Behavior determinant based cervical cancer early detection with machine learning algorithm

Article *in* Advanced Science Letters · October 2016

DOI: 10.1166/asl.2016.7980

3 authors, including:

Adi Wijaya
Sekolah Tinggi Ilmu Kesehatan Indonesia Maju
25 PUBLICATIONS **17** CITATIONS

Advanced Science Letters
Vol. 22, 3120-3123, 2016

Behavior Determinant based Cervical Cancer Early Detection with Machine Learning Algorithm

Sobar[1], Rizanda Machmud[1], Adi Wijaya[2]

[1]School of Public Health, University of Andalas, West Sumatera, Indonesia
[2]School of Information Technology, STMIK Eresha, Jakarta, Indonesia

Cervical cancer (Ca Cervix) is serious public health problem in women in the world. Fortunately, this disease is preventable. Current prevention method remains low both in the result and participation. So, prevention or early detection method still open and challenging. Behavior and its determinant are promising as Ca Cervix predictor and event as early detection. To date, there is still limited amount of research conducted in Ca Cervix detection based on behavior and machine learning in the combination of the field of public health and computer science. In this study, machine learning is used as classifier to detect the probability of Ca Cervix risk based on behavior and its determinant. Two popular machine learning, Naïve Bayes (NB) and Logistic Regression (LR) are used as classifier. From the experimental result, both NB and LR are promising as classifier to detect Ca Cervix risk based on behavior and its determinant with accuracy 91.67% and 87.5% respectively and with AUC 0.96 and 0.97 respectively.

Keywords: Ca Cervix, Behavior Determinant, Early Detection, Machine Learning.

1. INTRODUCTION

Cervical cancer (Ca Cervix) is a serious public health problem in women all over the world [1], it is the second most common cancer among women worldwide [2]. The estimate of global Ca Cervix prevalence is 11.7%, and is most prevalent in Sub-Saharan Africa (24.0%), Eastern Europe (21.4%), and Latin America (16.1%) [3]. It is evident that over 85% of cervical cancers occur in developing countries [4].

In Indonesia, the part of the population at risk of developing Ca Cervix, women aged 15 years and older, is 79.14 million. It estimates indicate that every year 13762 women are diagnosed with Ca Cervix and 7493 die from the disease [5]. Common problems in the middle low class is that they generally seek care only when they develop symptoms and by that time the cancer has advanced and is difficult to treat [6].

Fortunately, almost every case of cervical cancer is potentially preventable [7] because of its slow progression, cytological identifiable precursors, and effective treatments if detected early [8]. Recently, Ca Cervix prevention based on screening (pap smear test and VIA test) and Human papillomavirus (HPV) vaccination.

The scope of screening remains low in all regions because of the lack of basic knowledge women have regarding screening as an opportunity for the prevention of cervical cancer [9], lack of awareness, being un-comfortable with the procedure and not knowing where to go for a pap smear test [10]. HPV vaccination does not offer full protection either, because only some HPV types are covered by the vaccines and the long-term efficacy of the vaccines has not been determined yet [11].

Ca Cervix cannot be cured, so primary and secondary prevention provides the best choices for this health care issue [12]. Hence, primary prevention of Ca Cervix through culturally-informed personal health behavior become important [13] since sexual behavior factors increase the risk of Ca Cervix. The primary prevention of Ca Cervix through behavior prevention that the early

detection of malignancies are cost-effective in reducing cancer-specific mortality [14]. To date, the using of behavior for Ca Cervix early detection or primary prevention is limited and the using of machine learning (ML) is limited as well for cancer early detection especially in Ca Cervix. In this Study, ML is used to classify whether safe or risky from Ca Cervix disease based on behavior determinant as early detection or primary prevention.

This paper is organized as follows. In section 2, the behavior determinant based Ca Cervix early detection is given. In section 3, the machine learning classification as early detection is presented. The experimental results of evaluation from the proposed method for Ca Cervix early detection based on behavior determinant with machine learning are also presented in section 4. Finally, our work of this paper is summarized in the last section.

2. BEHAVIOR DETERMINANT BASED CA CERVIX EARLY DETECTION

There are important variables for determining a risk of Ca Cervix disease based on behavior. In social science theory including health science and psychology, behavior is widely studied. Common behavior related theory or model such as: The Health Belief Model (HBM), Protection Motivation Theory (PMT), Theory of Planned Behavior (TPB), Social Cognitive Theory (SCT), etc.

