signified a robust performance in the classification problem involving prostate tumors in cancer determination. This model has again proved that it can handle such complex datasets with complicated features as determinants due to its ensemble techniques that optimize accuracy. The extreme gradient boosting technique is also another method that showed a good performance in all metrics of evaluation. Its boosting techniques made it superior in all spaces of classification. Errors were optimally minimized, and as per the literature, it is yet again one of the best methods for classification and maybe a prediction problem. It shared almost the same accuracy as random forest, just being slightly higher, standing at a whopping 0.971. Although random forest was slightly higher in recall, XGBoost was best in classifying correctly the malignant cases.

Table 2. *Summary statistics.*

Model	Accuracy	95% CI
SVM	0.880	(0.822,0.924)
Random Forest	0.970	(0.951,0.997)
XGBoost	0.971	(0.935,0.991)
Blended Model	0.982	(0.950,0.999)

Support Vector Machine had an accuracy way lower than the first two base models, with 0.88, and a significantly lower recall and specificity. Though with lower metrics, SVM is still viable for medical classification. Finally, our main objective of the study was the blended model. It exhibited an extremely marvelous performance across all metrics of evaluation. With an accuracy of 0.98, this model is undoubtedly the best for this type of dataset and problem. This is due to the existing individual performance of our base models, which have individual sets of advantages. The model worked well by combining all these advantages into building a superior output for this problem. Table 2 gives a clearer summary on the performance metrics. The blended model is therefore more accurate and less complicated in computing for the case of tumor classification, a determinant in prostate cancer classification.

Table 3. *Evaluation Metrics.*

Model	Model	Specificity	F1-Score	ROC	Precision
SVM	0.900	0.853	0.88	0.975	0.88
Random Forest	0.980	0.967	0.976	0.998	0.98
XGBost	0.970	0.973	0.96	0.997	0.96
Blended Model	0.980	0.981	0.98	0.998	0.99

Table 3 presents the comparative performance of four machine learning models—Support Vector Machine, Random Forest, XGBoost, and the Blended Model. The results indicate that all models performed exceptionally well, with area under the ROC curve (AUC) values exceeding 0.97, reflecting strong discriminative power in classifying prostate tumor cases. Among the individual models, Random Forest and XGBoost demonstrated superior accuracy, achieving sensitivities and specificity above 0.96, which highlights their reliability in identifying both malignant and benign cases.

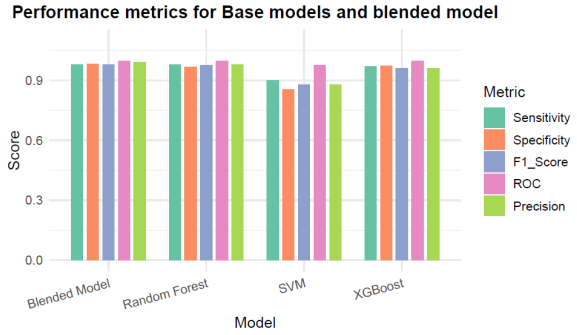

Figure 3. *Performance metrics showing the strengths of each of the three fitted ML algorithms and the Blended model.*

In contrast, the SVM model showed relatively lower sensitivity (0.90) and specificity (0.85), suggesting higher misclassification tendencies. The Blended Model, which integrates the predictions of the three base learners, achieved the highest overall performance with sensitivity = 0.98, specificity = 0.981, precision = 0.99, and F1-score = 0.98. These results confirm that blending effectively captures the strengths of the individual classifiers, yielding a more balanced and robust predictive model. The improved sensitivity ensures that fewer positive cases are missed, while high specificity minimizes false negatives, both critical in clinical decision support. Overall, the blended model provides the most reliable diagnostic performance, demonstrating the advantage of ensemble learning in prostate tumor classification. Figure 3 is a grouped graphical representation of the model's performance across metrics of study.

5. Conclusion

Through our study, we can conclude and recommend that the three base models can be used in prostate cancer tumor classification, although the support vector machine has a lower accuracy, it is still viable for this type of dataset. The blended model, which combines the power and strengths of our base models, is the optimal model based on its superior performance in accuracy, sensitivity, specificity, precision, and F1-score. The ensemble technique it employs renders its resilience relationships, which enhances overall performance.Also by balancing our dataset, the model achieved a strong and more accurate performance that can be trusted. A web base deployment incorporated during screening

process that is easy to interpret would be appropriate for less complicated and more accurate clinical diagnosis and interpretation. This study concludes that the blended logistic model is the optimal solution for this particular dataset and the set of problems associated with prostate tumor classification.

ORCID

0009-0009-9627-8990 (Kofuna Alfayo Odongo)
0000-0002-0099-5832 (Charity Wamwea)
0009-0005-9570-9112 (Susan Mwelu)

Abbreviations

SMOTE	Synthetic Minority Over-sampling Technique
PSA	Protein Specific Antigen
GLOBOCAN	Global Cancer Observatory
SVM	Support Vector Machine
RF	Random Forest
XGBOOST	Extreme Gradient Boosting
AUC	Area Under the Curve
ROC	Receiver Operating Characteristic

Acknowledgments

My deepest appreciation to all who supported me and got fully or partially involved in the success of this research. Special thanks to my advisors and mentors for the brainstorming sessions and constructive criticism during this study. I also appreciate the support and collaborations from organizations and stakeholders for their unwavering help. Special thanks to family, friends and classmates for support and encouragement throughout this entire study. Special one to to the Almighty and Al knowing for keeping this process for keeping this study steady despite the passing of Rt. Hon. Raila Amolo Odinga, may he rest peacefully.

Author Contributions

Kofuna Alfayo Odongo: Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing-Original draft , Writing-review & editing.

Conflicts of Interest

The authors declare that there are no conflicts of interest related to this study.

References

[1] Barrett, J. E., & Moore, S. E. (2019). Multiparametric MRI in prostate cancer diagnosis and management. Clinical Radiology, 74(11), 841–849. https://doi.org/10.1016/j.crad.2019.06.002

[2] Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 74(3), 229-263.

[3] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[4] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

[5] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

[6] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018

[7] Cruz, J. A., & Wishart, D. S. (2007). Applications of machine learning in cancer prediction and prognosis. Cancer Informatics, 2, 59–77. https://doi.org/10.1177/117693510700200031

[8] Raschka, S. (2018). Model stacking and blending: A practical guide. arXiv preprint arXiv: 1810.01806. https://doi.org/10.48550/arXiv.1810.01806

[9] Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

[10] World Health Organization. (2020). GLOBOCAN 2020: Global Cancer Observatory. International Agency for Research on Cancer (IARC). Retrieved from https://gco.iarc.fr/

[11] Qi, X., Wang, K., Feng, B., Sun, X., Yang, J., Hu, Z., Zhang, M., Lv, C., Jin, L., Zhou, L., et al. (2023). Comparison of machine learning models based on multi-parametric magnetic resonance imaging and ultrasound videos for the prediction of prostate cancer. Frontiers in Oncology, 13: 1157949.

[12] Singh, D., Kumar, V., Das, C. J., Singh, A., and Mehndiratta, A. (2022). Machine learning-based analysis of a semi-automated pi-rads v2. 1 scoring for prostate can- cer. Frontiers in Oncology, 12: 961985.

[13] Tan, Y. G., Pang, L., Khalid, F., Poon, R., Huang, H. H., Chen, K., Tay, K. J., Lau, W., Cheng, C., Ho, H., et al. (2020). Local and systemic morbidities of de novo metastatic prostate cancer in singapore: insight from 685 consecutive patients from a large prospective uro-oncology registry. BMJ open, 10(2): e034331.