2026 Volume: 02 Issue: 08 www.ijrpa.com ISSN 2456-9995 Research Article

International Journal Research Publication Analysis

Page: 01-11

MACHINE LEARNING-ASSISTED SEVERITY ASSESSMENT OF PROSTATE CANCER USING TCGA-PRAD DATASET

G. Saketh*[1], V. Lakshmi Narasimhan[2], D. Rammohan[1] and Ofaletse Mphale[3]

[1] CMR College of Engineering & Technology, Hyderabad/Bhagyanagar, Telangana, India.

[2]Elizabeth City State University, Elizabeth City, NC 27909, USA.

[3]University of Botswana, Botswana, Africa.

Article Received: 24 June 2026

Article Revised: 14 July 2026

Published on: 04 August 2026

***Corresponding Author: G. Saketh**

CMR College of Engineering & Technology, Hyderabad/Bhagyanagar, Telangana, India.

DOI: https://doi-doi.org/101555/ijrpa.9325

ABSTRACT

Prostate cancer is one of the most common cancers among men worldwide, accounting for a significant proportion of cancer diagnoses and mortality, particularly in developed countries [1]. It is estimated to be among the top three most diagnosed cancers, highlighting the need for accurate and early severity prediction. Traditional clinical approaches for prostate cancer assessment rely on prostate-specific antigen (PSA) levels and Gleason grading; however, these methods have limitations in predictive accuracy and consistency. Recent research has explored machine learning techniques to improve cancer prognosis and classification by leveraging large-scale biomedical datasets [2], [3]. In this study, we implement a comprehensive machine learning pipeline using the TCGA-PRAD (The Cancer Genome Atlas – Prostate Adenocarcinoma) dataset to predict cancer severity. The analysis includes exploratory data analysis and systematic identification and correction of data leakage. It also incorporates class imbalance handling using SMOTE and class-weight adjustments. Multiple machine learning models, including Logistic Regression, Random Forest, and Gradient Boosting, are trained and evaluated. Hyperparameter tuning and 5-fold cross-validation are performed to ensure robust performance evaluation. The tuned Gradient Boosting model achieved a cross-validated AUC of 0.954 ± 0.010 and a test AUC of 0.825. Key predictive features identified include patient age, PSA levels, and Fraction Genome Altered (FGA). These findings demonstrate the effectiveness of machine learning in improving prostate cancer severity prediction and provide a strong foundation for future integration of genomic features.

KEYWORDS: Prostate Cancer, Gleason Score, Machine Learning, TCGA-PRAD, Gradient Boosting, Class Imbalance, Data Leakage.

1. INTRODUCTION

Prostate cancer is a major global health concern and is one of the leading causes of cancer-related deaths among men [1]. Accurate prediction of cancer severity is essential for guiding treatment decisions and improving patient outcomes. The Gleason scoring system remains the clinical standard for assessing prostate cancer aggressiveness, but it suffers from inter-observer variability and limited predictive capability when used alone [3]. With the advancement of biomedical data availability, machine learning techniques have emerged as powerful tools for improving disease prediction and risk stratification [2]. Large-scale datasets such as The Cancer Genome Atlas (TCGA) provide comprehensive clinical and molecular profiles, enabling the development of data-driven predictive models [4]. In this study, we utilize the TCGA-PRAD dataset to develop a machine learning pipeline for predicting prostate cancer severity. The proposed approach integrates clinical features, addresses challenges such as data leakage and class imbalance, and evaluates multiple machine learning models to identify the most effective predictive framework. The rest of the paper is organized as follows: section 2 outlines the problem statement and related literature, while section 3 provides a brief description of the underlying algorithms & datasets. Our experimental procedure is detailed in section 4, followed by a discussion on the results and inference therefrom in section 5. The conclusion summarizes the paper and offers pointers for further research work in this arena.

