Optimization of Prostate Cancer Biopsy Decision-Making in Korean Patients Using Machine Learning

$Ji-Eun\ An^{1}, Minji\ Noh^{2}, Sanghwa\ Lee^{3}, Jaeheung\ Jeong^{4}, Su-Yeon\ Yu^{1}$

$Department\ of\ Medical\ Health, School\ of\ Nursing\ and\ Health, Kongju\ National\ University^{1}$

$Department\ of\ Big\ Data\ Medical\ Convergence, Kangwon\ National\ University^{2}$

$Department\ of\ Data\ Science, Kangwon\ National\ University^{3}$

$Department\ of\ Urology, Wonju\ Severance\ Christian\ Hospital^{4}$

Background

- According to the National Health Insurance statistics, prostate cancer incidence in Korea has tripled over the past decade, ranking as the second most common malignancy in men aged ≥65 years.
- Prostate-Specific Antigen (PSA) screening facilitates early detection but suffers from low specificity, leading to unnecessary biopsies and associated complications

Objective

- This study aims to identify prostate cancer biomarkers and develop predictive models for cancer diagnosis and biopsy-related complications in Korean patients.

Methods

- This retrospective cohort study analyzed data from 17,530 Korean male patients who underwent prostate biopsies (2011–2019) across six tertiary hospitals.
- Variables included sociodemographic (age, smoking, family history), clinical (PSA, Gleason Score, TNM stage, DRE, MRI), and outcome data (cancer diagnosis, biopsy complications).
- Statistical analyses (chi-square, t-test, ANOVA, LASSO regression) identified significant predictors.
- Predictive models (logistic regression, SVM, Random Forest, XGBoost, MLP, soft voting ensemble) were trained using an 80:20 dataset split with 10-fold cross-validation and hyperparameter tuning.

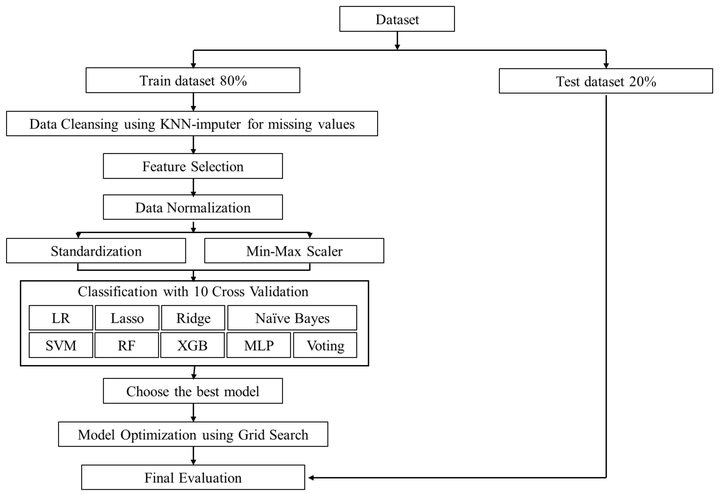

Figure 1. Flow chart of the development of predictive models

- The dashboard was developed using Flask, Chart.js, and ChatGPT APIs.
- The program used is SAS 9.4 and Python 3.11.

Results

(1) Biomarkers

- Age, Smoking, Family History of Other Cancers, Family History of Prostate Cancer, Comorbidity within 1 Year, Hypertension, Diabetes, Cancer, Other Diseases, Prostate Density, Total PSA were significant predictors.

Table 2. Significant Variables in LASSO Logistic Regression Model

Variable	Beta	SE	z	p-value
Intercept	1.377672	0.064309	21.42284	8.18E-102
Age	0.707930	0.021063	33.60999	1.19E-247
Smoking	0.130044	0.022340	5.82118	5.84E-09
Family History of Other Cancers	-0.278140	0.035056	-7.98419	2.11E-15
Family History of Prostate Cancer	0.176144	0.033002	5.337388	9.43E-08
Comorbidity within 1 Year	-0.149290	0.041898	-3.55554	3.77E-04
Hypertension	0.136887	0.031226	4.383721	1.67E-05
Diabetes	0.137715	0.037111	3.710084	2.06E-04
Cancer	-0.308870	0.055084	-5.5166	3.45E-08
Other Diseases	-0.154840	0.033200	-4.65831	3.10E-06
Prostate Density	7.337309	0.632895	11.59325	4.46E-31
Total PSA	13.129210	0.595190	22.05884	7.86E-108