HBM is determined by two cognitions perceptions of illness threat and evaluation of behaviors to counteract threat [15]. A study states that indicates low levels of Ca Cervix knowledge, perceived seriousness, and perceived susceptibility coupled with high risk sexual behaviors [16].

There are some theories in social and health psychology assume that intentions cause and determine behaviors [17]. Whereas PMT mention that the primary determinant of performing is protection motivation or intention to perform prevention behavior [18]. Motivation is one of determining factor of organizational prevention behavior [19].

In TPB, behavior is affected by intention, while intention is assumed to be determined by three factors: attitudes, subjective norms, and Perceived behavioral control (PBC) [18]. Attitude and subjective norm interacted with perceived control such that both were more potent predictors of intention [20].

In SCT that prevention behavior is held to be determined by three factors: goals, outcome expectancies, and self-efficacy [15]. Emphasizing social support among participants may improve cervical cancer prevention behavior [21]. Empowerment may be defined as a system of beliefs, referring to the ability to make decisions, access information, and use social and internal resources to cervical cancer prevention behavior [22].

From those theories perspective, there are seven determinants of behavior, i.e.: perception, intention, motivation, subjective norm, attitude, social support and empowerment. In this study, these eight variables (seven determinants and behavior itself) translated into questionnaire with nine questions for each variable. The questionnaire then distributed to 72 respondents with 22 respondents is Ca Cervix sufferers and 50 respondents is not Ca Cervix sufferers. All of respondent is urban citizen in Jakarta, Indonesia. This seven determinants and behavior itself are using as features or attributes to create a classification model with machine learning as early detection for Ca Cervix risk.

3. MACHINE LEARNING CLASSIFICATION AS EARLY DETECTION METHOD

Machine learning methods commonly used for early detection, such as: early detection of Alzheimer's disease [23], early detection of liver disease [24], and even for cancer prediction and prognosis [25] with classification as common technique. Some machine learning algorithm from classification technique, such as Naïve Bayes (NB), Logistic Regression (LR), C4.5, k-NN and Support Vector Machine are top 10 algorithm in data mining [26].

In this study, NB and LR are used as classification methods for Ca Cervix early detection based on behavior determinants. NB is often used to outlier or novelty detection [27], [28]. NB is fast, easy to implement with the simple structure and effective [29]. NB classifier continues to be a popular learning algorithm for data mining applications due to its simplicity and linear run-time [30].

While LR commonly used in health science research for classification since many statistical package has LR in their software such as SPSS. LR provides a mechanism for applying the techniques of linear regression to classification problems [31]. LR is linear classifier that proved as powerful probability statistic classifier and has ability to handle multi-class classification [32].

In this study, the evaluation method using the classifier’s effectiveness [33]. The results of this process will produce a confusion matrix that contains the value true positive (TP), true negative (TN), false positive (FP) and false negative (FN) as shown in Table 1. TP means when predicted value is risky while the actual value is risky as well. In case of safe status, when predicted value is safe and the actual value is safe, it is TN. FP is when predicted value is risky while the actual value is safe. With the opposite, FN, when predicted value is safe while the actual value is risky. The main evaluation of this study is accuracy and area under receiver operating characteristic (ROC) curve (AUC). AUC evaluates the performance of a scoring classifier on a test set, but ignores the magnitude of the scores and only takes their rank order into account. AUC is expressed on a scale of 0 to 1, where 0 means that all negatives are ranked before all positives, and 1 means that all positives are ranked before all negatives [31].

Table.1. Confusion Matrix

	Actual	
Predicted	Risky	Safe
Risky	TP	FP
Safe	FN	TN

Based on confusion matrix as shown in Table 1, the accuracy calculation as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

4 EXPERIMENTAL RESULTS

The experiments are conducted using a computing platform based on Intel Core 2 Duo 2.2 GHz CPU, 2 GB RAM, and Microsoft Windows 7 32-bit operating system and Rapidminer version 5.3 as data analytics tool. Rapidminer will produce both accuracy and AUC as the calculation output. In this study, we conduct two experiments to evaluate the behavior determinant as attributes to classify whether the respondent has safe from Ca Cervix disease or risky with NB and LR. Each experiment using 10-fold cross validation, so the dataset will split into 10 parts dataset, 1 part as testing dataset and the rest as training datasets and this process repeated 10 times.

The result of first experiment, taken from Rapidminer calculation output, is the model using NB as classifier is shown in Table 2 and Figure 1. From the confusion matrix as shown in Table 2, the accuracy of the model is 91.67% and as shown in Figure 1, the AUC is 0.96. This classification accuracy is excellent since it's above 90% and as preliminary research in this topic; this result is promising both from accuracy and AUC.

