Development of Breast Cancer Risk Prediction Model for Women in Indonesia: A Case-Control Study

Ricvan Dana Nindrea (ricvandana7@gmail.com)
Universitas Gadjah Mada https://orcid.org/0000-0002-1844-3323
Hari Kusnanto
Universitas Gadjah Mada Fakultas Kedokteran
Samuel Johny Haryono
Rumah Sakit Dharmais Pusat Kanker Nasional
Wirsma Arif Harahap
Universitas Andalas Fakultas Kedokteran
Iwan Dwiprahasto
Universitas Gadjah Mada Fakultas Kedokteran
Lutfan Lazuardi
Universitas Gadjah Mada Fakultas Kedokteran
Teguh Aryandono
Universitas Gadjah Mada Fakultas Kedokteran

Research article

Keywords: Breast Cancer, Risk Factors, Prediction, Classification, Machine Learning

Posted Date: April 24th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-24225/v1

Abstract

Background: In 2018, estimated that 11.6% or 2,088,849 new breast cancer cases and 6.6% or 626,679 cases are predicted to end with mortality due to this disease. One of the causes of the high mortality rate due to breast cancer in Indonesia is found in 60-70% of patients identified with advanced-stage breast cancer. It is related to the perception of breast cancer risk in Indonesian women. Therefore, it is necessary to calculate the risk factors for breast cancer risk to help increase public awareness in recognizing the risk of breast cancer in Indonesia. This study was held to construct risk factors calculation model for breast cancer risk based on machine learning in Indonesia.

Methods: This research was quantitative which was conducted using a case-control study design. Data were collected in Dr. M. Djamil General Hospital Padang, Sardjito General Hospital Yogyakarta and Dharmais Cancer Hospital Jakarta from July 2018-July 2019. The number of samples in this study were 1,000 women cases groups (breast cancer) and 1,000 women control groups (non-breast cancer) matching by age and sex. The sampling technique in this study was convenience sampling. Data were collected from medical records and primary data collection used a questionnaire. Chi-square test used for bivariate analysis and risk factors calculation were used machine learning algorithm Naive Bayes, decision tree, k-nearest neighbors, support vector machine and logistic regression. Determination of algorithm selection by comparing the highest accuracy, true positive rate, false-positive rate and Area Under Curve (AUC). STATA version 14.2 and the Waikato Environment for Knowledge Analysis (WEKA) version 3.6.4 were used to process the data.

Results: The model construction for calculating risk factors for breast cancer risk in Indonesia is based on predictors of menopause age, the first age of pregnancy, the first and second-degree family history of breast cancer, use of oral contraceptives, history of smoking, overweight, obesity, high-fat diet, high-calorie diets and physical activity. The cut-off point in classifying high risk of breast cancer and low risk of breast cancer is based on a total score of > 7 (high risk of breast cancer) and ≤ 7 (low risk of breast cancer). The accuracy of the breast cancer risk factor calculation model in Indonesia was 79.9% with a sensitivity of 76.90% and a specificity of 70.4%.

Conclusion: This breast cancer risk factor calculation can be categorized quite well in classifying breast cancer risk in Indonesia.

Background

Among all cancers in Indonesian women, breast cancer is the the first rank of cancer. In 2018, estimated that 11.6% or 2,088,849 new breast cancer cases and 6.6% or 626,679 cases are predicted to end with mortality due to this disease [1]. Three provinces with the highest prevalence of breast cancer in Indonesia are Yogyakarta (2.4% or 4,325 cases), East Kalimantan (1.0% or 1,879 cases) and West Sumatra (0.9% or 2,285 cases). About 60–70% of breast cancer patients who come for treatment are in stage III-IV (advanced stage) [2]. This condition is much different from western countries where almost

80% of breast cancer patients arrive at an early stage [3]. This shows that breast cancer survivors in Indonesia are late in detecting the cancer they have.

The problem of breast cancer in Indonesian women who have not been a concern is the cause of the high cases at the advanced stage. The problem of breast cancer in Indonesian women who have not been a concern is the cause of the high cases at the advanced stage. In the past 20 years, there has been a dramatic increase in the understanding of multistep carcinogenesis and the main role of genetic changes in the diagnosis, treatment and prevention of breast carcinoma. The problem of breast cancer in Indonesian women who have not been a concern is the cause of the high cases at the advanced stage. This causes an increase in prevention, detection and treatment strategies in patients with breast carcinoma. The incidence of breast cancer is multifactorial. These risk factors are divided into non-modifiable risk factors including age, gender, genetic factors (5–7%), age of menarche, age of menopause, age at the first pregnancy, parity, family history of breast cancer, history of previous cancer, proliferative breast disease. Modifiable risk factors include the use of hormone replacement therapy, lactation, use of oral contraceptives, alcohol, body mass index, smoking, physical activity and a high-fat diet and a high-calorie diet [4, 5, 6].

Data there are differences in risk factors between previous breast cancer risk prediction models; Gail model, Claus model, Tyrer-Cuzick model dan BOADICEA model with breast cancer risk in Indonesia include age at menarche, below the average of 13 year-old. Data from The Indonesian Ministry of Health in 2010 showed that the age of menarche 12–13 year-old in urban areas was 39.8% and in rural areas was 34.8%. Furthermore, the age of menopause in which the risk of breast cancer increases at the age of menopause > 45 year-old [7–11]. The proportion of women 30–49 year-old in Indonesia who are menopausal increases with age. The percentage of menopause is 11.4% in women aged 30–34 year-old, 22.6% in women aged 44–45 year-old and 44% in women age 48–49 year-old [10].

Other risk factors for breast cancer in Indonesia are smoking, alcohol, the use of hormone replacement therapy that is rarely found in people in Asia [12–14]. This illustrates that breast cancer patients in Indonesia have risk factors that could be different from breast cancer patients in western countries. Therefore, it is necessary to calculate the appropriate risk factors for Indonesians to help increase public awareness in recognizing the risk of breast cancer in Indonesia.

The calculations of breast cancer risk factors can be determined through algorithms or early detection models through useful predictors so that patients can detect breast cancer risk themselves. This is a preventive measure, using machine learning by classifying breast cancer risk from the predictor variables, making it easier to classify. Machine learning algorithm can classify groups at risk and not at risk of breast cancer, built based on existing predictor data so that the risk of breast cancer can be identified early.

Methods

Study design and research sample

This research was conducted using a case-control study design. Location of research in Dr. M. Djamil General Hospital Padang, Sardjito General Hospital Yogyakarta and Dharmais Cancer Hospital Jakarta from July 2018-July 2019. The number of samples in this study were 1,000 women in the case groups (breast cancer) and 1,000 patients in the control groups (non-breast cancer), with the inclusion criteria of the case group in this study, were patients with pathology examination showed positive breast cancer. The exclusion criteria for case groups are male breast cancer survivors and the required research data is not available in full. The control group was matched for ± 5 years of age and sex based on hospital control. The sampling technique in this study is convenience sampling.

Data collection technique

This study passed the ethical review by the ethics commiittee of the Faculty of Medicine, Public Health and Nursing, Universitas Gadjah Mada University, Yogyakarta, Indonesia (No.KE/FK/0717/EC/2018). Data was collected through medical records review and primary data collection using a research questionnaire. Data collection was carried out through interviews by doctors or team members who treated patients at Dr. M. Djamil General Hospital Padang, Sardjito General Hospital Yogyakarta and Dharmais Cancer Hospital Jakarta, by the written informed consent.

