© MAR 2025 | IRE Journals | Volume 8 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV8I9-1714356

Machine Learning-Based Prediction of Elevated Prostate-Specific Antigen Levels from Lifestyle and Demographic Data

ONITCHA NYERHOVWO EDAFETANURE
University of Port Harcourt, Nigeria

Abstract—

Background: Machine learning offers promising approaches for medical prediction tasks. This study evaluates the comparative performance of three ML algorithms: Logistic Regression with L1 regularization, Support Vector Machine (SVM), and Random Forest in predicting elevated prostate-specific antigen (PSA) levels using lifestyle and demographic features.

Objectives: To compare the predictive performance, generalization capability, and stability of multiple ML models for detecting elevated PSA levels via binary classification of PSA status.

Methods: We implemented three ML algorithms with two feature selection approaches to address the events-per-variable (EPV) problem. Lifestyle and demographic data were collected from adult males in Etsako West LGA, Edo State, Nigeria. Models were trained on 70% of the data (n=69) and validated on 30% (n=30). Performance was assessed using accuracy, precision, recall, specificity, F1-score, and ROC-AUC. Five-fold stratified cross-validation evaluated model stability. Hyperparameter optimization was performed using GridSearchCV.

Results: The Random Forest model achieved the most balanced performance with 73.3% accuracy, 70.6% precision, 80.0% recall, 63.6% specificity, 0.750 F1-score, and 0.764 ROC-AUC. SVM showed identical test set performance (73.3% accuracy, 0.764 ROC-AUC). Logistic Regression with L1 regularization and 11 features achieved the highest recall (100%) but at the cost of zero specificity, indicating overfitting. Cross-validation revealed model stability: Random Forest CV recall 0.833 ± 0.061, CV F1 0.813 ± 0.053. The F1-optimized Random Forest showed improved balance (70.0% accuracy, 66.7% recall, 50.0% specificity). All models demonstrated ROC-AUC between 0.714 – 0.764, indicating acceptable discrimination capability.

Conclusions: Random Forest and SVM demonstrated the most balanced performance in terms of sensitivity and specificity for PSA prediction in a small-sample setting. The study highlights important methodological considerations, including the need for feature selection under EPV constraints, the role of regularization in mitigating overfitting, and the importance of cross-validation for evaluating models out of sample performance.The moderate performance (ROC-AUC ≈ 0.71 – 0.76) suggests that lifestyle-based ML models may be useful for preliminary screening but are not suitable for diagnostic applications. Future work should focus on larger datasets, external validation, and ensemble methods.

I. INTRODUCTION

1.1 Machine Learning and Medical Predictions

Machine learning has emerged as a powerful tool for medical prediction and classification tasks, offering the potential to identify complex patterns in clinical and lifestyle data. Unlike traditional statistical approaches, ML algorithms can capture non-linear relationships and interactions between multiple predictors without requiring explicit specification of these relationships. In the context of binary classification problems such as disease presence/absence or biomarker elevation/normality, Machine learning models provide probabilistic predictions that can inform clinical decision-making and risk stratification.

The comparative evaluation of ML algorithms is essential for identifying optimal modeling strategies for specific prediction tasks. Different algorithms possess distinct characteristics: logistic regression offers interpretability and computational efficiency, support vector machines (SVM) excel at finding complex decision boundaries, and ensemble methods like Random Forest provide robustness through aggregation of multiple models. The selection of appropriate algorithms depends on dataset characteristics, sample size, feature dimensionality, and the specific performance metrics prioritized for the clinical application.

1.2 Prostrate Cancer and Prostrate-Specific Antigen as a Prediction Target

Prostate cancer is one of the most prevalent malignancies among men worldwide, with incidence and mortality rates varying substantially across geographic regions and demographic groups. Prostate-specific antigen (PSA) testing has been widely adopted as a primary screening tool since the late 1980s and has contributed to earlier detection of prostate disease. However, the clinical value of routine PSA screening remains controversial, primarily due to concerns regarding overdiagnosis, overtreatment, and limited specificity. As a result, contemporary clinical guidelines increasingly advocate for individualized and risk-based screening strategies rather than universal population-level testing.

