2023

Detection of Prostate Cancer Using Machine Learning Techniques: An Exploratory Study

Laxmi Manasa Gorugantu

Omar El-Gayar

Nevine Nawar

Detection of Prostate Cancer Using Machine Learning Techniques: An Exploratory Study

Laxmi Manasa Gorugantu
Dakota State University
laxmi.gorugantu@trojans.dsu.edu

Omar El-Gayar
Dakota State University
omar.el-gayar@dsu.edu

Nevine Nawar
Alexandria University
nevine.nawar@gmail.com

ABSTRACT

Prostate Cancer (PCa) is one of the most frequent cancers worldwide and the most common cancer in males. Testing for PCa remains problematic. Evidence is mounting that overdiagnosis and over-treatment can result in adverse side-effects yet have little impact in preventing death from PCa. Consequently, the importance of predictive tools that help physicians in the diagnosis of the condition cannot be understated. Though there exist several predictive models for the detection of clinically significant PCa, these models mainly depend on logistic regression. The objective of this research is to investigate the potential of various machine learning techniques to improve the sensitivity and specificity of detecting clinically significant PCa. Risk factors considered include prostate-specific antigen (PSA), digital rectal examination (DRE), as well as age, race/ethnicity, and family history. According to the results, Logistic Regression has outperformed all the models followed by Random Forest, SVM and XG Boost.

Keywords

Machine Learning Techniques, Data Analysis, Data pre-processing, Prostate Cancer, Prognosis.

INTRODUCTION

Prostate cancer is the cancer occurring in the prostate gland in men. It has been noted as the second most frequent cancer worldwide and the most common cancer in males around 84 countries, and is rapidly increasing in developed countries. The statistics show that every year around 240,000 US men are being diagnosed with prostate cancer and an average of 30,000 men succumb to prostate cancer every year (Liu et al., 2019). Prostate cancer can be a slow growing, low grade or insignificant cancer. Yet, significant numbers of positive cases and mortality rates indicate that PCa can also be devasting based on the aggressiveness of the disease (Liu et al., 2019). Significant prostate cancers that are aggressive have the ability to metastasize leading to higher death rates since predicting and treating metastatic cancer is complex and may not be effective (Wilbur, 2008). Therefore, it is necessary to detect prostate cancer while it is still confined to the prostate gland. PCa can be diagnosed through various screening methods where DRE and PSA tests are the most widely used, easy to administer, and relatively low-cost screening tools for early prediction (Wilbur, 2008). However, prostate biopsy remains the reference standard for PCa detection. Though biopsy is essential to confirm the presence of cancer, it also associated with different health complications such as infection, rectal bleeding, numbness, pain, and unnecessary financial expenses (Liu et al., 2019). Therefore, it is essential to minimize avoidable repeated biopsies by making screening tests efficient enough that they only lead to true positive prostate biopsy results. Many predictive tools of PCa are present in the literature that mainly rely on logistic regression (Chun et al., 2007; Finne et al., 2004; Karakiewicz et al., 2005; Roobol et al., 2010). In this research we aim to complement prior research by exploring various machine learning algorithms and their potential to accurately detect prostate cancer.

LITERATURE REVIEW

Prostate cancer may be the result of socioeconomic, genetic, dietary, medical, and environmental factors. (Lynch et al., 2020). However, the strongest risk factors include age, ethnicity, and family history of the disease (Tikkinen et al., 2018). For instance, black men have more than twice the risk of developing and dying from prostate cancer than white men. Moreover, black men with family history of prostate cancer have two-seven fold increased risk of developing the condition (Lynch et al., 2020). Age

