PREDICTION OF BREAST CANCER DIAGNOSIS USING MACHINE LEARNING IN MALAYSIAN WOMEN

by

TENGKU MUHAMMAD HANIS BIN TENGKU MOKHTAR

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

March 2024

ACKNOWLEDGEMENT

Praise to Allah, the All-knowing for His mercy and love, for giving me a chance, endurance, and strength to walk on this journey. This three-year journey has been an enjoyable and eye-opening yet challenging and humbling time for me.

My deepest appreciation to my main supervisor, Associate Professor Dr Kamarul Imran Musa for taking me under his wing. Without him, this journey will be a lot more difficult than it should be.

My deepest appreciation also to my co-supervisors, Associate Professor Dr Juhara Haron, and Professor Dr Rosni Abdullah. I thank them for their time and patience in guiding me during my PhD journey.

My heartfelt gratitude goes to all FRGS team members, Associate Professor Dr Wan Faiziah Wan Abdul Rahman, Dr Wan Nor Arifin Wan Mansor, Dr Nur Intan Raihana Ruhaiyem, and the late Associate Professor Dr Bakri Adam (May Allah reward him Jannah), as well as Dr Md Asiful Islam. I have learnt a lot from them.

To my fellow PhD mates, Nadia, Subirman, Hamid, Shakira, Dr Wira, Dr Hidayat, and Dr Zarudin, I thank them for their companionship and friendship.

I will always be grateful to all the lecturers of my master's program at the Biostatistics and Research Methodology Unit, USM, for opening the door to my academic pursuits.

Finally, I offer my sincerest appreciation to my parents, YM Tengku Mokhtar Tengku Zainal and Nor Malaysia Rose, for their immense support, tolerance, understanding, and never-ending dua throughout my journey. I am also eternally grateful to my wife, Nurul Asmaq Mazalan, for her unwavering love, support, and understanding, for me to complete this journey.

To myself, you made it, despite all the odds. Keep hustling and stay humble.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF SYMBOLS

%	Percentage
P	*P* value
=	Equal to
>	More than
<	Less than
≥	More than and equal to
≤	Less than and equal to
Δ	The change

LIST OF ABBREVIATIONS

AI	Artificial intelligence
ANN	Artificial neural network
AUC	Area under the curve
BCRAT	Breast Cancer Risk Assessment Tool
BestARi	Breast Cancer Awareness and Research Unit
BI-RADS	Breast Imaging-Reporting and Data System
BMI	Body mass index
BOADICEA	Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm
BRCA	Breast cancer gene
BSE	Breast self-examination
CBIS-DDSM	Curated Breast Imaging Subset-Digital Database for Screening Mammography
CI	Confidence interval
CLAHE	Contrast limited adapted histogram equalisation
CNN	Convolutional neural network
CYPD	Cytochrome P450 2D6 or CYP2D6
dAUC	Difference in the area under the curve
DBT	Digital breast tomosynthesis
DDSM	Digital Database for Screening Mammography
df	Degree of freedom
DL	Deep learning
DM	Digital mammogram
DOR	Diagnostic odds ratio
DT	Decision tree
EGFR	Epidermal growth factor receptor
ELM	Extreme learning machine
FDA	Food and Drug Administration
FISH	Fluorescence in situ hybridization
FN	False negative
FP	False positive
GAM	Generalised additive model

GLOBOCAN	Global Cancer Observatory
GMM	Gaussian mixture model
GVIF	Generalised variance inflation factor
HPE	Histopathological examination
HRT	Hormonal replacement therapy
HSROC	Hierarchical summary receiver operating characteristics
HUSM	Hospital Universiti Sains Malaysia
IHC	Immunohistochemistry
IQR	Interquartile range
JEPeM	Jawatankuasa Etika Penyelidikan Manusia
kNN	k-nearest neighbour
LASSO	Least absolute shrinkage and selection operator
LDA	Linear discriminant analysis
LOO-CV	Leave-one-out cross validation
LR	Logistic regression
MARS	Multivariate adaptive regression splines
MIAS	Mammographic image analysis society
ML	Machine learning
MLP	Multilayer perceptron
MMD	Mammographic Mass Database
MRI	Magnetic resonance imaging
NA	Not available
NB	Naïve bayes
NE	Not clearly explained
NN	Neural network
OCP	Oral contraceptive pills
OR	Odds ratio
OTC	Over the counter
PACS	Picture Archiving and Communication System
PALB	Partner and localiser of BRCA2
PLS	Partial least square
PRISMA-DTA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy
PR-AUC	Precision recall-area under the curve
QUADAS-2	The updated quality assessment of diagnostic accuracy studies

RF	Random forest
RNN	Recurrent neural network
ROC	Receiver operating curve
ROI	Region of interest
ROSE	Random over-sampling examples
SD	Standard deviation
SEDATE	Synthesising evidence from diagnostic accuracy tests
SVM	Support vector machine
TAHBSO	Total abdominal hysterectomy bilateral salpingo-oophorectomy
TN	True negative
TP	True positive
UK	United Kingdom
US	United State
USA	United State of America
USM	Universiti Sains Malaysia
vs.	Versus
WDBC	Wisconsin diagnostic breast cancer
WHO	World Health Organisation
XAI	Explainable artificial intelligence
XGBoost	Extreme gradient boosting

