

Article

Over-the-Counter Breast Cancer Classification Using Machine Learning and Patient Registration Records

Tengku Muhammad Hanis ^{1,*}, Nur Intan Raihana Ruhaiyem ², Wan Nor Arifin ³, Juhara Haron ^{4,5},
Wan Faiziah Wan Abdul Rahman ^{5,6}, Rosni Abdullah ² and Kamarul Imran Musa ^{1,*}

¹ Department of Community Medicine, School of Medical Sciences, Universiti Sains Malaysia, Kubang Kerian 16150, Kelantan, Malaysia

² School of Computer Sciences, Universiti Sains Malaysia, Gelugor 11800, Penang, Malaysia

³ Biostatistics and Research Methodology Unit, School of Medical Sciences, Universiti Sains Malaysia, Kubang Kerian 16150, Kelantan, Malaysia

⁴ Department of Radiology, School of Medical Sciences, Universiti Sains Malaysia, Kubang Kerian 16150, Kelantan, Malaysia

⁵ Breast Cancer Awareness and Research Unit, Hospital Universiti Sains Malaysia, Kubang Kerian 16150, Kelantan, Malaysia

⁶ Department of Pathology, School of Medical Sciences, Universiti Sains Malaysia, Kubang Kerian 16150, Kelantan, Malaysia

* Correspondence: tengkuhanismokhtar@gmail.com (T.M.H.); drkamarul@usm.my (K.I.M.)



Citation: Hanis, T.M.; Ruhaiyem, N.I.R.; Arifin, W.N.; Haron, J.; Wan Abdul Rahman, W.F.; Abdullah, R.; Musa, K.I. Over-the-Counter Breast Cancer Classification Using Machine Learning and Patient Registration Records. *Diagnostics* **2022**, *12*, 2826. <https://doi.org/10.3390/diagnostics12112826>

Academic Editor: Md Mohaimenul Islam

Received: 10 September 2022

Accepted: 15 October 2022

Published: 16 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: This study aims to determine the feasibility of machine learning (ML) and patient registration record to be utilised to develop an over-the-counter (OTC) screening model for breast cancer risk estimation. Data were retrospectively collected from women who came to the Hospital Universiti Sains Malaysia, Malaysia for breast-related problems. Eight ML models were used: k-nearest neighbour (kNN), elastic-net logistic regression, multivariate adaptive regression splines, artificial neural network, partial least square, random forest, support vector machine (SVM), and extreme gradient boosting. Features utilised for the development of the screening models were limited to information in the patient registration form. The final model was evaluated in terms of performance across a mammographic density. Additionally, the feature importance of the final model was assessed using the model agnostic approach. kNN had the highest Youden J index, precision, and PR-AUC, while SVM had the highest F2 score. The kNN model was selected as the final model. The model had a balanced performance in terms of sensitivity, specificity, and PR-AUC across the mammographic density groups. The most important feature was the age at examination. In conclusion, this study showed that ML and patient registration information are feasible to be used as the OTC screening model for breast cancer.

Keywords: Asian women; breast cancer; explainable artificial intelligence; machine learning; medical consultation delays; screening model; clinical decision support systems

1. Introduction

Breast cancer is the most common cancer among women in at least 140 countries [1]. The WHO aims to reduce global breast cancer mortality by 25% annually between 2020 and 2040, which is equivalent to 2.5 million breast cancer death worldwide [2]. Generally, breast cancer affects women above the age of 50 and the risk of having the disease increases with increased age [3–5]. The risk factors for breast cancer are mainly divided into two groups [6]. The inherent risk factors include a family history of breast cancer, age, and gender, while the extrinsic risk factors include diet and lifestyle. The risk factors differ according to the individual and population. One of the important risk factors for breast cancer is mammographic density which reflects the amount of dense and fatty tissue in the breast [7,8]. Women with denser breasts had four to six times higher chances of developing

breast cancer than those with less dense breasts [9]. Asian women or women with Asian ancestry had denser breasts compared to other populations [10].

Early detection of breast cancer is crucial in reducing the severity of the disease. Any delay in the diagnosis and prognosis may worsen the presentation of the disease. Generally, delay in the management of cancer is divided into two: patient delay and provider delay [11]. Patient delay is the delay during the period between the first discovery of the symptom and medical consultation. Provider delay is the delay between medical consultation and the beginning of the cancer treatment. Additionally, the combination of both types of delay is known as a total delay. However, a more complicated model of the total delay had also been proposed. For example, the total patient delay model detailed the total delay into five stages [12] while the total breast cancer delay model detailed it into eight stages [13]. Nonetheless, a total delay of more than 1 to 3 months has been observed to be associated with advanced stages of cancer and reduced survival of the patients [11,14,15]. Thus, there is a need to improve the efficiency of the medical workflow for breast cancer patients in arranging a medical consultation.

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 is expected to improve many areas including the fields of healthcare and medicine. AI had been studied to be used as a medical analytic tool including for drug discovery, genomic medicine, disease prognosis and diagnosis, and personalised healthcare [16,17]. For example, AI has been shown to aid the diagnosis of fibrotic lung diseases, tuberculosis, and diabetes in research studies [18–20]. AI also had been shown to track disease progression in diseases such as systemic sclerosis [21], osteoarthritis [22], and mild cognitive impairment [23], and predict disease complications in diseases such as diabetes [24], Crohn's disease [25], and atrial fibrillation [26]. However, the adoption of AI in healthcare and medicine is slower than in other fields [27]. Explainable AI (XAI) aims to make the AI more interpretable and understandable to the end-users. Thus, the use of XAI will further help in the successful implementation of AI in healthcare. Generally, the approaches used in the XAI can be divided into model-specific and model-agnostic approaches [28]. Model-specific approaches are limited to specific machine learning (ML) models. One of the main limitations of this approach is that a comparison between models is not appropriate. Model-agnostic approaches overcome this limitation and are applicable to any ML model. XAI had been researched for diagnosis and prediction of glioblastoma [28], colorectal cancer [29], thoracic cancer [30], renal cell carcinoma [31], COVID-19 [32], chronic wounds [33], and Alzheimer's disease [34]. The use of XAI in medicine is expected to provide insights and transparency into the AI models. Thus, XAI can further help in establishing trust and confidence among medical professionals in the utilisation and implementation of AI in clinical settings [35].

This study aims to develop an over-the-counter (OTC) ML model for breast cancer screening to be deployed in a breast clinic using patient registration records. The model can accelerate the medical workflow for breast cancer management and provide women with a high probability of breast cancer with a timely medical consultation. In other words, women predicted by the model to have a suspicious breast case can be given a high priority for medical consultation with clinicians. Additionally, the performance of the model will be evaluated across dense and non-dense cases. Lastly, we aim to determine the top influential features of the OTC screening model.

2. Materials and Methods

2.1. Data

Breast cancer data were collected retrospectively from the Breast Cancer Awareness and Research Unit (BestARi), Department of Radiology, and Department of Pathology at Hospital Universiti Sains Malaysia (HUSM). BestARi is a breast cancer resource centre in HUSM. BestARi receives women with breast-related problems from the northeast coast region of Malaysia, especially from the state of Kelantan. The breast cancer data records

in BestARi were limited to 1 January 2014 and 30 June 2021. Twenty-seven variables were collected in this study. Twenty-four features were collected from the BestARi including (1) date of examination; eight features related to sociodemographic and personal information: (1) age at examination, (2) race, (3) marital status, (4) number of children, (5) age at menarche, (6) weight, (7) height, (8) handedness; six features regarding the symptoms or patient complaints: (1) lump, (2) nipple discharge, (3) nipple retraction, (4) axillary mass, (5) pain, and (6) skin changes; and nine features regarding the medical history: (1) history of breast surgery or implant, (2) history of breast trauma, (3) history of birth control or hormone replacement therapy, (4) history of the previous mammography, (5) history of breast self-examination, (6) breastfeeding history, (7) history of total abdominal hysterectomy bilateral salpingo-oophorectomy (TAHBSO), (8) family history of breast cancer, and (9) menopausal status. All features were used in the ML model development except for the date of examination as the feature provided no information for the model development. Another two variables collected from the Department of Radiology, HUSM, were breast imaging-reporting and data system (BIRADS) classification information and BIRADS density (or mammographic density). Both variables were used to classify the cases into dense vs. non-dense groups and normal vs. suspicious groups. Finally, the last information collected from the Department of Pathology, HUSM, was histopathological examination (HPE) results. The latter three variables were used to determine the outcome variable.