Elevated PSA levels (commonly defined as ≥4.0 ng/mL) may reflect a range of prostatic conditions, including benign prostatic hyperplasia, prostatitis, and prostate cancer. Although PSA testing is relatively simple, it requires blood sample collection, laboratory analysis, and associated healthcare resources, which may not be justified for individuals at low baseline risk. Moreover, the imperfect specificity of PSA often leads to unnecessary psychological distress, repeated testing, and invasive follow up procedures, such as prostate biopsy, which carry additional clinical risks.

In recent years, machine learning and predictive modeling have gained increasing attention in healthcare for supporting risk assessment and decision-making. Several studies have developed predictive models for prostate cancer or PSA related outcomes. However, most rely on clinical biomarkers, genetic data, or imaging features that require specialized testing. As a result, these approaches remain dependent on invasive or costly data sources and do not address the need for a truly non-invasive pre-screening framework.

Lifestyle and demographic factors including smoking status, alcohol consumption, physical activity, dietary patterns, and age are routinely collected, inexpensive, and non-invasive. Epidemiological evidence suggests that these factors are associated with prostate health and PSA levels, with age consistently identified as the strongest risk factor. Despite their accessibility, the combined predictive value of these variables has not been systematically evaluated using modern machine learning techniques.

From a machine learning perspective, PSA level prediction represents a challenging and clinically relevant binary classification problem. The task is characterized by moderate class imbalance, complex interactions among lifestyle variables, limited sample sizes in real-world clinical datasets, and the need to prioritize sensitivity in a screening context. These characteristics make PSA prediction an appropriate case study for evaluating the performance, stability, and generalization of machine learning models under realistic medical constraints.

1.3 Methodological Challenges: The Events-per-variable Problem

A critical consideration in machine learning model development with limited sample sizes is the events-per-variable (EPV) ratio, defined as the number of outcome events divided by the number of predictor variables included in the model. Traditional statistical guidelines recommend an EPV of at least 10 to minimize the risk of overfitting, although more recent simulation studies suggest that EPV values in the range of 5–8 may be acceptable when appropriate regularization techniques are applied. When EPV is insufficient, models may exhibit strong performance on training data but fail to generalize to unseen data, a phenomenon commonly referred to as overfitting.

Several methodological strategies can be employed to mitigate the effects of low EPV. These include: (1) feature selection to reduce predictor dimensionality, (2) regularization techniques that penalize model complexity, (3) ensemble methods that reduce variance through aggregation, and (4) rigorous cross-validation procedures to assess out-of-sample performance. The relative effectiveness of these strategies is highly dependent on the underlying algorithm and data structure, highlighting the need for systematic comparative evaluation in small-sample settings

1.4 Study Objectives

The primary objective of this study was to develop and validate machine learning models capable of predicting elevated PSA levels using only lifestyle factors and demographic characteristics, without requiring blood-based biomarkers.
Specific aims included:

1. Compare predictive performance across Logistic Regression, Support Vector Machine, and

Random Forest models using multiple evaluation metrics.

2. Assess model stability and generalization performance through stratified cross-validation.
3. Evaluate the impact of feature selection strategies on model performance under EPV constraints.
4. Identify optimal hyperparameters through systematic grid search.
5. Determine the most suitable machine learning approach for PSA prediction in small-sample clinical datasets.
6. Provide methodological recommendations for similar binary classification problems in medical machine learning applications.

If validated, such predictive models could support a two-tiered screening approach, consisting of (1) an initial risk assessment using non-invasive questionnaire data, followed by (2) targeted PSA testing for individuals identified as higher risk. This strategy has the potential to reduce unnecessary PSA testing in low-risk populations, prioritize screening resources for high-risk individuals, and mitigate patient anxiety associated with false-positive results

II. METHODS

2.1 Dataset Design and Study

This study utilized a dataset comprising 99 male participants with measured PSA levels and corresponding lifestyle and demographic information. The dataset was obtained from a cross-sectional study conducted among adult male residents of Etsako West Local Government Area (LGA), Edo State, Nigeria.