LIST OF APPENDICES

RAMALAN DIAGNOSA KANSER PAYUDARA MENGGUNAKAN PEMBELAJARAN MESIN DALAM POPULASI WANITA MALAYSIA

ABSTRAK

Kanser payudara merupakan kanser yang paling banyak dikesan di seluruh dunia dan penyebab utama kematian kanser di dua belas rantau di dunia. Maka, terdapat keperluan untuk mewujudkan proses penyaringan dan diagnosis penyakit yang berkesan. Justeru itu, tesis ini bertujuan untuk meneroka penggunaan pembelajaran mesin (ML) untuk penilaian dan ramalan risiko kanser payudara. Tesis ini mengandungi enam kajian yang saling berkaitan, bermula daripada Bab 2 hingga Bab 7. Bab 2 memberi gambaran tentang penyelidikan kanser payudara di Malaysia. Analisis bibliometrik digunakan untuk menggambarkan aktiviti penyelidikan kanser payudara di Malaysia. Kajian ini menunjukkan bahawa tiada bidang kajian yang dominan dalam penyelidikan kanser payudara di Malaysia. Selain itu, kajian ini menemui dua tema kajian yang semakin popular berkaitan dengan kanser payudara di Malaysia. Bab 3 menyoroti penyelidikan global yang paling banyak dipetik berkaitan dengan kanser payudara dan ML. Kajian ini juga menggunakan analisis bibliometrik pada artikel penyelidikan yang paling banyak dipetik berkaitan dengan kanser payudara dan ML. Kajian ini telah menemukan bahawa terdapat minat yang mendalam berkenaan aplikasi ML pada kanser payudara dalam tiga dekad terakhir. Tiga algoritma ML yang kerap digunakan adalah pembelajaran mendalam, *support vector machine* (SVM), dan analisis gugusan. Bab 4 mengkaji faktor-faktor yang mempengaruhi kepadatan mammografi pada wanita Asia. Kajian ini menggunakan kaedah pelbagai imputasi untuk mengatasi masalah data yang hilang dan regresi logit untuk menganalisis data. Lima faktor yang mempengaruhi kepadatan mammografi

adalah umur pesakit, bilangan anak, indeks jisim tubuh, status menopaus, serta sistem pemeriksaan dan data imbasan payudara (BI-RADS). Kajian dalam Bab 5 meninjau penggunaan rekod pendaftaran pesakit dan ML untuk menganggar risiko kanser payudara. Model ML yang dihasilkan dalam bab ini boleh digunakan sebagai alat saringan pemula untuk wanita yang menghadiri klinik payudara. Lapan algoritma ML telah diterokai dalam kajian ini. Model k-jiran terdekat (kNN) mempunyai prestasi yang lebih baik berbanding tujuh model lain. Bab 6 pula membentangkan analisis meta pada model ML dalam klasifikasi kanser payudara. Kajian ini bertujuan untuk menetapkan ketepatan diagnostik bagi model ML yang digunakan pada data mammogram. Kajian ini telah menemui bahawa rangkaian neural, pembelajaran mendalam, model berasaskan pokok dan SVM memperlihatkan prestasi yang baik pada data mammogram untuk pengesanan kanser payudara. Kajian ini juga telah menentukan bahawa ML mempunyai ketepatan diagnostik yang baik, sekaligus menyokong penggunaan ML dalam bidang ini, terutamanya sebagai alat pemeriksaan dan diagnosis tambahan. Akhirnya, kajian dalam Bab 7 meninjau penggunaan kaedah gabungan rangkaian yang telah dipralatih dan data mammogram untuk klasifikasi payudara yang abnormal. Kajian ini meninjau tiga belas rangkaian yang dipralatih sebagai calon untuk membentuk model gabungan. Gabungan rangkaian yang dipralatih telah memperlihatkan prestasi yang baik dalam meramal mammogram yang normal dan meragukan. Kesimpulannya, tesis ini menonjolkan potensi ML dalam penilaian risiko dan ramalan kanser payudara. Hasil kajian dalam tesis ini menyumbang kepada karya sastera berkenaan ML dalam penyelidikan kanser payudara dan memberikan input yang berharga untuk penyelidikan pada masa akan datang.

PREDICTION OF BREAST CANCER DIAGNOSIS USING MACHINE LEARNING IN MALAYSIAN WOMEN

ABSTRACT

Breast cancer is the most prevalent cancer in the world and the main cause of cancer mortality in the twelve regions of the world. Thus, there is a need for efficient screening and diagnosis of the disease. Thus, this thesis aims to explore the use of machine learning (ML) for breast cancer risk estimation and prediction. This thesis included six interrelated projects starting from Chapter 2 to Chapter 7. Chapter 2 presents an overview of breast cancer research in Malaysia. A bibliometric analysis was used to describe the research activities of breast cancer research in Malaysia. This project revealed there was no dominant research area in breast cancer research in Malaysia. Additionally, the study found that two growing research themes related to breast cancer in Malaysia were precision medicine and deep learning. Chapter 3 explored the most cited global research related to breast cancer and ML. This project also utilised bibliometric analysis applied to the most cited papers related to breast cancer and ML. This project found that there was a strong interest in the application of ML to breast cancer in the last three decades. The three frequently used ML algorithms were deep learning, support vector machine (SVM), and cluster analysis. In Chapter 4, factors influencing mammographic density among Asian women including Malaysia women were investigated. The study utilised a multiple imputation approach to overcome a missing data issue and a logistic regression to analyse the data. Five factors affecting mammographic density were age, number of children, body mass index, menopause status, and breast imaging-reporting and data system (BI-RADS) classification. The study in Chapter 5 explored the use of patient registration

records and ML for breast cancer risk estimation. The ML model developed in this chapter could be used as an over-the-counter screening (OTC) model for women attending breast clinics. Eight ML algorithms were explored in this project. k-nearest neighbour (kNN) models had a significantly better performance compared to the other seven models. Additionally, Chapter 6 presents a meta-analysis of ML models on breast cancer classification. This project seeks to establish the diagnostic accuracy of ML used on mammographic data. This project found that neural network, deep learning, tree-based models, and SVM performed well on mammographic data for breast cancer detection. The study established the good diagnostic accuracy of ML in this area of research, thus, further supporting the use of ML in this area, especially for screening and supplementary diagnostic tools. Lastly, the study in Chapter 7 explored the use of an ensemble of pre-trained networks for breast abnormality classification using digital mammograms. This project explored thirteen pre-trained networks as candidates for the ensemble model. Each network was further fine-tuned, and the top networks were used to develop the ensemble model. The ensemble pre-trained network displayed a good performance in classifying the normal and suspicious mammograms. In conclusion, this thesis highlights the potential of ML in breast cancer risk estimation and prediction. The findings of this thesis contribute to the growing body of literature on ML in breast cancer research and provide valuable insights for future research in this area.