Table.2. Confusion Matrix of NB Classification

	Actual	
Predicted	Risky	Safe
Risky	17	2
Safe	4	49

Accuracy = (17+49) / (17+49+2+4) = 0.9167
= 0.9167 x 100% = 91.67%

As shown in Table 2, the accuracy is high since FP and FN are low, 2 and 4 respectively; while TP and TN are high, 17 and 49 respectively. From this result, NB almost recognize or predict all of actual status both risky and safe.

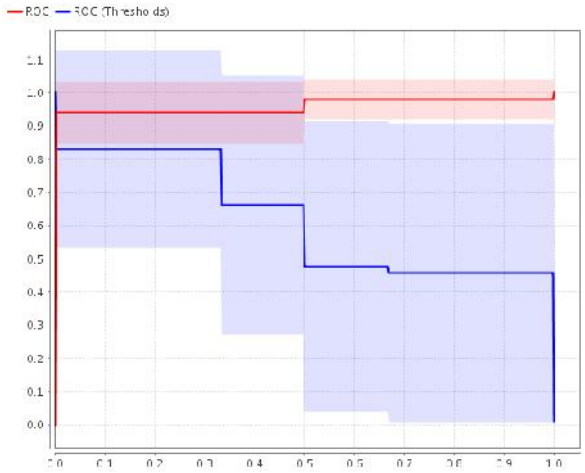

Fig.1. AUC of NB

The result of second experiment is the model using LR as classifier is shown in Table 3 and Figure 2. From the confusion matrix as shown in Table 3, the accuracy of the model is 87.5% and as shown in Figure 2, the AUC is 0.97. This classification accuracy is good enough since it almost 90% and the AUC is above 0.95.

Table.3. Confusion Matrix of LR Classification

	Actual	
Predicted	Risky	Safe
Risky	16	4
Safe	5	47

Accuracy = (16+47) / (16+47+4+5) = 0.875
= 0.875 x 100% = 87.5%

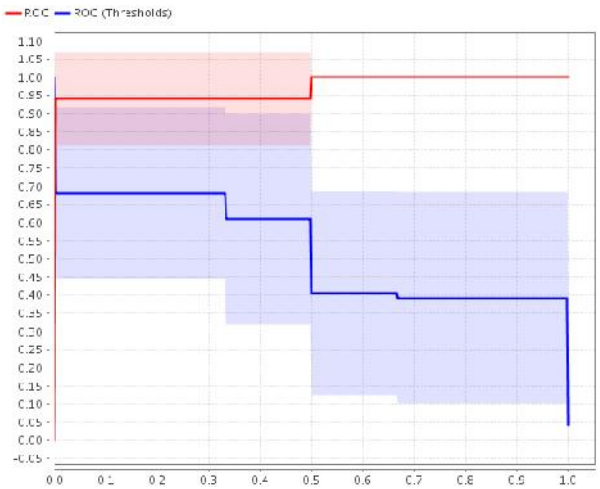

Fig.2. AUC of LR

From both experiment, we can see that NB is outperform LR in accuracy but LR is outperform in AUC. Both classifiers have good result since this is the primary research of this topic.

5. CONCLUSIONS

Ca Cervix is a serious public health problem in women throughout the world and fortunately preventable. The primary prevention of Ca Cervix through behavior prevention as early detection has many advantage, such as reduce mortality and increase effectiveness of treatments if detected early. To date, there is still limited amount of research conducted in Ca Cervix detection based on behavior and machine learning in the combination of the field of public health and computer science.

Based on the experimental result, the models of this study is promising since the classification accuracy above 90% and AUC is above 0.95. Behavior determinant become important aspect to involve in Ca Cervix primary prevention and with this promising result, behavior could be preemptive prevention. Behavior prevention approach is cheaper than other approach such as HPV vaccination, Pap smear test and VIA test. This study is still preliminary research and still open for further improvement; so, we establish a benchmark for next research comparison since we provide both accuracy and AUC.

Future research will be concerned with feature or attribute selection in order to increase the accuracy and the using meta-learning techniques such as boosting and stacking.

ACKNOWLEDGMENTS

This work was supported in part by Higher Education Directorate of Indonesia Ministry of Research, Technology and Higher Education under Grant No. 1865.34/E4.4/2014.