Data collection of risk factors through interviews with respondents includes modifiable risk factors consisting of: use of hormone replacement therapy (used; never used) [15], lactation (≥ 12 months; <12 months) [16], use of oral contraceptives (≥ 12 months; <12 months) [17], marital status (single/widowed; married) [18], alcohol (consuming alcohol, ≤ 5 drinks/week and > 5 drinks/week; non-drinkers for ≥ 10 years) [19], body mass index (normal, 18.5-23.49 kg/m2; overweight, 23.5-24.99 kg/m2; obesity, ≥ 25 kg/m2) [20], smoking (active smoker, if smoking routinely in the last 12 months; no smoker, never smoked or has stopped smoking for 12 months or more) [21], physical activity is measured by the instrument The Global Physical Activity Questionnaire (GPAQ), developed by the World Health Organization (WHO). The WHO STEP wise method was used to calculate physical activity and was expressed as Metabolic Equivalent minutes per day (METmins/day). The participants were classified as having “high activity” if they accumulated ≥ 3000 METmins/week, “moderate activity” if > 3000 MET ≥ 600 or “low activity” if < 600 METmins/week [22], Semi-Quantitative Food Frequency Questionnaire (SM-FFQ) was used to measure high-fat diets (excess, > 100% Recommended Dietary Allowance (RDA); sufficient, 100% RDA) [23], and high-calorie diets (excess,> 100% RDA; sufficient, 100% RDA) [23].

Non-modifiable risk factors consist of: age (≥ 50 year-old; <50 year-old) [24], menarhe (early age of menarche, 7–11 year-old; ideal age of menarche, 12–13 year-old; late age of menarche, > 13 year-old) [25], age of menopause (≥ 50 year-old; <50 year-old) [24], age at the first pregnancy (< 20 year-old; 20–29 year-old; >30 year-old) [24], parity (nulliparous; primiparous ; ≥ multiparous) [24], the first-degree family history (history: if there is a sister, mother, daughter diagnosed with breast cancer; no history: if there is no sister, mother, daughter diagnosed with breast cancer) [11], and second-degree family history of breast

cancer (there is a history, if there is cousin and aunt diagnosed with breast cancer; no history, if there is no cousin and aunt diagnosed with breast cancer) [11].

Data analysis

Data were analyzed bivariate using the chi-square test to select candidate variables. P value < 0.05 was stated as statistically significant and the variable that passed as a candidate variable with p value < 0.25. Analysis continued with Naive Bayes (NB) machine learning algorithm, Decision Tree (DT), k-nearest neighbors (KNN), Support Vector Machine (SVM) and logistic regression. The choice of algorithm is determined by comparing the value of accuracy, true positive rate, false-positive rate and Area Under Curve (AUC). Data analysis uses the STATA version 14.2 and the Waikato Environment for Knowledge Analysis (WEKA) version 3.6.4.

Results

Non-modifiable and modifiable risk factors

The relationship of non-modifiable and modifiable risk factors to the risk of breast cancer (Table 1).

Table 1 known non-modifiable risk factors in Table 1 shown the relationship of age at menopause, age at the first pregnancy, the first-degree family history of breast cancer and second-degree family history of breast cancer with breast cancer ($p < 0.05$). But no significant relationship was found between age, menarche age and parity with the occurrence of breast cancer ($p > 0.05$).

Based on modifiable risk factors known to have a significant relationship between the use of oral contraceptives, smoking history, body mass index, high-fat diet, high-calorie diet and physical activity with the occurrence of breast cancer ($p < 0.05$). However, no significant relationship was found between the use of hormone replacement therapy, lactation, marital status and history of alcohol consumption with breast cancer ($p > 0.05$).

Stratification of breast cancer risk based on predictor variables according to ethnic Minangkabau and Javanese

The calculation of risk of breast cancer is determined through selection in a bivariate analysis, candidate variables that have $p < 0.25$ are age of menopause, age at the first pregnancy, the first-degree family history of breast cancer, second-degree family history of breast cancer, use of oral contraceptives, smoking history, body mass index, high-fat diet, high-calorie diet and physical activity.

At the initial stage, before calculating the risk of breast cancer, an analysis was done using the Mantel-Haenszel test to observe uniformity and there were no confounding variables between the research groups on the ethnic Minangkabau and Javanese (Table 2).

Table 2 known ethnic variables do not affect the relationship of age at menopause (≥ 50 year-old), age at the first pregnancy (≥ 30 year-old), there is a the first-degree family history of breast cancer, there is a

second-degree family history of breast cancer, use of oral contraceptives (≥ 12 months), history of smoking (≥ last 12 months), overweight, obesity, high-fat diet, high-calorie diet and mild physical activity with breast cancer.

Multivariate factors analysis related to breast cancer in Indonesia

At this stage multivariate factors analysis related to breast cancer was performed (Table 3).

Table 3 known variable age of menopause (≥ 50 year-old), age of the first pregnancy (≥ 30 year-old), there is a the first-degree family history of breast cancer, there is a second-degree family history of breast cancer, use of oral contraceptives (≥ 12 months), smoking history (≥ last 12 months), overweight, obesity, high-fat diet, high-calorie diet and low physical activity were included in the analysis of determining breast cancer risk scoring ($p < 0.05$). The variables that meet the requirements above then enter the analysis stage using a machine learning algorithm.

Comparison of breast cancer risk calculations in Indonesia based on machine learning

Preprocessing stages to calculate breast cancer risk using machine learning (Fig. 1).

The calculations of breast cancer risk in Indonesia based on machine learning were compared based on 5 algorithms; Naive Bayes (NB), Decision Tree (DT), k-Nearest Neighbors (KNN), Support Vector Machine (SVM) and Logistic Regression (LR) (Table 4).

Table 4 known the machine learning algorithm that has the best accuracy is logistic regression with an accuracy of 74.7% and AUC 81.0% followed by a decision tree with an accuracy of 74.5% and AUC 80.8%. Comparison of Receiver Operating Characteristics (ROC) among machine learning algorithms in predicting breast cancer risk (Fig. 2).

Scoring determination through the calculation of breast cancer risk in Indonesia

The determination of scoring through cancer risk calculation (Table 5).

Table 5 known scores of each variable, which are predictors of breast cancer risk are menopause age (≥ 50 year-old, score = 3), age of the first pregnancy (< 20 year-old, score = 1), age of the first pregnancy (≥ 30 year-old, score = 3), have never been pregnant (score = 1), the first-degree family history of breast cancer (score = 2), second-degree family history of breast cancer (score = 2), oral contraceptives ≥ 12 months (score = 3), smoking ≥ last 12 months (score = 1), overweight (score = 2), obesity (score = 3), high-fat diet (score = 3) and high-calorie diet (score = 5) and low physical activity (score = 1). Total maximum score based on a predictor of breast cancer risk is 30.

In the modeling conducted, the total scoring formed is the right model to be used as a scoring model. The significance of the total score in breast cancer screening modeling in Indonesia (Table 6).

Table 6 found the breast cancer scoring model is a good in predicting breast cancer risk ($p < 0.05$). The regression equation obtained is -2.563 + 0.317 x total score. Because the regression equation has been obtained, the probability of a subject with a certain score is calculated to have a poor prognosis. The probability of a subject experiencing a poor prognosis can be calculated by the following equation.

$$p = \frac{1}{1 + \exp(-y)}$$

Information :

p = probability

exp = natural number

y = logistic equation = $a + b_1x_1 + b_2x_2 + ... + b_ix_i$

a = constant

b = coefficient

x = prognostic variable

Calculation of "y" is obtained from the results of multivariate analysis with the equation − 2.563 + 0.317 x total score. The possible range of subject scores is between 0 and 30. A score of 0 if the subject has no risk factors and a score of 30 if the subject has all risk factors. The subject's probability of having a poor prognosis for each score (Table 7).

Table 7 known the probability for the occurrence of poor prognosis of each score is known to be the lowest probability at a score of 0 at 7.16% and the highest is 30 with a poor probability of 99.90%. Poor prognosis probability curves for each score (Fig. 3).

Model of machine learning risk factor calculation in Indonesia

Based on the calculation of the scores obtained for each predictor variable, then an analysis using the Youden index J is obtained, a cut-off point or measurement threshold for the classification of the risk of breast cancer (Table 8).

Table 8 known several potential points were formed to determine the cut-off point or the measurement limit used to classify groups at risk or not at risk of breast cancer. The cut point value is known based on the cut point curve (Fig. 4).

The accuracy of breast cancer risk factor calculations can be seen from the results of the analysis using ROC (Fig. 5).

