Diagnosis of prostate cancer in a Chinese population by using machine learning methods

Guanjin Wang[1], Jeremy Yuen-Chun Teoh[2] and Kup-Sze Choi[3]

Abstract— **An early diagnosis of prostate cancer (PC) is key for the successful treatment. Although invasive prostate biopsies can provide a definitive diagnosis, the number of biopsies should be reduced to avoid side effects and risks especially for the men with the low risk of cancer. Therefore, an accurate model is in need to predict PC with the aim of reducing unnecessary biopsies. In this study, we developed predictive models using four machine learning methods including Support Vector Machine (SVM), Least Squares Support Vector Machine (LS-SVM), Artificial Neural Network (ANN) and Random Forest (RF) to detect PC cases using available prebiopsy information. The models were constructed and evaluated on a cohort of 1625 Chinese men with prostate biopsies from Hong Kong hospital. All the models have the excellent performances in detecting significant PC cases, with ANN achieving the highest accuracy of 0.9527 and the AUC value of 0.9755. RF outperformed the other three methods in classifying benign, significant and insignificant PC cases, with an accuracy of 0.9741 and a F1 score of 0.8290.**

I. INTRODUCTION

In the United States, prostate cancer (PC) is the most common malignancy in men[1]. Prostate-specific antigen (PSA) level is widely recognized as an early screening tool for the diagnosis of PC [2], [3]. Traditionally, the presence of an elevated PSA level or an abnormal digital rectal examination (DRE) finding is associated with the higher risk of PC, which leads to a decision to perform prostate biopsy [4], [5]. However, biopsies may bring the side effects and risks. Moreover, the majority of detected PCs were insignificant which did not affect patients' survival in long-run [6]. Therefore, it is important to reduce unnecessary biopsies and at the same time guarantee the most important PC cases can be detected. Several risk calculators have been developed in the Caucasian patient population, such as the European Randomized Study of Screening for Prostate Cancer (ERSPC) risk calculator [7], the Prostate Cancer Prevention Trial (PCPT) risk calculator [8] and the Sunnybrook risk calculator [9]. Nevertheless, Chinese men are genetically and physiologically different from the Caucasians. An accurate diagnostic model of PC for the Chinese population is in demand.

In this study, we attempted to construct four predictive models using traditional machine learning methods including Support Vector Machine (SVM), Least Squares Support Vector Machine (LS-SVM), Artificial Neural Network (ANN) and Random Forest (RF) to detect PC cases while avoiding unnecessary biopsies. In addition to PSA level, other available prebiopsy information is incorporated as inputs in the model construction. The prediction performances of four machine learning methods are evaluated and compared using several metrics.

II. METHODS

A. The real-world PC dataset

The PC cohort used in this study was retrieved from a transrectal ultrasound (TRUS)-guided prostate biopsy database in a Hong Kong hospital. The cohort consists of 1625 Chinese patient records that all men had TRUS biopsy performed. 258 (15.88%) of them are healthy, 539 (33.17%) have insignificant PC and 828 (50.95%) have significant PC. In addition to the PSA level, relevant prebiopsy information including age, results of DRE and TRUS, and prostate volume are also included. The missing data in the cohort were filled using the k-NN imputation method. Table I lists the baseline characteristics of the cohort in detail.

The aim of this study is to build predictive models to diagnose prostate cancer using available prebiopsy information in addition to PSA level. The experiments consist of two parts. The first part focuses on detecting significant PC patients. The second part focuses on distinguishing healthy men, insignificant PC patients and significant PC patients. All the experiments are implemented using 64-bit MATLAB R2014a on a computer with Intel Core i5-6300 2.40 GHz CPU and 8.00GB RAM.

B. Classification using machine learning methods

The use of machine learning has been rapidly spread beyond computer science and successfully applied in health care predictive analytics. In this study, four popular machine learning methods - SVM, LS-SVM, ANN and RF are adopted to detect PC associated with diagnostic predictors. A brief introduction of these methods are given below.

1) Support Vector Machine (SVM): It was proposed by Cortes and Vapnik in 1995 [10]. The main idea is to project the original data to a higher dimensional feature space in which an optimal hyperplane can be found to maximize the margin between classes. Moreover, such mapping can be implicitly achieved using kernel trick via simply computing the selected kernel function in the original space. Hence, SVM is also known as the famous kernel method for pattern analysis.

