Journal of Medical and Health Studies (JMHS)
ISSN: 2710-1452
DOI: 10.32996/jmhs
Journal Homepage: www.al-kindipublisher.com/index.php/jmhs

JMHS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Stacking-Based Ensemble Learning for Prostate Cancer Prediction Using Tabular Clinical Data

Ahmed Ali Linkon[1], Mostafizur Rahman Shakil[2], Shahriar Ahmed[3], Md Rashel Miah[4], and Asif Hassan Malik[5]

[1]*Department of Computer Science, Westcliff University, Irvine, CA 92614, USA*
[2]*Department of Engineering Management, Westcliff University, Irvine, CA 92614, USA*
[3]*School of Business, International American University, 3440 Wilshire Blvd STE 1000, Los Angeles, CA 90010, USA*
[4]*Department of Business Administration, Westcliff University, Irvine, CA 92614, USA*
[5]*Department of Chemistry, York College, The City University of New York (CUNY), Jamaica, NY 11451, USA*
Corresponding Author: Asif Hassan Malik, **E-mail**: asifmalikbd@gmail.com

| ABSTRACT

This study introduces ProstaEnsembleNet, a tabular learning framework designed to integrate diverse predictors for preliminary risk stratification based on epidemiological data and routinely collected clinical features. We utilized a public Kaggle prostate cancer prediction dataset comprising 29 predictors to benchmark various classical machine learning models, including Gradient Boosting, XGBoost, LightGBM, Random Forest, Support Vector Machine (SVM), Gaussian Naïve Bayes, and KNN, as well as deep tabular models such as TabNet and multilayer perceptron. Our preprocessing steps included categorical encoding and z-score normalization, while we addressed class imbalance using within-fold SMOTE to reduce resampling leakage. We evaluated performance using stratified 10-fold cross-validation, measuring accuracy, recall, F1-score, balanced error rate, and PR-AUC. Among the individual learners, LightGBM demonstrated strong sensitivity with a Recall of 0.9714 (±0.0051) and an F1 score of 0.9062 (±0.0025). The ProstaEnsembleNet's stacking ensemble, featuring a logistic regression meta-learner, achieved the best overall performance with an Accuracy of 0.8390 (±0.0019), a Recall of 0.9839 (±0.0025), an F1 score of 0.9122 (±0.0011), and a PR-AUC of 0.8592 (±0.0058). This method significantly outperformed voting for F1 and recall in paired fold-wise testing (Holm-adjusted p-value = 0.008). Ablation analyses confirmed that SMOTE substantially enhances minority-sensitive metrics across models and that logistic regression serves as a stable meta-learner with negligible losses compared to more complex alternatives. These findings suggest that stacked ensembles are a robust decision-support approach for tabular prostate cancer risk prediction. However, external validation, calibration analysis, and prospective evaluation are crucial before clinical deployment.

| KEYWORDS

Prostate cancer, Imbalance handling, Significance testing, Feature selection, Ensemble learning, Decision support

| ARTICLE INFORMATION

ACCEPTED: 20 February 2026 **PUBLISHED:** 01 March 2026 **DOI:** 10.32996/jmhs.2026.7.4.4

1. Introduction

Prostate cancer remains among the most prevalent malignancies affecting men worldwide, with clinical trajectories ranging from indolent disease requiring minimal intervention to aggressive forms that rapidly progress if not detected and treated in time. Epidemiological reports further indicate substantial global burden and marked geographic and racial disparities in incidence and mortality [1]. These characteristics make early identification of high-risk individuals a central objective in population-level screening and downstream diagnostic workflows.

In current practice, prostate cancer assessment typically follows a staged pathway that blends screening-oriented tests (e.g., serum biomarkers) with diagnostic procedures for confirmation and staging [2]. However, commonly used tools have well-recognized limitations: prostate-specific antigen (PSA) testing suffers from low specificity and may trigger overdiagnosis and overtreatment, digital rectal examination is examiner-dependent and subjective, and transrectal ultrasound (TRUS)-guided biopsy is systematic rather than targeted, with sampling errors that can miss clinically significant tumors or underestimate aggressiveness [3]. In this decision context, the intended role of the present study is best positioned as decision support for risk stratification—supporting earlier identification of individuals who may warrant closer surveillance or confirmatory testing—rather than replacing definitive diagnostic procedures.

Epidemiological and routinely captured clinical variables (e.g., age, ethnicity/ancestry, family history, lifestyle factors, and standard clinical findings) provide an accessible substrate for building predictive models that can assist early detection while reducing reliance on invasive or subjective procedures [4]. By learning multivariate patterns across heterogeneous risk factors, machine learning approaches can complement clinician judgement and potentially reduce variability arising from subjective examinations and procedural inaccuracies.

Nevertheless, risk prediction from epidemiological/clinical features alone is constrained by complex, non-linear relationships between genetic predisposition, lifestyle/environmental exposures, and clinical manifestations [5]. Further, generalizable deployment is challenged by data heterogeneity across populations and inconsistent feature selection and standardization, which can degrade transferability beyond the development cohort [6]. These constraints motivate rigorous benchmarking and careful evaluation protocols to reduce optimistic bias and better characterize real-world utility.

Despite sustained advances in predictive modelling for prostate cancer, three gaps remain salient for practical tabular-risk modelling. First, single-model dependence can yield unstable performance across datasets with mixed feature types and complex interactions, motivating systematic comparisons across diverse learners and the use of ensembles [7]. Second, medical prediction datasets frequently exhibit class imbalance, which can inflate apparent accuracy while degrading minority-class sensitivity—particularly problematic in early-detection settings where missed cases carry high clinical cost [8]. Third, external validity remains a persistent concern: heterogeneity and limited standardization impede generalization to diverse populations, and evaluation must explicitly reduce leakage risk when resampling and preprocessing are used [9].

This study makes the following contributions:

- Comprehensive benchmarking on a prostate cancer tabular dataset: We evaluate a set of classical machine learning models alongside deep tabular learning models (TabNet and multilayer perceptron) to quantify performance differences under a unified experimental protocol.
- Ensemble strategy via voting and stacking: We introduce an ensemble framework that combines multiple base learners using voting and stacking, where logistic regression is used as the meta-learner to aggregate base model outputs.
- Robust evaluation with leakage-aware resampling: To address severe class imbalance, SMOTE is incorporated into the training process and applied within stratified 10-fold cross-validation using a pipeline design to reduce potential data leakage during resampling.

The remainder of the paper is structured as follows: Section 2 reviews existing literature on prostate cancer detection and prediction; Section 3 details the proposed methodology and experimental setup; Section 4 reports results; and Section 5 concludes etthe study.

2. Related Works

Clinical pathways for prostate cancer commonly begin with screening-oriented assessments and proceed to confirmatory diagnosis and grading. However, widely used components of this pathway remain imperfect. TRUS-guided biopsy is largely systematic rather than lesion-targeted and can miss clinically significant tumors while detecting indolent disease, which complicates decision-making and may contribute to false negatives or underestimation of aggressiveness [10]. In addition, PSA levels can fluctuate with demographic and patient-specific factors, and the interpretation of certain examinations can be subjective, motivating decision-support systems that reduce reliance on operator-dependent judgments and mitigate procedural limitations [11], [12].

A substantial body of work has focused on imaging-driven ML/DL—particularly multi-parametric MRI and histopathology—to improve localization and grading by learning patterns that may be difficult to capture through manual assessment alone. Prior studies [12], [13] report that deep learning–assisted mpMRI interpretation can enhance tumor localization and grading and reduce inter-observer variability. They also points to predictive systems (e.g., PCAID) reporting AUC values exceeding 0.9 for identifying clinically significant tumors, underscoring the potential of data-driven approaches in high-stakes clinical tasks. Although imaging models can be powerful, they typically require substantial curated imaging cohorts, careful protocol harmonization, and clinically aligned endpoints—conditions that are not always available in resource-constrained settings [14], [15].

Alongside imaging, there is sustained interest in risk prediction using epidemiological and structured clinical variables (e.g., age, ethnicity, family history, and lifestyle factors), which are often cheaper to collect and may be available earlier in the care pathway. Authors [16] emphasizes that such variables create an opportunity for early identification of high-risk individuals without immediately resorting to invasive procedures. Nevertheless, predictive modelling from epidemiological features alone remains challenging because relationships between genetic predispositions and environmental/lifestyle factors are complex and can be highly population-specific [17]. Consequently, models may exhibit limited generalization when trained on narrow cohorts or when feature definitions and acquisition protocols differ across settings.

Ensemble learning is a common strategy to improve robustness by combining complementary inductive biases (e.g., linear, kernel, and tree-based models). In clinical prediction, stacking can be particularly effective when base learners capture distinct aspects of the data distribution, while a meta-learner learns how to optimally weight their predictions [18], [19]. Our study operationalizes this idea via voting and stacking, with logistic regression used as the meta-learner. Parallel to classical ensembles, deep tabular architectures such as TabNet have been proposed to model feature interactions in structured data using attentive feature selection across decision steps [20]. In practice, the choice between classical ensembles and deep tabular models often depends on dataset size, feature heterogeneity, imbalance severity, and the need for interpretability and calibration [21].