The data from the Department of Radiology and Department of Pathology were combined with BestARi’s data if both data were dated within a year after the date of BestARi’s data for each patient. The latest medical record was taken if patients had several records in the BestARi and a single record from the Department of Radiology or Department of Pathology. Afterwards, a body mass index (BMI) was further calculated from the individual weight and height and was added to the existing list of features. Each patient was classified as a normal or suspicious class. The normal class was patients with a BIRADS classification of 1 or who had a diagnosis of normal from the HPE result. The suspicious class was patients with a BIRADS classification of 2, 3, 4, 5, and 6 or who had a diagnosis of benign or malignant subtype of breast cancer from the HPE result. Patients with a BIRADS classification of 0 and missing BIRADS classification or mammographic density were excluded from the study. Additionally, non-dense breast women were those with BIRADS density of A and B, while dense breast women were those with BIRADS density of C and D. Table 1 presents the characteristics of the collected data.

Table 1. Characteristics of the features collected from Hospital Universiti Sains Malaysia.

Characteristic	Normal, n = 230 n (%)	Suspicious, n = 861 n (%)	Missing Values n (%)	Overall, n = 1091
Age at examination ^{1,2}	50.0 (8.1)	53.7 (9.6)	3 (0.3%)	52.9 (9.4)
Age at menarche ^{1,2}	13.0 (1.5)	13.1 (1.5)	97 (8.9%)	13.1 (1.5)
No of children ^{1,2}	3.8 (2.7)	3.6 (2.4)	85 (7.8%)	3.7 (2.5)
Weight (kg) ^{1,2}	64.2 (12.9)	63.5 (12.8)	263 (24.0%)	63.7 (12.8)
Height (cm) ¹	156.4 (5.5)	155.0 (6.4)	692 (63.0%)	155.2 (6.3)
BMI ^{1,2}	27.1 (5.7)	26.7 (5.6)	696 (64.0%)	26.8 (5.6)
Race ²			34 (3.1%)	
Chinese	21 (9.4%)	112 (13.4%)		133 (12.6%)
Indian	0 (0.0%)	4 (0.5%)		4 (0.4%)
Malay	201 (89.7%)	706 (84.8%)		907 (85.8%)
Others	0 (0.0%)	3 (0.4%)		3 (0.3%)
Siamese	2 (0.9%)	8 (1.0%)		10 (0.9%)
Marriage status ²			59 (5.4%)	
Divorced	0 (0.0%)	4 (0.5%)		4 (0.4%)
Married	208 (95.9%)	759 (93.1%)		967 (93.7%)
Single	8 (3.7%)	46 (5.6%)		54 (5.2%)
Widowed	1 (0.5%)	6 (0.7%)		7 (0.7%)

Table 1. Cont.

Characteristic	Normal, <i>n</i> = 230 <i>n</i> (%)	Suspicious, <i>n</i> = 861 <i>n</i> (%)	Missing Values <i>n</i> (%)	Overall, <i>n</i> = 1091
Breastfeeding ²			541 (50.0%)	
No	30 (24.4%)	131 (30.7%)		161 (29.3%)
Yes	93 (75.6%)	296 (69.3%)		389 (70.7%)
Lump ²			41 (3.8%)	
No	148 (67.0%)	588 (70.9%)		736 (70.1%)
Yes	73 (33.0%)	241 (29.1%)		314 (29.9%)
Nipple discharge ²			52 (4.8%)	
No	205 (94.9%)	793 (96.4%)		998 (96.1%)
Yes	11 (5.1%)	30 (3.6%)		41 (3.9%)
Nipple retraction ²			45 (4.1%)	
No	213 (97.3%)	784 (94.8%)		997 (95.3%)
Yes	6 (2.7%)	43 (5.2%)		49 (4.7%)
Axillary mass ²			55 (5.0%)	
No	203 (94.0%)	764 (93.2%)		967 (93.3%)
Yes	13 (6.0%)	56 (6.8%)		69 (6.7%)
Pain ²			54 (4.9%)	
No	172 (80.0%)	691 (84.1%)		863 (83.2%)
Yes	43 (20.0%)	131 (15.9%)		174 (16.8%)
Skin changes ²			55 (5.0%)	
No	204 (94.0%)	772 (94.3%)		976 (94.2%)
Yes	13 (6.0%)	47 (5.7%)		60 (5.8%)
Breast surgery / implant ²			76 (7.0%)	
No	143 (69.1%)	531 (65.7%)		674 (66.4%)
Yes	64 (30.9%)	277 (34.3%)		341 (33.6%)
Trauma ²			108 (9.9%)	
No	191 (94.6%)	754 (96.5%)		945 (96.1%)
Yes	11 (5.4%)	27 (3.5%)		38 (3.9%)
BC-HR ²			51 (4.7%)	
No	130 (59.1%)	554 (67.6%)		684 (65.8%)
Yes	90 (40.9%)	266 (32.4%)		356 (34.2%)
Previous mammogram ²			40 (3.7%)	
No	116 (52.5%)	348 (41.9%)		464 (44.1%)
Yes	105 (47.5%)	482 (58.1%)		587 (55.9%)
Breast self-examination ²			106 (9.7%)	
No	44 (20.9%)	149 (19.3%)		193 (19.6%)
Yes	167 (79.1%)	625 (80.7%)		792 (80.4%)
Handedness ²			667 (61.0%)	
Left	6 (7.4%)	20 (5.8%)		26 (6.1%)
Right	75 (92.6%)	323 (94.2%)		398 (93.9%)
TAHBSO ²			70 (6.4%)	
No	187 (86.6%)	720 (89.4%)		907 (88.8%)
Yes	29 (13.4%)	85 (10.6%)		114 (11.2%)
Family history ²			520 (48.0%)	
No	101 (80.2%)	352 (79.1%)		453 (79.3%)
Yes	25 (19.8%)	93 (20.9%)		118 (20.7%)
Menopause status ²			0 (0.0%)	
No	139 (60.4%)	385 (44.7%)		524 (48.0%)
Yes	91 (39.6%)	476 (55.3%)		567 (52.0%)
Mammographic density			0 (0.0%)	
Non-dense	124 (53.9%)	468 (54.4%)		592 (54.3%)
Dense	106 (46.1%)	393 (45.6%)		499 (45.7%)

Notes: BestARI = breast cancer awareness and research unit; Family history = family history of breast cancer; BC-HR = history of birth control or hormone replacement; TAHBSO = history of total abdominal hysterectomy bilateral salpingo-oophorectomy; ¹ mean (SD); ² Features included in the model development.

2.2. Pre-Processing Steps

Initially, all 24 features including the additional variable of BMI were included in the model development. Next, missing values in the data were imputed using a bagged tree model. Subsequently, numerical variables with absolute correlations above 0.8 with other numerical variables were removed. Then, the training dataset was balanced using a random over-sampling examples (ROSE) algorithm [36]. All numerical features were normalised and transformed using a Yeo-Johnson transformation [37]. A dummy coding variable was created for all categorical features for all ML models except for the random forest model.

The random forest model had been shown to have at least similar performance if not better when categorical features were used as factor variables as opposed to when the dummy variables were used in the model [38]. The ROSE algorithm was implemented using a **themis** package version 1.0.0 [39]. The remaining pre-processing steps were implemented using a **recipes** package 1.0.1 [40].

2.3. Machine Learning Models

Eight OTC screening models were developed from ML methods including k-nearest neighbour (kNN), elastic-net logistic regression, multivariate adaptive regression splines (MARS), artificial neural network (ANN), partial least square (PLS), random forest, support vector machine (SVM), and extreme gradient boosting (XGBoost). SVM was implemented using a radial basis function kernel which used a nonlinear class boundary to maximize the width margin between the class. All ML algorithms were implemented using the **parsnip** package version 1.0.1 [41] with the **kknn** package version 1.3.1 [42] as a backend for kNN, **glmnet** package version 4.1-4 [43] for elastic-net logistic regression, **earth** package version 5.3.1 [44] for MARS, **nnet** package version 7.3-17 [45] for ANN, **mixOmics** package version 6.16.3 [46] for PLS, **ranger** package version 0.14.1 [47] for random forest, **kernlab** package version 0.9-31 [48] as a backend for SVM, and **xgboost** package 1.6.0.1 [49] for XGBoost. R version 4.1.3 was used to develop all the screening models [50].

2.4. Model Comparison and Hyperparameter Tuning

The data were split into 80% development dataset and 20% validation dataset. The development dataset was further split into nested cross-validation groups for model comparison and hyperparameter tuning. The outer folds were split into 10-fold cross-validation groups of 80% training and 20% testing datasets. Each training dataset of each fold was further split into 25 bootstrap samples (inner folds). The validation dataset was further split into a dense breast dataset and a non-dense breast dataset. Thus, there were three validation datasets available: (4.1-41) the whole validation dataset, (2) the dense breast validation dataset, and (4.1-43) the non-dense breast validation dataset.

A random search with a Latin hypercube grid design of 500 combinations of hyperparameters was used for model comparison and hyperparameter tuning. Firstly, all the performance metrics from the results of the bootstrapped samples were summarised by the mean and standard deviation to obtain the descriptive result for each model. The performance metrics of each model were compared using a one-way ANOVA and subsequently pairwise independent *t*-test if the former test was significant. A *p*-value below 0.05 was considered significant. Additionally, the *p*-values for the post hoc pairwise independent *t*-test were adjusted using Bonferroni corrections. Once the best model was identified, the hyperparameters were chosen based on the highest performance metrics from the bootstrapped sample. Figure 1 elucidates the flow of the analysis for this study. Finally, the best model was re-fit using the chosen hyperparameters on the whole development dataset to obtain the final model.