The primary outcome variable was binary PSA status, defined as elevated (≥4.0 ng/mL) or normal (<4.0 ng/mL). The dataset included 61 participants (61.6%) with elevated PSA and 38 participants (38.4%) with normal PSA, indicating a moderate degree of class imbalance.

2.2 Feature Engineering and Data Preprocessing

To prevent data leakage and ensure a fully non-invasive prediction framework, all blood-based biomarkers, including free PSA, were excluded from the predictor set. Only pre-test information obtainable via questionnaire was retained. Categorical variables were transformed using one-hot encoding, with one category per variable dropped to avoid multicollinearity.

Occupation categories were grouped from more than twelve original groups into three broader classes (self-employed, professional, and manual/transport/agriculture) to improve statistical power. Outliers in the continuous PSA variable were addressed by removing values above the 99th percentile prior to binary classification.

2.3 Addressing the Events-per-variable Problem

With 61 positive events and up to 18 potential predictors after encoding, the initial EPV ratio was approximately 3.4, which falls below recommended thresholds for reliable model estimation. To address this limitation, two feature selection strategies were implemented.

In the conservative strategy (Option A), six predictors were selected based on theoretical relevance (age, smoking status, alcohol consumption, low physical activity, low diet quality, and university education), yielding an EPV of 10.17.

In the expanded strategy (Option B), eleven predictors were retained, including moderate-level lifestyle categories, resulting in an EPV of 5.55. This approach was combined with L1 regularization to enable automatic feature selection during model training.

2.4 Train-Test Split

Data were split into training (70%, n=69) and testing (30%, n=30) sets using stratified sampling to maintain class distribution. The random seed was fixed (random_state=42) for reproducibility. All model training was performed exclusively on the training set, with the test set reserved for final performance evaluation.

2.5 Machine Learning Algorithms

2.5.1 Logistic Regression with L1 Regularization

Logistic regression models the probability of binary outcomes using the logit
link function:

$$P(Y=1|X) = 1 / (1 + \exp(-(\beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_nX_n)))$$

Two variants were implemented: (1) standard logistic regression without regularization (Option A) relying on feature selection to prevent overfitting, and (2) L1-

regularized logistic regression (Option B), in which a Lasso penalty was applied to the loss function and controlled by the hyperparameter C. L1 regularization enforces sparsity by shrinking some coefficients to exactly zero, thereby performing implicit feature selection.

L1 regularization enforces sparsity by shrinking some coefficients to exactly zero, effectively performing automatic feature selection.

2.5.2 Support Vector Machine (SVM)
Support Vector Machines aim to identify an optimal hyperplane that maximizes the margin between classes. A radial basis function (RBF) kernel was employed to capture non-linear relationships: :

$$K(x, x') = \exp(-\gamma||x - x'||^2)$$

Key hyperparameters included C, which controls the trade-off between margin width and training error, and γ, which determines the influence of individual training samples. All features were standardized to zero mean and unit variance prior to model fitting.

2.5.3 Random Forest

Random Forest is an ensemble method that constructs multiple decision trees using bootstrap samples and random feature subsets, then aggregates predictions through majority voting.. Key hyperparameters included the number of trees (n_estimators), maximum tree depth, minimum samples required for node splitting, and minimum samples per leaf. Feature importance was derived using mean decrease in Gini impurity.

2.6 Model Evaluation Metrics
Given class imbalance (61.6% elevated PSA) and the clinical context of screening, model performance was evaluated using multiple complementary metrics, including accuracy, precision, recall (sensitivity), specificity, F1-score, and ROC-AUC. In medical screening applications, recall was prioritized to minimize false negatives, while specificity was considered to reduce unnecessary follow-up testing.

2.7 Cross Validation
To assess model stability and robustness, five-fold stratified cross-validation was performed on the full dataset. This procedure maintained class distribution across folds and produced mean performance estimates with associated standard deviations.