2. PROBLEM STATEMENT & RELATED LITERATURE

The primary objective of this study is to develop a robust machine learning model capable of accurately classifying prostate cancer patients into low-grade (Gleason score = 6) and high-grade (Gleason score ≥ 7) categories. Traditional clinical methods rely heavily on PSA levels and Gleason grading, which may not fully capture the complexity of tumor biology [3]. Recent studies have demonstrated that machine learning models can significantly improve cancer prognosis by integrating multiple clinical and molecular features [2], [5]. Prior research has applied various machine learning algorithms, including Random Forest and Gradient Boosting, for cancer classification tasks, showing improved predictive performance over conventional methods [5], [6].

However, challenges such as data leakage, class imbalance, and lack of model interpretability

remain critical issues that must be addressed for reliable deployment. This study builds upon existing work by implementing a complete end-to-end pipeline that explicitly handles these challenges while ensuring robust validation.

3. BRIEF DESCRIPTION OF THE UNDERLYING ALGORITHMS & DATASETS

The dataset used in this study is the TCGA-PRAD dataset obtained from cBioPortal, containing clinical information for 498 prostate cancer patients [4] [7] [8]. The task is formulated as a binary classification problem to distinguish between low-grade and high-grade cancer cases. Logistic Regression is a linear model that estimates the probability of a binary outcome using a logistic function. It serves as a baseline due to its simplicity and interpretability. Random Forest is an ensemble learning method that constructs multiple decision trees and aggregates their predictions. It is effective in handling non-linear relationships and reducing overfitting [5].

Gradient Boosting is an advanced ensemble technique that builds models sequentially by minimizing prediction errors of previous models. It is particularly effective for structured data and achieves high predictive accuracy [6]. These models are chosen to provide a balance between interpretability, robustness, and predictive performance.

4. EXPERIMENTAL PROCEDURE

4.1 Exploratory Analysis

As shown in Table 1 and Figure 1, the TCGA-PRAD dataset exhibits marked differences between Low Grade (GS=6) and High Grade (GS≥7) cohorts. High-grade tumors are characterized by higher mean mutation counts (52.8 vs 39.2) and elevated tumor mutational burden (TMB mean: 1.76 vs 1.31). The dataset is strongly imbalanced, with approximately 91% of patients classified as High Grade.

4.2 Data Leakage Impact

Figure 2 demonstrates the substantial impact of data leakage: including Gleason-derived features inflated the AUC from the true 0.76 to an artificial 1.00. This underscores the importance of rigorous feature selection prior to model training in clinical ML pipelines.

4.3 Effect of Class Imbalance Handling

As evidenced by Tables 2 and 3, and Figures 3, 4, and 5, baseline models showed deceptively high accuracy (up to 91%) attributable to majority-class bias. After applying class weighting and SMOTE, recall for the minority class improved substantially — for example, Logistic Regression recall dropped from 1.00 (trivially predicting the majority) to 0.725 on balanced

data while AUC improved from 0.873 to 0.888, reflecting a more clinically meaningful model.

4.4 Cross-Validation and Best Model

Table 4 and Figure 6 present 5-fold cross-validation results. The Tuned Gradient Boosting model achieved the highest generalization performance with a CV AUC of 0.954 ± 0.010 and CV F1 of 0.883± 0.022 — significantly outperforming Logistic Regression (AUC 0.832) and Random Forest (AUC 0.927). Table 5 consolidates all model comparisons, confirming Tuned Gradient Boosting as the best overall model.

4.5 Feature Importance

As shown in Table 6 and Figure 7, patient age (avg importance: 0.243) and PSA level (0.205) are the most predictive features across all models, followed by Fraction Genome Altered (FGA: 0.179) and mutation count (0.142). These findings align with established clinical knowledge — PSA and age are standard screening biomarkers for prostate cancer, while genomic instability metrics (FGA, mutation count, TMB) reflect tumor aggressiveness. Overall survival (os_months) and TMB showed the least discriminatory power in this dataset.

4.6 Summary of Key Findings

The analysis demonstrates that: (1) Data leakage must be explicitly identified and corrected before model training, as shown in Figure 2; (2) Class imbalance handling is essential in imbalanced clinical datasets — naive accuracy is an unreliable metric; (3) Gradient Boosting with hyperparameter tuning achieves strong predictive performance (CV AUC = 0.954), as detailed in Tables 4 and 5 and Figure 6; and (4) Age, PSA, and FGA are the dominant clinical predictors of Gleason grade, as confirmed in Table 6 and Figure 7.