2.5. Performance Metrics

Four performance metrics used for model comparison were precision, precision recall-area under the curve (PR-AUC), F2 score, and Youden J index. Once the final model was identified, four hyperparameter tuning results with the highest mean of the aforementioned performance metrics were determined. The best hyperparameters result was selected from the four tuning results based on the highest sensitivity value. The performance metrics were defined below:

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall/sensitivity} = \frac{TP}{TP + FN}$$

$$F2 \text{ score} = (1 + 2^2) \frac{\text{precision} \times \text{recall}}{2^2 \times \text{precision} + \text{recall}}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Youden } J \text{ index} = \text{sensitivity} + \text{specificity} - 1$$

A true positive (TP) case was defined as a suspicious case and predicted suspicious by the model, while a true negative (TN) case was a normal case and predicted normal by the model. A false negative (FN) case was a suspicious case but predicted normal by the model, while a false positive (FP) case was a normal case but predicted suspicious by the model.

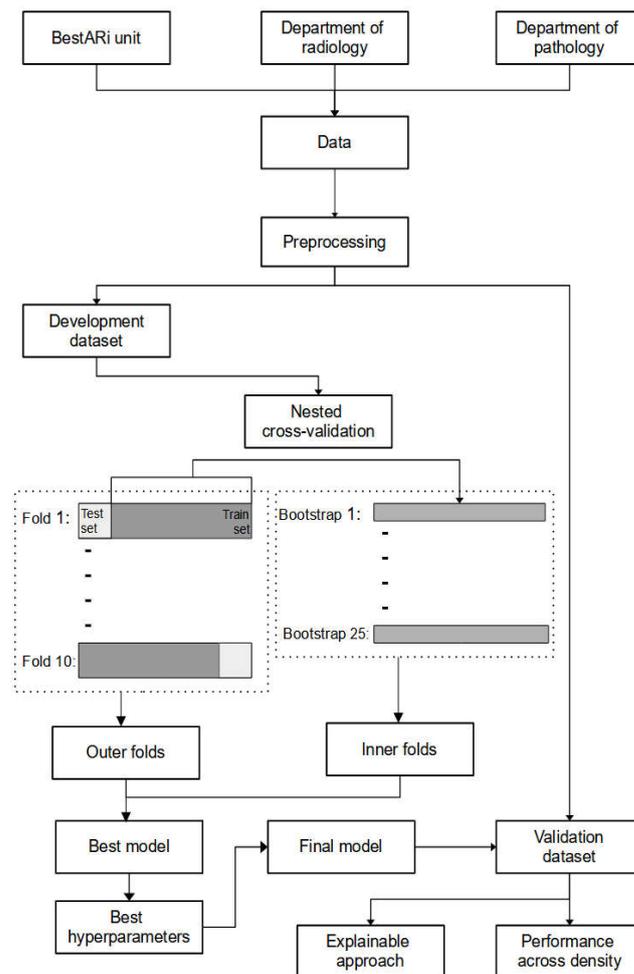


Figure 1. The flow of the analysis.

2.6. Explainable Approach

The model agnostic approach was used to estimate the variable importance for the final ML model. The variable importance was estimated as a mean change in the value of the loss function after variable permutations. The number of permutations was set to 50. The loss function was defined as 1—PR-AUC. The PR-AUC in the loss function reflected the performance of the ML model. Thus, if the feature was important, the performance of the ML model would worsen after permutating the feature. The worse performance of the ML model would in turn result in a high value of the loss function. Hence, the most important feature was the feature with the highest value of 1—PR-AUC. Only the top fifteen important variables were displayed in the variable importance plot. The explainable approach was applied using DALEX and DALEXtra packages versions 2.4.2 and 2.2.1 [51,52].

3. Related Works

Numerous research had been conducted related to breast cancer and ML. Previous studies had used different types of data including imaging modalities, genomic data, and clinical data. Most studies involving ML and breast cancer utilised imaging data especially mammograms and ultrasound [53], while only several studies utilised tabular data. Additionally, a public dataset such as Wisconsin diagnostic breast cancer (WDBC) dataset, despite the tabular nature of the data, the features were derived from the fine needle aspirate imaging of breast mass [54]. Other types of tabular data used for ML classification of breast cancer were sociodemographic, clinical, histological, and pathological data. These types of tabular data were used to predict breast cancer recurrence [55] and survival [56]. Additionally, for breast cancer risk estimation such as screening and diagnosis, imaging data and imaging-derived features were commonly utilised [53]. The use of imaging data in previous studies limited the utilisation of the ML model in the early phase of the screening stage prior to medical consultation.

Several ML algorithms had been used in previous studies that utilised tabular data for the prediction of breast cancer, breast cancer recurrence, and survival of breast cancer patients. Table 2 presents the summary of the previous research related to machine learning classification and breast cancer that utilised tabular data such as sociodemographic, medical history, clinical, pathological, histological, molecular, and genomic data. SVM had been shown to outperform other ML models in several studies involving the prediction of breast cancer recurrence and distant recurrence with the best accuracy at 0.96 [57–59]. However, other studies found ANN and random forest had the best performance in predicting breast cancer recurrence [60,61]. Moreover, for the prediction of the survival of breast cancer patients, naïve Bayes, deep learning, and multilayer perceptron (MLP) had the best accuracy at 0.80, 0.83, and 0.88, respectively [60,62,63]. All the aforementioned studies utilised different datasets which may contribute to the difference in the model performance. Additionally, for breast cancer prediction, random forest showed a promising result with accuracy and an area under the curve (AUC) of 0.98 [64]. Other studies showed that XGBoost and MLP had better performance and outperformed random forest in their respective studies [65,66]. However, all three studies except for Hout et al. [66] used clinical data such as the level of glucose, insulin, leptin, and adiponectin which was beyond the initial screening stage of breast cancer. Additionally, a meta-analysis study had shown that SVM outperformed the other classifier such as ANN, decision tree, naïve Bayes, and kNN in breast cancer risk estimation [67]. This meta-analysis was limited to ML models performed on imaging data, thus, the performance of the aforementioned ML models as an initial breast cancer screening model utilising a tabular dataset have yet to be explored.

Table 2. Summary of the previous works related to machine learning classification and breast cancer that utilised tabular data.

Study	Dataset	ML Classifier	Purpose	Performance Metrics ¹
Kim 2012 [57]	Clinical, histological, and pathological data	SVM ² , ANN, Cox regression	Breast cancer recurrence	Accuracy = 0.85 AUC = 0.85 Sensitivity = 0.89 Specificity = 0.73
Ahmad 2013 [58]	Sociodemographic, clinical, and pathological data	DT, SVM ² , ANN	Breast cancer recurrence	Accuracy = 0.96 Sensitivity = 0.97 Specificity = 0.95 Accuracy = 0.93
Cirkovic 2015 [60]	Clinical, histological, and molecular data	ANN ² , SVM, LR, DT, NB	Breast cancer recurrence	AUC = 0.95 Sensitivity = 0.96 Specificity = 0.83

Table 2. *Cont.*

Study	Dataset	ML Classifier	Purpose	Performance Metrics ¹
		ANN, SVM, LR, DT, NB ²	Breast cancer survival	Accuracy = 0.80 AUC = 0.83 Sensitivity = 0.65 Specificity = 0.85
Sun 2018 [62]	Clinical and genomic data	DL ² , SVM, RF, LR	Breast cancer survival	Accuracy = 0.83 Sensitivity = 0.20 Specificity = 0.95 Precision = 0.75
Kalafi 2019 [63]	Sociodemographic, clinical, and pathological data	MLP ² , DT, RF, SVM	Breast cancer survival	Accuracy = 0.88 Sensitivity = 0.96 Specificity = 0.83 Precision = 0.79 F1 score = 0.87 AUC = 0.87
Zeng 2019 [59]	Sociodemographic, clinical, histological, and pathological data	SVM ²	Breast cancer distant recurrence	Sensitivity = 0.47 Precision = 0.68 F1 score = 0.56
Hou 2020 [66]	Sociodemographic and medical history	XGBoost ² , RF, DL, LR	Breast cancer prediction	Accuracy = 0.67 AUC = 0.74 Sensitivity = 0.66 Specificity = 0.69 Accuracy = 0.75
Kabiraj 2020 [61]	Sociodemographic and clinical data	RF ² , XGBoost	Breast cancer recurrence	Sensitivity = 0.94 Specificity = 0.32 Precision = 0.72 F1 score = 0.64 AUC = 0.89
Khatun 2021 [65]	Sociodemographic and clinical data	NB, RF, MLP ² , LR	Breast cancer prediction	Sensitivity = 0.85 Precision = 0.85 F1 score = 0.84
Anisha 2021 [64]	Sociodemographic and clinical data	RF ²	Breast cancer prediction	Accuracy = 0.98 AUC = 0.98

AUC = area under the curve, SVM = support vector machine, ANN = artificial neural network, DT = decision tree, LR = logistic regression, NB = naive Bayes, DL = deep learning, RF = random forest, MLP = multilayer perceptron. ¹ Performance metrics of the best or final model in the study. ² Model with best performance metrics/selected as the final model in the study.