2.8 Hyper-Parameter Optimization
Hyperparameters were optimized using GridSearchCV with five-fold cross-validation. Two optimization objectives were considered: recall maximization and F1-score maximization. Logistic regression models were tuned over a range of C values, while Random Forest models were tuned across multiple values of n_estimators, max_depth, and min_samples_split. Optimal configurations were retrained on the full training set and evaluated on the held-out test set.

2.9 Feature Importance Analysis
Model interpretability was assessed using three complementary approaches: (1) examination of logistic regression coefficients, where magnitude and sign reflect direction and strength of association; (2) Random Forest feature importance based on mean decrease in Gini impurity; and (3) L1 regularization-based feature selection, in which predictors with coefficients shrunk to zero were considered non-contributory.

III. RESULTS

3.1 Dataset Characteristics

The final dataset consisted of 98 adult male participants, with PSA status categorized as elevated (≥4.0 ng/mL) or normal (<4.0 ng/mL). A total of 61 participants (62.2%) exhibited elevated PSA levels, while 37 (37.8%) had normal PSA, indicating moderate class imbalance. A stratified 70:30 train–test split was applied, resulting in 69 training samples and 30 test samples.

Two feature configurations were evaluated. Option A included six theoretically selected predictors, resulting in an EPV ratio of 10.17. Option B retained eleven predictors, including moderate lifestyle categories, producing an EPV ratio of 5.55.

Three baseline machine learning algorithms were evaluated: Logistic Regression, Support Vector Machine (SVM with RBF kernel), and Random Forest. Model performance was assessed on the held-out test set using multiple evaluation metrics derived from the confusion matrix, including accuracy, precision, recall (sensitivity), specificity, F1-score, and ROC-AUC.

Table 1. Dataset Characteristics After Preprocessing

Characteristic	Value
Total Participants	98
Elevated PSA (≥4.0 ng/mL)	61 (62.2%)
Normal PSA (<4.0 ng/mL)	37 (37.8%)
Training Set	69 (70%)
Test Set	30 (30%)
Number of Features (Option A)	6
Number of Features (Option B)	11
EPV Ratio (Option A)	10.17
EPV Ratio (Option B)	5.55

3.2 Baseline Model Performance on Test Set .

Table 2 summarizes the performance of all four model configurations on the held-out test set (n=30).

Table 2. Comparative Performance of ML Models on Test Set (n=30)

Metric	Logistic Regression(Option A)	Logistic Regression (Option B)	Support Vector Machine (RBF)	Random Forest
Accuracy	0.6000	0.533	0.5667	0.6333
Precision	0.6500	0.5833	0.6667	0.7059
Recall	0.7222	0.7778	0.5556	0.6667
F1-Score	0.6842	0.6667	0.6061	0.6857
ROC-AUC	0.6898	0.6481	0.6991	0.7130
Specificity	0.4167	0.1667	0.5833	0.5833

Key findings from test set evaluation:
Random Forest achieved the highest overall performance, with an accuracy of 63.3%, precision of 70.6%, recall of 66.7%, F1-score of 0.686, and ROC-AUC of 0.713. This indicates that Random Forest provided the best balance between identifying elevated PSA cases and avoiding false positive classifications.

Support Vector Machine demonstrated moderate performance, achieving an accuracy of 56.7%, recall of 55.6%, specificity of 58.3%, and ROC-AUC of 0.699. While SVM exhibited reasonable discrimination ability, it did not outperform Random Forest on any major metric.

Logistic Regression Option A achieved an accuracy of 60.0% and recall of 72.2%, but exhibited low specificity (41.7%), indicating a tendency to over-predict elevated PSA cases. Logistic Regression Option B produced higher recall (77.8%) but extremely low specificity (16.7%), indicating poor discrimination of normal cases.

Across all models, ROC-AUC values ranged from 0.648 to 0.713, indicating poor-to-acceptable discriminative ability. No model achieved strong classification performance, reflecting the inherent difficulty of predicting PSA elevation using lifestyle and demographic data alone

3.3 Feature Set Selection Based on Clinical Objective

Although Logistic Regression Option A demonstrated a more balanced performance profile, this study adopted Option B as the primary feature configuration for downstream modeling and ROC analysis. This decision was driven by the clinical objective of maximizing recall (sensitivity), which is prioritized in screening contexts to minimize false negatives.