CHAPTER 1

INTRODUCTION

1.1 Breast cancer

Breast cancer is defined as uncontrolled cell growth in the region of the breast. In 2020, 2.3 million women were diagnosed with breast cancer, leading to 685,000 deaths worldwide (Arnold *et al.*, 2022). Moreover, breast cancer is the most prevalent form of cancer among women in at least 140 countries (DeSantis *et al.*, 2015). Thus, the World Health Organisation (WHO) considers the disease the most frequent cancer globally (WHO, 2021a). Based on the Global Cancer Observatory (GLOBOCAN) online database, breast cancer incidence in Asia amounted to 1 million cases which constituted 45.4% of the total global incidence of breast cancer in 2020 (Lei *et al.*, 2021). The study further revealed that about 50.5% of the global mortality of breast cancer occurred in the Asian region. In fact, breast cancer has been recognised as the most prevalent cancer among women in several Asian countries such as Japan, Singapore, Malaysia, Indonesia, China, South Korea, and Iran (J. H. Park *et al.*, 2018; Sinaga *et al.*, 2018; Tolou-Ghamari, 2018; National Cancer Registry Department, 2019a; Yap *et al.*, 2019; D. Sun *et al.*, 2020). Thus, these statistics make a systematic approach to breast cancer screening, diagnosis, and management in Asian countries a regional concern.

Southeast Asia is comprised of eleven countries including Malaysia, Singapore, Brunei, Indonesia, Thailand, Laos, Myanmar, Philippines, Vietnam, Cambodia, and East Timor. The incidence and mortality of breast cancer in Southeast Asia in 2019 were estimated at 139,786 and 67,276 cases, respectively (Xu *et al.*, 2021). Additionally, Xu *et al.* (2021) reported that the relative change in breast cancer incidents and deaths from 1990 to 2019 was 211% and 130%, respectively in Southeast

Asia. Furthermore, Arnold *et al.* (2022) reported that breast cancer incidence increased to 158,939 but breast cancer mortality dropped to 58,670 in 2020. Overall, breast cancer incidence in Southeast Asia comprised about 12.7% of the total breast cancer incidence in Asia in 2020 (International Agency for Research on Cancer, 2020).

1.1.1 Breast cancer incidence in Malaysia

In Malaysia, a report from the Malaysia National Cancer Registry divulged an 18.8% increase in breast cancer cases among females from 18,206 cases between 2007 and 2011 to 21,634 cases between 2012 and 2016 (National Cancer Registry Department, 2019b). The report further specified that breast cancer was the top most common cancer in Malaysia between 2012 and 2016. Based on the GLOBOCAN report in 2021, breast cancer was ranked first in terms of cancer incidence and second in terms of cancer mortality in Malaysia (WHO, 2021b).

1.1.2 Breast cancer presentation in Malaysia

Diagnosis of breast cancer at an early stage leads to a better prognosis and more simplified management (Ahmad, 2019). In Malaysia, about 34.5% of women are diagnosed at stage II, 25.1% at stage III, 22.8% at stage IV, and 17.5% at stage I (National Cancer Registry Department, 2019b). The majority of breast cancer women in Malaysia presented with infiltrating ductal carcinoma (Hanis *et al.*, 2019; Tan *et al.*, 2021).

1.1.3 Risk factors of breast cancer

Risk factors are any factors influencing the risk of having the disease. Identifying the risk factors of breast cancer especially the modifiable risk factors help in screening, diagnosing, and management of breast cancer.

1.1.3(a) Age at diagnosis

Generally, breast cancer incidence increases with age. The majority of women were diagnosed at the age of 50 years and older (Coughlin, 2019). However, the majority of breast cancer women in Asia were diagnosed at an earlier age of 40 years (Mubarik *et al.*, 2019). In Malaysia, the age-specific incidence rate of breast cancer had been observed to increase at the age of 25 years, peak at the age of 60 to 64 years, and reduce after the age of 65 years (National Cancer Registry Department, 2019b).

1.1.3(b) Age at menarche

Age at menarche was not significantly associated with breast cancer risk (Rojas & Stuckey, 2016; Tan *et al.*, 2018). Several studies had found otherwise (Kamińska *et al.*, 2015; World Cancer Research Fund, 2018). Additionally, a meta-analysis study reported that early menarche was associated with an increased risk of breast cancer (Hamajima *et al.*, 2012).

1.1.3(c) Menopausal status

A Malaysian study found that women who menopause had a higher risk of breast cancer (Tan *et al.*, 2018). The study further revealed that there was no significant difference in terms of the risk of breast cancer between those who menopause before and after the age of 50 years. The finding coincided with a previous meta-analysis study that concluded postmenopausal women had a higher risk of developing breast cancer (Nindrea, Aryandono & Lazuardi, 2017).

1.1.3(d) Ethnicity

The effect of ethnicity on breast cancer incidence differs according to country. In Malaysia, the age-standardised rate of breast cancer was higher among Chinese,

followed by Indian, Malay, and other ethnicities (National Cancer Registry Department, 2019b).

1.1.3(e) Physical activities

Physical activities were associated with a lower risk of breast cancer (Tan *et al.*, 2018; Xuyu Chen *et al.*, 2019). World Cancer Research Fund (2018) further detailed the effect of physical activities on premenopausal and postmenopausal breast cancer. The report concluded that vigorous activity had a high evidence to protect against premenopausal breast cancer and total physical activity had a high evidence to protect against postmenopausal breast cancer.

1.1.3(f) Body mass index

A study in Malaysia found that a lower body mass index (BMI) increased the risk of breast cancer (Tan *et al.*, 2018). However, an Asia- and Southeast Asia-based meta-analysis reported a different finding that a higher BMI was associated with a higher risk of breast cancer (Nindrea, Aryandono & Lazuardi, 2017; Nindrea *et al.*, 2019). Additionally, a breast cancer report from World Cancer Research Fund (2018) elucidated the effect of BMI on premenopausal and postmenopausal breast cancer women. A higher BMI was protective against both breast cancer with more convincing evidence supporting the protective effect on post-menopausal breast cancer.