presents as a strong risk factor for prostate cancer. Studies related to age-specific prostate cancer show that the likelihood of developing PCa before the age of 40 is extremely low, starts to rise around the age of 55, spikes between 70-74 and declines thereafter. The rate of being diagnosed with PCa increases approximately by 9th – 10th power of age, which is greater than any other cancer (Lichtenstein, 2000). Family history is another important risk factor for prostate cancer. It involves a combination of inherited genetics and shared environmental exposure (Albright et al., 2015). Lichtenstein (2000) study from Sweden, Denmark and Finland mentions that genetic risk accounts for 42% of complete familial risk in PCa, which is much higher than any other cancer. Further, Albright (2015) states that people with first-degree family history have an elevated risk of developing the disease. However, the evidence for being diagnosed with PCa based on any specific factor has not been very consistent and risk factors differ for various population groups (Giovannucci et al., 2007). Therefore, along with considering the potential risk factors, it is necessary to diagnose the cancer at early stages through standard prostate cancer screening methods for a definite result. Prostate biopsy, which confirms the presence of cancer based on Gleason Score is considered the reference standard for prostate cancer detection. But without efficient screening, prostate biopsy has proved to be an overdiagnosis and overtreatment approach (Vos et al., 2013). There are several prediction models using PSA and other risk factors for detecting clinically significant PCa. These models tend to predominantly rely on multivariate logistic regression (Chun et al., 2007; Finne et al., 2004; Karakiewicz et al., 2005; Roobol et al., 2010). Alternately, this research aims to explore various machine learning techniques and their potential to improve the prediction of prostate cancer.

METHODOLOGY

Selection of Biomarkers

Based on the Literature review, the finalized set of biomarkers that increases the risk of incidence of prostate cancer include, age, race, personal and family history of PCa, personal and family history of any cancer, PSA and DRE results (Chun et al., 2007; Finne et al., 2004; Karakiewicz et al., 2005; Nam et al., 2007; Thompson et al., 2006).

Exploratory Data Analysis

The Prostate, Lung, Colorectal, and Ovarian cancer (PLCO) was a randomized controlled screening trial that was designed and sponsored by National Cancer Institute (NCI), a unit of the National Institute of Health (NIH). Around 155,000 participants were registered for the trial between November 1993 and July 2001, these participants were randomized on a 50/50 basis into control arm and intervention arms. Participants in the control arm received regular care whereas those in the intervention arm received trial-provided screening tests for prostate, lung, colorectal and ovarian cancers. A total of (n= 76,678) participants were registered under the prostate cancer trial. Out of which, (n=38,340) participants were randomized to the intervention arm and (n=38,338) participants to the control arm. Those in the intervention arm received Digital Rectal Examination (DRE) and Prostate Specific Antigen (PSA) test at scheduled intervals, while those in the control arm did not receive DRE or PSA screening examinations. The further analysis of this study included all the participants from the intervention arm as they underwent prostate cancer screening exams and eliminated the participants belonging to control group from further analysis because the individuals in this arm did not undergo any screening. Accordingly, From the intervention group (n=38,340) participants, a total of (n=4579) participants were diagnosed with positive prostate cancer. Whereas, remaining (n= 33,761) participants were diagnosed with no prostate cancer. For the purposes of prostate cancer risk modeling, pros_cancer is considered as dependent variable, that depicts the presence the prostate cancer and classified into two categories namely, 0 representing “No Cancer” and 1 representing “Cancer”. Further, the independent variables from the PLCO dataset according to the biomarkers from the literature, are as follows, age, race, family history of PCa, personal history of PCa, family history of any cancer, personal history of any cancer, DRE and PSA results. We further performed data pre-processing in order to deal with categorical and continuous variables and the final dataset now consists of 15 columns and 34735 rows namely, pros_cancer, age, PSA, race7_2, race7_3, race7_4, race7_5, race7_6, pros_fh_1, pros_fh_9, fh_cancer_1, ph_any_trial_1, DRE_2.0, DRE_3.0.

Model Development

We investigated six predictive models to detect the presence of PCa, namely, Logistic Regression, K-nearest neighbors, Naïve Bayes, Support Vector Machines, Random Forest and XG Boost. The dataset was divided into 80% - 20% as train and test splits respectively, out of the 34,735 rows of the whole dataset, the training dataset included 27,788 rows. Most (87.9%) of the training dataset consisted of “Class 0” category for the dependent variable, which was 24,414 rows and only 12.1% consisted of “Class 1” category, which was only 3,374 rows. In order to overcome the skewed proportions of the dataset we employed Synthetic Minority Oversampling Technique (SMOTE) to address the class imbalance problem. As the dataset was highly imbalanced, we also implemented Stratified K- Cross Validation, where (k=5) to ensure that the training and validation datasets

had the same proportions of the feature interest as in the original dataset. Moreover, we also performed hyperparameter tuning on the training dataset to enhance the learning and performance of the model by choosing optimal hyper parameters. As the implementation of SMOTE and hyper parameter is an iterative process for each fold in Stratified K-Cross Validation, we further employed the pipeline function of scikit-learn to integrate the steps of the machine learning workflow.