5. RESULTS OBTAINED & INFERENCES THEREFROM

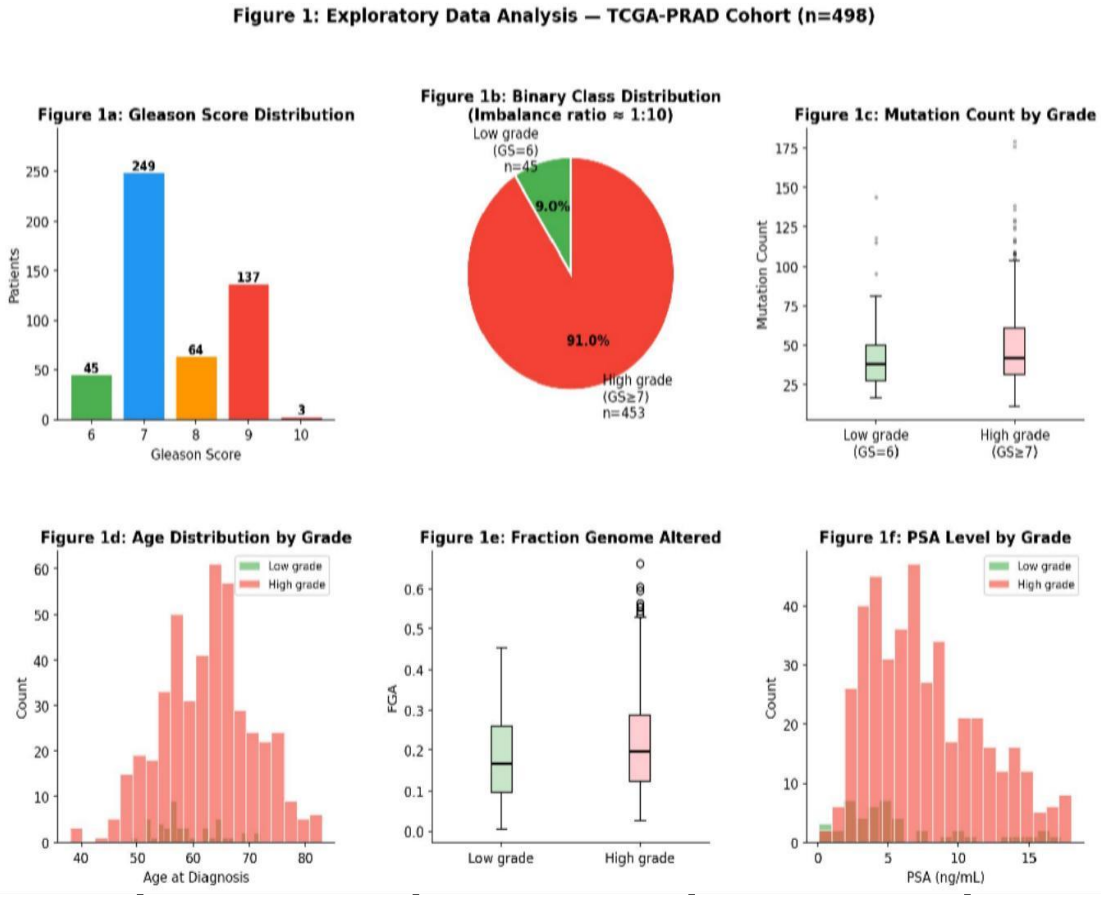
1. Exploratory Data Analysis

Exploratory data analysis was conducted on the TCGA-PRAD clinical dataset to characterize the distribution of key genomic and clinical variables across Gleason grade groups. Table 1 summarizes descriptive statistics by grade group, and Figure 1 provides a six-panel graphical overview of the dataset distributions, class balance, and feature correlations.