[1] Guanjin Wang is with Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hong Kong, China guanjin.br.wang@connect.polyu.hk

[2] Jeremy Yuen-Chun Teoh is with Division of Urology, Department of Surgery, Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong, China jeremyteoh@surgery.cuhk.edu.hk

[3] Kup-Sze Choi is with Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hong Kong, China thomasks.choi@polyu.edu.hk

TABLE I: Baseline characteristics of the cohort

	Value	Percentage
Total number of patients	1625	
Number and percentage of patients with to PSA level (ng ml^{-1})		
<4	84	5.17
4-10	570	35.08
10.1-20	343	21.11
20.1-50	220	13.54
>50	408	25.11
Age(year, mean±s.d.)	70±8	
Estimated prostate volume on TRUS (ml, mean±s.d.)	49.69±26.33	
PSA level (ng ml^{-1})	42.45±274.26	
DRE finding (number of patients)		
Normal	1313	80.80
Abnormal	312	19.20
TRUS finding (number of patients)		
Normal	703	43.26
Abnormal	922	56.74

2) Least Squares Support Vector Machine (LS-SVM): It is a variant of the standard SVM, which was proposed by Suykens and Vandewalle in 1999 [11]. Its learning process is much more simplified compared with SVM, which is to solve a set of linear equations rather than a quadratic programming (QP) problem in SVM. Several empirical studies [12], [13] have proved that LS-SVM and SVM can have comparative generalization performances.

3) Artificial Neural Network (ANN): It is originally inspired by the structure of the biological neural network in neuroscience. In 1943, McCulloch and Pitts[14] proposed the first artificial neuron called McCulloch-Pitts (MCP) model, which performs like a linear threshold gate. In 1957, the simplest neural network - perceptron [15] was invented by Rosenblatt, which consists of two layers of nodes to learn a binary classifier. In order to solve more complex non-linear separable problems, the traditional ANN usually contains several hidden layers to adequately model the underlying behavior of the inputs.

4) Random Forest (RF): It is an ensemble learning method which retains multiple decision trees forming a 'forest' to jointly determine output class [16]. Specifically, for classification tasks, every decision tree in the forest performs a classification of the new input, and the output class is the one which has the highest votes made by all the trees. RF is one of the most accurate machine learning methods, and run fast particularly on big datasets.

C. Performance metrics

1) cross validation: 10-fold cross validation is employed to compare the classification performances of four machine learning methods. The adopted dataset is randomly split into ten folds in which one of them is retained as the validation data for testing the model, and the remaining nine folds are used for training model. This process is repeated ten times such that each fold can be used as validation data for once. After that, the validation results from the ten constructed models can be averaged to produce a single estimation.

2) F1 score: F1 score is a harmonic mean between precision and sensitivity, which tends to get closer to the smaller value of the two measures. Therefore, a higher F1 score indicates that both measures are comparatively higher.

3) Receiver Operating Characteristic curve: The receiver operating characteristic (ROC) curve and the corresponding area under the ROC (AUC) value reflect the trade-off between the sensitivity and (1-specificity) at different threshold settings of a diagnostic test. The larger the AUC value is, the better performance the model achieves.

4) Confusion matrix: The confusion matrix is a table visually describing the classification performance of a model, in which the row represents the samples in a predicted class and the column represents the number of samples in a target class. Since confusion matrix can deeply look into each pair of classes which is suitable for performance evaluation on multi-class classification.

III. Experimental results

A. Classification between significant cancer vs. benign and insignificant cancer

In the first experiment, four machine learning methods were used to construct prediction models for the detection of significant PC cases, and their classification performances were compared with respect to accuracy, sensitivity, specificity, F1 score and AUC. Table II lits the performance results of four methods on the adopted dataset. SVM, LS-SVM, ANN and RF achieved excellent performances with an average accuracy of 0.9506, 0.9363, 0.9527 and 0.9416, respectively. Among all the methods, ANN achieved the highest accuracy, sensitivity and F1 score. The ROC curves of four machine learning methods with their AUC values are demonstrated in Figure 1. ANN remained the advantage over the other methods with the highest AUC of 0.9755.

B. Classification between benign vs. insignificant cancer vs. significant cancer

In the second experiment, the classification performances of four machine learning methods were evaluated on the same dataset but for the diagnosis of benign versus insignificant PC versus significant PC. The experimental results were listed in Table III. RF achieved the highest accuracy (0.7941), sensitivity (0.8277), specificity (0.8771) and F1-score (0.8290) among all the methods. We also used the confusion matrix to display the classification performance of each class using four methods in Fig. 2. In this three-class classification, we assume that it is comparatively more significant to detect all the patients with significant PC. We observed that RF outperformed the other methods which distinguished 210 out of 247 significant PC cases with an accuracy of 0.8500. RF also achieved the highest accuracy (0.6800) of classifying insignificant PC. In addition, all the healthy cases can be accurately classified by using these four methods.