Discussion

Non-modifiable risk factors found that there was a significant relationship between menopausal age, age at the first pregnancy, the first-degree family history of breast cancer and second-degree family history of breast cancer with breast cancer in Indonesia. Modifiable risk factors indicate a significant relationship between the use of oral contraceptives, smoking history, body mass index, high-fat diets, high-calorie diets and physical activity with the occurrence of breast cancer in Indonesia. The construction of a breast cancer risk factor calculation model can identify the findings of breast cancer risk with an accuracy of 79.9%, a sensitivity of 76.9% and a specificity of 70.4%.

In previous studies, models for detecting breast cancer risk have been developed, including the Gail model, to estimate breast cancer risk based on a case control study called The Breast Cancer Detection Demonstration Project (BCDDP) called Gail model 1 (GM1). This study was conducted to predict the probability of invasive ductal carcinoma, ductal carcinoma in situ, and lobular carcinoma in situ [7].

Furthermore, other models are used in different studies such as the Claus model, which was developed to assess the family risk of developing breast cancer from a case-control study conducted by the Centers for Disease Control [8]. Parmigiani in 1998, developed a BRCAPRO model in which hereditary factors were considered primarily in assessing the risk of BRCA 1 and BRCA 2 mutations in the family [9]. Several previous studies were conducted to detect the risk of breast cancer in western countries such as the United States and the United Kingdom. However, there is currently no assessment of breast cancer risk based on conditions and risk factors that exist in Asian countries in general, particularly Southeast Asia and specifically in Indonesia.

Gail determined the predictor variables in detecting breast cancer risk were age, age at menarche, age at birth of the the first child, breast biopsy and the first-degree family history of breast cancer [7]. Claus, in 1991, stated the predictor variables that could be used in classifying breast cancer risk were age, age at menarche, age at birth of the the first child and family history of the first and second-degree breast cancer [8]. The BRCAPRO model states that breast cancer risk predictors are age, family history (the first level, second level and third level) suffering from breast cancer, age at onset of breast cancer, bilateral breast cancer, ovarian cancer and family of men suffering from breast cancer [9].

The difference in the model obtained by researchers with the previous model is in terms of the inclusion of predictor variables. Researchers found a link between risk factors for body mass index (overweight and obesity), smoking history, oral contraceptives or pills, high-fat diets, high-calorie diets and physical activity. The inclusion of variables in calculation calculations is based on several possibilities; the phenomenon of obesity prevalence has increased since 2007 to the present. In 2007, the prevalence of overweight in the adult group was 8.6%. That number increased in 2013 and 2018 to 11.5% and 13.6%. Based on its contribution, the proportion of women experiencing obesity is greater than men with a ratio of 15% vs 11%. Although the increase in cigarette consumption in women in Indonesia is still below the number of male smokers, the smoking trend has increased quite high. Riskesdas data for 2018 showed an increase in cigarette consumption among Indonesian women by 4.8%. In 1995 or 20 years ago, it was

found that 4 out of 100 Indonesian women were smokers. Data in 2016 shows that 7 out of 100 Indonesian women are active smokers. Oral contraception or pill is the second-highest contraceptive method after injection that is widely used by women of childbearing age. Pill use was 26.60% and injections were 48.56% [23, 26]. Besides, the high consumption of high-fat diets and high-calorie diets is also a predictor variable in the calculation of breast cancer risk in this study. This is caused by the high consumption of fat and calories, diet, and varied food processing in Indonesia. Physical activity in women who usually do housework, but lack physical cardio activities such as jogging, swimming, cycling, aerobics and regular light walking every week.

A comparison of sensitivity to detect breast cancer risk seen based on Gail model is 5.0% with a specificity of 97.1%. The sensitivity of the Health Risk Appraisal (HRA) model conducted in China is 70.0% and the specificity is 60.6%. The accuracy of the Gail model is 0.542 (95% CI, 0.426–0.658) and the HRA model (AUC, 0.734 (95% CI, 0.643–0.825) [7, 27]. If the Gail model is compared with the model created by researchers it has an accuracy of 79.9% with a sensitivity of 76.90% and a specificity of 70.4% These results conclude that the tools produced by researchers are quite accurate and the early detection algorithm for breast cancer risk is good enough to classify subjects at risk of breast cancer or not at risk of breast cancer.

Based on the results of this study, the model used can help to determine whether a person has a risk of developing breast cancer. This model is very important in helping early detection and preference for breast cancer. Calculating the risk by using scoring can help the community to be able to carry out routine checks for early detection of breast cancer and assist health care providers in finding people at risk of breast cancer. Measurement of breast cancer risk can determine whether a person has a safe risk of breast cancer, adequate for the prevention of breast cancer or dangerous against the occurrence of breast cancer. If someone in the dangerous category, the action that must be hastened is to do screening to ascertain whether a person has breast cancer or not. If someone in the prevention behavior is sufficient, they are advised to keep their behavior to avoid breast cancer. Included in the safe category, a person is recommended to maintain behavior and avoid breast cancer risk factors.

The problem currently faced is that breast cancer in women is not diagnosed early. It can be caused by ignorance of the patient (patient delay), ignorance of the doctor or medical staff (doctor delay), or hospital delay. The low perception of risk of breast cancer causes the majority of women to underestimate their risk which may have an important influence on the practice of early detection and attention to medical symptoms so that it can affect the delayed discovery of breast cancer [2, 3].

Early detection of breast cancer contributes to a reduction in the number of deaths from breast cancer. Another beneficial value of early detection of breast cancer is the reduced cost effect. If someone is identified at an advanced stage of breast cancer that requires further treatment, he would certainly spend a lot of money for treatment. But someone who knows the risk of breast cancer early and makes prevention efforts as early as possible, he will save more costs. Previous research by Nguyen et al., Explained the value of early detection of breast cancer risk in incremental cost-effectiveness ratios (ICER)

or the cost-effectiveness ratio per year of life obtained from the the first year after screening mammography in Vietnam was $ 3,647.06 and $ 4,405.44 for women aged 50–54 years and 55–59 years [28]. Roshidian et al stated the cost-effectiveness per year of life obtained when screening for breast cancer risk ranged from $ 1,634 (age 50 years in India) to $ 65,000 (younger screening age in Australia). Biennial screening tests for those aged 50–70 years have cost-effectiveness ($ 2,685) [29]. Okonkwo et al have known the cost-effectiveness per year of life that can be obtained if one screened at the age of 50 years and twice in the 50–70 year age range respectively: $ 1,634 and $ 3,308; cost per death avoided: $ 22,220 and $ 36,731 [30].

Based on the results of calculating the risk of breast cancer, the researchers applied the scoring calculation through learning media outputs or manual-based applications; scorecard and smartphone for easy and flexible use, practical, open access and without charge. Utilization of this application is expected to be used anywhere, anytime, easy to carry, easy to obtain or load. Smartphone-based applications in this study are made use of because smartphone is a necessity and almost owned by the whole community. Not only smartphone is used for entertainment, but it also begins to use for learning. The development of learning media using this smartphone is expected to facilitate the public in conducting initial screening in detecting breast cancer risk. And if breast cancer is found to be at high risk, the public can immediately consult a medical professional that is a tumor specialist (oncologist) to take measures to reduce the risk of breast cancer, including further examination. This application is expected to increase public awareness of the importance of early detection or screening and increase public knowledge of breast cancer risk. This research is expected to be able to contribute to reducing the high number of breast cancer communities at an advanced stage.

Conclusion

This study proves the relationship of non-modifiable risk factors; menopause age, age at the first pregnancy, the first-degree family history of breast cancer and second-degree family history of breast cancer with breast cancer in women in Indonesia. There is a significant relationship between modifiable risk factors and the use of oral contraceptives, smoking history, body mass index, high-fat diets, high-calorie diets and physical activity with the occurrence of breast cancer in Indonesia. The construction of a breast cancer risk factor calculation model can identify breast cancer risk with an accuracy of 79.9%, sensitivity of 76.9% and specificity of 70.4%.