4. Results

4.1. Model Comparison

Eight OTC screening models were developed from ML. kNN had the highest Youden J index, precision, and PR-AUC, while the ML model with the highest F2 score was SVM. Table 3 presents the descriptive performance of all ML models, while Figure 2 further illustrates the performance comparison of all models.

Table 3. Descriptive performance of all machine learning models.

Models	Youden J Index Mean (SD)	F2 Score Mean (SD)	Precision Mean (SD)	PR-AUC Mean (SD)
k-nearest neighbour	0.58 (0.06)	0.75 (0.03)	0.83 (0.04)	0.86 (0.02)
Elastic-net logistic regression	0.17 (0.05)	0.62 (0.06)	0.59 (0.03)	0.63 (0.03)
MARS	0.21 (0.05)	0.60 (0.04)	0.62 (0.02)	0.65 (0.03)
Artificial neural network	0.25 (0.05)	0.62 (0.04)	0.64 (0.03)	0.67 (0.03)
Partial least square	0.19 (0.01)	0.59 (0.01)	0.61 (0.01)	0.62 (0.01)
Random forest	0.35 (0.04)	0.66 (0.03)	0.69 (0.02)	0.74 (0.03)
Support vector machine	0.08 (0.16)	0.79 (0.09)	0.55 (0.08)	0.64 (0.06)
XGBoost	0.17 (0.07)	0.62 (0.09)	0.60 (0.04)	0.65 (0.03)

MARS = multivariate adaptive regression splines, XGBoost = extreme gradient boosting, PR-AUC = precision recall-area under the curve.

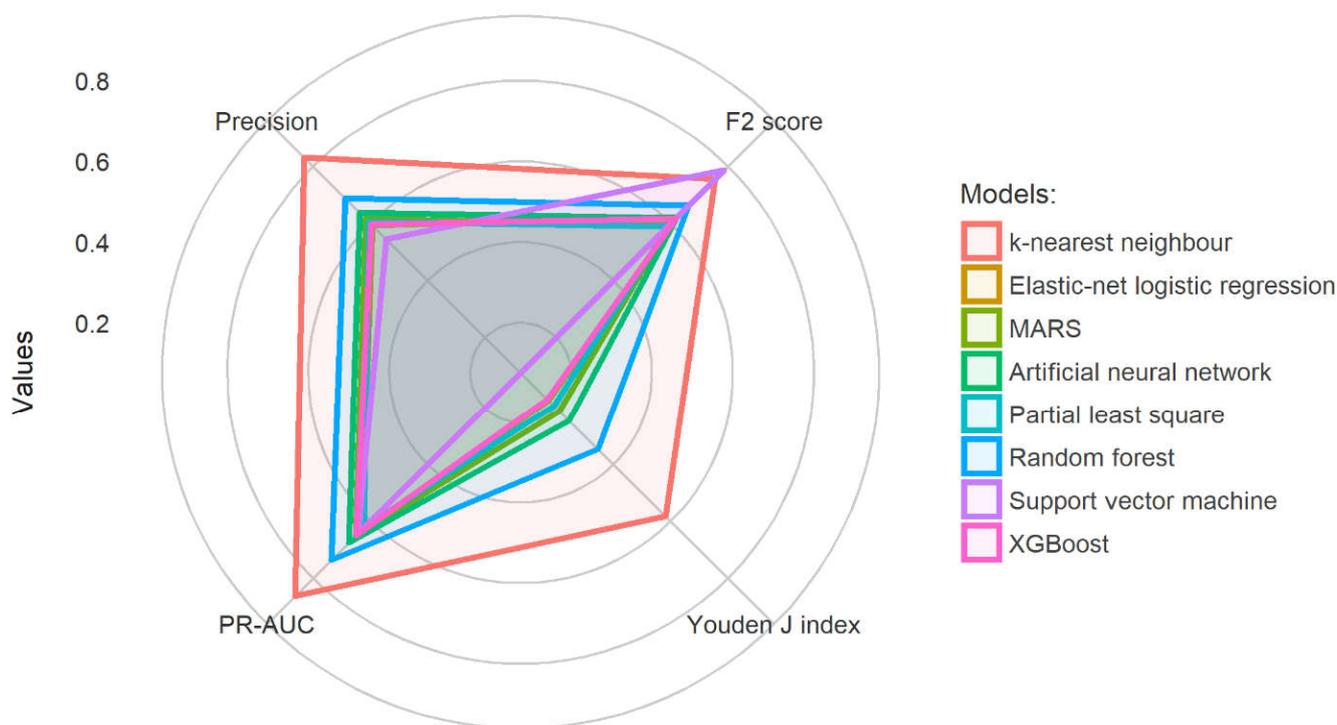


Figure 2. Model comparison across four performance metrics.

One-way ANOVA showed that there was a significant difference between the mean of Youden J index, F2 score, precision, and PR-AUC among all ML models (Table 4). Further post hoc pairwise comparison using *t*-test indicated all pairwise comparisons were significant after Bonferroni correction except for XGBoost vs. elastic-net logistic regression for Youden J index and XGBoost vs. elastic-net logistic regression, ANN vs. elastic-net logistic regression, and XGBoost vs. ANN for F2 score (Figure 3). Thus, kNN was identified to be the best ML model for the purpose of OTC breast cancer screening in this study.

Table 4. Model comparison using one-way ANOVA test.

Models	n	Youden J Index		F2 Score		Precision		PR-AUC	
		F-Statistics (df1, df2)	p-Value						
kNN	5000								
EN-LR	5000								
MARS	3140								
ANN	5000	21,471	<i>p</i> < 0.01	8511	<i>p</i> < 0.01	24,768	<i>p</i> < 0.01	27,694	<i>p</i> < 0.01
PLS	5000	(7, 38,132)		(7, 38,132)		(7, 38,132)		(7, 38,132)	
RF	5000								
SVM	5000								
XGBoost	5000								

kNN = k-nearest neighbour, EN-LR = elastic-net logistic regression, MARS= multivariate adaptive regression splines, ANN = artificial neural network, PLS = partial least square, RF = random forest, SVM = support vector machine, XGBoost = extreme gradient boosting, PR-AUC = precision recall-area under the curve.