TABLE II: Classification performances on the adopted dataset (significant cancer vs. benign and insignificant cancer)

	Data set	SVM	LS-SVM	ANN	RF
Accuracy	training	0.9528±0.0032	0.9491±0.0025	0.9569±0.0033	0.9998±3.7051e-04
	testing	0.9506±0.0078	0.9363±0.0058	**0.9527±0.0079**	0.9416±0.0063
Sensitivity	training	0.9155±0.0053	0.9088±0.0049	1.0000±0.0000	0.9997±7.2577e-04
	testing	0.9112±0.0125	0.8895±0.0117	**0.9996±0.0013**	0.9062±0.0131
Specificity	training	1.0000±0.0000	1.0000±0.0000	0.9124±0.0068	1.0000±0.0000
	testing	1.0000±0.0000	1.0000±0.0000	0.9035±0.0163	0.9858±0.0075
F1 score	training	0.9559±0.0029	0.9522±0.0027	0.9594±0.0030	0.9998±3.6320e-04
	testing	0.9535±0.0069	0.9415±0.0065	**0.9558±0.0071**	0.9451±0.0063
AUC	training	0.9821±0.0033	0.9795±0.0022	0.9827±0.0037	1.0000±0.0000
	testing	0.9578±0.0117	0.9706±0.0044	**0.9755±0.0073**	0.9702±0.0030

TABLE III: Classification performances on the adopted dataset (benign vs. insignificant cancer vs. significant cancer)

	Data set	SVM	LS-SVM	ANN	RF
Accuracy	training	0.7662±0.0093	0.7879±0.0095	0.7787±0.0151	0.9985±9.3171e-04
	testing	0.7684±0.0118	0.7725±0.0082	0.7594±.0.0208	**0.7941±0.0130**
Sensitivity	training	0.8170±0.0060	0.8273±0.0067	0.8199±0.0110	0.9989±6.9356e-04
	testing	0.8169±0.0092	0.8097±0.0102	0.8064±0.0173	**0.8277±0.0122**
Specificity	training	0.8673±0.0052	0.8755±0.0048	0.8719±0.0079	0.9992±5.0620e-04
	testing	0.8695±0.0078	0.8684±0.0052	0.8604±0.0108	**0.8771±0.0092**
F-score	training	0.8115±0.0064	0.8262±0.0070	0.8178±0.0115	0.9988±7.7380e-04
	testing	0.8114±0.0081	0.8090±0.0089	0.8034±0.0191	**0.8290±0.0115**

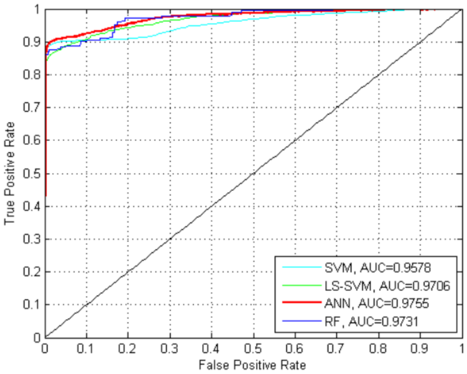

Fig. 1: ROC curves of four machine learning methods with AUC values

IV. Conclusions and future work

Prostate cancer is one of the most common malignancies in the urological cancers in the male worldwide. A thorough and accurate diagnosis of PC serves an important role for the successful individualized treatment for the patients. This study aims to construct a reliable diagnostic model for early PC detection while reducing unnecessary biopsies on the Chinese population. We used four machine learning methods - SVM, LS-SVM, ANN and RF on a Chinese cohort after TRUS-guided prostate biopsy with the inputs of age, PSA level, prostate volume, DRE and TRUS findings. The classification performances are evaluated and compared using several metrics. To detect the significant PC, all the methods performed well while ANN achieved the highest accuracy of 0.9527 and the highest AUC value of 0.9755, showing its outstanding capability in the diagnosis of PC. To classify between benign, significant PC and insignificant PC, RF exhibited the superior advantage in the classification with the highest accuracy of 0.7940 and F1 score of 0.8290. More specific, it successfully identified 210 out of 247 (0.850) significant cases and 115 out of 170 (0.680%) insignificant cases. In addition, all the healthy cases can be distinguished using these four methods.