1.1.3(g) Parity

Parity was not significantly associated with breast cancer risk (Lee *et al.*, 2014; Tan *et al.*, 2018). However, another study found that nulliparous women had a 30% higher risk of breast cancer compared to multiparous women (Nindrea, Aryandono & Lazuardi, 2017). This finding was aligned with other studies that found lower parity increased the risk of breast cancer (Nguyen *et al.*, 2016).

1.1.3(h) Breastfeeding

Breastfeeding lowered the risk of breast cancer (Tan *et al.*, 2018; World Cancer Research Fund, 2018). World Cancer Research Fund (2018) further concluded an increased duration of breast feeding was significantly associated with a reduced risk of breast cancer. In contrast, a meta-analysis published in 2017 reported a contradictory finding in which breastfeeding was not significantly associated with breast cancer risk (Nindrea, Aryandono & Lazuardi, 2017). Another study reported a similar result (Lee *et al.*, 2014).

1.1.3(i) Dietary habit

Most studies that explored dietary habits and breast cancer risk provided inconsistent results (Coughlin, 2019). For example, soy milk and soy product intake reduced the risk of breast cancer (Fritz *et al.*, 2013; Tan *et al.*, 2018). However, the effect of soy was only observed among the Asian-based studies (Coughlin, 2019). Many studies had been conducted to study the effect of alcohol intake on breast cancer risk. Several studies found alcohol consumption was not associated with breast cancer risk (Lee *et al.*, 2014). In contrast, other studies found alcohol consumption reduced the risk of breast cancer (Coughlin, 2019). However, a dose-response meta-analysis from World Cancer Research Fund (2018) found an increased alcohol consumption increased the risk of breast cancer. Additionally, two meta-analysis studies reported a similar finding (Chen *et al.*, 2016; Q. Sun *et al.*, 2020). However, J.-Y. Chen *et al.* (2016) further concluded that low doses of alcohol consumption protected against breast cancer.

1.1.3(j) Family history

A first-degree family history of breast cancer was linked to an elevated risk of the disease (Tan *et al.*, 2018). Other studies found a similar relationship between

family history and breast cancer risk (Lee *et al.*, 2014; Rojas & Stuckey, 2016; Nindrea, Aryandono & Lazuardi, 2017; World Cancer Research Fund, 2018).

1.1.3(k) Oral contraceptive pills

Oral contraceptive pills (OCP) were not associated with breast cancer risk (Kamińska *et al.*, 2015; Nguyen *et al.*, 2016; Tan *et al.*, 2018). However, other studies suggested otherwise (Karim *et al.*, 2015; Soroush *et al.*, 2016; Nindrea, Aryandono & Lazuardi, 2017; Wahidin, Djuwita & Adisasmita, 2018; World Cancer Research Fund, 2018). Wahidin *et al.* (2018) and Karim *et al.* (2015) further concluded that a longer duration of using OCP was associated with a higher breast cancer risk.

1.1.3(l) Smoking

Smoking status was not associated with the risk of breast cancer (Lee *et al.*, 2014; Tan *et al.*, 2018). However, another study found otherwise. Rojas and Stuckey (2016) reported that smoking increased the relative risk of breast cancer by about 10%. This study coincided with the finding from a meta-analysis published in 2015. The meta-analysis study concluded that there was a moderate increase in the risk of breast cancer in smoking women (Macacu *et al.*, 2015). Additionally, the study further concluded that passive smoking moderately increased the risk of breast cancer.

1.1.3(m) Hormonal replacement therapy

The use of hormonal replacement therapy (HRT) reduced the risk of breast cancer (Tan *et al.*, 2018). Other studies reported contradictory results (Rojas & Stuckey, 2016; World Cancer Research Fund, 2018; Collaborative Group on Hormonal Factors in Breast Cancer, 2019). The use of HRT which included both estrogen and progesterone associated with a higher risk of breast cancer compared to HRT which only included estrogen (Coughlin, 2019).

1.1.3(n) Mammographic density

Mammographic density or breast density is the amount of dense tissue visible on the mammogram. Mammographic density was linked to an increased risk of breast cancer (Mokhtary, Karakatsanis & Valachis, 2021). Additionally, mammographic density is affected by other factors. For example, Asian women had a higher breast density compared to other races (Nazari & Mukherjee, 2018).

Table 1.1 Summary of risk factors of breast cancer.

Risk factors	Findings	Authors, Year
Age at diagnosis	Majority of breast cancer women were diagnosed at the age of 50 years and older	(Coughlin, 2019)
	Majority of breast cancer women in Asia diagnosed at the age of 40 years	(Mubarik *et al.*, 2019)
	Age-specific incidence rate of breast cancer in Malaysia increased at the age of 25 years, peaked at the age of 60 to 64 years, and reduced after the age of 65 years	(National Cancer Registry Department, 2019b)
Age at menarche	Age at menarche was not significantly associated with breast cancer risk	(Rojas & Stuckey, 2016; Tan *et al.*, 2018)
	Age at menarche was significantly associated with breast cancer risk	(Kamińska *et al.*, 2015; World Cancer Research Fund, 2018)
	Early menarche associated with increased risk of breast cancer	(Hamajima *et al.*, 2012)
Menopausal status	Menopaused women had a higher risk of breast cancer	(Nindrea, Aryandono & Lazuardi, 2017; Tan *et al.*, 2018)
Ethnicity	Age-standardised rate of breast cancer in Malaysia was higher among Chinese, followed	(National Cancer Registry Department, 2019b)