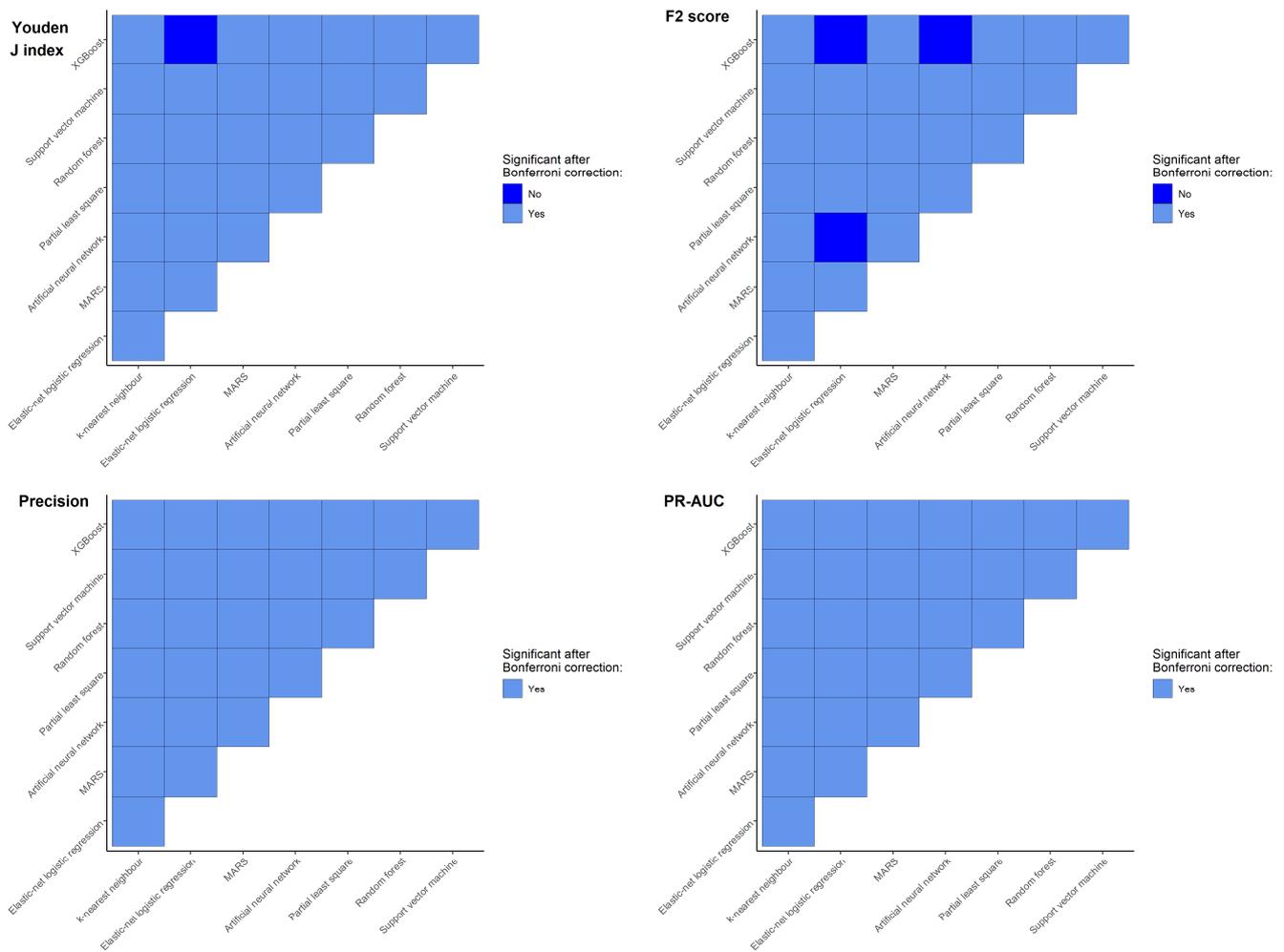


Figure 3. Post hoc pairwise comparison using *t*-test.

4.2. Hyperparameter Tuning

Table 5 presents the four results of hyperparameter tuning with the highest Youden J index, F2 score, precision, and PR-AUC. Models 1, 2, and 4 had lower specificity than sensitivity, while model 3 had it otherwise. kNN model 3 was selected as the best hyperparameters tuning result as it had the highest sensitivity.

Table 5. Top four hyperparameter tuning results of k-nearest neighbour with the highest Youden J index, F2 score, precision, and precision recall-area under the curve.

Model	Fold	Neighbours	Distance Weighting Function	Minkowski Distance	Sensitivity Mean (SD)	Specificity Mean (SD)	Youden J Index Mean (SD)	F2 Score Mean (SD)	Precision Mean (SD)	PR-AUC Mean (SD)
1	1	1	Inversion	1.24	0.77 (0.03)	0.88 (0.03)	0.65 (0.04)	0.78 (0.03)	0.87 (0.03)	0.88 (0.02)
2	3	10	Triweight	1.92	0.76 (0.03)	0.87 (0.03)	0.63 (0.03)	0.78 (0.02)	0.86 (0.02)	0.89 (0.02)
3	10	3	Rank	1.99	0.82 (0.02)	0.79 (0.03)	0.62 (0.03)	0.82 (0.02)	0.81 (0.03)	0.88 (0.02)
4	10	4	Triweight	1.97	0.79 (0.02)	0.87 (0.03)	0.66 (0.03)	0.80 (0.02)	0.87 (0.03)	0.88 (0.02)

PR-AUC = precision recall-area under the curve.

4.3. Explainable Approach

Table 6 displays the performance of the final kNN model on the validation dataset across mammographic density. The model had a higher sensitivity on the non-dense cases and a higher specificity on the dense cases. Additionally, the performance differences across the mammographic density were very minimal as shown in Table 5. Furthermore,

Figure 4 indicated that there was no difference between PR-AUC of non-dense and dense breast women for the final kNN model as both lines were overlapped.

Table 6. Performance metrics across mammographic density on the validation dataset.

Performance Metrics	Validation Dataset		
	Overall	Non-Dense	Dense
Sensitivity	0.74	0.76	0.71
Specificity	0.34	0.25	0.43
Youden J index	0.08	0.01	0.15
F2 score	0.75	0.77	0.73
Precision	0.80	0.80	0.81
PR-AUC	0.82	0.83	0.82

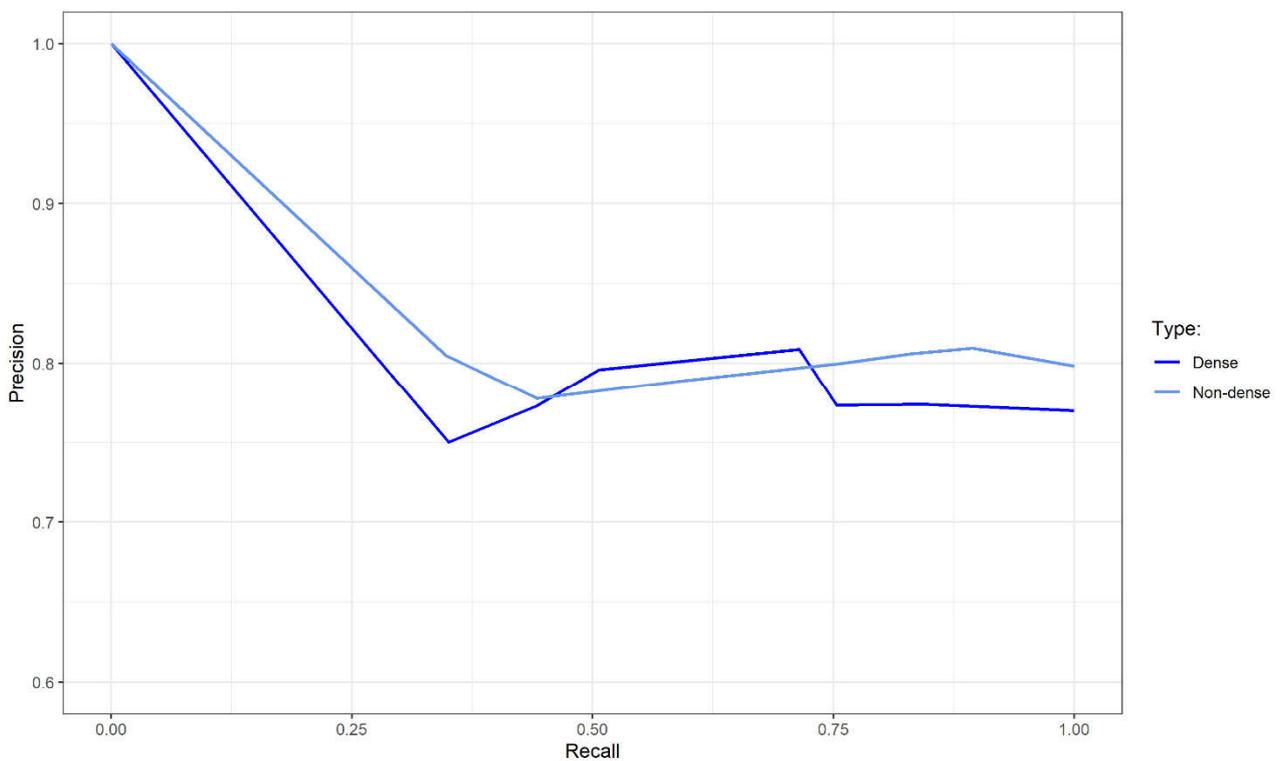


Figure 4. Precision recall-area under the curve for the final machine learning model across mammographic density on the validation dataset.

Figure 5 illustrates the top fifteen influential features of the final ML model. The top three most influential variables were age at examination, birth control/hormone replacement, and race. In terms of patient complaints, breast pain, breast lump, and breast trauma were the most important factors that influence the model’s prediction as opposed to the other complaints.