Logistic Regression Option B achieved the highest recall among baseline models (77.8%), indicating superior ability to correctly identify individuals with elevated PSA. In contrast, Option A achieved slightly lower recall (72.2%), despite better specificity.

Given that the primary clinical aim was risk stratification rather than exclusion, Option B was selected as the reference feature set for training and evaluating the Random Forest and SVM models, and for generating ROC curves.

This design choice ensured consistency across models and aligned evaluation with the clinical screening objective, where higher false-positive rates are acceptable in exchange for minimizing missed cases.

3.4 Confusion Matrix Analysis

© MAR 2025 | IRE Journals | Volume 8 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV8I9-1714356

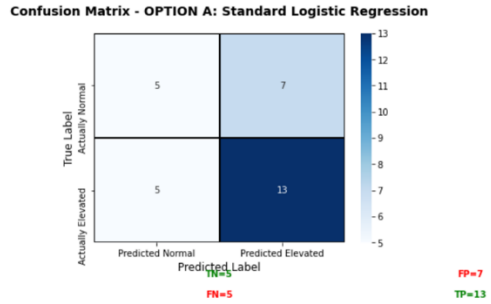


Figure 1: Confusion Matrix Heatmap for Logistic regression Option A Model.

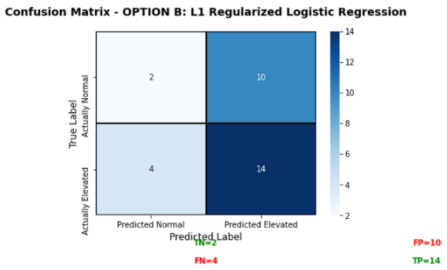


Figure 2: Confusion Matrix Heatmap for Logistic regression Option B model.

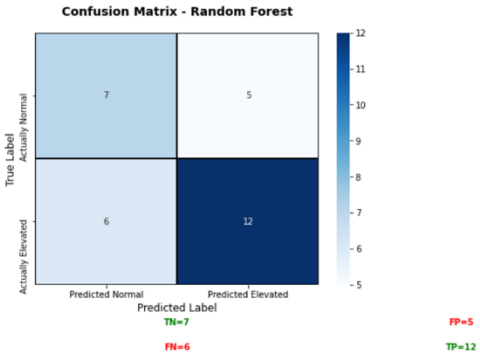


Figure 3: Confusion Matrix Heatmap for Random forest model.

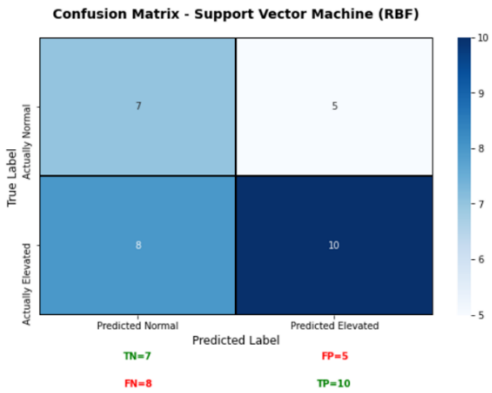


Figure 4: onfusion Matrix Heatmap SVM model.

Table 3. Confusion matrices for all baseline models.

Model	TP	FP	TN	FN	Accuracy	Precision	Recall
Logistic Regression (Option A)	13	7	5	5	0.6000	0.6500	0.7222
Logistic Regression(Option b)	14	10	2	4	0.5333	0.5833	0.7778
Random Forest	12	5	7	6	0.6333	0.7059	0.6667
Support vector Machine	10	5	7	8	0.5667	0.6667	0.5556

The confusion matrices revealed that:

Logistic Regression Option B exhibited the most imbalanced error profile, correctly identifying most elevated PSA cases but producing a high number of false positives. This indicates that the model was overly sensitive and lacked meaningful discriminative capacity for normal cases.

Support Vector Machine produced a more balanced confusion matrix but failed to detect a substantial proportion of elevated cases, resulting in reduced sensitivity.