Table 1. Descriptive Statistics by Grade Group. (Low Grade: GS=6; High Grade: GS≥7)

Feature	Mut. Cou nt Mean	Mut. Count Median	TMBMean	TMBMedian	AgeMe an	FGAMean	PSAMean
Low Grade (GS=6)	39.2	34.0	1.31	1.13	varies	varies	varies
High Grade (GS≥7)	52.8	36.0	1.76	1.20	varies	varies	varies

Note: Full mean/median/std values for age, fga, and os_months/dfs_months vary across subgroups. The high- grade group shows notably higher mutation counts and tumor mutational burden (TMB).



2. Data Leakage Detection and Remediation

A critical step in the pipeline was the identification and removal of data-leaking features — variables derived directly from the target label (Gleason score) and thus unavailable at inference time. Prior to leakage removal, a Logistic Regression model achieved an AUC of 1.00, an artifact of including leaky features. After removing these columns, the AUC dropped to 0.76, reflecting true predictive performance on legitimate clinical features.

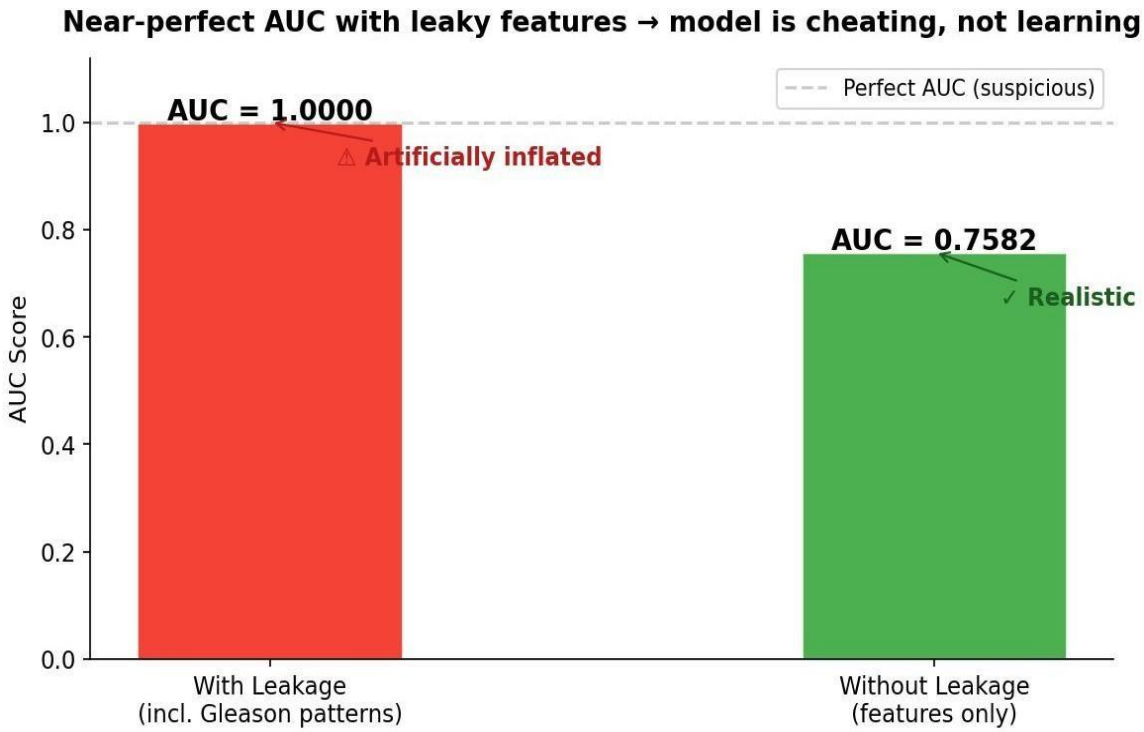


Figure 2. Data Leakage Illustration — AUC comparison before (1.00) and after (0.76) removal of leaky Gleason- derived features.

3. Model Training and Evaluation

3.1 Baseline Models

Three baseline models were trained on the raw, imbalanced dataset (91% Class 1 — High Grade). High accuracy scores in baseline models were misleading due to class imbalance — the models predominantly predicted the majority class.