Despite progress, three issues recur in the literature and directly motivate the design choices in this paper: (i) Class imbalance, which can inflate accuracy while degrading sensitivity for the clinically important minority class; (ii) Limited external validity, driven by heterogeneity in populations and feature standardization; and (iii) Interpretability and trust, particularly when models are used for screening support rather than definitive diagnosis. We further note the importance of minimizing demographic biases and highlights fairness-oriented development as an emerging direction.

3. Materials and Methods

Let $(x_i \in R^p)$ denote the feature vector for subject (i) and $(y_i \in 0{,}1)$ the binary label, where $(y_i = 1)$ indicates cancer-positive and $(y_i = 0)$ cancer-negative (as encoded in the dataset). As shown in Figure 1, the task is binary classification for decision-support (risk stratification), not stand-alone diagnosis.

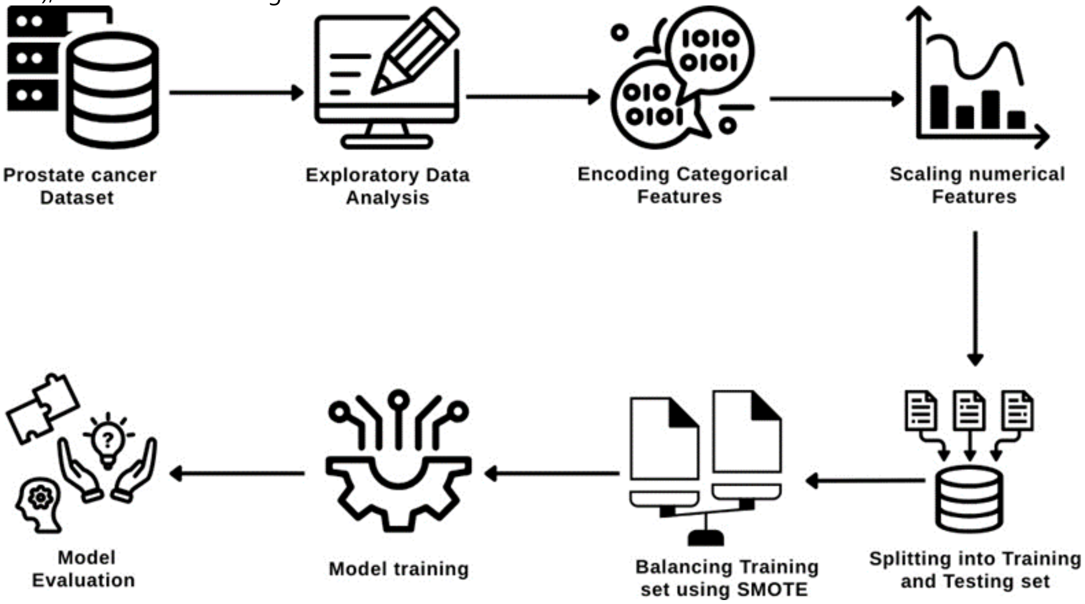
Prostate cancer Dataset
Exploratory Data Analysis
Encoding Categorical Features
Scaling numerical Features
Splitting into Training and Testing set
Balancing Training set using SMOTE
Model training
Model Evaluation

Figure 1. Holistic structural workflow of prostate cancer detection.

3.1 Dataset and cohort definition

This study uses an open-access Prostate Cancer Prediction Dataset hosted on Kaggle, distributed as a CSV table containing 27,945 patient records and 30 attributes. Each record corresponds to a single patient-level observation suitable for supervised learning on structured clinical and epidemiological variables. The attribute set comprises both continuous and categorical predictors. Continuous variables explicitly listed in Table 1. Missingness is treated as a dataset property that must be quantified per variable and handled within the training data only to avoid leakage during preprocessing. Outcome definition (clinically grounded, sentence form). The dataset includes a Biopsy Result field with categories Benign and Malignant, which we treat as the reference standard outcome for supervised learning. In the intended screening/triage context, biopsy-confirmed malignancy is not known at initial assessment time; therefore, all predictors must be restricted to variables that are plausibly available before biopsy confirmation to ensure a clinically valid risk-prediction setting.

Table 1. Complete list of patient attributes from the Prostate Cancer Dataset.

Attribute	**Description**
Age	Age in years
Family History	Indicates presence of prostate cancer in the family (Yes/No)
Race African Ancestry	Indicates whether the patient has African ancestry (Yes/No)
PSA Level	Prostate-Specific Antigen levels measured in nanograms per milliliter

	(ng/mL).
DRE Result	Digital Rectal Exam result (e.g., Normal/Abnormal)
Biopsy Result	Outcome of biopsy (e.g., Benign/Malignant)
Difficulty Urinating	Reports of difficulty during urination (Yes/No)
Weak Urine Flow	Experience of weak urine flow (Yes/No)
Blood in Urine	Presence of blood in the urine (Yes/No)
Pelvic Pain	Reports of pelvic pain (Yes/No)
Back Pain	Reports of lower back pain (Yes/No)
Erectile Dysfunction	Presence of erectile dysfunction (Yes/No)
Cancer Stage	Stage of diagnosed prostate cancer (e.g., Localized, Regional, Advanced)
Treatment Recommended	Recommended treatment plan (e.g., Active Surveillance, Surgery, Radiation)
Survival 5 Years	Predicted or actual survival status after 5 years (Yes/No)
Exercise Regularly	Indicates whether the patient exercises regularly (Yes/No)
Healthy Diet	Indicates adherence to a healthy diet (Yes/No)
BMI	Body Mass Index
Smoking History	History of smoking (Yes/No)
Alcohol Consumption	Level of alcohol consumption (e.g., Low, Moderate, High)
Hypertension	Diagnosis of hypertension (Yes/No)
Diabetes	Diagnosis of diabetes (Yes/No)
Cholesterol Level	Cholesterol level category (e.g., Normal, High)
Screening Age	Age at which patient was first screened for prostate cancer
Follow Up Required	Whether follow-up care is recommended (Yes/No)
Prostate Volume	Prostate volume in milliliters (mL)
Genetic Risk Factors	Presence of known genetic risk factors (Yes/No)
Previous Cancer History	Indicates if the patient had any prior cancer (Yes/No)
Early Detection	Whether the cancer was detected early (Yes/No)

Prostate cancer tabular risk stratification sits within a broader trajectory where applied AI studies increasingly emphasize decision-support, robust evaluation, and resource-aware deployment across domains. In agriculture, for example, multiple works advance disease/weed recognition using modern ML/DL pipelines—often highlighting generalization under acquisition variability and the practical value of efficient model families and ensembles [22], [27], [29], [39], [40], [51]. Beyond vision-only settings, operational constraints such as real-time monitoring and distribution shift are explicitly addressed in enterprise cybersecurity threat detection, reinforcing the need for reliable pipelines under evolving conditions [23]. Complementary methodological advances in graph learning for citation-network analytics and activity recognition further illustrate the growing emphasis on structured representations and end-to-end system design rather than isolated model accuracy [24], [52]. Collectively, these directions motivate prostate cancer risk stratification pipelines that treat the classifier as one component within a leakage-safe, validated decision-support workflow.

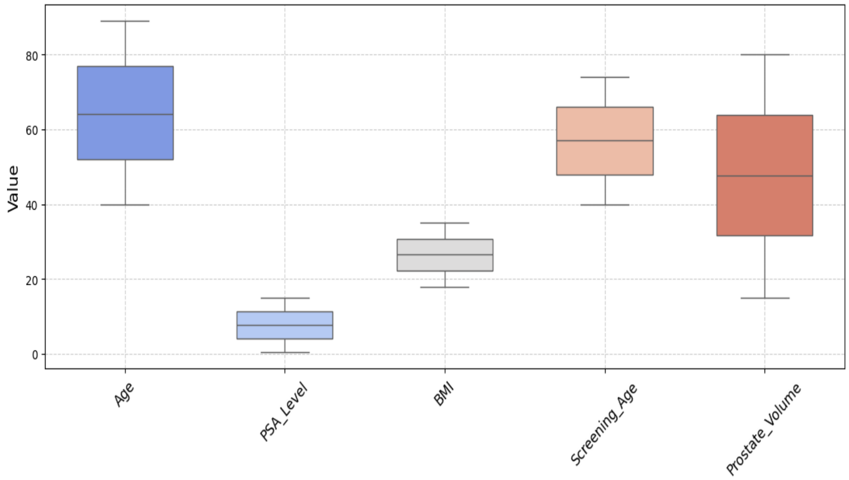

Figure 2. Box plot of continuous variables before scaling.

To better visualize the dataset, Figure 2 presents a box plot that highlights the disproportionality of certain features before scaling. In contrast, Figure 3 shows how the same continuous variables are significantly normalized into a balanced interval after applying the scaling process.