Abbreviations

AUC
Area Under Curve
BCDDP
The Breast Cancer Detection Demonstration Project
CIs
Confidence Intervals

DT
Decision Tree
GM1
Gail Model 1
GPAQ
The Global Physical Activity Questionnaire
HRA
Health Risk Appraisal
ICER
Incremental Cost-Effectiveness Ratios
KNN
k-nearest neighbors
LR
Logistic Regression
NB
Naive Bayes
OR
Odds Ratios
RDA
Recommended Dietary Allowance
ROC
Receiver Operating Characteristics
SM-FFQ
Semi-Quantitative Food Frequency Questionnaire
SVM
Support Vector Machine
WEKA
Waikato Environment for Knowledge Analysis
WHO
World Health Organization

Declarations

Ethics approval and consent to participate

This study passed the ethical review by the ethics commiittee of the Faculty of Medicine, Public Health and Nursing, Universitas Gadjah Mada University, Yogyakarta, Indonesia (No.KE/FK/0717/EC/2018).

Consent for publication

Not applicable.

Availability of data and materials

The datasets analyzed of the present study could be obtained from the corresponding author upon reasonable request.

Competing interest

The authors declared no potential conflicts of interest.

Funding

Not applicable

Authors' contributions

RDN, TA, LL and ID conceived and designed the study. RDN, WAH, SJH collected the data. RDN and HK performed analysis and interpretation. RDN wrote the first draft with critical feedback from TA, LL, ID, HK, SJH and WAH. All authors read, reviewed and edited the draft and approved the final version of the manuscript.

Acknowledgments

The authors would like to thank Syiva Uwa Rahmah for collecting data. Mac Arif Hamdanas, MA for translating.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68:394–424.
2. Harahap WA, Ramadhan, Khambri D, Haryono S, Nindrea RD. Outcomes of trastuzumab therapy for 6 and 12 months in Indonesian national health insurance system clients with operable HER2-positive breast cancer. Asian Pac J Cancer Prev. 2017;18:1151–7.
3. Nindrea RD, Aryandono T, Lazuardi L. Breast cancer risk from modifiable and non–modifiable risk factors among women in Southeast Asia: A meta–analysis. Asian Pac J Cancer Prev. 2017;18:3201–6.
4. Maas P, Barrdahl M, Joshi AD, Auer PL, Gaudet MM, Milne RL, et al. Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States. JAMA Oncol. 2016;2(10):1295–302.
5. Willet WC. Diet and breast cancer. J Intern Med. 2001;249(5):395–411.
6. Clemons M, Goss P. Estrogen and the risk of breast cancer. N Engl J Med. 2001;344:276–85.

7. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst. 1989;81:1879–86.
8. Claus EB, Risch N, Thompson WD. Genetic analysis of breast cancer in the cancer and steroid hormone study. Am J Hum Genet. 1991;48:232–42.
9. Parmigiani G, Berry DA, Aguilar O. Determining Carrier Probabilities for Breast Cancer–Susceptibility Genes BRCA1 and BRCA2. Am J Hum Genet. 1998;62(1):145–58.
10. Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. Stat Med. 2004;23:1111–30.
11. Evans DG, Howell A. Breast cancer risk-assessment models. Breast Cancer Res. 2007;9(5):213.
12. Wu JH, Chang YK, Hou YC, Chiu WJ, Chen JR, Chen ST, et al. Meat-fat dietary pattern may increase the risk of breast cancer—A case–control study in Taiwan. Tzu Chi Medical Journal. 2013;25(4):233–8.
13. Lee AJ, Cunningham AP, Kuchenbaecker KB, Mavaddat N, Easton DF, Antoniou AC. BOADICEA breast cancer risk prediction model: updates to cancer incidences, tumour pathology and web interface. Br J Cancer. 2004;110(2):535–45.
14. Li H, Sun X, Miller E, Wang Q, Tao P, Liu L, et al. BMI, reproductive factors, and breast cancer molecular subtypes: A case-control study and meta-analysis. J Epidemiol. 2017;27(4):143–51.
15. Ritte R, Lukanova A, Berrino F, Dossus L, Tjonneland A, Olsen A, et al. Adiposity, hormone replacement therapy use and breast cancer risk by age and hormone receptor status: a large prospective cohort study. Breast Cancer Res. 2012;14:R76.
16. Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50302 women with breast cancer and 96973 women without the disease. Lancet. 2002;360(9328):187–95.
17. Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and hormonal contraceptives: collaborative reanalysis of individual data on 53297 women with breast cancer and 100 239 women without breast cancer from 54 epidemiological studies. Lancet. 1996;347:1713–27.
18. Hinyard L, Wirth LS, Clancy JM, Schwartz T. The effect of marital status on breast cancer-related outcomes in women under 65: A SEER database analysis. Breast. 2017;32:13–7.
19. Strumylaite L, Sharp SJ, Kregzdyte R, Poskiene L, Bogusevicius A, Pranys D. The association of low-to-moderate alcohol consumption with breast cancer subtypes defined by hormone receptor status. PloS One. 2015;10(12):e0144680.
20. Expert Consultation WHO. Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies. Lancet. 2004;363:157–63.
21. Passarelli MN, Newcomb PA, Hampton JM, Trentham-Dietz A, Titus LJ, Egan KM, et al. Cigarette smoking before and after breast cancer diagnosis: mortality from breast cancer and smoking-related diseases. J Clin Oncol. 2016;34(12):1315–22.

22. Armstrong T, Bull F. Development of the world health organization global physical activity questionnaire (GPAQ). J Public Health. 2006;14:66–70.
23. Ministry of Health Republic of Indonesia. Basic Health Research in 2013. Jakarta. Ministry of Health Republic of Indonesia. 2013.
24. Chen HL, Zhou MQ, Tian W, Meng KX, He HF. Effect of age on breast cancer patient prognoses: a population-based study using the SEER 18 database. PLoS One. 2016;11(10):e0165409.
25. Collaborative Group on Hormonal Factors in Breast Cancer. Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118†964 women with breast cancer from 117 epidemiological studies. Lancet Oncol. 2012;13(11):1141–51.
26. Ministry of Health Republic of Indonesia. Basic Health Research in 2018. Jakarta. Ministry of Health Republic of Indonesia. 2018.
27. Zhou W, Ding Q, Pan H, Wu N, Liang M, Huang Y, et al. Risk of breast cancer and family history of other cancers in first-degree relatives in Chinese women: a case control study. BMC Cancer. 2014;14:662.
28. Nguyen CP, Adang EMM. Cost effectiveness of breast cancer screening using mammography in Vietnamese women. PLoS One. 2018;13(3):e0194996.
29. Rashidian A, Barfar E, Hosseini H, Nosratnejad S, Barooti E. Cost effectiveness of breast cancer screening using mammography; a systematic review. Iran J Publ Health. 2013;42(4):347–57.
30. Okonkwo QL, Draisma G, der Kinderen A, Brown ML, de Koning HJ. Breast cancer screening policies in developing countries: a cost-effectiveness analysis for India. J Natl Cancer Inst. 2008;100(18):1290–300.