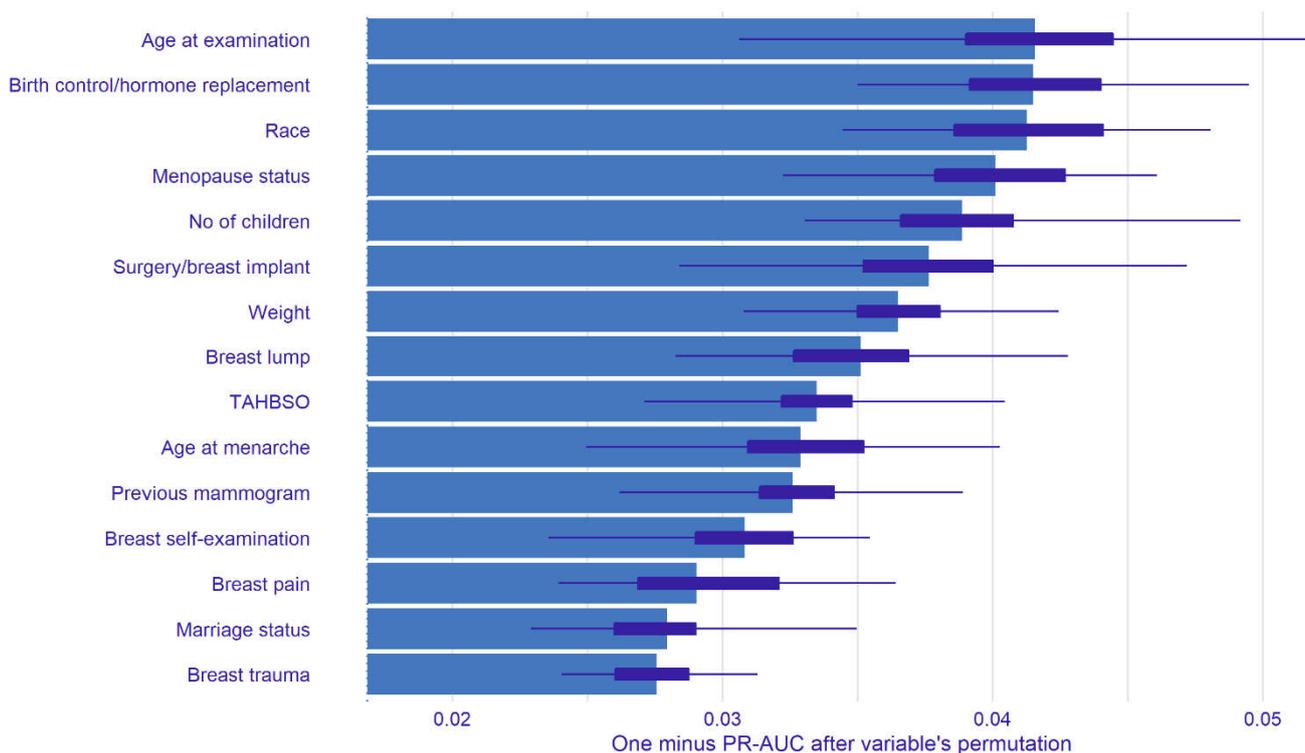


Figure 5. Top fifteen influential features for the k-nearest neighbour model. The bar indicates the mean values of one minus PR-AUC, and the box plot reflects the distribution of the values of one minus PR-AUC.

5. Discussion

In this study, we evaluated the feasibility of OTC breast cancer screening models developed from ML. The model was aimed to predict women with suspicious breast problems or women with a high probability of developing breast cancer. The screening model used the information obtained during patient registration prior to a medical consultation with the clinician. Thus, patients with a suspicious breast issue would be prioritised at the screening stage and referred to a breast cancer specialist for timely consultation. Previous studies showed that early detection of breast cancer reduces its mortality [68,69]. Additionally, one of the factors of severe breast cancer presentation and poor survival among breast cancer patients was a delay in seeking medical treatment [70–73]. The development of the OTC screening model would be beneficial in minimising the time between a woman first noticing a symptom and arranging a medical consultation. At least about 17% of women with breast cancer symptoms in European countries had a delayed medical consultation of at least 3 months or more [74]. In southeast Asian countries such as Malaysia, a delay in medical consultation was estimated at 2 months [75]. In general, shortening the delay in arranging medical consultations would be helpful for the prognosis of breast cancer women.

OTC models were developed from eight ML models in this study. The kNN models were significantly better than the other seven models in terms of the Youden J index, precision, and PR-AUC. Additionally, in terms of F2 score SVM had the highest performance value. Thus, the best model based on the four-performance metrics was kNN followed by random forest and ANN. The SVM model had the lowest Youden J index and precision and one of the lowest PR-AUC, despite having the highest F2 score. SVM was believed to work well with imbalanced datasets compared to other ML models, however, this was not the case in our study [76]. Additionally, the final kNN model had a balanced performance between sensitivity and specificity (Table 4). In the hyperparameters tuning stage, we prioritised ML models with a higher sensitivity value. The OTC model aimed to be deployed in the breast clinic during the registration prior to the medical consultation.

The model with high sensitivity would prioritise women with a suspicious breast issue which in turn accelerates the needed process for those with medical urgency.

The features used for the development of ML screening models were sociodemographic information, medical history, and patient complaints. A study conducted to develop ML models to predict breast cancer in Chinese women included ten risk factors that achieved the best sensitivity and specificity of 0.66 and 0.69 using XGBoost [66]. This study achieved the sensitivity and specificity of 0.82 and 0.79, respectively, using kNN. Therefore, our study showed that adding patient symptoms or complaints to the features used in the development of the screening model improved the predictive performance of the screening model. Another study conducted to predict breast cancer using laboratory data showed the best precision performance at 0.85 using ANN [65] while the precision for our final kNN model was at 0.81. Although the performance of our model was slightly lower, however, obtaining laboratory data before medical consultation was unfeasible and impractical in our study.

Mammographic density is a known risk factor for developing breast cancer [77]. Asian women had a higher mammographic density than non-Asian women [78,79], thus, having a higher risk of getting breast cancer. For example, in Malaysia, Chinese women had been shown to have denser breasts than the other races [80,81]. A few studies denoted that the proportion of women who attended mammogram procedures in Malaysia was at least half of them were women with dense breasts [82,83]. An ML screening model aimed to be applied to this population should take this information into account. However, it was inappropriate to include the mammographic density as one of the features in the screening model as the density was known at a later stage after medical examination. The final kNN model had a slightly higher sensitivity and specificity in a non-dense and dense group, respectively (Table 6). However, the comparison of the PR-AUC of the model indicated that there was no performance difference between the two groups. Additionally, the explainable ML revealed the most significant feature in the final model was the age at examination. The incidence of breast cancer had been shown to increase with age [84]. However, breast cancer presented at a younger age tended to be more aggressive and at a higher stage of cancer [84–86]. Thus, in developing the ML screening model, misclassification of suspicious cases as normal cases especially in younger women could be a catastrophic error. Moreover, there were two modifiable features which were weight and breast self-examination (BSE). Weight control had been suggested to reduce breast cancer risk [87,88]. Although BSE did not relate to breast cancer risk, frequent BSE led to an increased incidence of breast cancer [87]. Additionally, there were three influential features related to patient complaints including breast pain, breast lump and breast trauma.

This study used secondary data collected from a university- and research-based hospital in Kelantan, Malaysia. The data was further validated by a radiologist and pathologist to ensure the good quality of the data. However, our study still had a few limitations. One of the main limitations of this study was the size of the data to develop our screening models. The lack of data was a prevalent issue in the application of ML in healthcare [89]. However, this issue was worsened in our study as the dataset had missing values and imbalanced outcome classification. Subsequently, we used a bagged tree model and ROSE algorithm to overcome these issues, and undeniably larger data will further improve our model. Additionally, we only included one hospital in our study as we utilised information from patient registration records which were specific to the BestARi, HUSM at the time this study was conducted. Including more hospitals in the study was not feasible due to the lack of standardisation in the patient registration record among the hospitals. However, future studies should aim to include more hospitals, if possible, thus increasing the size of the data. Nonetheless, the challenges and approaches presented in the study reflected a real workflow in the development and application of the OTC ML model for breast cancer screening.

6. Conclusions

We evaluated eight ML to be developed as an OTC screening model for breast cancer. We used patient registration records including sociodemographic, medical history, and patient complaints as features for the development of the screening models. This study found that the OTC screening models developed from the ML and patient registration records show promising performance. The screening models can be deployed in a breast clinic and improve the workflow of breast cancer management. Thus, the deployment of the model will reduce patient delays in arranging investigations and consultations from the breast cancer team.

Author Contributions: Conceptualization, T.M.H., N.I.R.R., W.N.A. and K.I.M.; data curation, T.M.H., N.I.R.R. and W.N.A.; formal analysis, T.M.H., N.I.R.R. and K.I.M.; funding acquisition, K.I.M.; investigation, J.H. and W.F.W.A.R.; methodology, T.M.H., N.I.R.R. and W.N.A.; project administration, K.I.M.; resources, J.H. and W.F.W.A.R.; supervision, J.H. and R.A.; validation, J.H., W.F.W.A.R. and R.A.; visualization, T.M.H. and K.I.M.; writing—original draft, T.M.H.; writing—review and editing, N.I.R.R., W.N.A., J.H. and K.I.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Fundamental Research Grant Scheme (FRGS), Ministry of Higher Education, Malaysia (FRGS/1/2019/SKK03/USM/02/1).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the human research ethics committee of Universiti Sains Malaysia (JEPeM) on 19 November 2019 (USM/JEPeM/19090536).

Informed Consent Statement: Patient consent was waived due to the retrospective nature of this study and the use of secondary data.

Data Availability Statement: The data are available upon reasonable request to the corresponding author.