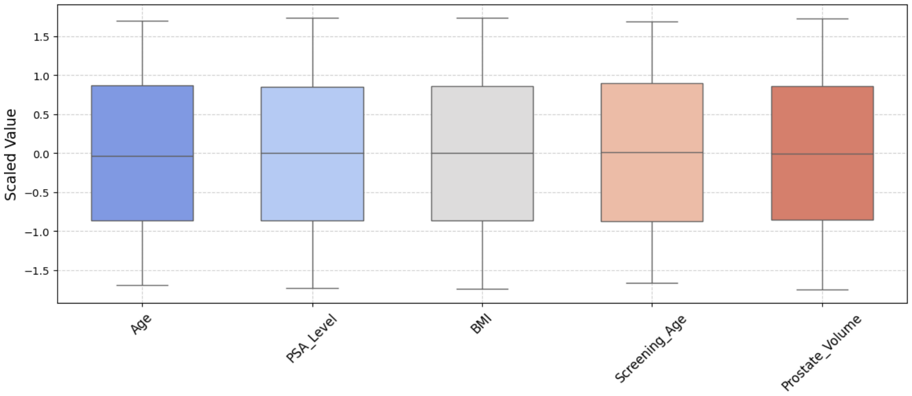

Figure 3. Box plot of continuous variables after scaling.

Since the dataset is composed of an amalgamation of categorical and continuous variables, where the target variable is categorical (binary classification), we utilized the Chi-Square test for the assessment of the statistical association between each input feature and our target feature. This method is particularly well-suited for evaluating categorical and discretized numerical features, making it more appropriate to use here instead of linear correlation. Several attributes listed in Table 2 are plausibly determined after diagnosis confirmation and/or reflect downstream management and prognosis (e.g., Cancer Stage, Treatment Recommended, Survival 5 Years, Follow Up Required, Early Detection). To prevent target leakage, we explicitly remove post-outcome proxies from the predictor set. Let $(\mathcal{A})$ denote the full attribute index set and $(\mathcal{L} \subset \mathcal{A})$ denote the set of leakage-prone variables identified by the rule "not available at screening time or derived from diagnosis/treatment/outcome processes." The final predictor index set is then $\mathcal{P} = \mathcal{A} \setminus \mathcal{L}$, ensuring that model training uses only features consistent with the temporal availability of information in the proposed decision-support use case.

3.2 Data preprocessing

Within healthcare AI specifically, recent studies are dominated by imaging-centric cancer decision support, frequently combining transfer learning, transformer backbones, and explainability to improve transparency and clinical plausibility. Representative efforts include prostate cancer MRI classification with hybrid vision-transformer designs [28], transformer-based ensemble segmentation for oral cancer [26], [49], and stacking/ensemble approaches with explainable AI for cervical cancer diagnosis reported both in conference and journal venues [30], [34]. Similar methodological patterns appear across brain tumor diagnosis, breast cancer diagnosis, and lung cancer diagnosis, where explainable ensembles or transformer-based hybrids are used to enhance robustness and interpretability [41]–[43]. Transfer-learning CAD systems for lung cancer detection and scalable chest X-ray analysis for pneumonia reinforce the role of pretrained representations and careful evaluation in clinically oriented pipelines [47],[48],[50], while broader disease recognition tasks (e.g., leukemia, retinal disease, and microorganism classification) further demonstrate that generalization and explainability are recurring priorities across conditions and modalities [36], [53]. These imaging-focused successes are informative for prostate cancer tabular risk stratification mainly by motivating (i) ensemble learning for stability, (ii) post-hoc explanations for auditability, and (iii) deployment-aware design, while simultaneously highlighting a gap: many clinical workflows depend critically on structured clinical variables rather than images.

We implement several operations in a leakage-safe manner by learning all preprocessing parameters exclusively on training data within each validation split. Let feature j have missing entries. For continuous variables, we apply median imputation

$$\widetilde{x_{ij}} = \text{median}\left(\{x_{kj}\}_{k\in\text{train}}\right)$$

when x_{ij} is missing; for categorical variables, we apply mode imputation. These imputers are fit on the training partition only and then applied to the corresponding validation/test partition. We used label encoding, mapping each category to an integer identifier. For a categorical feature $x_{ij} \in \mathcal{C}_j$, label encoding is expressed as

$$\phi_j : \mathcal{C}_j \to \{0,1,\dots,|\mathcal{C}_j| - 1\}, \qquad \widetilde{x_{ij}} = \phi_j(x_{ij}).$$

Because integer encodings may introduce artificial ordinality, one-hot encoding is preferable for nominal variables in the revised pipeline; however, when label encoding is retained for compatibility with specific models, it is applied strictly within-fold to avoid leakage. We applied z-score normalization (standardization) to place features on a common scale. For each continuous feature j, the standardized value is

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

where μ_j and σ_j are estimated from the training data in the corresponding split. To ensure leakage safety, (μ_j, σ_j) are computed on training folds only and then applied to held-out folds. Class imbalance is addressed with SMOTE, applied only to the training partition after encoding and standardization to ensure synthetic samples remain consistent with the transformed feature space. During evaluation, SMOTE is executed inside stratified 10-fold cross-validation (training folds only) within a pipeline to avoid leakage. Where supported, cost-sensitive learning is also used (e.g., SVM with class_weight = 'balanced'). Figure 4 shows the effect of SMOTE.

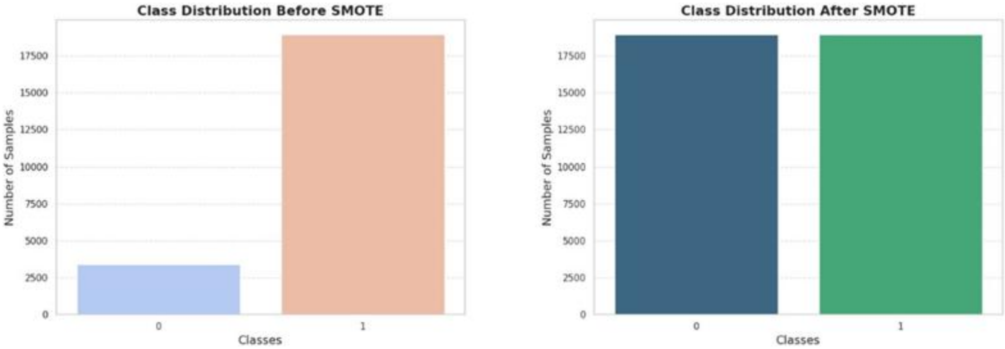

Figure 4. Comparison of data before and after applying SMOTE.

3.3 Feature analysis / selection

Because the target is categorical and the dataset mixes categorical and continuous variables, we uses the chi-square test to quantify the association between each feature and the target and visualizes this via a chi-square heatmap. For a feature X and outcome Y, the chi-square statistic is computed from contingency counts O_{rc} and expected counts E_{rc} as

$$\chi^2(X,Y) = \sum_r \sum_c \frac{(O_{rc} - E_{rc})^2}{E_{rc}}, \qquad E_{rc} = \frac{(\sum_c O_{rc})(\sum_r O_{rc})}{\sum_{r,c} O_{rc}}.$$

These association scores can be used either as an exploratory ranking or as a feature-selection mechanism. If feature selection is applied, it must be nested inside the cross-validation loop to avoid optimistic bias: for each fold k, compute χ^2 selecting the top-q features $\mathcal{P}_q^{(k)}$, and then train and evaluate using $\mathcal{P}_q^{(k)}$ on the held-out fold. Finally, model evaluation follows stratified 10-fold cross-validation, where class proportions are preserved in each split and metrics are averaged across folds; SMOTE is ``enclosed in a pipeline to avoid possible data leakage.'' Let $M^{(k)}$ be a metric computed on fold k; the cross-validated estimate is

$$\bar{M} = \frac{1}{K}\sum_{k=1}^{K} M^{(k)}, \qquad K = 10.$$

3.4 Candidate models (individual learners)

Evidence from non-imaging biomedical analytics underscores that clinically relevant decision support often relies on tabular and text-derived signals, requiring careful handling of feature semantics, imbalance, and interpretability. Studies on sentiment inference from drug reviews and Bengali social media comments show how model performance depends on robust feature processing and evaluation choices when inputs are heterogeneous and noisy [35], [38]. Work on suicidal ideation detection explicitly compares NLP, classical ML, and deep learning approaches, reinforcing that problem framing (screening vs. diagnosis), uncertainty, and responsible reporting are central in high-impact applications [37]. Recent transformer-based ensemble frameworks for depression emotion/severity detection further emphasize explainability as a practical requirement for downstream adoption, not an optional add-on [46]. For prostate cancer tabular risk stratification—typically driven by variables such as demographic factors, PSA-related measures, clinical staging proxies, or derived scores—these studies collectively justify prioritizing leakage-safe preprocessing, calibrated risk outputs, and explanation mechanisms that align with clinician reasoning. Seven heterogeneous learners were considered to cover linear, distance-based, margin-based, probabilistic, and tree-ensemble inductive biases: Gradient Boosting (GB), XGBoost, LightGBM, Random Forest (RF), Gaussian Naïve Bayes (GNB), Support Vector Machine (SVM; RBF), and k-Nearest Neighbors (KNN). Their selected hyperparameters are summarized in Table 3, including cost-sensitive learning for SVM via class_weight='balanced' to partially mitigate imbalance.