Tables

Table 1. Relationship between non-modifiable and modifiable risk factors with breast cancer risk

Variables	Group		p-value
	Cases (f/%) (n=1,000)	Control (f/%) (n=1,000)	
A. Non-modifiable			
Age (year-old)			1.000
≥ 50	489 (48.9)	489 (48.9)	
< 50	511 (51.1)	511 (51.1)	
Age of menarche (year-old)			0.269
7-11	67 (6.7)	67 (6.7)	
12-13	479 (47.9)	444 (44.4)	
>13	454 (45.4)	489 (48.9)	
Age of menopause (year-old)			<0.001*[a]
≥ 50	634 (63.4)	440 (44.0)	
< 50	366 (36.6)	560 (56.0)	
Age of the first pregnancy (year-old)			<0.001*[a]
< 20	158 (15.8)	105 (10.5)	
20-29	634 (63.4)	803 (80.3)	
> 30	164 (16.4)	54 (5.4)	
Never been pregnant	44 (4.4)	38 (3.8)	
Parity			0.548
Nulipara	44 (4.4)	38 (3.8)	
Primipara	128 (12.8)	142 (14.2)	
≥ Multipara	828 (82.8)	820 (82.0)	
The first-degree family history of breast cancer			<0.001*[a]
Yes	119 (11.9)	3 (0.3)	
No	881 (88.1)	997 (99.7)	
Second-degree family history of breast cancer			<0.001*[a]
Yes	88 (8.8)	19 (1.9)	
No	912 (91.2)	981 (98.1)	
B. Modifiable			
Hormone replacement therapy			0.625
Yes	3 (0.3)	1 (0.1)	
No	997 (99.7)	999 (99.9)	
Lactation			0.305
≥ 12 months	850 (85.0)	867 (86.7)	
< 12 months	150 (15.0)	133 (13.3)	
Oral contraceptives			<0.001*[a]
≥ 12 months	332 (33.2)	156 (15.6)	
< 12 months	668 (66.8)	844 (84.4)	
Marital Status			0.268
Single/widowed	58 (5.8)	46 (4.6)	
Married	942 (94.2)	954 (95.4)	
Alcohol consumption			0.385
Yes	8 (0.8)	4 (0.4)	
No	992 (99.2)	996 (99.6)	
Smoking			0.002*[a]
Yes	22 (2.2)	5 (0.5)	
No	978 (97.8)	995 (99.5)	
BMI			<0.001*[a]
Normal	388 (38.8)	601 (60.1)	
Overweight	177 (17.7)	112 (11.2)	
Obesity	435 (43.5)	287 (28.7)	
High-fat diet			<0.001*[a]
High	648 (64.8)	542 (54.2)	
Normal	352 (35.2)	458 (45.8)	
High-calorie diet			<0.001*[a]
High	486 (48.6)	185 (18.5)	
Normal	514 (51.4)	815 (81.5)	
Physical activity			0.021*[a]
Light	257 (25.7)	210 (21.0)	

Moderate	88 (8.8)	78 (7.8)
Hard	655 (65.5)	712 (71.2)

*, significant at p<0.05

[a], p<0.25 entered the candidate variable and multivariate analysis

Table 2. Mantel-Haenszel test for breast cancer risk based on stratification according to ethnic Minangkabau and Javanese

Variables	Category	Minangkabau (n-1,200)		Javanese (n=800)		p value
		Cases(f/%)	Control (f/%)	Cases(f/%)	Control (f/%)	
Menopause	<50 (year-old)	241 (40.2)	374 (62.3)	125 (31.3)	186 (46.5)	ref
	≥ 50 (year-old)	359 (59.8)	226 (37.7)	275 (68.8)	214 (53.5)	<0.001*
Age of the first pregnancy	20-29 (year-old)	431 (71.8)	493 (82.2)	203 (50.8)	310 (77.5)	ref
	<20 (year-old)	85 (14.2)	73 (12.2)	73 (18.3)	32 (8.0)	<0.001*
	≥ 30 (year-old)	75 (12.5)	34 (5.7)	89 (22.3)	20 (5.0)	<0.001*
	Never	9 (1.5)	0	35 (8.8)	38 (9.5)	0.002*
The first-degree family history	No	536 (89.3)	598 (99.7)	345 (86.3)	399 (99.8)	ref
	Yes	64 (10.7)	2 (0.3)	55 (13.8)	1 (0.3)	<0.001*
Second-degree family history	No	538 (89.7)	590 (98.3)	374 (93.5)	391 (97.8)	ref
	Yes	62 (10.3)	10 (1.7)	26 (6.5)	9 (2.3)	<0.001*
Oral contraceptives	< 12 months/ none	459 (76.5)	501 (83.5)	209 (52.3)	343 (85.8)	ref
	≥12 months	141 (23.5)	99 (16.5)	191 (47.8)	57 (14.3)	<0.001*
Smoking	No	584 (97.3)	597 (99.5)	394 (98.5)	398 (99.5)	ref
	Yes	16 (2.7)	3 (0.5)	6 (1.5)	2 (0.5)	0.002*
BMI	Normal	250 (41.7)	347 (57.8)	138 (34.5)	254 (63.5)	ref
	Overweight	101 (16.8)	65 (10.8)	76 (19.0)	47 (11.8)	<0.001*
	Obesity	249 (41.5)	188 (31.3)	186 (46.5)	99 (24.8)	<0.001*
High-fat diet	Normal	55 (9.2)	274 (45.7)	297 (74.3)	184 (46.0)	<0.001*
	High	545 (90.8)	326 (54.3)	103 (25.8)	216 (54.0)	<0.001*
High-calorie diet	Normal	401 (66.8)	463 (77.2)	113 (28.3)	352 (88.0)	ref
	High	199 (33.2)	137 (22.8)	287 (71.8)	48 (12.0)	<0.001*
Physical activity	Heavy	341 (56.8)	436 (72.7)	314 (78.5)	276 (69.0)	ref
	Moderate	88 (14.7)	43 (7.2)	0	35 (8.8)	0.003
	Low	171 (28.5)	121 (20.2)	86 (21.5)	89 (22.3)	<0.001*

*, p<0.05 homogeneous in the Mantel-Haenszel test

Table 3. Multivariate analysis of breast cancer

Variables	Category	OR Unadjusted		OR Adjusted	
		OR (95%CI)	p value	OR (95%CI)	p value
Menopause	<50 (year-old)	ref	ref	ref	ref
	≥ 50 (year-old)	2.21 (1.84-2.64)	<0.001*	2.49 (2.01-3.09)	<0.001*
Age of the first pregnancy	20-29 (year-old)	ref	ref	ref	ref
	<20 (year-old)	1.91 (1.46-2.49)	<0.001*	1.74 (1.27-2.39)	0.001*
	≥ 30 (year-old)	3.85 (2.78-5.32)	<0.001*	4.86 (3.36-7.05)	<0.001*
	Never	1.47 (0.94-2.29)	0.093	2.54 (1.49-4.34)	0.001*
The first-degree family history	No	ref	ref	ref	ref
	Yes	44.89 (14.22-141.68)	<0.001*	33.92 (10.56-109.01)	<0.001*
Second-degree family history	No	ref	ref	ref	ref
	Yes	4.98 (3.01-8.25)	<0.001	5.34 (3.04-9.38)	<0.001*
Oral contraceptives	< 12 months/ none	ref	ref	ref	ref
	≥12 months	2.69 (2.17-3.34)	<0.001*	2.61 (2.02-3.37)	<0.001*
Smoking	No	ref	ref	ref	ref
	Yes	4.48 (1.69-11.87)	0.003	3.96 (1.32-11.90)	0.014*
BMI	Normal	ref	ref	ref	ref
	Overweight	2.45 (1.87-3.20)	<0.001*	2.26 (1.64-3.11)	<0.001*
	Obesity	2.35 (1.93-2.86)	<0.001*	2.25 (1.79-2.84)	<0.001*
High-fat diet	Normal	ref	ref	ref	ref
	High	1.56 (1.30-1.86)	<0.001*	2.42 (1.93-3.04)	<0.001*
High-calorie diet	Normal	ref	ref	ref	ref
	High	4.17 (3.40-5.10)	<0.001*	3.735 (2.97-4.70)	<0.001*
Physical activity	Heavy	ref	ref	ref	ref
	Moderate	1.23 (0.89-1.69)	0.215	1.15 (0.78-1.70)	0.479
	Light	1.33 (1.08-1.64)	0.008*	1.42 (1.10-1.83)	0.007*

*, significant at p<0.05; ref, reference; BMI

Table 4. Comparison of breast cancer risk calculations in Indonesia based on machine learning

Algorithms	Accuracy (%)	TP Rate (%)	FP Rate (%)	AUC (%)
NB	74,0	73,8	26,2	80,7
DT	74,5	74,5	25,6	80,8
KNN	74,1	73,6	26,5	79,9
SVM	73,9	73,8	26,2	73,8
LR	74,7	74,7	25,3	81,0