RESULTS AND DISCUSSIONS

Table 1 represents the best hyperparameters for each classifier on which the model was trained and Table 2 represents the results of performance measures of the classifiers on the test dataset. Logistic Regression was the standard model for the prediction of prostate cancer in the literature, and according to the results, Logistic Regression outperformed all other models followed by XG Boost based on the AUC scores.

Model	Parameters
Logistic Regression	C=0.01, penalty = L2, solver = newton-cg
K-NN	Metric = Manhattan, n_neighbors = 4, weights = uniform
Naïve Bayes	Var_smoothing = 0.0002651083601908539
SVM	C=0.01, gamma = 0.001, kernel = rbf
Random Forest	Max_depth = 3, min_samples_leaf = 10, n_estimators = 1000
XGBoost	Learning_rate = 0.1, max_depth = 6, n_estimators = 780

Table 1. Best Hyper parameters

Classifier	CV Score	Accuracy	Precision	Recall	F1-Score	Specificity	Sensitivity	AUC Score
Logistic Regression	89.6%	88.8%	0.532	0.856	0.657	0.892	0.856	0.934
K-NN	89.3%	88.7%	0.887	0.887	0.887	0.912	0.711	0.873
Naïve Bayes	90.8%	91.0%	0.728	0.449	0.556	0.975	0.449	0.909
SVM	90.5%	89.4%	0.553	0.808	0.658	0.906	0.808	0.922
Random Forest	89.6%	88.8%	0.534	0.843	0.654	0.894	0.843	0.919
XG Boost	90.9%	89.8%	0.567	0.789	0.660	0.913	0.789	0.931

Table 2. Evaluation metrics for each classifier

Further, figure 1 represents the confusion matrices for the classifiers. Considering the False Negative (FN) results from the confusion matrices which depict positive cancer patients being classified as negative, Logistic Regression had the lowest FN rates, followed by Random Forest, and SVM classifier. On the other hand, false positive rates for classifiers LR, SVM and RF which represent negative prostate cancer patients being predicted as positive, were relatively quite high in comparison to other classifiers. Moreover, Naïve Bayes classifier had the least False Positive rate and the highest False Negative rate. Though the classifier had the lowest False positive rate, it cannot be considered as one of the best classifiers due to high FN rate. Further, the AUC scores and accuracies for classifier – LR, XGB, RF and SVM were almost similar. Though, the AUC score for XG Boost was the second highest after Logistic Regression, Random Forest and SVM classifier are more preferrable due to lower FN rates. Therefore, the analysis proves that Logistic Regression followed by Random Forest, SVM and XG Boost classifiers as the best models.