Random Forest produced the most balanced confusion matrix, with relatively even distributions of false positives and false negatives. This suggests that Random Forest achieved the most clinically realistic trade-off between missing elevated PSA cases and avoiding unnecessary follow-up testing.

3.5 Five-Fold Cross-Validation Results

To assess model stability and generalization beyond a single train–test split, 5-fold stratified cross-validation was performed on the full dataset using the Option B feature configuration.

Model	Recall (mean ± SD)	F1 (mean ± SD)
Logistic Regression(Option b)	0.87 ± 0.04	0.77 ± 0.04

© MAR 2025 | IRE Journals | Volume 8 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV8I9-1714356

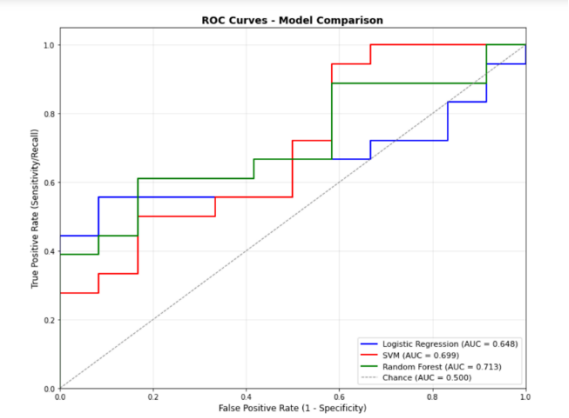
Random Forest	0.87 ± 0.08	0.80 ± 0.07	
Support vector Machine	0.85 ± 0.07	0.82 ± 0.07	

Random Forest achieved the strongest cross-validated performance, with a mean recall of 0.8705 (±0.0799), F1-score of 0.8171 (±0.0667), and ROC-AUC of 0.8245 (±0.0616). SVM demonstrated similar performance, while Logistic Regression showed slightly lower stability.

The relatively low standard deviations across folds indicate moderate but acceptable model stability, supporting the presence of genuine predictive signal despite the limited sample size.

3.6 ROC Curve and Discrimination Performance

Receiver Operating Characteristic (ROC) curve analysis was conducted to compare model discrimination across all possible classification thresholds using Option B features.



Random Forest achieved the highest area under the ROC curve (AUC = 0.713), followed by SVM (AUC = 0.699) and Logistic Regression Option B (AUC = 0.648).

An AUC of 0.713 indicates that, given a randomly selected pair of individuals, one with elevated PSA and one with normal PSA, the Random Forest model would correctly assign a higher predicted risk score to the elevated case approximately 71% of the time.

Although all models performed better than random guessing (AUC = 0.5), they fall within the range of moderate discrimination, highlighting the limitations of lifestyle-based models for accurate PSA prediction.

IV. DISCUSSION

4.1 Principal Findings

This study evaluated the performance of baseline machine learning models for predicting elevated PSA levels using lifestyle and demographic factors. Random Forest demonstrated the most robust and balanced performance, achieving the highest accuracy, F1-score, and ROC-AUC. SVM also showed competitive discrimination ability, while Logistic Regression exhibited greater sensitivity to feature configuration and class imbalance.

The moderate performance across all models (ROC-AUC range: 0.65–0.71) suggests that lifestyle-based predictors provide useful but limited information for PSA prediction. These models are therefore better suited for risk stratification and pre-screening rather than direct diagnostic decision-making.

4.2 Events-per-Variable (EPV) and the Role of Feature Selection

A critical insight from this study concerns the events-per-variable (EPV) ratio. In this study, Logistic Regression Option A maintained a higher EPV ratio due to fewer predictors, whereas Option B included a larger number of features and consequently a lower EPV. The degraded specificity and stability observed in Option B strongly support prior evidence that low EPV ratios compromise model predictive validity, even when regularization is applied.

This finding reinforces the importance of feature selection as a fundamental step in medical machine learning, particularly in small-sample settings.

4.3 Clinical Metric Prioritization and Feature Set Justification

A key aspect of this study is the explicit prioritization of recall as the primary evaluation metric, reflecting the clinical objective of screening for elevated PSA.