AUC, Area Under Curve
TP, True Positive
FP, False-*Positive*

Table 5. Determination of scoring through the calculation of breast cancer risk

Variables	B	S.E	B/S.E	B/S.E/2.45	Score
Menopause (≥ 50 year-old)	0,911	0,110	8,283	3,381	3
The first pregnancy (< 20 year-old)	0,555	0,161	3,452	1,409	1
The first pregnancy (≥ 30 year-old)	1,582	0,189	8,352	3,409	3
Never been pregnant	0,932	0,273	3,416	1,394	1
The first-degree family history	3,524	0,596	5,916	2,415	2
Second-degree family history	1,675	0,288	5,820	2,376	2
Oral contraseptive (≥12 months)	0,960	0,130	7,375	3,010	3
Smoking (≥ 12 months)	1,376	0,561	2,453	1,001	1
Overweight	0,814	0,164	4,981	2,033	2
Obesity	0,812	0,118	6,867	2,803	3
High-fat diet	0,885	0,116	7,653	3,124	3
High-calorie diet	1,318	0,118	11,200	4,571	5
Low physical activity	0,350	0,129	2,711	1,106	1
Total score					**30**

Table 6. The significance of the total score in breast cancer screening modeling in Indonesia

	B	S.E	p value	OR (95% CI)
Breast cancer scoring model	0,317	0,016	<0,001	1,373 (1,331-1,415)
a constant	**-2,563**			

Table 7. The poor prognosis probability for each score

Subject score	a constant	Coefficient	y = -2.563 + 0.317 x total score	$p = \frac{1}{1+\exp(-y)}$	p (%)
0	-2.563	0.317	-2.563	0.072	7.156
1	-2.563	0.317	-2.246	0.096	9.570
2	-2.563	0.317	-1.929	0.127	12.686
3	-2.563	0.317	-1.612	0.166	16.631
4	-2.563	0.317	-1.295	0.215	21.501
5	-2.563	0.317	-0.978	0.273	27.329
6	-2.563	0.317	-0.661	0.341	34.052
7	-2.563	0.317	-0.344	0.415	41.484
8	-2.563	0.317	-0.027	0.493	49.325
9	-2.563	0.317	0.29	0.572	57.200
10	-2.563	0.317	0.607	0.647	64.726
11	-2.563	0.317	0.924	0.716	71.586
12	-2.563	0.317	1.241	0.776	77.574
13	-2.563	0.317	1.558	0.826	82.607
14	-2.563	0.317	1.875	0.867	86.704
15	-2.563	0.317	2.192	0.900	89.953
16	-2.563	0.317	2.509	0.925	92.477
17	-2.563	0.317	2.826	0.944	94.406
18	-2.563	0.317	3.143	0.959	95.863
19	-2.563	0.317	3.46	0.970	96.953
20	-2.563	0.317	3.777	0.978	97.762
21	-2.563	0.317	4.094	0.984	98.360
22	-2.563	0.317	4.411	0.988	98.800
23	-2.563	0.317	4.728	0.991	99.123
24	-2.563	0.317	5.045	0.994	99.360
25	-2.563	0.317	5.362	0.995	99.533
26	-2.563	0.317	5.679	0.997	99.659
27	-2.563	0.317	5.996	0.998	99.752
28	-2.563	0.317	6.313	0.998	99.819
29	-2.563	0.317	6.63	0.999	99.868
30	-2.563	0.317	6.947	0.999	99.904

Table 8. Potential cut off points in classifying breast cancer risk

Potential Cut off points	Sensitivity	95% CI	Specificity	95% CI	+LR	-LR
≥0	100.00	99.6 - 100.0	0.00	0.0 - 0.4	1.00	
>0	99.90	99.4 - 100.0	7.50	5.9 - 9.3	1.08	0.013
>1	99.60	99.0 - 99.9	10.00	8.2 - 12.0	1.11	0.040
>2	99.40	98.7 - 99.8	12.80	10.8 - 15.0	1.14	0.047
>3	96.10	94.7 - 97.2	30.60	27.8 - 33.6	1.38	0.13
>4	94.10	92.5 - 95.5	37.30	34.3 - 40.4	1.50	0.16
>5	91.80	89.9 - 93.4	41.80	38.7 - 44.9	1.58	0.20
>6	83.40	80.9 - 85.7	62.20	59.1 - 65.2	2.21	0.27
>7	76.90	74.2 - 79.5	70.40	67.5 - 73.2	2.60	0.33
>8	68.60	65.6 - 71.5	75.40	72.6 - 78.0	2.79	0.42
>9	56.50	53.4 - 59.6	84.50	82.1 - 86.7	3.65	0.51
>10	49.10	46.0 - 52.2	86.60	84.3 - 88.7	3.66	0.59
>11	37.00	34.0 - 40.1	92.50	90.7 - 94.1	4.93	0.68
>12	27.80	25.0 - 30.7	95.50	94.0 - 96.7	6.18	0.76
>13	22.50	19.9 - 25.2	96.40	95.1 - 97.5	6.25	0.80
>14	12.60	10.6 - 14.8	98.90	98.0 - 99.4	11.45	0.88
>15	8.50	6.8 - 10.4	99.30	98.6 - 99.7	12.14	0.92
>16	4.90	3.6 - 6.4	99.60	99.0 - 99.9	12.25	0.95
>17	2.10	1.3 - 3.2	99.90	99.4 - 100.0	21.00	0.98
>18	0.80	0.3 - 1.6	100.00	99.6 - 100.0		0.99
>19	0.20	0.02 - 0.7	100.00	99.6 - 100.0		1.00
>20	0.00	0.0 - 0.4	100.00	99.6 - 100.0		1.00