Acknowledgments: We thank all staff and workers in the Department of Radiology, Department of Pathology, and BestARi unit in HUSM for facilitating the data collection and extraction process.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. DeSantis, C.E.; Bray, F.; Ferlay, J.; Lortet-Tieulent, J.; Anderson, B.O.; Jemal, A. International variation in female breast cancer incidence and mortality rates. *Cancer Epidemiol. Biomark. Prev.* **2015**, *24*, 1495–1506. [[CrossRef](#)] [[PubMed](#)]
2. WHO Breast Cancer. Available online: <https://www.who.int/news-room/fact-sheets/detail/breast-cancer> (accessed on 24 May 2022).
3. Parks, R.M.; Derks, M.G.M.; Bastiaannet, E.; Cheung, K.L. Breast Cancer Epidemiology. In *Breast Cancer Management for Surgeons*; Springer: Cham, Switzerland, 2018; pp. 5615–5623. ISBN 9783319566733.
4. Anders, C.K.; Johnson, R.; Litton, J.; Phillips, M.; Bleyer, A. Breast Cancer Before Age 40 Years. *Semin. Oncol.* **2009**, *36*, 237–249. [[CrossRef](#)] [[PubMed](#)]
5. Momenimovahed, Z.; Salehiniya, H. Epidemiological characteristics of and risk factors for breast cancer in the world. *Breast Cancer Targets Ther.* **2019**, *11*, 151–164. [[CrossRef](#)]
6. Kamińska, M.; Ciszewski, T.; Łopacka-Szatan, K.; Miotła, P.; Starosławska, E. Breast cancer risk factors. *Prz. Menopauzalny* **2015**, *14*, 196–202. [[CrossRef](#)]
7. Warner, E.T.; Rice, M.S.; Zeleznik, O.A.; Fowler, E.E.; Murthy, D.; Vachon, C.M.; Bertrand, K.A.; Rosner, B.A.; Heine, J.; Tamimi, R.M. Automated percent mammographic density, mammographic texture variation, and risk of breast cancer: A nested case-control study. *npj Breast Cancer* **2021**, *7*, 68. [[CrossRef](#)]
8. Burton, A.; Maskarinec, G.; Perez-Gomez, B.; Vachon, C.; Miao, H.; Lajous, M.; López-Ridaura, R.; Rice, M.; Pereira, A.; Garmendia, M.L.; et al. Mammographic density and ageing: A collaborative pooled analysis of cross-sectional data from 22 countries worldwide. *PLoS Med.* **2017**, *14*, e1002335. [[CrossRef](#)] [[PubMed](#)]
9. Sherratt, M.J.; McConnell, J.C.; Streuli, C.H. Raised mammographic density: Causative mechanisms and biological consequences. *Breast Cancer Res.* **2016**, *18*, 45. [[CrossRef](#)]
10. Nazari, S.S.; Mukherjee, P. An overview of mammographic density and its association with breast cancer. *Breast Cancer* **2018**, *25*, 259–267. [[CrossRef](#)] [[PubMed](#)]

11. Unger-Saldaña, K. Challenges to the early diagnosis and treatment of breast cancer in developing countries. *World J. Clin. Oncol.* **2014**, *5*, 465–477. [[CrossRef](#)]
12. Andersen, B.L.; Cacioppo, J.T.; Roberts, D.C. Delay in seeking a cancer diagnosis: Delay stages and psychophysiological comparison processes. *Br. J. Soc. Psychol.* **1995**, *34*, 33–52. [[CrossRef](#)]
13. Taib, N.A.; Yip, C.H.; Low, W.Y. A grounded explanation of why women present with advanced breast cancer. *World J. Surg.* **2014**, *38*, 1676–1684. [[CrossRef](#)]
14. McKenzie, F.; Zietsman, A.; Galukande, M.; Anele, A.; Adisa, C.; Parham, G.; Pinder, L.; Cubasch, H.; Joffe, M.; Kidaaga, F.; et al. Drivers of advanced stage at breast cancer diagnosis in the multicountry African breast cancer—Disparities in outcomes (ABC-DO) study. *Int. J. Cancer* **2018**, *142*, 1568–1579. [[CrossRef](#)] [[PubMed](#)]
15. Ren, S.; Zhang, Y.; Qin, P.; Wang, J. Factors Influencing Total Delay of Breast Cancer in Northeast of China. *Front. Oncol.* **2022**, *12*, 10–15. [[CrossRef](#)] [[PubMed](#)]
16. Toh, T.S.; Dondelinger, F.; Wang, D. Looking beyond the hype: Applied AI and machine learning in translational medicine. *EBioMedicine* **2019**, *47*, 607–615. [[CrossRef](#)] [[PubMed](#)]
17. Blasiak, A.; Khong, J.; Kee, T. CURATE.AI: Optimizing Personalized Medicine with Artificial Intelligence. *SLAS Technol.* **2020**, *25*, 95–105. [[CrossRef](#)] [[PubMed](#)]
18. Raghu, G.; Remy-Jardin, M.; Myers, J.L.; Richeldi, L.; Ryerson, C.J.; Lederer, D.J.; Behr, J.; Cottin, V.; Danoff, S.K.; Morell, F.; et al. Diagnosis of idiopathic pulmonary fibrosis An Official ATS/ERS/JRS/ALAT Clinical practice guideline. *Am. J. Respir. Crit. Care Med.* **2018**, *198*, e44–e68. [[CrossRef](#)]
19. Hwang, E.J.; Park, S.; Jin, K.N.; Kim, J.I.; Choi, S.Y.; Lee, J.H.; Goo, J.M.; Aum, J.; Yim, J.J.; Park, C.M.; et al. Development and Validation of a Deep Learning-based Automatic Detection Algorithm for Active Pulmonary Tuberculosis on Chest Radiographs. *Clin. Infect. Dis.* **2019**, *69*, 739–747. [[CrossRef](#)]
20. Zou, Q.; Qu, K.; Luo, Y.; Yin, D.; Ju, Y.; Tang, H. Predicting Diabetes Mellitus With Machine Learning Techniques. *Front. Genet.* **2018**, *9*, 515. [[CrossRef](#)]
21. van Leeuwen, N.M.; Maurits, M.; Liem, S.; Ciaffi, J.; Ajmone Marsan, N.; Ninaber, M.; Allaart, C.; Gillet van Dongen, H.; Goekoop, R.; Huizinga, T.; et al. New risk model is able to identify patients with a low risk of progression in systemic sclerosis. *RMD Open* **2021**, *7*, e001524. [[CrossRef](#)]
22. Tiulpin, A.; Klein, S.; Bierma-Zeinstra, S.M.A.; Thevenot, J.; Rahtu, E.; van Meurs, J.; Oei, E.H.G.; Saarakkala, S. Multimodal Machine Learning-based Knee Osteoarthritis Progression Prediction from Plain Radiographs and Clinical Data. *Sci. Rep.* **2019**, *9*, 20038. [[CrossRef](#)]
23. Ansart, M.; Epelbaum, S.; Bassignana, G.; Bône, A.; Bottani, S.; Cattai, T.; Couronné, R.; Faouzi, J.; Koval, I.; Louis, M.; et al. Predicting the progression of mild cognitive impairment using machine learning: A systematic, quantitative and critical review. *Med. Image Anal.* **2021**, *67*, 101848. [[CrossRef](#)] [[PubMed](#)]
24. Dagliati, A.; Marini, S.; Sacchi, L.; Cogni, G.; Teliti, M.; Tibollo, V.; De Cata, P.; Chiovato, L.; Bellazzi, R. Machine Learning Methods to Predict Diabetes Complications. *J. Diabetes Sci. Technol.* **2018**, *12*, 295–302. [[CrossRef](#)] [[PubMed](#)]
25. Ungaro, R.C.; Hu, L.; Ji, J.; Nayar, S.; Kugathasan, S.; Denson, L.A.; Hyams, J.; Dubinsky, M.C.; Sands, B.E.; Cho, J.H. Machine learning identifies novel blood protein predictors of penetrating and stricturing complications in newly diagnosed paediatric Crohn’s disease. *Aliment. Pharmacol. Ther.* **2021**, *53*, 281–290. [[CrossRef](#)] [[PubMed](#)]
26. Lip, G.Y.H.; Genaidy, A.; Tran, G.; Marroquin, P.; Estes, C. Incidence and Complications of Atrial Fibrillation in a Low Socioeconomic and High Disability United States (US) Population: A Combined Statistical and Machine Learning Approach. *Int. J. Clin. Pract.* **2022**, *2022*, 8649050. [[CrossRef](#)]
27. Poon, A.I.F.; Sung, J.J.Y. Opening the black box of AI-Medicine. *J. Gastroenterol. Hepatol.* **2021**, *36*, 581–584. [[CrossRef](#)]
28. Adadi, A.; Berrada, M. Explainable AI for Healthcare: From Black Box to Interpretable Models. In *Advances in Intelligent Systems and Computing*; Springer: Singapore, 2020; Volume 1076, pp. 327–337. ISBN 9789811509469.
29. Sabol, P.; Sinčák, P.; Hartono, P.; Kočan, P.; Benetinová, Z.; Blichárová, A.; Verbóová, L.; Štammová, E.; Sabolová-Fabianová, A.; Jašková, A. Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopathological images. *J. Biomed. Inform.* **2020**, *109*, 103523. [[CrossRef](#)]
30. Cozma, G.V.; Onchis, D.; Istin, C.; Petrache, I.A. Explainable Machine Learning Solution for Observing Optimal Surgery Timings in Thoracic Cancer Diagnosis. *Appl. Sci.* **2022**, *12*, 6506. [[CrossRef](#)]
31. Kim, H.M.; Jeong, C.W.; Kwak, C.; Song, C.; Kang, M.; Seo, S.I.; Kim, J.K.; Lee, H.; Chung, J.; Hwang, E.C.; et al. A Machine Learning Approach to Predict the Probability of Brain Metastasis in Renal Cell Carcinoma Patients. *Appl. Sci.* **2022**, *12*, 6174. [[CrossRef](#)]
32. Wang, L.; Lin, Z.Q.; Wong, A. COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Sci. Rep.* **2020**, *10*, 19549. [[CrossRef](#)]
33. Sarp, S.; Kuzlu, M.; Wilson, E.; Cali, U.; Guler, O. The enlightening role of explainable artificial intelligence in chronic wound classification. *Electron.* **2021**, *10*, 1406. [[CrossRef](#)]
34. El-Sappagh, S.; Alonso, J.M.; Islam, S.M.R.; Sultan, A.M.; Kwak, K.S. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer’s disease. *Sci. Rep.* **2021**, *11*, 2660. [[CrossRef](#)] [[PubMed](#)]
35. Holzinger, A.; Langs, G.; Denk, H.; Zatloukal, K.; Müller, H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2019**, *9*, e1312. [[CrossRef](#)] [[PubMed](#)]