REFERENCES

[1] A. Saha, A. N. Chaudhury, P. Bhowmik, and R. Chatterjee, “Awareness of cervical cancer among female students of premier colleges in Kolkata, India.,” *Asian Pac. J. Cancer Prev.*, vol. 11, no. 4, pp. 1085–90, Jan. 2010.

[2] F. H. Zhao, S. M. Tiggelaar, S. Y. Hu, L. N. Xu, Y. Hong, M. Niyazi, X. H. Gao, L. R. Ju, L. Q. Zhang, X. X. Feng, X. Z. Duan, X. L. Song, J. Wang, Y. Yang, C. Q. Li, J. H. Liu, J. H. Liu, Y. B. Lu, L. Li, Q. Zhou, J. F. Liu, N. Zhao, J. E. Schmidt, and Y. L. Qiao, “A multi-center survey of age of sexual debut and sexual behavior in Chinese women: Suggestions for optimal age of human papillomavirus vaccination in China,” *Cancer Epidemiol.*, vol. 36, pp. 384–390, 2012.

[3] M. Urasa and E. Darj, “Knowledge of cervical cancer and screening practices of nurses at a regional hospital in Tanzania,” *Afr. Health Sci.*, vol. 11, no. 2, pp. 48–57, 2011.

[4] A. F. Rositch, A. Gatuguta, R. Y. Choi, B. L. Guthrie, R. D. Mackelprang, R. Bosire, L. Manyara, J. N. Kiarie, J. S. Smith, and C. Farquhar, “Knowledge and acceptability of Pap smears, self-sampling and HPV vaccination among adult women in Kenya,” *PLoS One*, vol. 7, no. 7, 2012.

[5] L. Jaspers, S. Budiningsih, R. Wolterbeek, F. C. Henderson, and a a W. Peters, “Parental acceptance of human papillomavirus (HPV) vaccination in Indonesia: a cross-sectional study.,” *Vaccine*, vol. 29, no. 44, pp. 7785–93, Oct. 2011.

[6] L. Nuranna, M. F. Aziz, S. Cornain, G. Purwoto, S. Purbadi, S. Budiningsih, B. Siregar, and A. A. W. Peters, “Cervical cancer prevention program in Jakarta, Indonesia: See and Treat model in developing country.,” *J. Gynecol. Oncol.*, vol. 23, no. 3, pp. 147–52, Jul. 2012.

[7] C. Banura, F. M. Mirembe, A. R. Katahoire, P. B. Namujju, and E. K. Mbidde, “Universal routine HPV vaccination for young girls in Uganda: a review of opportunities and potential obstacles.,” *Infect. Agent. Cancer*, vol. 7, no. 1, p. 24, Jan. 2012.

[8] M. R. Balogun, O. O. Odukoya, M. a Oyediran, and P. I. Ujomu, “Cervical cancer awareness and preventive practices: a challenge for female urban slum dwellers in Lagos, Nigeria.,” *Afr. J. Reprod. Health*, vol. 16, no. 1, pp. 75–82, Mar. 2012.

[9] E. J. Domingo, R. Noviani, M. R. M. Noor, C. a. Ngelangel, K. K. Limpaphayom, T. Van Thuan, K. S. Louie, and M. a. Quinn, “Epidemiology and Prevention of Cervical Cancer in Indonesia, Malaysia, the Philippines, Thailand and Vietnam,” *Vaccine*, vol. 26, 2008.

[10] E. Yanikkerem, A. Goker, N. Piro, S. Dikayak, and F. M. Koyuncu, “Knowledge about cervical cancer, Pap test and barriers towards cervical screening of women in Turkey,” *J. Cancer Educ.*, vol. 28, pp. 375–383, 2013.

[11] N. Ghotbi and A. Anai, “Assessment of the knowledge and attitude of female students towards cervical cancer prevention at an international university in Japan.,” *Asian Pac. J. Cancer Prev.*, vol. 13, no. 3, pp. 897–900, Jan. 2012.

[12] M. E. Vance and B. Keele, “Development and validation of the cervical cancer knowledge and beliefs of Appalachian women questionnaire.,” *J. Nurs. Meas.*, vol. 21, no. 3, pp. 477–501, Jan. 2013.

[13] J. Gregg, C. K. Y. Nguyen-Truong, P. Wang, and A.

Kobus, “Prioritizing prevention: culture, context, and cervical cancer screening among Vietnamese American women.,” *J. Immigr. Minor. Health*, vol. 13, no. 6, pp. 1084–9, Dec. 2011.