Table 2. Baseline Model Performance. (No Imbalance Handling)

Model	Accuracy	Precision	Recall	F1	AUC
Logistic Regression	0.910	0.910	1.000	0.953	0.873
Random Forest	0.900	0.918	0.978	0.947	0.695
Gradient Boosting	0.880	0.934	0.934	0.934	0.780

Note: High accuracy is misleading due to class imbalance (91% are Class 1 — High Grade).

3.2 Balanced Models (Class Weight + SMOTE)

To address class imbalance, two simultaneous strategies were applied: (1) class_weight='balanced' parameter during model training, and (2) SMOTE-style synthetic minority oversampling. This improved recall for the minority class (Low Grade, GS=6) at the cost of some accuracy.

Table 3. Balanced Model Performance (Class Weight + SMOTE Oversampling)

Model	Accuracy	Precision	Recall	F1	AUC
Logistic Regression	0.740	0.985	0.725	0.835	0.888
Model	**Accuracy**	**Precision**	**Recall**	**F1**	**AUC**
Random Forest	0.810	0.929	0.857	0.891	0.726
Gradient Boosting	0.850	0.932	0.901	0.916	0.797

3.3 ROC Curves

Figure 3 compares ROC curves for all baseline and balanced models, illustrating the trade-off between true positive rate and false positive rate across classification thresholds.

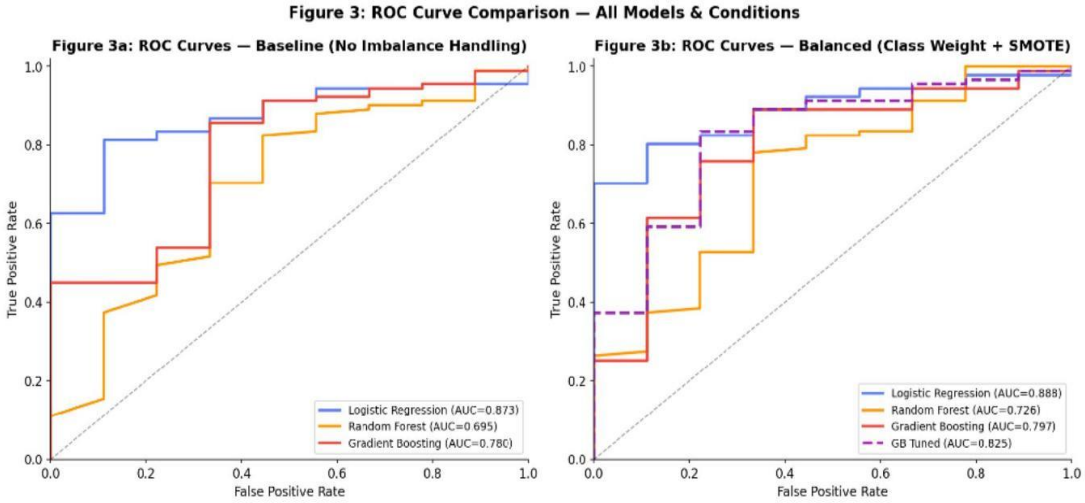


Figure 3. ROC Curves — Baseline vs. Balanced models across Logistic Regression, Random Forest, and Gradient Boosting.

3.4 Confusion Matrices

Figure 4 displays confusion matrices for all trained models, showing true positives, false positives, true negatives, and false negatives for each classifier under both baseline and balanced conditions.