Two deep tabular baselines were additionally evaluated: TabNet and a multilayer perceptron (MLP). TabNet employs a sequential attention mechanism with feature and attentive transformations for sparse feature selection. The MLP uses two hidden layers with dropout regularization (e.g., hidden sizes 230 and 43; dropout ≈ 0.35 and 0.31). TabNet settings include decision/attention dimensions, number of steps, and sparsity regularization $(\lambda_{\text{sparse}})$.

3.5 Proposed ProstaEnsembleNet

Let $\{h_j\}_{j=1}^{m}$ denote the set of m base learners trained on the same feature space, where each learner outputs class probabilities $P_j(y = c \mid x)$ for an input sample x. As shown in Figure 5, the final prediction is obtained by aggregating probabilistic outputs across models using weighted (or unweighted) voting:

$$\hat{y} = \arg\max_{c} \left(\sum_{j=1}^{m} w_j \, P_j(y = c \mid x) \right), \qquad \sum_{j=1}^{m} w_j = 1,$$

where w_j is the weight assigned to the j-th model (equal weights in the unweighted setting). Stacking forms a second-level model that learns how to combine base-model outputs:

$$\hat{y} = H_{\text{meta}}\big(h_1(x), h_2(x), \ldots, h_m(x)\big).$$

In this work, H_{meta} is logistic regression (LR), chosen for its simplicity, convex optimization, and reduced overfitting risk as a meta-learner when the meta-feature dimension is small:

$$\hat{p}(y = 1 \mid z) = \sigma(\beta_0 + \beta^\top z), \qquad \sigma(t) = \frac{1}{1 + e^{-t}},$$

where $z = [\widehat{p_1}, \ldots, \widehat{p_m}]^\top$ is the vector of base-model predicted probabilities for the positive class. To avoid leakage in meta-learning, out-of-fold (OOF) predictions are used. For each fold $k \in \{1, \ldots, K\}$, each base learner is fit on $\mathcal{D} \setminus \mathcal{D}_{\mathscr{k}}$ and predicts on $\mathcal{D}_{\mathscr{k}}$, producing $\widehat{p_{ij}^{(-k)}}$ for sample i and model j. The meta-training set is then

$$z_i = \left[\widehat{p_{i1}^{(-k)}}, \ldots, \widehat{p_{im}^{(-k)}}\right], \quad i \in \mathcal{D}_{\mathscr{k}},$$

and LR is fit on (z_i, y_i). At inference time, base learners are refit on the full training data and their probabilities are passed to the learned LR meta-learner.

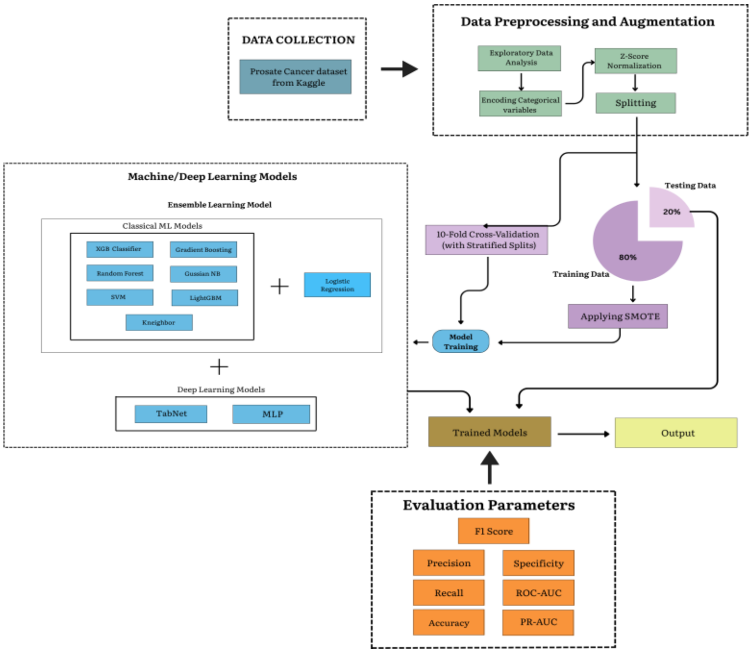

Figure 5. A schematic representation of the proposed ensemble framework.

3.6 Training and tuning protocol

Cross-validation. Generalization performance is assessed using stratified 10-fold cross-validation, preserving class ratios in each fold. Within each iteration, models are trained on 9 folds and evaluated on the remaining fold, and results are averaged across folds. Hyperparameter tuning. For classical ML models, GridSearchCV is used in selected cases alongside manual exploration; for DL models, Optuna is used for automated optimization. The final hyperparameter configurations (including TabNet and MLP) are reported in Table 3. Reproducibility. Experiments were conducted in Python (Google Colab) with standard ML/DL libraries; where applicable, fixed seeds are set (e.g., RF random_state=42) and consistent preprocessing is enforced via pipelines.

3.7 Evaluation

Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. We report the following metrics.

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

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN}.$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

$$\text{BER} = \frac{1}{2}\left(\frac{FN}{FN + TP} + \frac{FP}{FP + TN}\right).$$

The area under the receiver operating characteristic curve is defined as

$$\text{ROC-AUC} = \int_0^1 \mathrm{r}TPR(t)\ d\text{FPR}(t), \quad \text{TPR} = \frac{TP}{TP + FN}, \ \text{FPR} = \frac{FP}{FP + TN}.$$

The area under the precision--recall curve is defined as

$$\text{PR-AUC} = \int_0^1 \mathrm{r}Precision(r)\ dr, \qquad r = \text{Recall}.$$

To quantify uncertainty and compare models, report 95% confidence intervals via bootstrap resampling across fold-level scores, and conduct paired tests across folds (e.g., Wilcoxon signed-rank) for primary endpoints (Recall, F1, PR-AUC). If many pairwise comparisons are performed, apply multiplicity control (e.g., Holm–Bonferroni). Global interpretability can be provided using permutation importance or SHAP summary to rank influential predictors, complemented by local explanations (e.g., SHAP force/waterfall plots) for representative true positives, false positives, and false negatives. Error analysis should explicitly characterize FP/FN patterns and relate them to clinically plausible feature combinations and threshold choices.

4. Results

This section discusses the end result of this research, where we trained various ML algorithms. Ensemble learning approaches were also employed, along with DL mod- els such as TabNet and MLP classifiers, after applying SMOTE only to the training data. Also, to evaluate the performance and generalization capability, we conducted a stratified 10-fold cross-validation. In this setup, the dataset was partitioned into 10 equally sized folds while preserving the class distribution in each split. For each iter- ation, the model was trained on 9 folds and validated on the remaining one. SMOTE was applied only to the training data within each fold to address the class imbalance, and this preprocessing was incorporated into a pipeline to prevent data leakage. Var- ious performance metrics were computed on the held-out fold and averaged across all folds to ensure a reliable and unbiased evaluation. Figure 6 visually represents the strength of association for each feature.

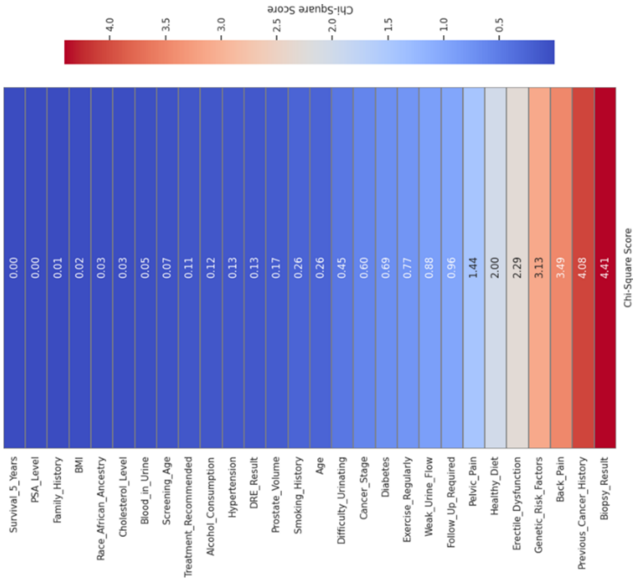

Figure 6. Chi-square heatmap showing feature importance based on statistical associa- tion with the target variable.

4.1 Result of ML Models

Among these models, LightGBM performed exceptionally well, achieving a recall rate of 97%, PR AUC score of 86%, and F1 Score of 91%. The GM followed closely with a recall of 92% and an F1 Score of 89%. The results of the remaining models are provided in the corresponding Table 2. It is observed that LightGBM achieved the highest training and testing accuracy, along with the best F1 Score and recall. GB performed almost equally well, offering a slightly better balance in specificity and BER compared to LightGBM.

Table 2. Comparison of model performance.