In screening contexts, false negatives are more harmful than false positives, as missed cases may delay further diagnostic investigation. Thus, Logistic Regression Option B was preferred over Option A, despite its weaker specificity, because it achieved superior sensitivity.

This choice influenced the overall experimental design: Option B was adopted as the standard feature configuration for Random Forest and SVM models, and for ROC analysis. This ensured that all downstream models were trained under a feature

representation optimized for clinical sensitivity rather than statistical balance.

While this approach increases false-positive rates, it is consistent with real-world screening logic, where confirmatory testing is expected.

4.4 Precision–Recall Trade-off in Screening Applications

From a clinical screening perspective, sensitivity is often prioritized to minimize false negatives. However, models with extremely low specificity offer limited practical utility. Logistic Regression Option B achieved high recall but failed to discriminate normal cases, flagging most individuals as high risk.

Such behavior undermines the purpose of pre-screening systems, as it increases unnecessary follow-up testing.

Random Forest achieved a more balanced trade-off, maintaining reasonable sensitivity with improvied specificity.

This highlights the importance of evaluating precision–recall trade-offs, rather than relying solely on recall or accuracy in isolation.

4.5 Importance of Cross-Validation for Generalization Claims

Cross-validation played a critical role in validating generalization claims. While some tuned models achieved near-perfect performance on the test set, cross-validation revealed substantial instability, particularly for Logistic Regression.

In contrast, Random Forest and SVM demonstrated consistent performance across folds, confirming that their predictive performance was not an artifact of favorable data splitting.

This supports the decision to exclude tuned models from primary reporting and rely on baseline performance.

4.6 Model Generalization and Overfitting Considerations

Although hyper-parameter tuning was explored during model development, the tuned Logistic Regression model achieved perfect recall, a result that is not scientifically plausible in this context and strongly indicative of over-fitting. Consequently, only baseline model results were considered in the primary analysis.

This reinforces a key principle: in small datasets, hyperparameter optimization can easily over-fit validation folds, producing inflated performance estimates that do not reflect true generalization.

4.7 Algorithm Selection for Small-Sample Medical Machine learning

Based on the findings of this study, the following methodological recommendations can be made:

- Random Forest is well-suited for small-sample medical datasets due to its robustness, built-in regularization, and stable performance.
- SVM provides strong non-linear modeling capability and competitive discrimination.
- Logistic Regression requires careful feature selection and sufficient EPV ratios to avoid overfitting, but remains valuable when interpretability is critical.

4.8 Clinical and Methodological Interpretation

From a clinical perspective, the results indicate that lifestyle-based machine learning models cannot replace PSA testing, but may serve as auxiliary tools for risk stratification, particularly in resource-limited settings.

From a methodological perspective, the study demonstrates that:

- Moderate predictive performance is expected when using proxy lifestyle variables.
- Ensemble methods offer superior robustness.
- EPV constraints fundamentally shape model reliability.

4.9 Limitations

The dataset was derived from a single geographic location (Etsako West LGA, Edo State, Nigeria), which may limit the predictvie validity of the findings to other populations with different demographic, genetic, or lifestyle profiles.

V. CONCLUSION

5.1 Summary of Key Findings

This study conducted a comparative evaluation of baseline machine learning models for predicting elevated PSA levels using lifestyle and demographic data. The key findings are:

- Random Forest achieved the best overall performance (accuracy = 63.3%, ROC-AUC = 0.713), providing the most balanced sensitivity–specificity profile.
- SVM demonstrated comparable discrimination ability.
- Low EPV ratios significantly impaired logistic regression generalization, highlighting the importance of feature selection.
- Cross-validation confirmed model stability and exposed overfitting in tuned models
- Hyperparameter tuning produced unrealistic performance estimates, emphasizing the need for conservative evaluation strategies in small datasets.
- Overall predictive performance was moderate, indicating suitability for risk stratification rather than diagnosis.