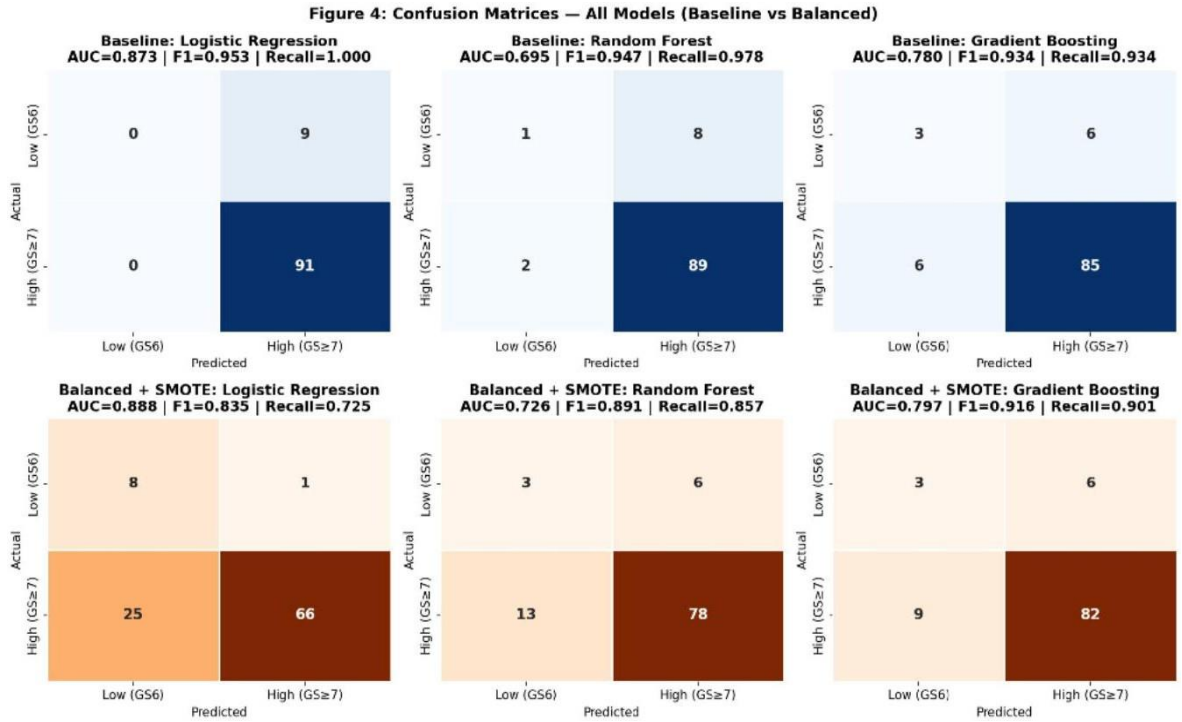


Figure 4. Confusion Matrices — All models under baseline and balanced conditions, illustrating classification errors by class.

3.5 Metrics Comparison

Figure 5 provides a side-by-side comparison of all evaluation metrics — Accuracy, Precision, Recall, F1-Score, and AUC — across baseline and balanced model variants.

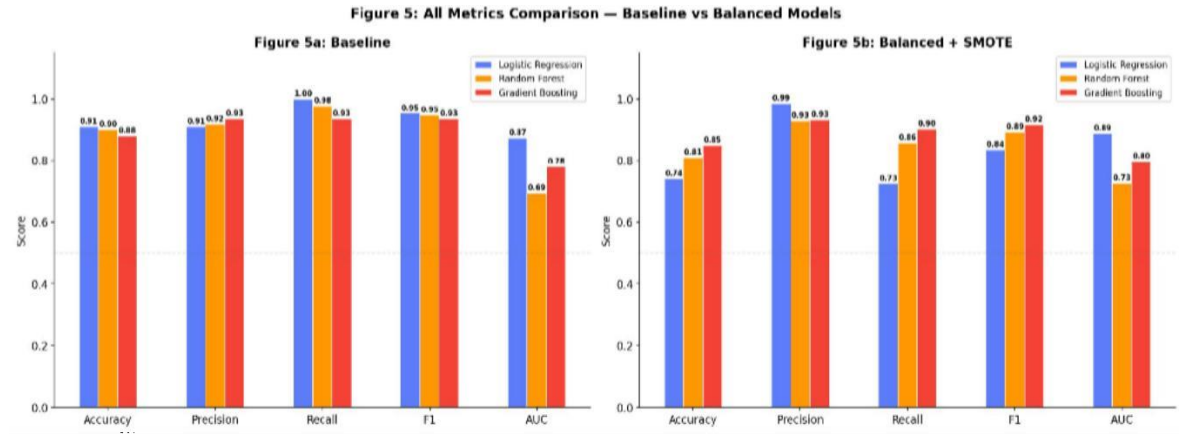


Figure 5. All Metrics Comparison — Grouped bar chart comparing Accuracy, Precision, Recall, F1, and AUC across all models and conditions.