In conclusion, we found that ANN is a successful machine learning method to distinguish significant PC patients while RF is a more appropriate method to distinguish benign, insignificant and significant PC compared with the other three methods on the Chinese population. Further studies are needed to investigate how to improve the classification performances between the significant and insignificant PC. Moreover, since other predictive models and risk scores have been generated on different populations to detect PC, we shall validate these models using the same cohort and compare the results with those using machine learning methods.

V. Acknowledgment

The work was supported by the Research Grants Council of the Hong Kong SAR (PolyU 152040/16E) and G. Wang is supported by the YC Yu Scholarship for the Centre for Smart Health.

References

[1] R. Siegel, K. Miller, and A. Jemal, “Cancer statistics, 2017,” *CA: A Cancer Journal for Clinicians*, vol. 67, no. 2, pp. 7–30, 2017.

[2] R. Ablin, L. Pfeiffer, M. Gonder, and W. Soanes, “Precipitating antibody in the sera of patients treated cryosurgically for carcinoma of the prostate,” *Experimental Medicine and Surgery*, vol. 27, no. 4, pp. 406–410, 1968.

[3] M. Nadji, S. Z. Tabei, A. Castro, T. M. Chu, G. P. Murphy, M. C. Wang, and A. R. Morales, “Prostatic-specific antigen: An immunohistologic marker for prostatic neoplasms,” *Cancer*, vol. 48, no. 5, pp. 1229–1232, 1981.

[4] W. Catalona, M. Hudson, P. Scardino, J. Richie, F. Ahmann, R. Flanigan, J. DeKernion, T. Ratliff, L. Kavoussi, and B. Dalkin, “Selection of optimal prostate specific antigen cutoffs for early detection of prostate cancer: receiver operating characteristic curves.” *The Journal of Urology*, vol. 152, no. 6 Pt 1, pp. 2037–2042, 1994.