Table 1.1 Continued

	by Indian, Malay, and other ethnicities	
Physical activities	Physical activities were associated with a lower risk of breast cancer	(Tan *et al.*, 2018; Xuyu Chen *et al.*, 2019)
	A vigorous activity was more likely to protect against premenopausal breast cancer and total physical activity was more likely to protect against postmenopausal breast cancer	(World Cancer Research Fund, 2018)
Body mass index	Lower body mass index increased the risk of breast cancer	(Tan *et al.*, 2018)
	A higher BMI was associated with a higher risk of breast cancer	(Nindrea, Aryandono & Lazuardi, 2017; Nindrea *et al.*, 2019)
	A higher BMI was protective against premenopausal and postmenopausal breast cancer	(World Cancer Research Fund, 2018)
Parity	Parity was not significantly associated with breast cancer risk	(Lee *et al.*, 2014; Tan *et al.*, 2018)
	Nulliparous women had a 30% higher risk of breast cancer compared to multiparous women	(Nindrea, Aryandono & Lazuardi, 2017)
	Lower parity increased the risk of breast cancer	(Nguyen *et al.*, 2016)
Breastfeeding	Breastfeeding lowered the risk of breast cancer	(Tan *et al.*, 2018; World Cancer Research Fund, 2018)
	Breastfeeding was not significantly associated with breast cancer risk	(Lee *et al.*, 2014; Nindrea, Aryandono & Lazuardi, 2017)
Dietary habit	Inconsistent results among the studies that reported dietary habits and breast cancer risk	(Coughlin, 2019)

Table 1.1 Continued

	Soy milk and soy product intake reduced the risk of breast cancer	(Fritz *et al.*, 2013; Tan *et al.*, 2018)
	The effect of soy was only observed among the Asian-based studies	(Coughlin, 2019)
	Alcohol consumption was not associated with breast cancer risk	(Lee *et al.*, 2014)
	Alcohol consumption reduced the risk of breast cancer	(Coughlin, 2019)
	Increased alcohol consumption increased the risk of breast cancer	(Chen *et al.*, 2016; World Cancer Research Fund, 2018; Q. Sun *et al.*, 2020)
	A low dose of alcohol consumption protected against breast cancer	(Chen *et al.*, 2016)
Family history	First-degree family history was associated with an increased risk of breast cancer	(Lee *et al.*, 2014; Rojas & Stuckey, 2016; Nindrea, Aryandono & Lazuardi, 2017; Tan *et al.*, 2018; World Cancer Research Fund, 2018)
Oral contraceptive pills	Oral contraceptive pills were not associated with breast cancer risk	(Kamińska *et al.*, 2015; Nguyen *et al.*, 2016; Tan *et al.*, 2018)
	Oral contraceptive pills were associated with breast cancer risk	(Karim *et al.*, 2015; Soroush *et al.*, 2016; Nindrea, Aryandono & Lazuardi, 2017; Wahidin, Djuwita & Adisasmita, 2018; World Cancer Research Fund, 2018)
	Long duration of using oral contraceptive pills was associated with a higher breast cancer risk	(Karim *et al.*, 2015; Wahidin, Djuwita & Adisasmita, 2018)
Smoking	Smoking status was not associated with the risk of breast cancer	(Lee *et al.*, 2014; Tan *et al.*, 2018)

Table 1.1 Continued

	Smoking increased the relative risk of breast cancer by about 10%	(Rojas & Stuckey, 2016)
	A moderate increase in the risk of breast cancer in smoking women	(Macacu *et al.*, 2015)
	Passive smoking moderately increased the risk of breast cancer	
Hormonal replacement therapy	The use of hormonal replacement therapy reduced the risk of breast cancer	(Tan *et al.*, 2018)
	The use of hormonal replacement therapy increased the risk of breast cancer	(Rojas & Stuckey, 2016; World Cancer Research Fund, 2018; Collaborative Group on Hormonal Factors in Breast Cancer, 2019)
	Estrogen- and progesterone-based hormonal replacement therapy was associated with a higher risk of breast cancer compared to estrogen-based hormonal replacement therapy only	(Coughlin, 2019)
Mammographic density	Increased mammographic density was associated with an increased risk of breast cancer	(Mokhtary, Karakatsanis & Valachis, 2021)
	Asian women had a higher breast density compared to other races	(Nazari & Mukherjee, 2018)

1.1.4 Screening and diagnosis of breast cancer

Breast cancer screening aims to detect the disease as early as possible, thus, improving the prognosis and management of the disease. As the risk of breast cancer among Malaysian women increases after the age of 35 years, a clinical breast examination is recommended at this age (Ministry of Health Malaysia, 2019).

Additionally, the clinical practice guideline for breast cancer management (2019) recommends a screening mammogram to be done every two years for women with average risk and annually for women with moderate to high risk of breast cancer. Furthermore, an annual screening mammogram and magnetic resonance imaging (MRI) are recommended for women suspecting of carriers of pathogenic variants such as breast cancer gene 1 (BRCA1), breast cancer gene 2 (BRCA2), and partner and localiser of BRCA2 (PALB).

There are several risk assessment tools available for breast cancer such as Tyre-Cuzick, Claus, breast and ovarian analysis of disease incidence and carrier estimation algorithm (BOADICEA), BRCAPRO, Pen II and breast cancer risk assessment tool (BCRAT) or Gail (Sciaraffa *et al.*, 2020). However, some of the tools especially the recently developed tools have yet to be validated externally and calibrated to a specific population such as Asian women (Cintolo-Gonzalez *et al.*, 2017). Furthermore, certain risk assessment tools such as BOADICEA had been observed to overestimate the risk of breast cancer in the Asian population (Ministry of Health Malaysia, 2019).

The established method for the diagnosis of breast cancer is the triple assessment which includes clinical assessment, imaging (ultrasound and/or mammography), and pathology (histology and/or cytology) (Ministry of Health Malaysia, 2019). These assessments complement each other and should be interpreted collectively.

1.1.4(a) Mammography

Mammography is the most effective screening tool for breast cancer detection (Lehman *et al.*, 2017). Generally, there are two types of mammography: (1) 2D mammography or digital mammogram and (2) 3D mammography or digital breast tomosynthesis (DBT). The former is more widely utilised in Malaysia compared to the

latter. The utilisation of both types of mammography increases the overall rates of breast cancer detection (Giampietro *et al.*, 2020).

1.1.4(b) BI-RADS

Breast imaging reporting and data system (BI-RADS) was developed by the American College of Radiology and was first published in 1993 (Spak *et al.*, 2017). BI-RADS is a standardised reporting system of mammography, ultrasound, and MRI for breast pathology. Table 1.2 describes the four categories of breast density composition, while Table 1.3 describes the BI-RADS assessment categories based on the BI-RADS 5th edition (D'Orsi *et al.*, 2014; Spak *et al.*, 2017; Ministry of Health Malaysia, 2019).