Figures

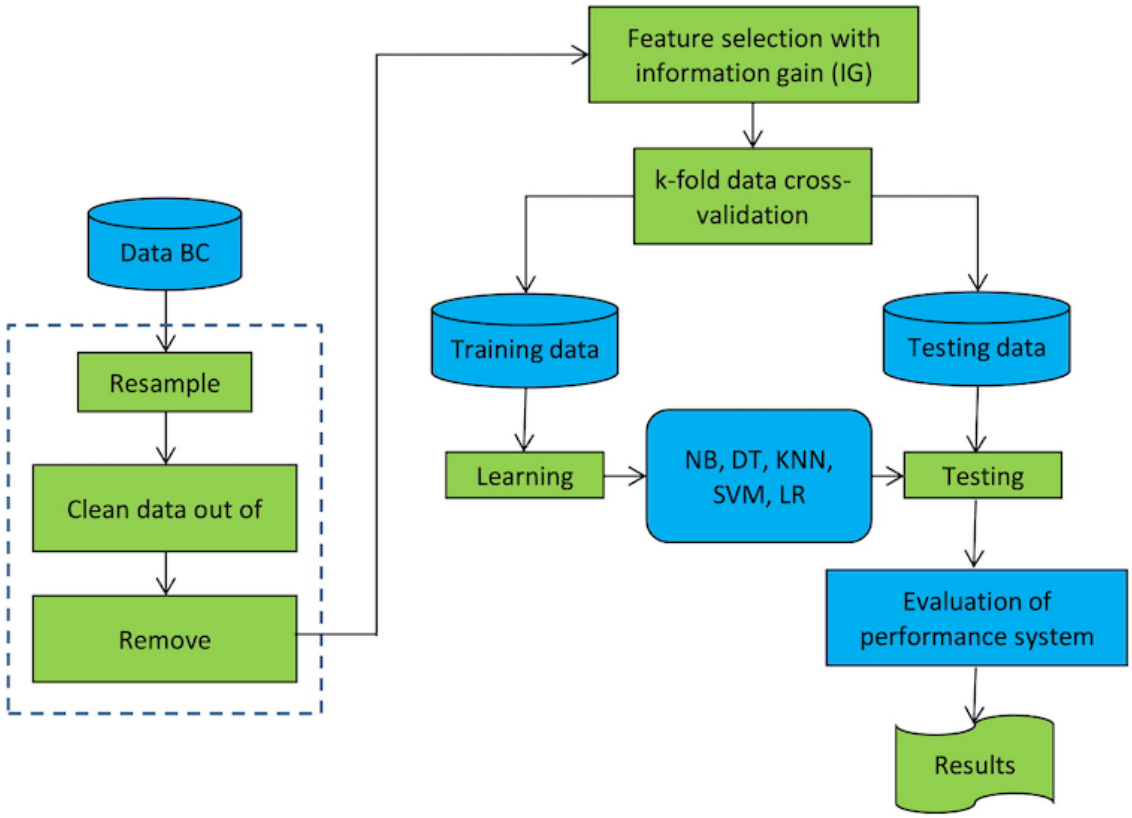

Figure 1

Preprocessing stages to calculate breast cancer risk using machine learning

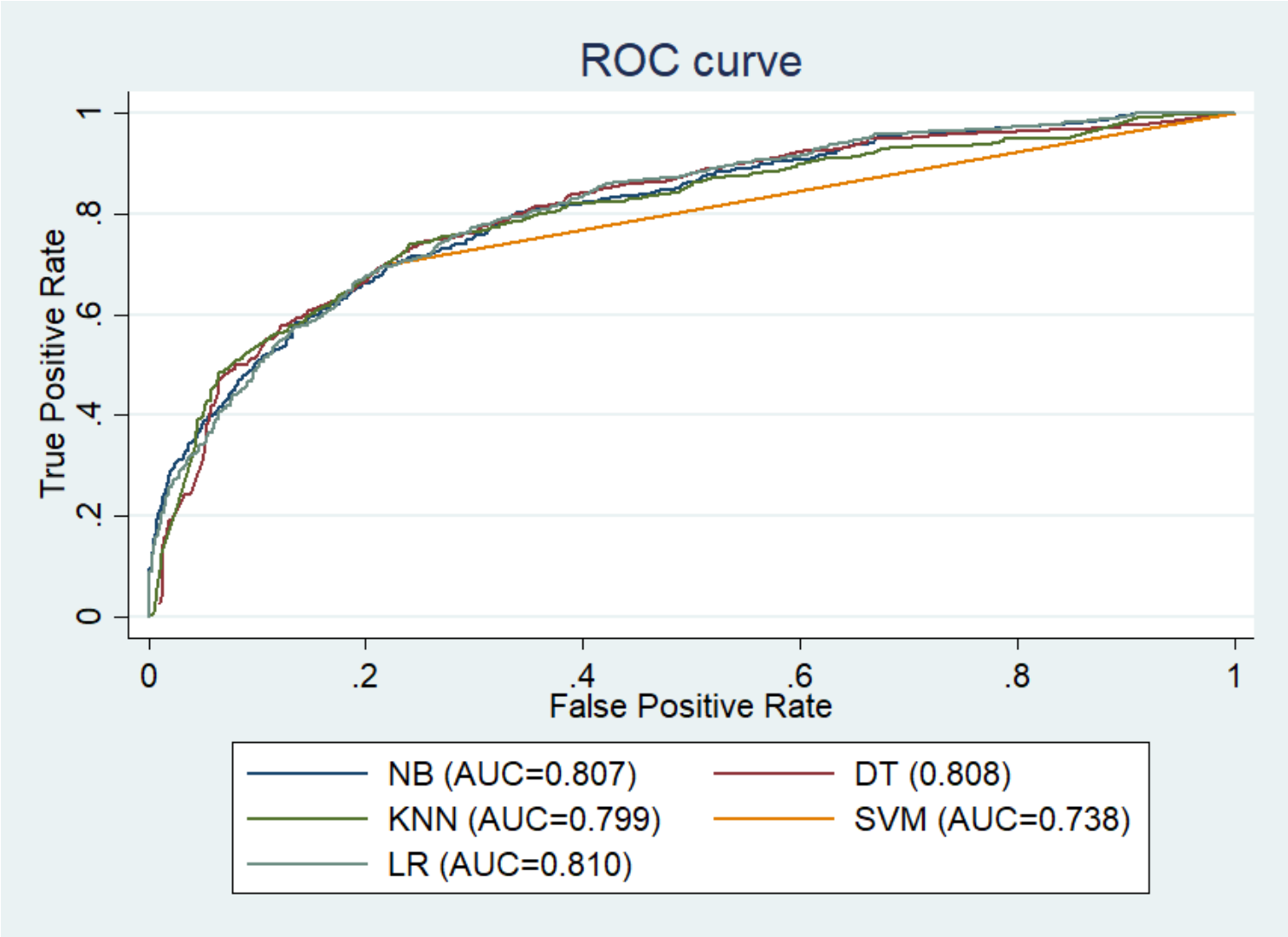

Figure 2

known the highest AUC value based on machine learning algorithm is logistic regression with AUC = 0.810, so the determination of breast cancer risk scoring was used logistic regression algorithm.