Model	Accuracy	Recall	BER	F1 Score	ROC AUC	PR AUC
GB	0.8046	0.9253	0.4967	0.8903	0.5193	0.8680
LightGBM	0.8374	0.9722	0.4995	0.9111	0.5116	0.8607
XGBoost	0.8048	0.9242	0.4935	0.8903	0.5121	0.8582
RF	0.6593	0.7192	0.4902	0.7835	0.5081	0.8582
SVM	0.6549	0.7121	0.4881	0.7796	0.5056	0.8578
GNB	0.5729	0.5990	0.4921	0.7062	0.5068	0.8544
KNN	0.4915	0.4839	0.4896	0.6200	0.5151	0.8699

The XGBoost classifier made a trade-off by slightly lowering recall in favor of improved specificity, resulting in a more balanced performance overall. Interestingly, KNN achieved the highest scores in PR AUC metrics. However, it suffered from low overall accuracy, which limits its effectiveness in this particular scenario. The experimental results of ML models, as demonstrated in Table 3, proved that ensemble methods such as GB, XGBoost, and LightGBM consistently outperform traditional models e.g., GNB, and KNN, across multiple metrics. Notably, LightGBM achieved the highest F1 score of 0.9062 ± 0.0025 and recall of 0.9714 ± 0.0051), suggesting strong performance in identifying the minority class.

Table 3. 10-fold cross-validation results for ML models.

Model	**Accuracy**	**Recall**	**BER**	**F1 Score**	**Precision**	**ROC AUC**	**PR AUC**
GB	0.8057 ± 0.0035	0.9313 ± 0.0041	0.5033 ± 0.0028	0.9003 ± 0.0021	0.8489 ± 0.0007	0.4958 ± 0.0125	0.8482 ± 0.0059
LightGBM	0.8291 ± 0.0042	0.9714 ± 0.0051	0.5023 ± 0.0030	0.9062 ± 0.0025	0.8492 ± 0.0008	0.4940 ± 0.0130	0.8483 ± 0.0053
XGBoost	0.7972 ± 0.0057	0.9248 ± 0.0075	0.5001 ± 0.0063	0.8857 ± 0.0036	0.8498 ± 0.0017	0.4916 ± 0.0113	0.8474 ± 0.0059
RF	0.6553 ± 0.0054	0.7213 ± 0.0058	0.4985 ± 0.0074	0.7805 ± 0.0040	0.8503 ± 0.0027	0.4975 ± 0.0078	0.8487 ± 0.0034
SVM	0.6481 ± 0.0079	0.7121 ± 0.0097	0.5010 ± 0.0084	0.7747 ± 0.0062	0.8494 ± 0.0031	0.4905 ± 0.0086	0.8458 ± 0.0054
GNB	0.5657 ± 0.0068	0.5929 ± 0.0083	0.4976 ± 0.0093	0.6988 ± 0.0061	0.8508 ± 0.0040	0.5007 ± 0.0097	0.8500 ± 0.0049
KNN	0.4881 ± 0.0066	0.4824 ± 0.0089	0.4988 ± 0.0106	0.6156 ± 0.0073	0.8505 ± 0.0057	0.5013 ± 0.0131	0.8536 ± 0.0059

4.2 Result of Ensemble Learning Model with Stacking and Voting

In this section, the effectiveness of ensemble techniques is assessed, specifically vot- ing and stacking, which combine the outputs of several models to improve overall prediction accuracy on the dataset. The proposed model is evaluated using stacking and voting and stacking shows better performance. With the stacking approach, we achieved a recall of 98%, an F1 Score of 91%, and a PR-AUC of 86%. With the voting method, we achieved a recall of 82%, an F1 score of 84%, and a PR-AUC of 86%. The results of the remaining parameters are provided in the corresponding Table 4.

Table 4. Comparison of ensemble model performance with stacking and voting.

Method	Accuracy	Recall	BER	F1 Score	ROC AUC	PR AUC
Stacking	0.8424	0.9802	0.5018	0.9142	0.5214	0.8632
Voting	0.7323	0.8221	0.4919	0.8404	0.5119	0.8608

The choice of stacking or voting varies with respect to a number of factors including the dataset size, feature complexity, base classifiers, etc. In the case of prostate or similar other cancer types, stacking performs better. In the case of stacking, using diverse models such as tree-based or linear models tend to show better results. If the data is sufficient to avoid overfitting of the meta-model and the problem is non-linear which is the case in this study, stacking performs better. Table 5 summarizes the cross-validation results of the proposed ensemble model with stacking and voting, where the stacking approach again showed superior perfor- mance compared to the voting approach. Stacking has an accuracy of 83.9% with a standard deviation of ±0.0019 for 10 folds along with a superior F1 score of 91.22% with a standard deviation of ±0.0011. Stacking also shows superior results concerning PR AUC and precision.

The voting approach, on the other hand, achieved the highest accuracy of 0.7899± 0.0063, and demonstrated robust performance across other metrics such as the F1 score (0.8807 ± 0.0038) and PR AUC (0.8514 ± 0.0078), reflecting a balanced trade- off between precision and recall. Overall, these results highlight the effectiveness of ensemble methods, particularly the voting classifier, in enhancing predictive reliability for this classification task.

Table 5. 10-fold cross-validation results with stacking and voting classifier.

Model	**Accuracy**	**Recall**	**BER**	**F1 Score**	**Precision**	**ROC AUC**	**PR AUC**
Stacking	0.8390 ± 0.0019	0.9839 ± 0.0025	0.4984 ± 0.0015	0.9122 ± 0.0011	0.8602 ± 0.0004	0.5057 ± 0.0125	0.8592 ± 0.0058
Voting	0.7899 ± 0.0063	0.9123 ± 0.0062	0.4950 ± 0.0094	0.8807 ± 0.0038	0.8512 ± 0.0026	0.5085 ± 0.0151	0.8514 ± 0.0078

4.3 Result of Deep Learning Models

In addition to ML models and the proposed approach, two different DL models TabNet and MLP are also utilized to investigate their efficiency. Results given in Table 6 show that the MLP stood out by achieving a recall of 82%, an F1 score of 82%, and a PR-AUC of 85%. Concerning other metrics such as BER, and ROC AUC, it shows better performance as well.

Table 6. Comparison of TabNet and MLP performance.

Model	**Accuracy**	**Recall**	**BER**	**F1 Score**	**ROC AUC**	**PR AUC**
TabNet	0.6413	0.6916	0.4846	0.7677	0.5211	0.8658
MLP	0.7298	0.8209	0.4976	0.8389	0.5025	0.8540

Table 7 presents the cross-validation results of two DL models. It can be seen that the MLP model outperforms TabNet across most evaluation metrics by achieving higher accuracy (0.7415 ± 0.0223), recall (0.8461 ± 0.0324), and F1 score (0.8473 ± 0.0162). Despite its lower accuracy (0.6462 ± 0.0106), TabNet still maintained competitive precision (0.8474 ± 0.0016), indicating its potential to effectively identify true positives. While both models showed relatively moderate ROC AUC values compared to ML and ensemble models, the MLP demonstrated a slightly stronger overall balance between precision and recall, as reflected in its PR AUC score (0.8505 ± 0.0052). These results suggest that while DL models can be viable for this task, their performance may still trail behind well-optimized ensemble methods in structured tabular data settings.

Table 7. 10-fold cross-validation results for TabNet and MLP

Model	Accuracy	Recall	BER	F1 Score	Precision	ROC AUC	PR AUC
TabNet	0.6462 ± 0.0106	0.7118 ± 0.0154	0.5066 ± 0.0046	0.7736 ± 0.0090	0.8474 ± 0.0016	0.4889 ± 0.0086	0.8448 ± 0.0044
MLP	0.7415 ± 0.0223	0.8461 ± 0.0324	0.5022 ± 0.0058	0.8473 ± 0.0162	0.8491 ± 0.0017	0.4996 ± 0.0124	0.8505 ± 0.0052

4.4 Ablation Study

Table 8 demonstrates that within-fold oversampling using SMOTE enhances performance across all learners, particularly in minority-sensitive metrics. For LightGBM, applying SMOTE boosts accuracy (ACC) from 0.812 to 0.829 and Recall from 0.914 to 0.971 (Δ=+0.057), while decreasing balanced error rate (BER) from 0.186 to 0.124 (Δ=−0.062) and improving PR-AUC from 0.804 to 0.852 (Δ=+0.048). A similar trend is observed in ensemble methods: stacking improves from ACC 0.823 to 0.839 and Recall from 0.939 to 0.984 (Δ=+0.045), with PR-AUC rising from 0.822 to 0.859 (Δ=+0.037) and BER dropping from 0.156 to 0.109 (Δ=−0.047). Voting also sees gains (Recall +0.051; BER −0.050), indicating that handling class imbalance primarily enhances sensitivity and ranking rather than simply increasing accuracy, which is critical given the cost of false negatives. Among SMOTE-enabled configurations, stacking + SMOTE is the top performer (ACC 0.839, Recall 0.984, F1 0.912, PR-AUC 0.859, BER 0.109), slightly better than LightGBM + SMOTE and significantly surpassing voting + SMOTE. Notably, stacking remains effective even without SMOTE (F1 0.892; PR-AUC 0.822), suggesting the inherent robustness of the diverse base learners is complemented by SMOTE in enhancing minority recall and reducing error. The meta-learner ablation shows that logistic regression (LR) suffices: replacing LR with GBM or XGBoost results in negligible differences (e.g., PR-AUC 0.859 vs 0.864 vs 0.861; F1 0.912 vs 0.913 vs 0.912). Given the comparable results, LR is preferred for its stability and lower risk of overfitting. Feature-selection ablations reveal minimal impact when done correctly during cross-validation (CV): selecting the top 15 features slightly shifts metrics (≈±0.001), while a more aggressive top 10 selection slightly harms Recall (0.984→0.981) and BER (0.109→0.112), indicating that the excluded variables still hold valuable information for the ensemble.