Table 1.2 Breast density composition based on the BI-RADS 5th edition for mammography.

Breast density	**Description**
A	The breast is almost entirely fatty
B	There are scattered areas of fibroglandular density
C	The breast is heterogeneously dense which may obscure small masses
D	The breast is extremely dense

Table 1.3 BI-RADS assessment categories of the breast.

Categories	**Description**	**Likelihood of cancer**
0	Incomplete	Not applicable
1	Negative	Essentially 0%
2	Benign	Essentially 0%
3	Probably benign	>0% but ≤2%
4	Suspicious	4a: low suspicion of malignancy (>2% to ≤10%)
		4b: moderate suspicion of malignancy (>10% to ≤50%)
		4c: high suspicion of malignancy (>50% to <95%)

Table 1.3 Continued

5	Highly suggestive of malignancy	≥95%
6	Proven malignancy	Not applicable

1.2 Artificial intelligence

Artificial intelligence (AI) is a subfield of computer science that aims to develop a system capable of performing a task that usually requires human intelligence. The rise of AI has the potential to enhance numerous fields, including the healthcare and medicine.

1.2.1 Machine learning

Machine learning is a branch of AI. It is one of the most thriving subfield areas of AI due to an advance in computer electronic and digital information. The main aim of machine learning is to make predictions and classifications based on a pattern learned from a dataset. Subsequently, this prediction and classification are used to support the decision-making process.

All machine learning algorithms can be classified into supervised, unsupervised, and semi-supervised machine learning (Rhys, 2020). Supervised machine learning utilises a labelled dataset, while unsupervised machine learning utilises an unlabelled dataset. Breast cancer data with a known label of benign and malignant is an example of labelled data. Semi-supervised machine learning combines both supervised and unsupervised approaches. Supervised machine learning can be further categorised into regression and classification algorithms. The former predicts a number, and the latter predicts a category (binary or multiclass). Additionally, unsupervised machine learning can be further classified into clustering and dimension reduction algorithms. The former aims to group the data into a cluster based on

similarity, while the latter aims to simplify and reduce the features or the variables of the data.

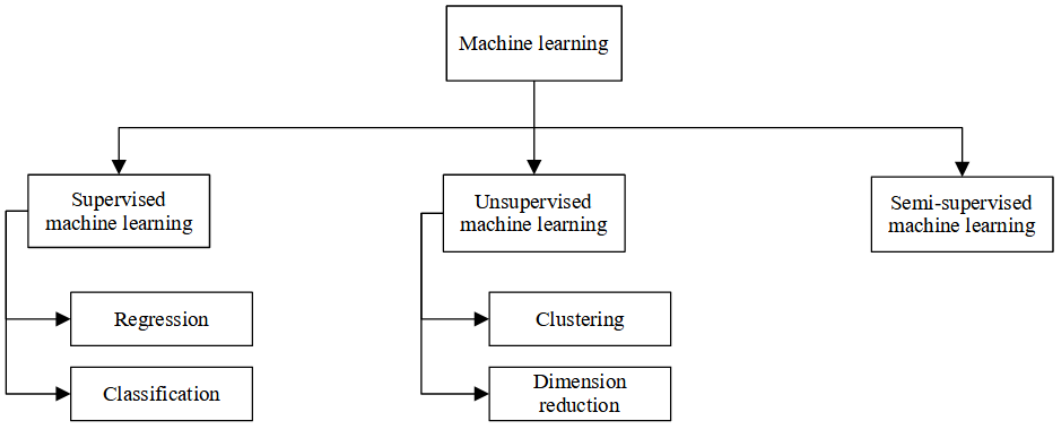

Figure 1.1 Type of machine learning algorithms.

1.2.2 Deep learning

Deep learning is a subfield area of machine learning which enables a machine to learn a representation from raw data and provides a better prediction and classification (Lecun, Bengio & Hinton, 2015). Since 2006, deep learning or representative learning had gained popularity and has been used in many areas, including medicine and healthcare (Al-Jarrah *et al.*, 2015). Deep learning utilised a successive layer of representation to map the input such as texts, images and videos to targets or labels of the data (Chollet, Kalinowski & Allaire, 2022). Hence, deep learning is also known as layered representations learning or hierarchical representations learning. Convolutional neural network (CNN) is one of the types of deep learning commonly utilised in computer vision and image classification. Another type of deep learning is a recurrent neural network (RNN) which is commonly utilised in natural language processing. One of the exciting developments in the area of deep learning is the introduction of transfer learning. Transfer learning or pre-trained network have been very popular in the recent years including in AI-based medical imaging research.

1.2.3 Explainable artificial intelligence

There is a trade-off between interpretability and explainability and the complexity of a machine learning model. As the model becomes more complex, the model becomes less transparent on how it produces a prediction. Thus, these complex models such as ensemble machine learning, neural network, and deep learning are knowns as black-box models, while interpretable models such as generalised additive model (GAM), least absolute shrinkage and selection operator (LASSO) model and decision tree are knowns as white-box models. The terms “explainability” and “interpretability” have been used interchangeably. Although the difference between both terms is not well established, “explainability” in regards to machine learning is related to the understanding of how the machine learning model reaches its prediction, while “interpretability” is related to the understanding of the internal logic and mechanics of the machine learning model (Linardatos, Papastefanopoulos & Kotsiantis, 2021). This thesis utilised both terminologies interchangeably.

Explainable artificial intelligence (XAI), generally, helps in gaining insight into the machine learning model. Thus, the model becomes more transparent and adoptable to the end-users. Additionally, the use of XAI makes the machine learning model accessible to domain knowledge experts with little to no machine learning background, thus, further helping in the validation of the machine learning model. In addition, XAI functions as a safety measure for the machine learning model in identifying potential problems and biases that may be present in the training data. Consequently, the adoption of XAI in the model development and validation ensures the model performs as expected.

XAI approaches can be divided into model-specific and model-agnostic methods. The model-specific approach is limited to a specific machine learning model.