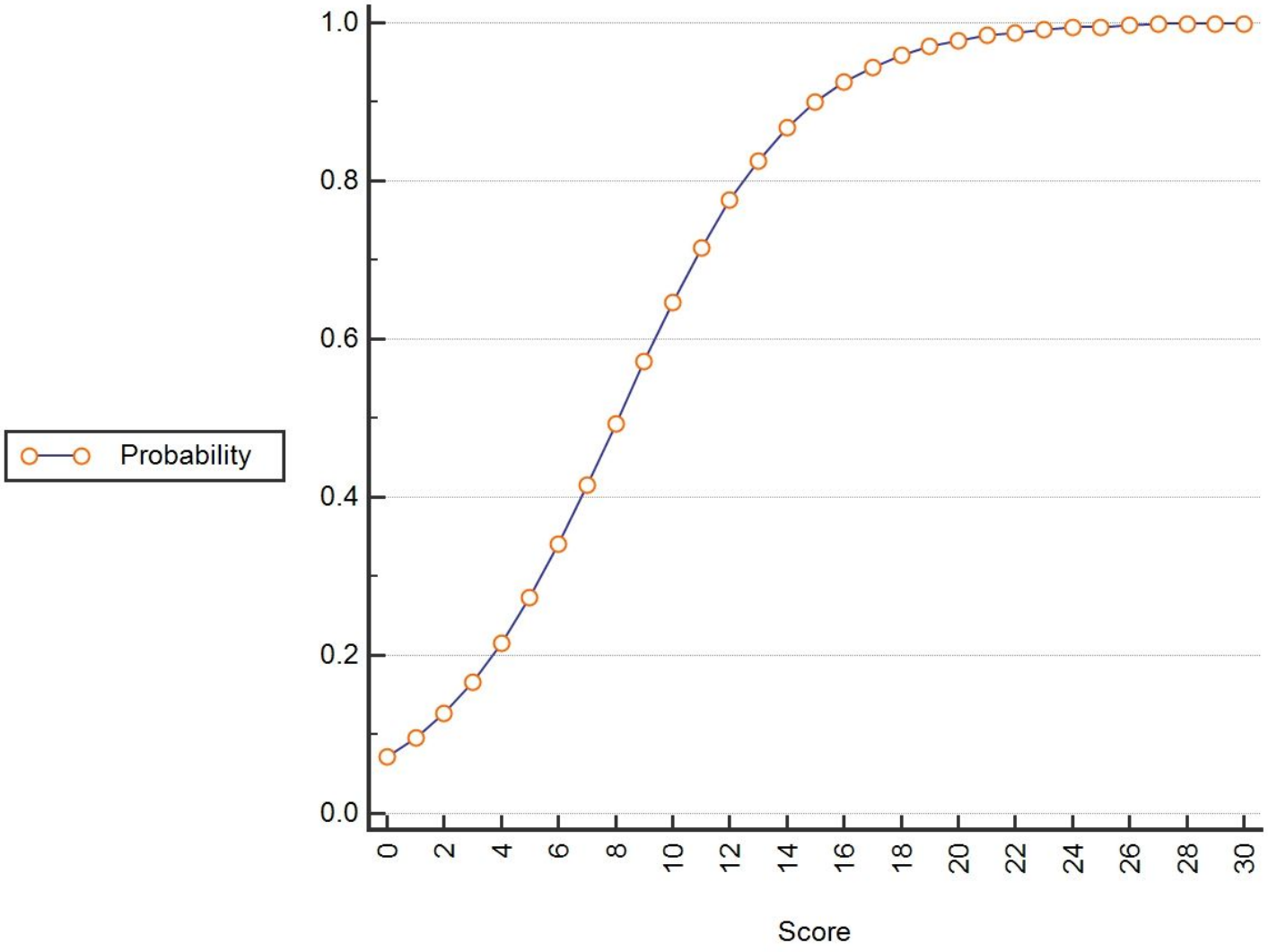

Figure 3

Poor prognosis probability curves for each score

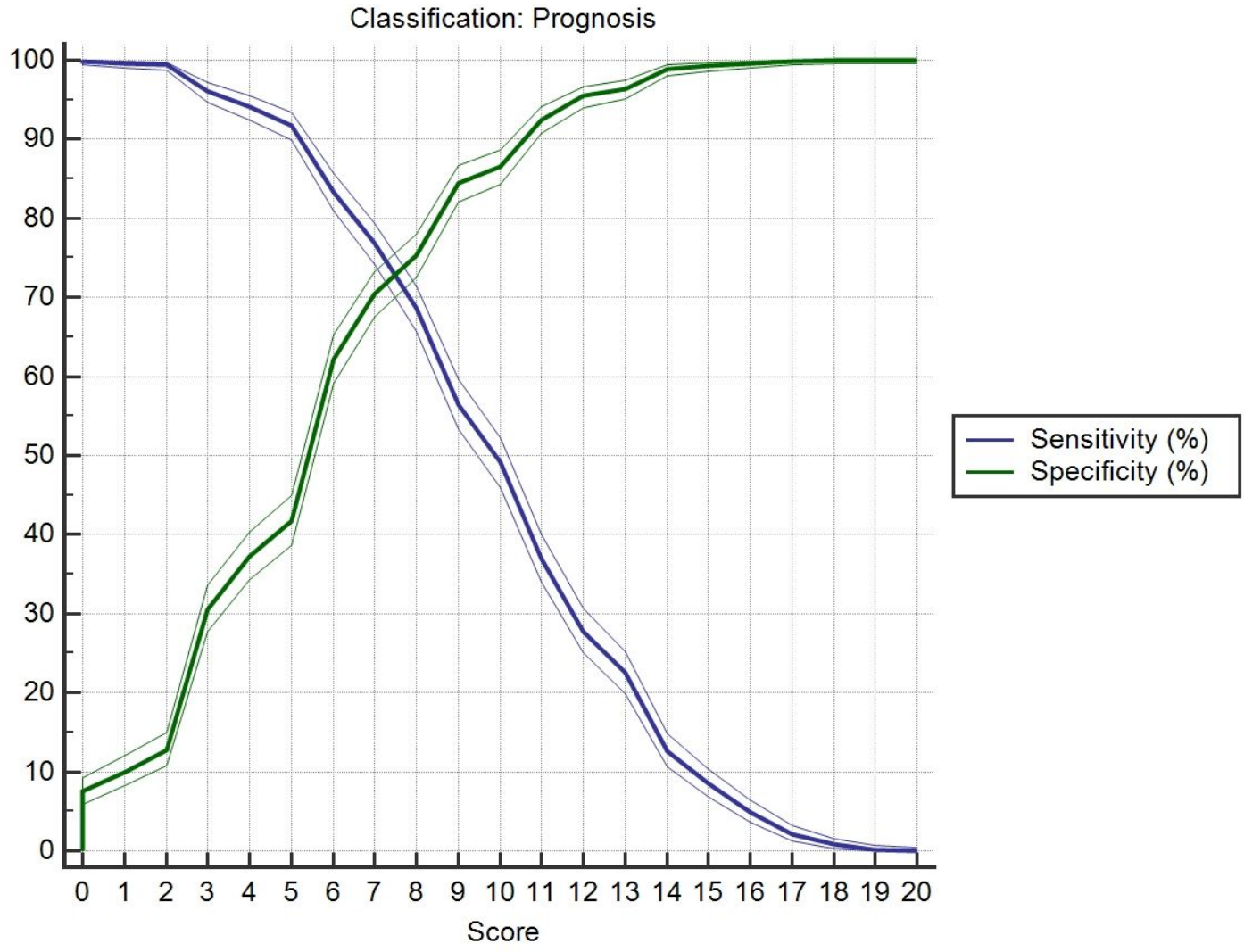

Figure 4

known the cut-off point for classifying breast cancer risk and non-risk groups is at a cutoff score> 7 with a sensitivity of 76.90% and a specificity of 70.4%. It can be concluded that there is a risk of breast cancer if the total score from the accumulation of predictor variables is >7 and there is no risk of breast cancer if the total score from the accumulated predictor variable is ≤7.

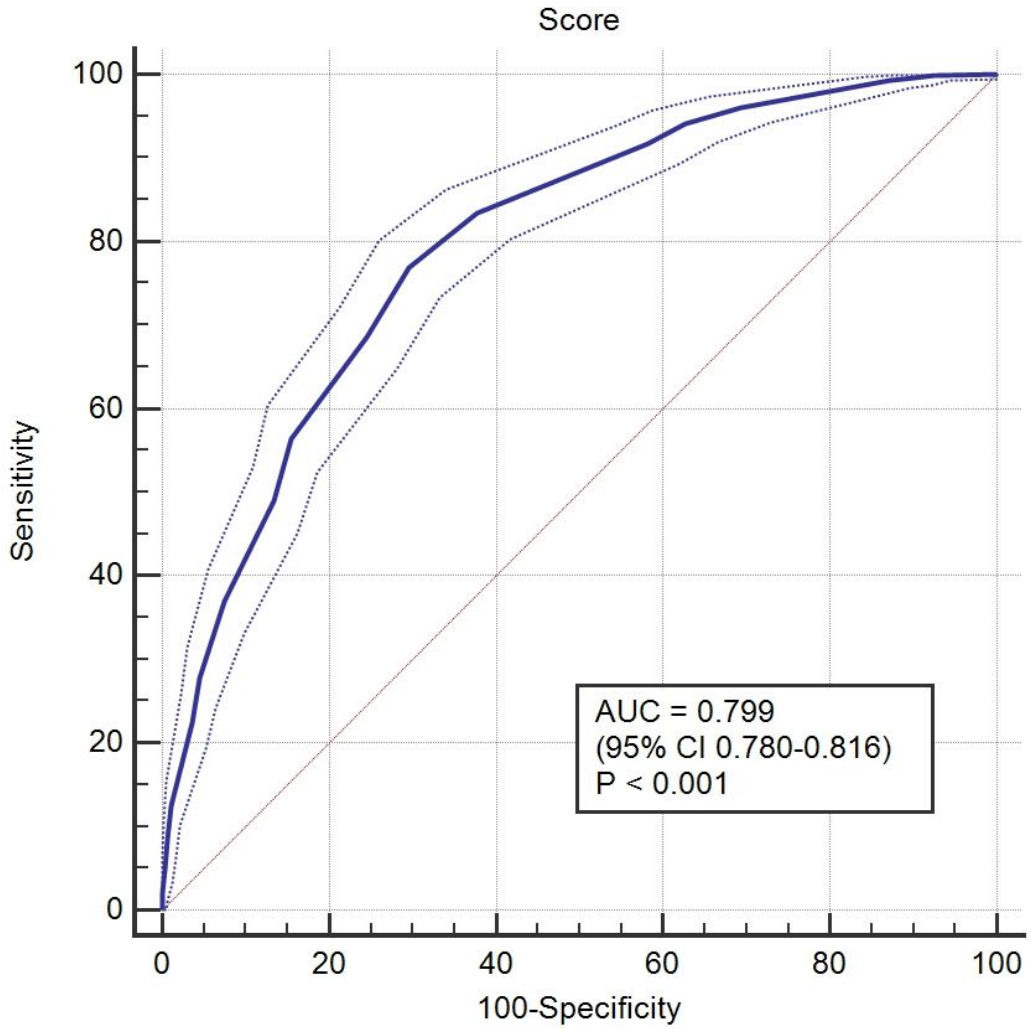

Figure 5

known the risk factor model for breast cancer in Indonesia has an AUC of 0.799 (79.9%). This model is categorized quite well in classifying breast cancer risk in Indonesia.