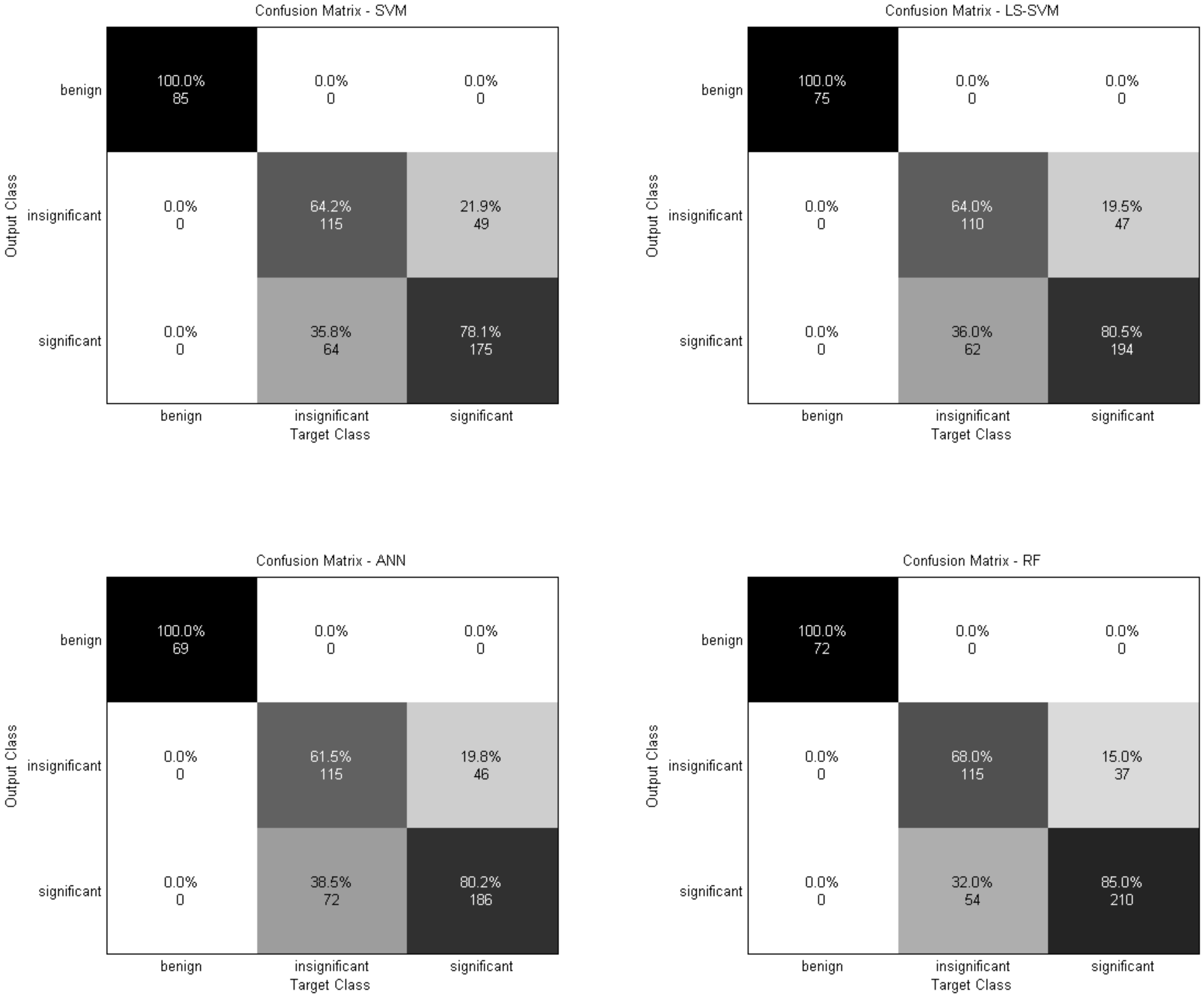

Fig. 2: Confusion matrix of four machine learning methods

[5] W. Catalona, J. Richie, F. Ahmann, M. Hudson, P. Scardino, R. Flanigan, J. Dekernion, T. Ratliff, L. Kavoussi, and B. Dalkin, “Comparison of digital rectal examination and serum prostate specific antigen in the early detection of prostate cancer: results of a multicenter clinical trial of 6,630 men,” *The Journal of Urology*, vol. 151, no. 5, pp. 1283–1290, 1994.

[6] F. C. Hamdy, J. L. Donovan, J. A. Lane, M. Mason, C. Metcalfe, P. Holding, M. Davis, T. J. Peters, E. L. Turner, R. M. Martin *et al.*, “10-year outcomes after monitoring, surgery, or radiotherapy for localized prostate cancer,” *New England Journal of Medicine*, vol. 375, no. 15, pp. 1415–1424, 2016.

[7] R. Kranse, M. Roobol, and F. H. Schröder, “A graphical device to represent the outcomes of a logistic regression analysis,” *The Prostate*, vol. 68, no. 15, pp. 1674–1680, 2008.

[8] D. P. Ankerst, J. Hoefler, S. Bock, P. J. Goodman, A. Vickers, J. Hernandez, L. J. Sokoll, M. G. Sanda, J. T. Wei, R. J. Leach *et al.*, “Prostate cancer prevention trial risk calculator 2.0 for the prediction of low-vs high-grade prostate cancer,” *Urology*, vol. 83, no. 6, pp. 1362–1368, 2014.

[9] R. K. Nam, A. Toi, L. H. Klotz, J. Trachtenberg, M. A. Jewett, S. Appu, D. A. Loblaw, L. Sugar, S. A. Narod, and M. W. Kattan, “Assessing individual risk for prostate cancer,” *Journal of Clinical Oncology*, vol. 25, no. 24, pp. 3582–3588, 2007.

[10] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[11] J. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least Squares Support Vector Machine Classifiers*. Singapore: World Scientific, 2002.

[12] T. Van Gestel, J. A. Suykens, B. Baesens, S. Viaene, J. Vanthienen, G. Dedene, B. De Moor, and J. Vandewalle, “Benchmarking least squares support vector machine classifiers,” *Machine Learning*, vol. 54, no. 1, pp. 5–32, 2004.

[13] P. Zhang and J. Peng, “SVM vs regularized least squares classification,” in *Proceedings of the 17th International Conference on Pattern Recognition*, vol. 1. IEEE, 2004, pp. 176–179.

[14] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *The Bulletin of Mathematical Biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[15] F. Rosenblatt, “The perceptron: A probabilistic model for information storage and organization in the brain.” *Psychological Review*, vol. 65, no. 6, p. 386, 1958.

[16] A. Liaw, M. Wiener *et al.*, “Classification and regression by randomForest,” *R News*, vol. 2, no. 3, pp. 18–22, 2002.