[14] C. M. van der Aalst, R. J. van Klaveren, and H. J. de Koning, “Does participation to screening unintentionally influence lifestyle behaviour and thus lifestyle-related morbidity?,” *Best Pract. Res. Clin. Gastroenterol.*, vol. 24, no. 4, pp. 465–78, Aug. 2010.

[15] M. Conner and P. Sparks, “Theory of planned behaviour and health behaviour,” *Predict. Heal. Behav. Res. Pract. with Soc. Cogn. Model.*, pp. 170–222, 2005.

[16] B. A. Ingledue Kimberly, Cottrell Randall, “COLLEGE WOMEN ’ S KNOWLEDGE , PERCEPTIONS , AND PREVENTIVE BEHAVIORS REGARDING HUMAN PAPILLOMAVIRUS INFECTION AND CERVICAL CANCER,” *Am. J. Health Stud.*, 2004.

[17] T. L. Webb and P. Sheeran, “Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence.,” *Psychol. Bull.*, vol. 132, no. 2, pp. 249–268, 2006.

[18] M. Conner, “Cognitive Determinants of Health Behavior,” pp. 19–31, 2010.

[19] J. L. Fitch, “Willpower and perceived behavioral control: influences on the intention-behavior relationship and postbehavior attributions,” vol. 33, no. 2, 2005.

[20] J. P. Dillard, “An Application of the Integrative Model to Women’s Intention to Be Vaccinated Against HPV: Implications for Message Design.,” *Health Commun.*, vol. 26, no. January, pp. 479–486, 2015.

[21] L. Larkey, “Las Mujeres Saludables: Reaching latinas for breast, cervical and colorectal cancer prevention and screening,” *J. Community Health*, vol. 31, no. 1, pp. 69–77, 2006.

[22] A. Luszczynska, A. B. Durawa, U. Scholz, and N. Knoll, “Empowerment Beliefs and Intention to Uptake Cervical Cancer Screening: Three Psychosocial Mediating Mechanisms,” *Women Health*, vol. 52, no. January, pp. 162–181, 2012.

[23] M. López, J. Ramírez, J. M. Górriz, I. Álvarez, D. Salas-Gonzalez, F. Segovia, R. Chaves, P. Padilla, M. Gómez-Río, and A. D. N. Initiative, “Principal component analysis-based techniques and supervised classification schemes for the early detection of Alzheimer’s disease,” *Neurocomputing*, vol. 74, no. 8, pp. 1260–1271, 2011.

[24] X. Zhou, Y. Zhang, M. Shi, H. Shi, and Z. Zheng, “Early detection of liver disease using data visualisation and classification method,” *Biomed. Signal Process. Control*, vol. 11, pp. 27–35, 2014.

[25] J. A. Cruz and D. S. Wishart, “Applications of machine learning in cancer prediction and prognosis,” *Cancer Inform.*, vol. 2, pp. 59–77, 2006.

[26] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, and S. Y. Philip, “Top 10 algorithms in data mining,” *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, 2008.

[27] O. Alan and C. Catal, “Thresholds based outlier detection approach for mining class outliers: An empirical case study on software measurement datasets,” *Expert Syst. Appl.*, vol. 38, no. 4, pp. 3440–3445, Apr. 2011.

[28] N. F. Lepora, M. J. Pearson, B. Mitchinson, M. Evans, C. Fox, A. Pipe, K. Gurney, and T. J. Prescott, “Naive Bayes novelty detection for a moving robot with whiskers,” *2010 IEEE Int. Conf. Robot. Biomimetics*, pp. 131–136, Dec. 2010.

[29] S. Taheri and M. Mammadov, “Learning the naive Bayes classifier with optimization models,” *Int. J. Appl. Math. Comput. Sci.*, vol. 23, no. 4, pp. 787–795, Jan. 2013.

[30] M. Hall, “A decision tree-based attribute weighting filter for naive Bayes,” *Knowledge-Based Syst.*, vol. 20, no. 2, pp. 120–126, Mar. 2007.

[31] C. Sammut and G. I. Webb, *Encyclopedia of machine learning*. Springer Science & Business Media, 2011.

[32] P. Karsmakers, K. Pelckmans, and J. A. K. Suykens, “Multi-class kernel logistic regression: a fixed-size implementation,” in *Neural Networks, 2007. IJCNN 2007. International Joint Conference on*, 2007, pp. 1756–1761.

[33] G. Oberreuter and J. D. Velásquez, “Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style,” *Expert Syst. Appl.*, vol. 40, no. 9, pp. 3756–3763, Jul. 2013.