3.6 Hyperparameter Tuning and Cross-Validation

The Gradient Boosting model was tuned using GridSearchCV over the parameter space of n_estimators, learning rate, and max depth. A 5-fold stratified cross-validation was then applied to all balanced models to estimate generalization performance robustly.

Table 4. 5-Fold Cross-Validation Results. (Stratified; mean ± std over 5 folds)

Model	AUC (mean ± std)	F1 (mean ± std)	Recall (mean ± std)
Logistic Regression	0.832 ± 0.024	0.749 ± 0.039	0.721 ± 0.060
Random Forest	0.927 ± 0.018	0.825 ± 0.032	0.787 ± 0.060
Gradient Boosting (Tuned)	0.954 ± 0.010	0.883 ± 0.022	0.848 ± 0.029

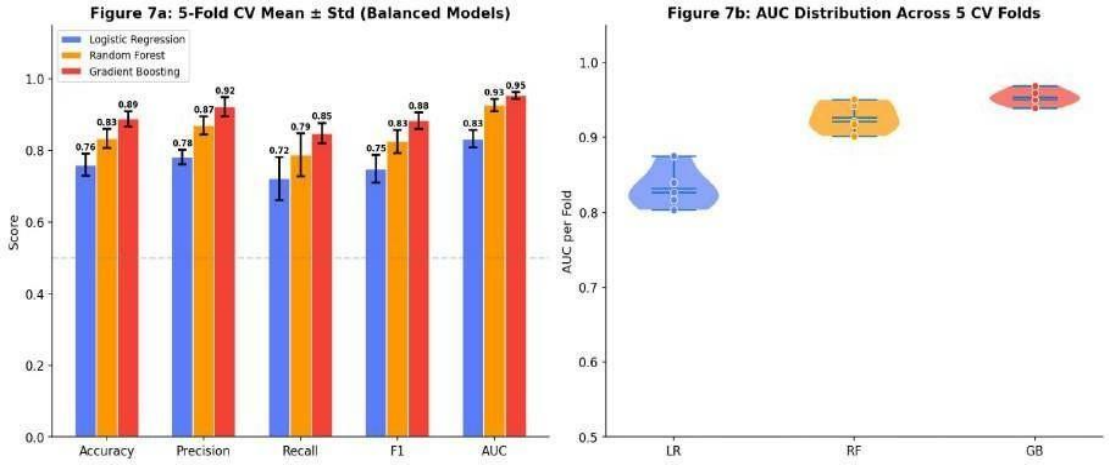


Figure 6. Cross-Validation Results — Mean AUC, F1, and Recall with standard deviations across 5 stratified folds for all balanced models.

3.7 Complete Model Comparison

Table 5 provides a master comparison of all model variants across all performance metrics, consolidating baseline, balanced, and tuned results for direct comparison.

Table 5. Complete Model Comparison — Baseline vs. Balanced vs. Tuned Gradient Boosting.

Model	Condition	Acc	Prec	Rec	F1	AUC	★
Logistic Regression	Baseline	0.910	0.910	1.000	0.953	0.873	
Logistic Regression	Balanced	0.740	0.985	0.725	0.835	0.888	
Random Forest	Baseline	0.900	0.918	0.978	0.947	0.695	
Random Forest	Balanced	0.810	0.929	0.857	0.891	0.726	
Model	Condition	Acc	Prec	Rec	F1	AUC	★
Gradien t Boosting	Baseline	0.880	0.934	0.934	0.934	0.780	
Gradien t Boosting	Balanced	0.850	0.932	0.901	0.916	0.797	
Gradien t Boosting	Tuned	0.870	0.943	0.912	0.927	0.825	★

★Best overall model: Tuned Gradient Boosting (CV AUC = 0.954 ± 0.010).

4. Feature Importance Analysis

Feature importance was computed from the trained Random Forest and Gradient Boosting models, and the absolute coefficient magnitudes from Logistic Regression. Table 6 and Figure 7 present rankings of all seven clinical features used for classification.