For example, the weights or coefficients in the regression model are specific to the regression model only. Thus, this approach is not appropriate for the comparison of different machine learning models. However, the model-agnostic approach is more robust and applicable to any machine learning model. This approach does not assume the internal structure of the model and is only implemented to the machine learning model post hoc. Additionally, XAI approaches can be classified into a local and global explanations. The former approach explains the behaviour of an individual prediction while the latter approach explains the behaviour of the entire model.

1.2.4 Application of artificial intelligence in medicine

AI has the potential to transform and improve many areas of healthcare and medicine. In fact, AI had been studied to be used as a medical analytic tool for drug discovery, genomic medicine, disease prognosis and diagnosis, and personalised healthcare (Toh, Dondelinger & Wang, 2019; Blasiak, Khong & Kee, 2020). For example, AI had been shown to aid in the diagnosis of fibrotic lung diseases, tuberculosis, and diabetes in research studies (Raghu *et al.*, 2018; Zou *et al.*, 2018; Hwang *et al.*, 2019). AI also was able to track disease progression in diseases such as systemic sclerosis (van Leeuwen *et al.*, 2021), osteoarthritis (Tiulpin *et al.*, 2019), and mild cognitive impairment (Ansart *et al.*, 2021), and predict disease complications in diseases such as diabetes (Dagliati *et al.*, 2018), Crohn's disease (Ungaro *et al.*, 2021), and atrial fibrillation (Lip *et al.*, 2022).

Between 2015 and 2020, about 222 and 240 AI-based medical devices were approved in the US and Europe, respectively (Muehlematter, Daniore & Vokinger, 2021). About 55.1% of the approved AI-based medical devices were for radiological purposes. Another study by Benjamens *et al.* (2020) reported about 46.9% of the US Food and Drug Administration (FDA) approved AI-based medical devices were for

the field of radiology and mostly focused on image reading software. Additionally, the study reported only six algorithms were for oncology and only three were focused on mammograms. Another study by Luchini *et al.*(2022) reported that out of 71 FDA-approved AI-based medical devices, about 54.9% and 19.7% of the devices were related to cancer radiology and pathology, respectively.

1.3 Rationale of the study

Asian women population including Malaysian women were affected by mammographic density at two levels. Firstly, mammographic density affects the risk of breast cancer in Asian women. Dense breast women had a 4-6 times higher risk of breast cancer compared to non-dense women (B. Park *et al.*, 2018; Bell, 2020). Secondly, mammographic density affects the mammogram reading of Asian women. Mammographic density reduces the accuracy and sensitivity of mammograms through the masking effect (Freer, 2015). A fibroglandular tissue appears white on the mammogram which is the same colour as cancerous tissue. Thus, this colour similarity between the two tissues makes the mammogram reading more challenging.

Early diagnosis of breast cancer improves the prognosis of the disease (Sun *et al.*, 2017). The time delay between the discovery of the symptoms and the first medical consultation is known as patient delay (Unger-Saldaña, 2014). There are several factors influencing the patient delay such as lack of knowledge, socio-cultural perception of the medical treatment, and the use of complementary and alternative medicine (Khan *et al.*, 2015a). Additionally, breast clinics receive numerous cases related to breast-related problems. Therefore, the abundance of cases in the clinics contributes to the delay in the diagnosis of breast cancer patients. A patient delay of

more than three months was associated with more advanced and metastatic tumours (Caplan, 2014; Tesfaw *et al.*, 2020).

Additionally, a second type of care delay is known as a provider delay. The provider delay is the time delay between the first medical consultation and the start of the treatment (Unger-Saldaña, 2014). The provider delay is suggested to be less significant than the patient delay as the provider which is the physician is capable of distinguishing between more and less aggressive breast tumours (Caplan, 2014). Nonetheless, it is important to reduce both types of care delay to ensure a better prognosis for breast cancer patients.

The use of AI is expected to improve healthcare and services. However, the practical implementation of the AI remains unclear. Different implementations of ML models have been shown to produce different performance results including in the area of breast cancer. Thus, there is a need to estimate the overall performance of ML models on breast cancer data.

1.4 Research objectives

1.4.1 General objective

To explore and develop machine learning-based predictive models for breast cancer risk estimation and prediction in Malaysian women using socio-demographic, clinical, and mammographic data.

1.4.2 Specific objectives

1. To map research activities on breast cancer in Malaysia, and further identify the trend of breast cancer research in Malaysia.

2. To assess the research output in breast cancer and machine learning, and further determine the themes and trends in breast cancer and machine learning research.
3. To determine factors affecting mammographic density and describe the impact of mammographic density on the severity of breast cancer in Malaysian women.
4. To develop an over-the-counter screening model using machine learning and evaluate its performance across mammographic density.
5. To determine the overall diagnostic accuracy of the machine learning model on mammographic images.
6. To develop an ensemble of pre-trained networks for breast cancer abnormality detection using digital mammograms, and further assess the performance of the models across mammographic density.

1.5 Thesis overview

This thesis consisted of six inter-related projects. This thesis began with bibliometric studies to explore the patterns and trends related to breast cancer research in Malaysia and global breast cancer research and machine learning. Bibliometric analysis is useful for evaluating the impact and trends of scholarly publications within a specific field, aiding researchers in identifying influential works, emerging topics, and collaborative networks. These projects are presented in Chapter 2 and Chapter 3 of this thesis. Both projects utilised past literature to achieve the objectives of the projects.

The third project studied the factors influencing mammographic density. The mammographic density or breast density affects the risk of breast cancer and the

accuracy of mammogram readings (Freer, 2015; B. Park *et al.*, 2018; Bell, 2020). The result of this project could help physicians and radiologists identify cases that are most likely dense breast cases, therefore, ensuring timely and appropriate management is delivered.

Next, the fourth project developed an over-the-counter (OTC) screening model for women who attended a breast clinic. The OTC model would predict women who attended a breast clinic as suspicious or normal cases. The suspicious cases were the cases with a high probability of breast cancer. Additionally, the model developed in this project would be validated across mammographic density to ensure the model produces a reliable and accurate prediction.