(2) Predictive model

- The soft voting ensemble achieved the best diagnostic performance (AUC 0.840, precision 0.863, F1-score 0.647).
- For borderline PSA patients, performance declined but improved with additional features.
- Biopsy complication prediction was limited by class imbalance, yet combining Random Over Sampling with focal loss improved recall (0.476) and AUC (0.660).

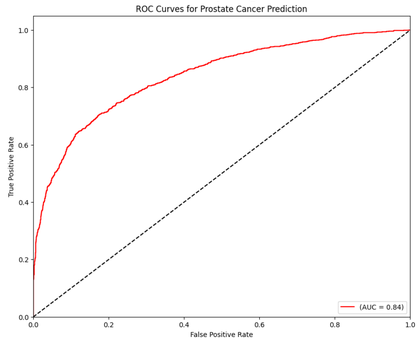

Figure 2. Prostate Cancer Diagnostic Prediction Model ROC Curve

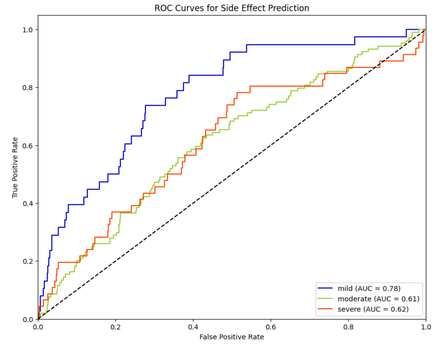

Figure 3. Biopsy Side Effects Prediction Model ROC Curve

(3) Dash board

- A web-based dashboard integrated SHAP analysis and ChatGPT for interpretability.

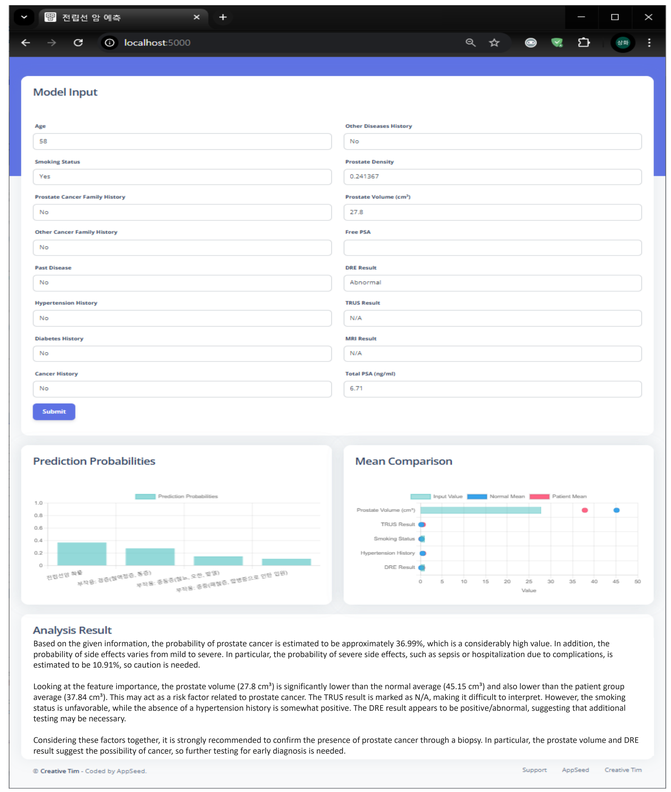

Figure 4. Dashboard Screen

Conclusion & Discussion

- This study supports prostate cancer diagnosis and biopsy decision-making in Korean patients.
- Future work should refine predictive models for complication severity stratification.

This research was supported by a grant of Patient-Centered Clinical Research Coordinating Center (PACEN) funded by the Ministry of Health & Welfare, Republic of Korea (grant number : RS-2025-02304176).

E-mail: je.an@kangwon.ac.kr