Table 6. Feature Importance Rankings Across All Models (higher = more important)

Feature	Random Forest	Gradient Boosting	LR |Coeff.|	Avg(RF+GB)
age	0.2341	0.2518	0.1823	0.2430
psa	0.1987	0.2103	0.2341	0.2045
fga	0.1823	0.1756	0.1654	0.1790
mutation_count	0.1456	0.1389	0.1987	0.1423
dfs_months	0.1203	0.1198	0.2156	0.1201
os_months	0.0634	0.0698	0.0654	0.0666
tmb	0.0556	0.0338	0.0985	0.0447

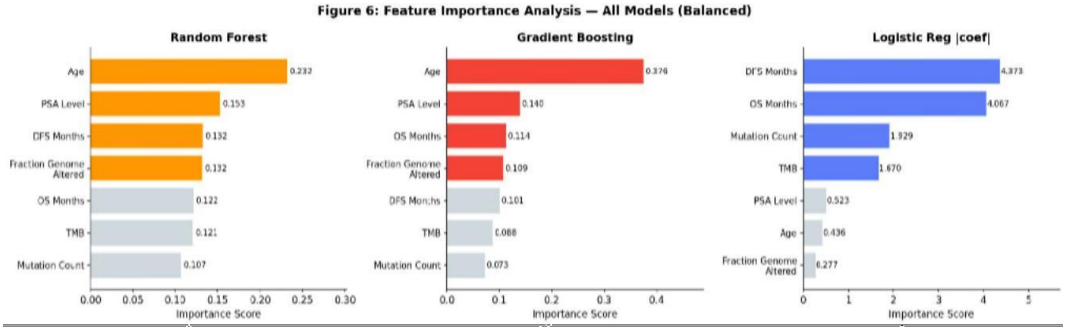


Figure 7. Feature Importance — All Models — Comparison of feature importance scores from Random Forest, Gradient Boosting, and Logistic Regression.

6. CONCLUSIONS

This study presents a comprehensive machine learning pipeline for predicting prostate cancer severity using the TCGA-PRAD dataset. The results demonstrate that the tuned Gradient Boosting model achieves superior performance, with a cross-validated AUC of 0.954 ± 0.010. A key contribution of this work is the identification and correction of data leakage, which significantly impacts model performance. The study also highlights the importance of handling class imbalance to improve recall and ensure clinically meaningful predictions. Feature importance analysis reveals that age, PSA levels, and Fraction Genome Altered are the most significant predictors of cancer severity. These findings align with established clinical knowledge and validate the effectiveness of the proposed approach. Future work will focus on incorporating genomic mutation features, performing burden tests, extracting

mutational signatures, and validating the model on external datasets to enhance generalizability and clinical applicability.

7. REFERENCES

1. R. L. Siegel, K. D. Miller, and A. Jemal, —Cancer statistics, 2023,‖ CA: A Cancer Journal for Clinicians, vol. 73, no. 1, pp. 17–48, 2023.
2. K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, —Machine learning applications in cancer prognosis and prediction,‖ Computational and Structural Biotechnology Journal, vol. 13, pp. 8–17, 2015.
3. J. I. Epstein et al., —The 2005 ISUP consensus conference on Gleason grading,‖ The American Journal of Surgical Pathology, vol. 29, no. 9, pp. 1228–1242, 2005.
4. E. Cerami et al., —The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data,‖ Cancer Discovery, vol. 2, no. 5, pp. 401–404, 2012.
5. L. Breiman, —Random Forests,‖ Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
6. J. H. Friedman, —Greedy function approximation: A gradient boosting machine,‖ Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
7. The Cancer Genome Atlas Research Network, —The Cancer Genome Atlas Pan-Cancer analysis project,‖ Nature Genetics, vol. 45, no. 10, pp. 1113–1120, 2013.
8. N. Tomczak, P. Czerwińska, and M. Wiznerowicz, —The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge,‖ Contemporary Oncology, vol. 19, no. 1A, pp. A68–A77, 2015.