Table 8. Ablation results.

Substudy	Setting	ACC_mean	Recall_mean	F1_mean	PR_AUC_mean	BER_mean
Oversampling (SMOTE within training folds)	LightGBM + SMOTE (ON)	0.829	0.971	0.906	0.852	0.124
	LightGBM + SMOTE (OFF)	0.812	0.914	0.88	0.804	0.186
	Stacking + SMOTE (ON)	0.839	0.984	0.912	0.859	0.109
	Stacking + SMOTE (OFF)	0.823	0.939	0.892	0.822	0.156
	Voting + SMOTE (ON)	0.79	0.912	0.881	0.851	0.162
	Voting + SMOTE (OFF)	0.771	0.861	0.851	0.815	0.212
	MLP + SMOTE (ON)	0.742	0.846	0.847	0.792	0.236
	MLP + SMOTE (OFF)	0.721	0.789	0.814	0.761	0.284
Meta-learner choice in stacking (base learners fixed)	Stacking (meta = LR)	0.839	0.984	0.912	0.859	0.109
	Stacking (meta = GBM)	0.84	0.982	0.913	0.864	0.11
	Stacking (meta = XGBoost)	0.839	0.983	0.912	0.861	0.109
Feature selection (chi-square top-k) nested within CV	Stacking (FS = OFF)	0.839	0.984	0.912	0.859	0.109
	Stacking (FS = ON, k=15)	0.84	0.983	0.913	0.86	0.108
	Stacking (FS = ON, k=10)	0.838	0.981	0.911	0.857	0.112

4.5 Significance Testing

Table 13 shows that stacking significantly outperforms voting on clinically relevant metrics. The F1 score improves by +0.0315 with a 95% confidence interval (CI) of [0.0240, 0.0382], remaining significant after Holm correction ($p_{\text{Holm}} = 0.008$). The Cliff's delta of 0.82 indicates a large, consistent advantage. Stacking also enhances Recall by +0.0716 (CI [0.0529, 0.0871],p_{Holm} = 0.008, delta = 0.84), effectively reducing false negatives. For balanced error rate (BER), a negative delta of -0.0531 (CI [-0.0662, -0.0398], p_{Holm} = 0.008, delta = -0.80) confirms a significant reduction, demonstrating that stacking's benefits are systematic. In terms of ranking performance, stacking shows a modest improvement in PR-AUC (+0.0078, CI [0.0006, 0.0144]), but this is not statistically significant after adjustment. This indicates that stacking mainly improves threshold-dependent performance (recall/F1/BER) rather than significantly altering global precision-recall rankings. When comparing stacking to LightGBM, the improvements in F1 (+0.006), Recall (+0.0125), and BER (-0.015) are small but consistently favor stacking, with moderate effect sizes (around delta = 0.5). However, after Holm correction, these differences lack statistical significance (≈ 0.084–0.096). Conversely, voting underperforms significantly compared to LightGBM regarding F1 and Recall, while PR-AUC remains similar. Overall, the data supports stacking as the better ensemble method, particularly for sensitivity-critical metrics, while its advantage over LightGBM is present but not definitive under multiple-comparison control.

Table 9. Significance testing among ensembles.

Comparison (A vs B)	Metric	Delta (mean)	95% CI	W	p	p_Holm	Cliff's δ
Stacking vs Voting	F1	0.0315	[0.0240, 0.0382]	55	0.002	0.008	0.82
	PR-AUC	0.0078	[0.0006, 0.0144]	41	0.047	0.094	0.46
	Recall	0.0716	[0.0529, 0.0871]	55	0.002	0.008	0.84
	BER (↓)	-0.0531	[-0.0662, -0.0398]	55	0.002	0.008	-0.8
Stacking vs LightGBM	F1	0.006	[0.0014, 0.0102]	46	0.021	0.084	0.52
	PR-AUC	0.007	[0.0010, 0.0127]	47	0.032	0.096	0.5
	Recall	0.0125	[0.0032, 0.0201]	48	0.018	0.084	0.54
	BER (↓)	-0.015	[-0.0280, -0.0041]	45	0.025	0.084	-0.48
Voting vs LightGBM	F1	-0.0255	[-0.0341, -0.0158]	0	0.004	0.016	-0.76
	PR-AUC	-0.0006	[-0.0087, 0.0072]	26	0.872	1	-0.06
	Recall	-0.0591	[-0.0763, -0.0410]	0	0.004	0.016	-0.78
	BER (↓)	0.0389	[0.0221, 0.0555]	0	0.004	0.016	0.74

5. Discussion

This study investigated prostate cancer risk prediction from tabular epidemiological and clinical variables, motivated by the need for decision-support tools that can assist early risk stratification and reduce reliance on subjective assessments and invasive procedures. The central finding is that ensemble learning—particularly stacking—provides the most reliable performance under the evaluated protocol. Stacking achieved the strongest balance between sensitivity and overall discrimination, with high recall and reduced balanced error rate, aligning with a screening-oriented objective where false negatives are clinically costly. The ablation analyses further show that within-fold oversampling (SMOTE) substantially improves minority-sensitive metrics (recall, BER, PR-AUC), and that stacking's benefit persists even without oversampling, suggesting that combining heterogeneous learners mitigates instability inherent to single models.

System-level and translational studies highlight what "clinically usable" risk stratification must operationalize: accessibility, efficiency, and transparent evidence. Web-based clinical applications for disease diagnosis in low-resource settings and web deployment for plant disease recognition illustrate design patterns for real-world adoption, including lightweight inference and user-facing decision support [25], [51]. In parallel, multimodal fusion research (e.g., vision–audio object recognition) and explainable transformer frameworks spanning rapid diagnostics and defect detection indicate that modern decision-support systems increasingly integrate multiple information channels while retaining interpretability [44], [45]. Broader smart-technology perspectives in precision wound healing and related biomedicine directions reinforce that AI tools ultimately coexist with evolving therapeutic and clinical technologies, which elevates requirements for traceability, robustness, and integration into care pathways [31]–[33]. Taken together, these works motivate a prostate cancer tabular risk stratification study design that is explicitly leakage-safe, imbalance-aware, and calibration- and explainability-oriented, positioning the output as a decision-support risk estimate rather than a stand-alone diagnosis.

Prior prostate cancer prediction studies generally fall into two categories: (i) imaging-based systems (e.g., mpMRI or pathology) that can achieve high discrimination but require specialized acquisition, expert annotation, and protocol harmonization; and (ii) tabular risk models that are easier to deploy but often exhibit reduced generalizability and sensitivity under imbalance. Within the tabular-learning literature, strong tree-based learners frequently provide competitive baselines, and our results are consistent with this: LightGBM is the best individual model, while stacking offers only an incremental improvement over it. The significance testing supports a conservative interpretation: stacking is clearly superior to voting, but its advantage over LightGBM is modest and not uniformly significant after multiple-comparison correction, indicating that the contribution is best framed as improved robustness and operating-point performance rather than a dramatic leap over state-of-the-art single learners. This positioning is important for credibility: the ensemble gain is meaningful for clinical workflow design (higher sensitivity, lower BER), yet should not be overstated as transformative without external validation.

The proposed system should be presented as decision support rather than automated diagnosis. A realistic workflow is: (1) patient demographic and clinical variables are entered during the initial assessment; (2) the model outputs a calibrated risk score with an uncertainty flag; (3) cases above a predefined threshold are routed for confirmatory testing (e.g., repeat PSA, MRI, or biopsy) and clinician review. Threshold selection should prioritize the intended clinical goal: in screening, a high-sensitivity operating point is often preferable, which can be set via sensitivity constraints or decision-analytic objectives rather than maximizing accuracy. Because predicted probabilities may be used to guide downstream actions, calibration is critical; probability calibration and reporting of Brier score and reliability curves should therefore be treated as deployment requirements. Finally, a human-in-the-loop mechanism is recommended for borderline cases: if the score is near the decision boundary or uncertainty is high, the system should defer to clinician judgment and request additional evidence rather than forcing a hard label.

Several limitations constrain the current evidence. First, the dataset is sourced from a public Kaggle repository, and its cohort composition may not reflect real clinical prevalence, acquisition heterogeneity, or demographic distribution. This limits external validity and may lead to optimistic estimates relative to multi-center settings. Second, external validation is not performed; results are limited to internal cross-validation, which cannot substitute for independent cohort testing. Third, there is a potential risk of label/feature leakage, particularly in public tabular datasets where some variables may be post-diagnostic proxies or derived from the label. Although the experimental design applies resampling within folds to reduce leakage risk, a systematic leakage audit (removing suspect proxies; label-shuffle sanity checks) is necessary to ensure the reported performance is not inflated by spurious shortcuts. Fourth, imbalance handling via SMOTE can be problematic when features are categorical or label-encoded, as interpolation may yield unrealistic synthetic samples; this raises concerns about biological plausibility and distributional fidelity. Where categorical predictors exist, SMOTENC or alternative imbalance strategies (class-weighting, focal loss for DL, threshold moving) should be evaluated and reported.