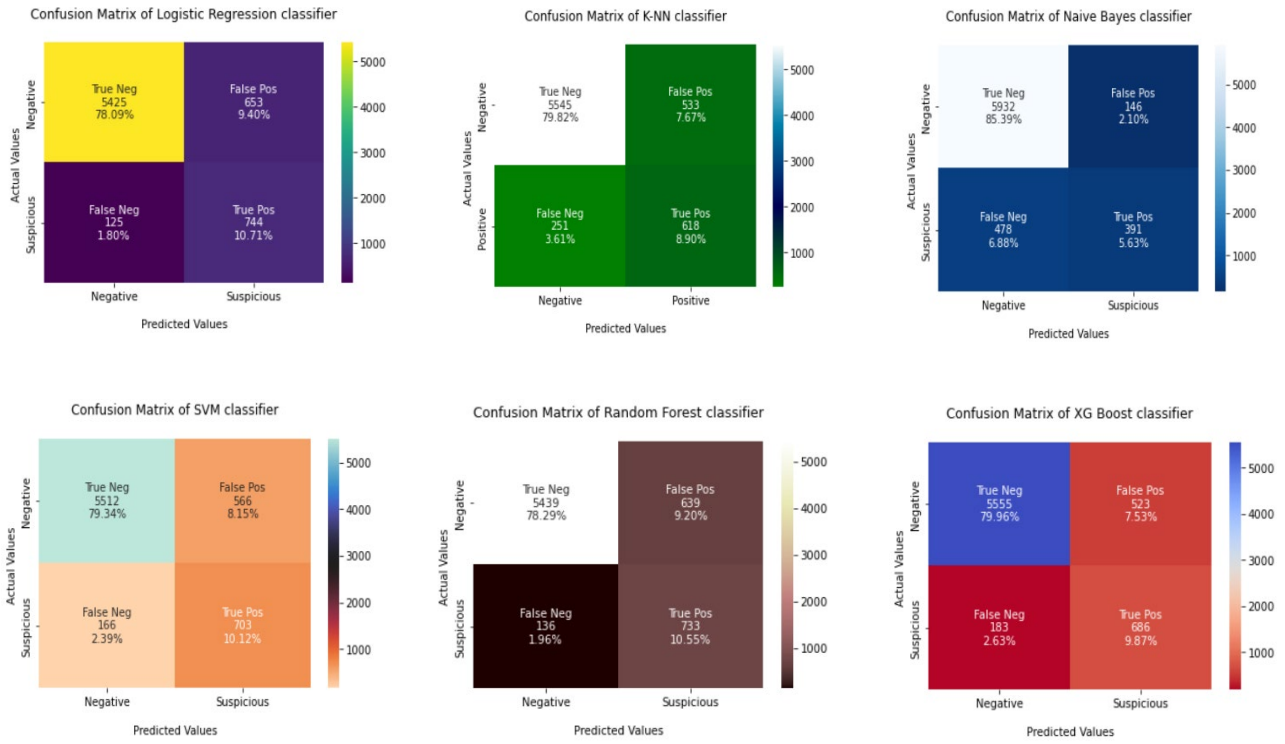

Figure 1. Confusion Matrices

CONCLUSION

The findings of this study are consistent with the literature, Logistic Regression performed better than all the models followed by Random Forest, SVM and XG Boost. Accordingly, the literature has also employed Logistic Regression as the standard model for predicting the presence of prostate cancer. Future research is needed to develop models to predict the presence of PCa not only with a limited set of predictors but with more variables that are believed to be important risk factors such as underlying health conditions and lifestyle habits (Saran et al., 2021).

REFERENCES

1. Albright, F., Stephenson, R. A., Agarwal, N., Teerlink, C. C., Lowrance, W. T., Farnham, J. M., & Albright, L. A. C. (2015). Prostate cancer risk prediction based on complete prostate cancer family history. *The Prostate*, *75*(4), 390–398. https://doi.org/10.1002/pros.22925

2. Chun, F. K.-H., Briganti, A., Graefen, M., Montorsi, F., Porter, C., Scattoni, V., Gallina, A., Walz, J., Haese, A., Steuber, T., Erbersdobler, A., Schlomm, T., Ahyai, S. A., Currlin, E., Valiquette, L., Heinzer, H., Rigatti, P., Huland, H., & Karakiewicz, P. I. (2007). Development and External Validation of an Extended 10-Core Biopsy Nomogram. *European Urology*, *52*(2), 436–445. https://doi.org/10.1016/j.eururo.2006.08.039

3. Finne, P., Finne, R., Bangma, C., Hugosson, J., Hakama, M., Auvinen, A., & Stenman, U.-H. (2004). Algorithms based on prostate-specific antigen (PSA), free PSA, digital rectal examination and prostate volume reduce false-postitive PSA results in prostate cancer screening. *International Journal of Cancer*, *111*(2), 310–315. https://doi.org/10.1002/ijc.20250

4. Gann, P. H. (2002). Risk Factors for Prostate Cancer. *Reviews in Urology*, *4*(Suppl 5), S3–S10.

5. Giovannucci, E., Liu, Y., Platz, E. A., Stampfer, M. J., & Willett, W. C. (2007). Risk factors for prostate cancer incidence and progression in the health professionals follow-up study. *International Journal of Cancer*, *121*(7), 1571–1578.