The fifth project aimed to establish the overall accuracy of machine learning models on mammography for breast cancer classification. Numerous previous studies were conducted utilising mammographic data for breast cancer classification. Various machine learning models were used in those studies. Therefore, it is imperative to establish the overall accuracy of machine learning models in previous studies to gauge the benefit of machine learning in this area of research. The details of this work are presented in Chapter 6.

Lastly, this thesis developed a supplementary diagnostic tool for radiologists to estimate the risk of breast cancer. This last project utilised mammographic data and ensemble transfer learning to develop the supplementary diagnostic tool. This tool was further validated across the mammographic density to ensure an accurate and reliable prediction produced by the tool.

Overall, the main focus of this thesis was to develop machine learning-based predictive models for breast cancer risk estimation and prediction using socio-demographic, clinical, and mammographic data. The main focuses were reflected in

the fourth and sixth projects. However, additional projects were incorporated to augment the scope of research, providing a comprehensive understanding of breast cancer risk estimation and prediction specifically in the context of Malaysian women.

CHAPTER 2

MAPPING BREAST CANCER RESEARCH IN MALAYSIA: A SCIENTOMETRIC ANALYSIS

2.1 Abstract

The purpose of this project is to provide an overview of breast cancer research in Malaysia. Besides, this project aims to identify the trends of breast cancer research in Malaysia. This project retrieved 343 related publications from the Scopus database. After removing one duplicated publication and another two publications that did not meet the study criteria, the remaining 340 publications were analysed using a bibliometric analysis and trending keywords analysis. This project found that the annual growth rate of publications was 7.4%. The majority of the publications were research articles and multi-author. The most productive author was Yip CH with 69 publications, and the University of Malaya was the top institution in Malaysia related to this research area. For the last five years, there were no dominant themes in this research area. However, this project found two emerging clusters of breast cancer research in Malaysia, which are related to medical data analytics and precision medicine in genomic breast cancer. Breast cancer research in Malaysia shows promising advancements, yet there remains room for enhancement. As the funding in this research area is scarce, proper allocation of the resources is needed.

2.2 Background

Breast cancer is the most common cancer in Malaysia, and it affected 34.1 per 100,000 Malaysian population between 2012 and 2016 (National Cancer Registry Department, 2019a). This statistic was almost similar to the South-East Asian statistic in which 34.8 per 100,000 people were affected by this cancer (Fan, Goss & Strasser-Weippl, 2015). In terms of breast cancer mortality, Malaysia had the highest age-

standardised mortality rate among South-East Asian countries in 2012 (Yip, 2016). Despite the concerning statistics, public awareness about breast cancer was relatively low (Mohamad & Kok, 2019; Lee *et al.*, 2022).

Scientometrics covers various quantitative approaches used to assess scientific literature productions and their practices (Leydesdorff & Milojević, 2015). Bibliometrics, a subset of scientometrics, provides a set of analyses to assess the research output of any research area. The bibliometric approach had been used to gauge the impact of the research area by evaluating the related researchers and publications within the field (Cobo *et al.*, 2011). The main advantage of scientometric analysis is that it can assess an unlimited amount of publication in any field of study. Thus, this quantitative approach provides insight into general publication patterns, which is beneficial for the researchers in the research area.

This project aims to provide an overview of research activities on breast cancer in Malaysia in terms of the distribution of publications and journals, top authors and institutional collaborations. Also, this project intends to identify the trend of breast cancer research in Malaysia.

2.3 Methodology

2.3.1 Data extraction and collection

All scientific publications related to breast cancer research in Malaysia were retrieved from the Scopus database on 24th May 2021. The following query was used:

(TITLE ("breast cancer" OR "breast carcinoma" OR tumour AND breast OR tumor AND breast OR "mammary cancer" OR "ductal carcinoma" OR "invasive carcinoma" OR "breast mass" OR "breast lesion" OR "breast malignancy") AND TITLE-ABS-

KEY(Malaysia)) AND (LIMIT-TO(DOCTYPE , "ar") OR LIMIT-TO(DOCTYPE, "cp") OR LIMIT-TO(DOCTYPE, "re")) AND (LIMIT-TO (LANGUAGE , "English"))

The searched outputs were limited to the English language and the type of documents was restricted to research articles, conference papers, and reviews. The metadata, title, and abstract for each article were downloaded in BibTeX file format.

2.3.2 Statistical analysis and software

Data cleaning and data management were done using R software version 4.1.0 (R Core Team, 2021). Data were checked for duplicates using the title and the DOI number. There were one duplicate publication, one response letter, and one case study. Thus, the three publications were removed from the data. Figure 2.1 presents the flow of analysis in this project.

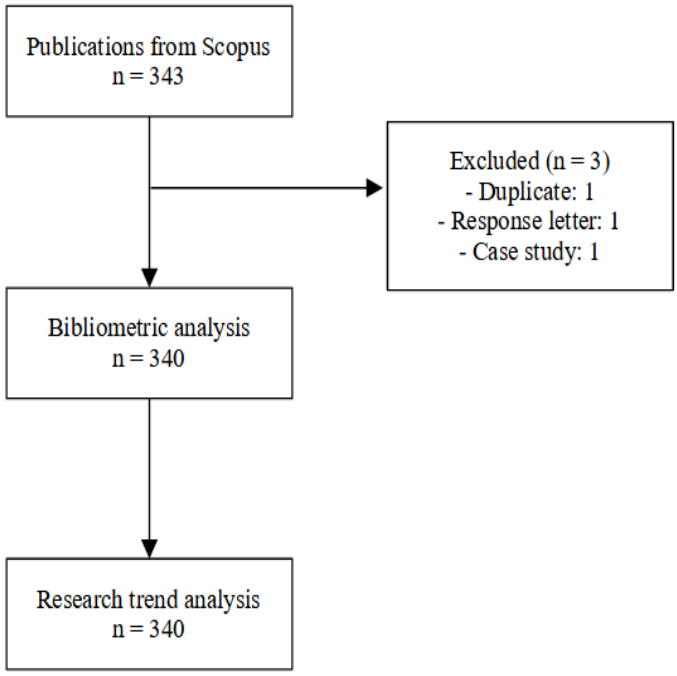

Figure 2.1 The flow of the analysis in this project.