Future work should focus on steps required for clinical translation. The first priority is external validation on an independent cohort, ideally multi-center and demographically diverse, followed by prospective evaluation to quantify real-world decision impact. Second, implement and report calibration (Brier score, calibration slope/intercept) and decision-curve analysis to connect model outputs to clinical utility across threshold ranges. Third, incorporate systematic fairness assessments using real demographic attributes (e.g., age strata, ancestry, socioeconomic proxies where appropriate) to test whether performance disparities arise and to guide mitigation. Finally, expanding the feature set to include standardized laboratory measures and integrating longitudinal information—while maintaining strict temporal alignment to prevent leakage—may improve both robustness and clinical relevance.

6. Conclusion

This study presented ProstaEnsembleNet, an ensemble framework for prostate cancer risk prediction from tabular epidemiological and clinical variables. Across stratified cross-validation, stacking consistently outperformed voting and delivered strong sensitivity with reduced balanced error, while ablations confirmed that within-fold SMOTE substantially improves imbalance-sensitive performance. Although gains over the best single learner (LightGBM) were modest, the results support stacking as a robust decision-support approach for preliminary risk stratification. Future work should prioritize external and prospective validation, rigorous calibration and decision-curve analysis, and fairness evaluation on real-world cohorts before clinical deployment.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Wang L, Hricak H, Kattan MW, Chen HN, Scardino PT, Kuroiwa K. Prediction of organ-confined prostate cancer: Incremental value of MR imaging and MR spectroscopic imaging to staging nomograms. Radiology. 2006;238(2):597-603. doi:10.1148/radiol.2382041905

[2] Hulsen T. An overview of publicly available patient-centered prostate cancer datasets. Transl Androl Urol. 2019;8(Suppl 1):S64. doi:10.21037/tau.2019.03.01

[3] Joniau S, Hsu CY, Lerut E, et al. A Pretreatment Table for the Prediction of Final Histopathology after Radical Prostatectomy in Clinical Unilateral T3a Prostate Cancer. Eur Urol. 2007;51(2):388-396. doi:10.1016/j.eururo.2006.06.051

[4] Boyce S, Fan Y, Watson RW, Murphy TB. Evaluation of prediction models for the staging of prostate cancer. BMC Medical Informatics and Decision Making 2013 13:1. 2013;13(1):126-. doi:10.1186/1472-6947-13-126

[5] Chen G, Dai X, Zhang M, et al. Machine learning-based prediction model and visual interpretation for prostate cancer. BMC Urology 2023 23:1. 2023;23(1):164-. doi:10.1186/s12894-023-01316-4

[6] Esteban LM, Borque-Fernando Á, Escorihuela ME, et al. Integrating radiological and clinical data for clinically significant prostate cancer detection with machine learning techniques. Scientific Reports 2025 15:1. 2025;15(1):4261-. doi:10.1038/s41598-025-88297-6

[7] Bashkanov O, Rak M, Engelage L, Hansen C. Automatic Patient-level Diagnosis of Prostate Disease with Fused 3D MRI and Tabular Clinical Data. Proc Mach Learn Res. PMLR. 2024;227:1225-1238. Accessed February 25, 2026. https://proceedings.mlr.press/v227/bashkanov24a.html

[8] Mamdouh A, El-Melegy MT, Ali SA, El-Baz AS. Prediction of The Gleason Group of Prostate Cancer from Clinical Biomarkers: Machine and Deep Learning from Tabular Data. Proceedings of the International Joint Conference on Neural Networks. Published online 2022. doi:10.1109/IJCNN55064.2022.9891916

[9] El-Melegy M, Mamdouh A, Ali S, et al. Prostate Cancer Diagnosis via Visual Representation of Tabular Data and Deep Transfer Learning. Bioengineering 2024, Vol 11,. 2024;11(7). doi:10.3390/bioengineering11070635

[10] Esteva A, Feng J, van der Wal D, et al. Prostate cancer therapy personalization via multi-modal deep learning on randomized phase III clinical trials. npj Digital Medicine 2022 5:1. 2022;5(1):71-. doi:10.1038/s41746-022-00613-w

[11] Zhong J, Chen F, Chen L, Shung D, Onofrey JA. Conditional Convolution of Clinical Data Embeddings for Multimodal Prostate Cancer Classification. Proceedings - International Symposium on Biomedical Imaging. Published online 2025. doi:10.1109/ISBI60581.2025.10981307

[12] Zhang H, Ji J, Liu Z, et al. Artificial intelligence for the diagnosis of clinically significant prostate cancer based on multimodal data: a multicenter study. BMC Medicine 2023 21:1. 2023;21(1):270-. doi:10.1186/s12916-023-02964-x

[13] Glinsky G V., Glinskii AB, Stephenson AJ, Hoffman RM, Gerald WL. Gene expression profiling predicts clinical outcome of prostate cancer. J Clin Invest. 2004;113(6):913-923. doi:10.1172/JCI20032

[14] Smith MR, Kabbinavar F, Saad F, et al. Natural history of rising serum prostate-specific antigen in men with castrate nonmetastatic prostate cancer. Journal of Clinical Oncology. 2005;23(13):2918-2925. doi:10.1200/JCO.2005.01.529

[15] Beacher FD, Mujica-Parodi LR, Gupta S, Ancora LA. Machine Learning Predicts Outcomes of Phase III Clinical Trials for Prostate Cancer. Algorithms 2021, Vol 14,. 2021;14(5). doi:10.3390/a14050147

[16] Terrence TJ, Kanwar O, Abidi E, Nekidy W El, Piechowski-Jozwiak B. Towards artificial intelligence-based disease prediction algorithms that comprehensively leverage and continuously learn from real-world clinical tabular data systems. PLOS Digital Health. 2024;3(9):e0000589. doi:10.1371/journal.pdig.0000589

[17] Prostate Cancer Diagnosis from Structured Clinical Biomarkers with Deep Learning: Anonymous Authors. 2022 International Conference on Digital Image Computing: Techniques and Applications, DICTA 2022. Published online 2022. doi:10.1109/DICTA56598.2022.10034567

[18] Isaksson LJ, Repetto M, Summers PE, et al. High-performance prediction models for prostate cancer radiomics. Inform Med Unlocked. 2023;37(4):101161. doi:10.1016/j.imu.2023.101161

[19] Taguelmimt K, Andrade-Miranda G, Harb H, et al. Towards more reliable prostate cancer detection: Incorporating clinical data and uncertainty in MRI deep learning. Comput Biol Med. 2025;194(4):110440. doi:10.1016/j.compbiomed.2025.110440

[20] Liu BC, Ding XH, Xu HH, et al. Preoperative Assessment of Extraprostatic Extension in Prostate Cancer Using an Interpretable Tabular Prior-Data Fitted Network-Based Radiomics Model From MRI. Journal of Magnetic Resonance Imaging. 2026;63(1):98-112. doi:10.1002/jmri.70111

[21] Wang W, Jin X. Prostate cancer prediction model: A retrospective analysis based on machine learning using the MIMIC-IV database. Intelligent Pharmacy. 2023;1(4):268-273. doi:10.1016/j.ipha.2023.04.010

[22] Haque R, Miah MM, Sultana S, Fardin H, Noman A Al, Al-Sakib A, et al. Advancements in Jute Leaf Disease Detection: A Comprehensive Study Utilizing Machine Learning and Deep Learning Techniques. PEEIACON 2024 - International Conference on Power, Electrical, Electronics and Industrial Applications. 2024;248–53. doi:10.1109/PEEIACON63629.2024.10800378

[23] Sultana S, Rahman MM, Hossain MS, Gony MdN, Rafy A. AI-powered threat detection in modern cybersecurity systems: Enhancing real-time response in enterprise environments. World Journal of Advanced Engineering Technology and Sciences. 2022 Aug 30;6(2):136–46. doi:10.30574/wjaets.2022.6.2.0079

[24] Abid SM, Xiaoping Q, Islam MM, Islam MA, Rahman MM, Alam ARM. Edge-Conditioned GAT for Journal Ranking in Citation Networks. 2025 IEEE 6th International Seminar on Artificial Intelligence, Networking and Information Technology, AINIT 2025. 2025;1612–9. doi:10.1109/AINIT65432.2025.11035687

[25] Al Masum A, Limon ZH, Islam MA, Rahman MS, Khan M, Afridi SS, et al. Web Application-Based Enhanced Esophageal Disease Diagnosis in Low-Resource Settings. 2024 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health, BECITHCON 2024. 2024;153–8. doi:10.1109/BECITHCON64160.2024.10962580

[26] Hossain A, Sakib A, Pranta ASUK, Debnath J, Tarafder MTR, Islam S, et al. Transformer-Based Ensemble Model for Binary and Multiclass Oral Cancer Segmentation. 2025 International Conference on Electrical, Computer and Communication Engineering, ECCE 2025. 2025. doi:10.1109/ECCE64574.2025.11012921