6. Karakiewicz, P. I., Benayoun, S., Kattan, M. W., Perrotte, P., Valiquette, L., Scardino, P. T., Cagiannos, I., Heinzer, H., Tanguay, S., Aprikian, A. G., Huland, H., & Graefen, M. (2005). Development and validation of a nomogram predicting the outcome of prostate biopsy based on patient age, digital rectal examination and serum prostate specific antigen. *Journal of Urology*, *173*(6), 1930–1934. https://doi.org/10.1097/01.ju.0000158039.94467.5d

7. Lichtenstein, P. (2000, July 13). *Environmental and Heritable Factors in the Causation of Cancer—Analyses of Cohorts of Twins from Sweden, Denmark, and Finland | NEJM*. https://www.nejm.org/doi/full/10.1056/NEJM200007133430201

8. Liu, B., Cheng, J., Guo, D. J., He, X. J., Luo, Y. D., Zeng, Y., & Li, C. M. (2019). Prediction of prostate cancer aggressiveness with a combination of radiomics and machine learning-based analysis of dynamic contrast-enhanced MRI. *Clinical Radiology*, *74*(11), 896.e1-896.e8. https://doi.org/10.1016/j.crad.2019.07.011

9. Lynch, S. M., Handorf, E., Sorice, K. A., Blackman, E., Bealin, L., Giri, V. N., Obeid, E., Ragin, C., & Daly, M. (2020). The effect of neighborhood social environment on prostate cancer development in black and white men at high risk for prostate cancer. *PLOS ONE*, *15*(8), e0237332. https://doi.org/10.1371/journal.pone.0237332

10. Nam, R. K., Toi, A., Klotz, L. H., Trachtenberg, J., Jewett, M. A. S., Appu, S., Loblaw, D. A., Sugar, L., Narod, S. A., & Kattan, M. W. (2007). Assessing Individual Risk for Prostate Cancer. *Journal of Clinical Oncology*, *25*(24), 3582–3588. https://doi.org/10.1200/JCO.2007.10.6450

11. Roobol, M. J., Steyerberg, E. W., Kranse, R., Wolters, T., van den Bergh, R. C. N., Bangma, C. H., & Schröder, F. H. (2010). A Risk-Based Strategy Improves Prostate-Specific Antigen–Driven Detection of Prostate Cancer. *European Urology*, *57*(1), 79–85. https://doi.org/10.1016/j.eururo.2009.08.025

12. Saran, U., Chandrasekaran, B., Kolluru, V., Tyagi, A., Nguyen, K. D., Valadon, C. L., Shaheen, S. P., Kong, M., Poddar, T., Ankem, M. K., & Damodaran, C. (2021). Diagnostic molecular markers predicting aggressive potential in low-grade prostate cancer. *Translational Research*, *231*, 92–101. https://doi.org/10.1016/j.trsl.2020.11.014

13. Thompson, I. M., Ankerst, D. P., Chi, C., Goodman, P. J., Tangen, C. M., Lucia, M. S., Feng, Z., Parnes, H. L., & Coltman, C. A., Jr. (2006). Assessing Prostate Cancer Risk: Results from the Prostate Cancer Prevention Trial. *JNCI: Journal of the National Cancer Institute*, *98*(8), 529–534. https://doi.org/10.1093/jnci/djj131

14. Tikkinen, K. A. O., Dahm, P., Lytvyn, L., Heen, A. F., Vernooij, R. W. M., Siemieniuk, R. A. C., Wheeler, R., Vaughan, B., Fobuzi, A. C., Blanker, M. H., Junod, N., Sommer, J., Stirnemann, J., Yoshimura, M., Auer, R., MacDonald, H., Guyatt, G., Vandvik, P. O., & Agoritsas, T. (2018). Prostate cancer screening with prostate-specific antigen (PSA) test: A clinical practice guideline. *BMJ*, *362*, k3581. https://doi.org/10.1136/bmj.k3581

15. Vos, E. K., Litjens, G. J. S., Kobus, T., Hambrock, T., Kaa, C. A. H. de, Barentsz, J. O., Huisman, H. J., & Scheenen, T. W. J. (2013). Assessment of Prostate Cancer Aggressiveness Using Dynamic Contrast-enhanced Magnetic Resonance Imaging at 3 T. *European Urology*, *64*(3), 448–455. https://doi.org/10.1016/j.eururo.2013.05.045

16. Wilbur, J. (2008). Prostate Cancer Screening: The Continuing Controversy. *American Family Physician*, *78*(12), 1377–1384.