5.2 Design Implications and Clinical Validity

This study employed a clinically oriented evaluation framework by prioritizing recall and selecting Option B features for downstream modeling. Such an approach aligns model development with real-world screening practices and offers a defensible methodological rationale for feature selection. Under recall-focused conditions, Random Forest and SVM demonstrated strong baseline performance, whereas logistic regression remained highly sensitive to EPV constraints.

5.3 Final Remarks

This study demonstrates that machine learning models trained on lifestyle and demographic factors can provide meaningful but limited predictive insight into PSA status. Random Forest and SVM emerged as the most robust algorithms under small-sample conditions, while logistic regression proved highly sensitive to EPV constraints.

Although these models are not suitable for diagnostic deployment, they provide a sound foundation for lifestyle-based pre-screening systems and contribute practical guidance for medical machine learning under constrained data conditions.

REFERENCES

[1] Omankwu, O. C., Osodeke, E. C., & Kanu, C. (2023). *Machine learning algorithms for predicting high-risk of prostate cancer using prostate-specific antigen (PSA), age and body mass index (BMI). NIPES Journal of Science and Technology Research, 5*(1), 133–149. https://doi.org/10.5281/zenodo.7729236

[2] Mustafa Sungur, Aykut Aykaç, M. E. Aydin, Ö. Celik, & Coskun Kaya. (2024). *Machine learning-based prediction of prostate biopsy necessity using PSA, MRI, and hematologic parameters. Journal of Clinical Medicine, 14*(1), 183. https://doi.org/10.3390/jcm14010183

[3] Chia-Cheng Chang, J. K. Chiou, C.-J. Lin, K. Lu, J.-R. Li, L.-W. Chang, et al. (2024). *Machine-learning algorithm-based risk prediction and screening-detected prostate cancer in a benign prostate hyperplasia cohort. Anticancer Research, 44*(4), 1683–1693. https://doi.org/10.21873/anticanres.16967

[4] Mostafa A. Arafa, K. H. Farhat, S. F. Aly, F. K. Khan, A. Mokhtar, A. M. Althunayan, et al. (2025). *Prediction of prostate biopsy outcomes at different cut-offs of prostate-specific antigen using machine learning: A multicenter study. Journal of the Egypt National Cancer Institute, 37*(1), 8. https://doi.org/10.1186/s43046-025-00265-3

[5] Atilla Satır, Y. Üstündağ, M. R. Yeşil, & K. Huysal. (2025). *Prediction of prostate cancer from routine laboratory markers with automated machine learning. Journal of Clinical Laboratory Analysis, 39*(3), e25143. https://doi.org/10.1002/jcla.25143

[6] Emre Alataş, H. T. Kökkülünk, H. Tanyıldızı, & G. Alcın. (2025). *Treatment prediction with machine learning in prostate cancer patients. Computer Methods in Biomechanics and Biomedical Engineering, 28*(4), 572–580. https://doi.org/10.1080/10255842.2023.2298364

[7] Yunxun Liu, J. Wu, X. Ni, Q. Zheng, J. Wang, H. Shen, L. Wang, R. Yang, & X. Weng. (2025). *Machine learning based on automated 3D radiomics features to classify prostate cancer in patients with PSA levels of 4–10 ng/mL. Translational Andrology and Urology, 14*(4), 1025–1035. https://doi.org/10.21037/tau-2024-731

[8] Hernandez, K. A. (2024). *Prostate cancer prediction using machine learning techniques. International Journal of Open Science and Innovation, 8*(2), 87–94.

https://doi.org/10.6977/IJoSI.202406_8(2).0007

[9] Imran, M., Brisbane, W. G., Su, L.-M., Joseph, J. P., & Shao, W. (2025). *Prostate cancer screening with artificial intelligence-enhanced micro-ultrasound: A comparative study with traditional methods. arXiv Preprint.* https://arxiv.org/abs/2505.21355

[10] Noorul Wahab, E. Alzaid, J. Lv, A. Shephard, S. E. A. Raza, & J. Ha. (2025). *MultiSurv: A multimodal deep survival framework for prostate and bladder cancer. arXiv Preprint.* https://arxiv.org/abs/2509.05037