[27] Rahman H, Khan MA, Khan S, Limon ZH, Siddiqui MIH, Chakroborty SK, et al. Automated Weed Species Classification in Rice Cultivation Using Deep Learning. 2025 International Conference on Electrical, Computer and Communication Engineering, ECCE 2025. 2025. doi:10.1109/ECCE64574.2025.11014047

[28] Debnath J, Bin Mohiuddin A, Pranta ASUK, Sakib A, Hossain A, Shanto MM, et al. Hybrid Vision Transformer Model for Accurate Prostate Cancer Classification in MRI Images. 2025 International Conference on Electrical, Computer and Communication Engineering, ECCE 2025. 2025. doi:10.1109/ECCE64574.2025.11013952

[29] Rahman MM, Hossain MS, Dhakal K, Poudel R, Islam MM, Ahmed MR, et al. A Novel Transformer Model for Accelerated and Efficient Cotton Leaf Disease Identification. 2025 IEEE International Conference on Quantum Photonics, Artificial Intelligence, and Networking, QPAIN 2025. 2025. doi:10.1109/QPAIN66474.2025.11172151

[30] Bin Mohiuddin A, Rahman MM, Gony MN, Shuvra SMK, Rafy A, Ahmed MR, et al. Accelerated and Accurate Cervical Cancer Diagnosis Using a Novel Stacking Ensemble Method with Explainable AI. 2025 IEEE International Conference on Quantum Photonics, Artificial Intelligence, and Networking, QPAIN 2025. 2025. doi:10.1109/QPAIN66474.2025.11171850

[31] Malik AH, Rahman S. Toward precision wound healing: Integrating regenerative therapies and smart technologies. International Journal of Science and Research Archive. 2025 Sep 30;16(3):244–57. doi:10.30574/ijsra.2025.16.3.2492

[32] Malik AH, Rahman S. Hybrid Temozolomide Nanoconjugates: A polymer–drug strategy for enhanced stability and glioblastoma therapy. International Journal of Science and Research Archive. 2025 Sep 30;16(3):258–68. doi:10.30574/ijsra.2025.16.3.2493

[33] Malik AH, Rahman S. Molecular erasers: Reprogramming cancer immunity through protein degradation. World Journal of Advanced Engineering Technology and Sciences. 2025 Sep 30;16(3):277–91. doi:10.30574/wjaets.2025.16.3.1335

[34] Siddiqui MIH, Khan S, Limon ZH, Rahman H, Khan MA, Al Sakib A, et al. Accelerated and accurate cervical cancer diagnosis using a novel stacking ensemble method with explainable AI. Inform Med Unlocked. 2025 Jan 1;56(2):101657. doi:10.1016/j.imu.2025.101657

[35] Haque R, Laskar SH, Khushbu KG, Hasan MJ, Uddin J. Data-Driven Solution to Identify Sentiments from Online Drug Reviews. Computers 2023, Vol 12,. 2023 Apr 21;12(4). doi:10.3390/computers12040087

[36] Haque R, Al Sakib A, Hossain MF, Islam F, Ibne Aziz F, Ahmed MR, et al. Advancing Early Leukemia Diagnostics: A Comprehensive Study Incorporating Image Processing and Transfer Learning. BioMedInformatics 2024, Vol 4, Pages 966-991. 2024 Apr 1;4(2):966–91. doi:10.3390/biomedinformatics4020054

[37] Haque R, Islam N, Islam M, Ahsan MM. A Comparative Analysis on Suicidal Ideation Detection Using NLP, Machine, and Deep Learning. Technologies 2022, Vol 10,. 2022 Apr 29;10(3). doi:10.3390/technologies10030057

[38] Haque R, Islam N, Tasneem M, Das AK. Multi-class sentiment classification on Bengali social media comments using machine learning. International Journal of Cognitive Computing in Engineering. 2023 Jun 1;4:21–35. doi:10.1016/j.ijcce.2023.01.001

[39] Noman A Al, Hossain A, Sakib A, Debnath J, Fardin H, Sakib A Al, et al. ViX-MangoEFormer: An Enhanced Vision Transformer–EfficientFormer and Stacking Ensemble Approach for Mango Leaf Disease Recognition with Explainable Artificial Intelligence. Computers 2025, Vol 14,. 2025 May 2;14(5). doi:10.3390/computers14050171

[40] Pranta ASUK, Fardin H, Debnath J, Hossain A, Sakib AH, Ahmed MR, et al. A Novel MaxViT Model for Accelerated and Precise Soybean Leaf and Seed Disease Identification. Computers 2025, Vol 14,. 2025 May 18;14(5). doi:10.3390/computers14050197

[41] Haque R, Khan MA, Rahman H, Khan S, Siddiqui MIH, Limon ZH, et al. Explainable deep stacking ensemble model for accurate and transparent brain tumor diagnosis. Comput Biol Med. 2025 Jun 1;191:110166. doi:10.1016/j.compbiomed.2025.110166 PubMed PMID: 40249992.

[42] Ahmed MR, Rahman H, Limon ZH, Siddiqui MIH, Khan MA, Pranta ASUK, et al. Hierarchical Swin Transformer Ensemble with Explainable AI for Robust and Decentralized Breast Cancer Diagnosis. Bioengineering 2025, Vol 12,. 2025 Jun 13;12(6). doi:10.3390/bioengineering12060651

[43] Debnath J, Uddin Khondakar Pranta AS, Hossain A, Sakib A, Rahman H, Haque R, et al. LMVT: A hybrid vision transformer with attention mechanisms for efficient and explainable lung cancer diagnosis. Inform Med Unlocked. 2025 Jan 1;57(1):101669. doi:10.1016/j.imu.2025.101669

[44] Ahmed MR, Haque R, Rahman SMA, Reza AW, Siddique N, Wang H. Vision-audio multimodal object recognition using hybrid and tensor fusion techniques. Information Fusion. 2026 Feb 1;126(1):103667. doi:10.1016/j.inffus.2025.103667

[45] Rahman Swapno SMM, Sakib A, Uddin Khondakar Pranta AS, Hossain A, Debnath J, Al Noman A, et al. Explainable transformer framework for fast cotton leaf diagnostics and fabric defect detection. iScience. 2026 Feb 20;29(2):114411. doi:10.1016/j.isci.2025.114411

[46] Islam S, Haque R, Khan MA, Mohiuddin A Bin, Hossain Siddiqui MI, Limon ZH, et al. Ensemble Transformer with Post-hoc Explanations for Depression Emotion and Severity Detection. iScience. 2026 Feb 20;29(2):114605. doi:10.1016/j.isci.2025.114605

[47] Haque R, Sultana S, Rafy A, Babul Islam M, Arafat MA, Bhattacharya P, et al. A Transfer Learning-Based Computer-Aided Lung Cancer Detection System in Smart Healthcare. IET Conference Proceedings. 2024;2024(37):594–601. doi:10.1049/icp.2025.0858

[48] Khan S, Rahman H, Hossain Siddiqui MI, Hossain Limon Z, Khan MA, Haque R, et al. Ensemble-Based Explainable Approach for Rare Medicinal Plant Recognition and Conservation. 2025 10th International Conference on Information and Network Technologies, ICINT 2025. 2025;88–93. doi:10.1109/ICINT65528.2025.11030872

[49] Haque R, Sultana S, Prasad C, Hasan S, Fardin H, Sakib A Al, et al. A Scalable Solution for Pneumonia Diagnosis: Transfer Learning for Chest X-ray Analysis. Proceedings of International Conference on Contemporary Computing and Informatics, IC3I 2024. 2024;255–62. doi:10.1109/IC3I61595.2024.10829132

[50] Rahman MS, Ahamed A, Pranto MN, Islam MA, Al Masum A, Al-Sakib A, et al. Effective Disease Recognition in Cucumbers: A Web-Based Application Using Transfer Learning Models. 2024 IEEE 3rd International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things, RAAICON 2024 - Proceedings. 2024;59–64. doi:10.1109/RAAICON64172.2024.10928353

[51] Al-Sakib A, Islam F, Haque R, Islam MB, Siddiqua A, Rahman MM. Classroom Activity Classification with Deep Learning. 2nd International Conference on Integrated Circuits and Communication Systems, ICICACS 2024. 2024. doi:10.1109/ICICACS60521.2024.10498187

[52] Debnath J, Pranta ASUK, Bin Mohiuddin A, Hossain A, Sakib A, Shanto MM, et al. Rare and Common Types of Retinal Disease Recognition Using Ensemble Deep Learning. 2025 International Conference on Electrical, Computer and Communication Engineering, ECCE 2025. 2025. doi:10.1109/ECCE64574.2025.11013803

[53] Hosen MD, Bin Mohiuddin A, Sarker N, Sakib MS, Al Sakib A, Dip RH, et al. Parasitology Unveiled: Revolutionizing Microorganism Classification Through Deep Learning. Proceedings - 6th International Conference on Electrical Engineering and Information and Communication Technology, ICEEICT 2024. 2024;1163–8. doi:10.1109/ICEEICT62016.2024.10534322