36. Lunardon, N.; Menardi, G.; Torelli, N. ROSE: A package for binary imbalanced learning. *R J.* **2014**, *6*, 79–89. [[CrossRef](#)]
37. Yeo, I.-K.; Johnson, R.A. A new family of power transformations to improve normality or symmetry. *Biometrika* **2000**, *87*, 954–959. [[CrossRef](#)]
38. Kuhn, M.; Johnson, K. *Feature Engineering and Selection: A Practical Approach for Predictive Models*, 1st ed.; CRC Press: Boca Raton, FL, USA, 2020; ISBN 9781315108230.
39. Hvitfeldt, E. Themis: Extra Recipes Steps for Dealing with Unbalanced Data. 2022. Available online: <https://themis.tidymodels.org> (accessed on 13 October 2022).
40. Kuhn, M.; Wickham, H. Recipes: Preprocessing and Feature Engineering Steps for Modeling. 2022. Available online: <https://rdrr.io/cran/recipes/> (accessed on 13 October 2022).
41. Kuhn, M.; Vaughan, D. Parsnip: A Common API to Modeling and Analysis Functions. 2022. Available online: <https://rdrr.io/cran/parsnip/> (accessed on 13 October 2022).
42. Schliep, K.; Hechenbichler, K. kkn: Weighted k-Nearest Neighbors 2016. Available online: <https://github.com/KlausVigo/kkn> (accessed on 13 October 2022).
43. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **2010**, *33*, 1–22. [[CrossRef](#)]
44. Milborrow, S. Earth: Multivariate Adaptive Regression Splines. 2021. Available online: <http://www.milbo.users.sonic.net/earth/> (accessed on 13 October 2022).
45. Ripley, B.; Venables, W. Nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models. 2022. Available online: <https://rdrr.io/cran/nnet/> (accessed on 13 October 2022).
46. Rohart, F.; Gautier, B.; Singh, A.; Lê Cao, K.-A. mixOmics: An R package for ‘omics feature selection and multiple data integration. *PLOS Comput. Biol.* **2017**, *13*, e1005752. [[CrossRef](#)]
47. Wright, M.N.; Ziegler, A. Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Stat. Softw.* **2017**, *77*, 1–17. [[CrossRef](#)]
48. Karatzoglou, A.; Smola, A.; Hornik, K.; Zeileis, A. kernlab—An S4 Package for Kernel Methods in R. *J. Stat. Softw.* **2004**, *11*, 1–20. [[CrossRef](#)]
49. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K.; Mitchell, R.; Cano, I.; Zhou, T.; et al. Xgboost: Extreme Gradient Boosting. 2022. Available online: <https://cran.utstat.utoronto.ca/web/packages/xgboost/vignettes/xgboost.pdf> (accessed on 13 October 2022).
50. R Core Team. *R: A Language and Environment for Statistical Computing*; R Core Team: Vienna, Austria, 2022.
51. Biecek, P. DALEX: Explainers for Complex Predictive Models in R. *J. Mach. Learn. Res.* **2018**, *19*, 1–5.
52. Maksymiuk, S.; Gosiewska, A.; Biecek, P. Landscape of R packages for eXplainable Artificial Intelligence. *arXiv* **2020**, arXiv:2009.13248. [[CrossRef](#)]
53. Yassin, N.I.R.; Omran, S.; El Houbay, E.M.F.; Allam, H. Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. *Comput. Methods Programs Biomed.* **2018**, *156*, 25–45. [[CrossRef](#)] [[PubMed](#)]
54. Breast Cancer Wisconsin (Diagnostic) Data Set. Available online: [https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+\(diagnostic\)](https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)) (accessed on 6 July 2022).
55. Richter, A.N.; Khoshgoftaar, T.M. A review of statistical and machine learning methods for modeling cancer risk using structured clinical data. *Artif. Intell. Med.* **2018**, *90*, 1–14. [[CrossRef](#)] [[PubMed](#)]
56. Li, J.; Zhou, Z.; Dong, J.; Fu, Y.; Li, Y.; Luan, Z.; Peng, X. Predicting breast cancer 5-year survival using machine learning: A systematic review. *PLoS ONE* **2021**, *16*, e0250370. [[CrossRef](#)] [[PubMed](#)]
57. Kim, W.; Kim, K.S.; Lee, J.E.; Noh, D.-Y.; Kim, S.-W.; Jung, Y.S.; Park, M.Y.; Park, R.W. Development of novel breast cancer recurrence prediction model using support vector machine. *J. Breast Cancer* **2012**, *15*, 230–238. [[CrossRef](#)]
58. LG, A.; AT, E. Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence. *J. Health Med. Inform.* **2013**, *4*, 2–4. [[CrossRef](#)]
59. Zeng, Z.; Yao, L.; Roy, A.; Li, X.; Espino, S.; Clare, S.E.; Khan, S.A.; Luo, Y. Identifying Breast Cancer Distant Recurrences from Electronic Health Records Using Machine Learning. *J. Healthc. Inform. Res.* **2019**, *3*, 283–299. [[CrossRef](#)]
60. Cirkovic, B.R.A.; Cvetkovic, A.M.; Ninkovic, S.M.; Filipovic, N.D. Prediction models for estimation of survival rate and relapse for breast cancer patients. In Proceedings of the 2015 IEEE 15th International Conference on Bioinformatics and Bioengineering (BIBE), Belgrade, Serbia, 2–4 November 2015; pp. 1–6.
61. Kabiraj, S.; Railhan, M.; Alvi, N.; Afrin, M.; Akter, L.; Sohagi, S.A.; Podder, E. Breast cancer risk prediction using XGBoost and random forest algorithm. In Proceedings of the 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 1–3 July 2020; pp. 1–4.
62. Sun, D.; Wang, M.; Li, A. A Multimodal Deep Neural Network for Human Breast Cancer Prognosis Prediction by Integrating Multi-Dimensional Data. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **2019**, *16*, 841–850. [[CrossRef](#)]
63. Kalafi, E.Y.; Nor, N.A.M.; Taib, N.A.; Ganggayah, M.D.; Town, C.; Dhillon, S.K. Machine Learning and Deep Learning Approaches in Breast Cancer Survival Prediction Using Clinical Data. *Folia Biol.* **2019**, *65*, 212–220.
64. Anisha, P.R.; Kishor Kumar Reddy, C.; Apoorva, K.; Meghana Mangipudi, C. Early Diagnosis of Breast Cancer Prediction using Random Forest Classifier. *IOP Conf. Ser. Mater. Sci. Eng.* **2021**, *1116*, 012187. [[CrossRef](#)]

65. Khatun, T.; Utsho, M.M.R.; Islam, M.A.; Zohura, M.F.; Hossen, M.S.; Rimi, R.A.; Anni, S.J. Performance Analysis of Breast Cancer: A Machine Learning Approach. In Proceedings of the 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2–4 September 2021; pp. 1426–1434.
66. Hou, C.; Zhong, X.; He, P.; Xu, B.; Diao, S.; Yi, F.; Zheng, H.; Li, J. Predicting breast cancer in chinese women using machine learning techniques: Algorithm development. *JMIR Med. Inform.* **2020**, *8*, e17364. [[CrossRef](#)]
67. Nindrea, R.D.; Aryandono, T.; Lazuardi, L.; Dwiprahasto, I. Diagnostic accuracy of different machine learning algorithms for breast cancer risk calculation: A meta-analysis. *Asian Pac. J. Cancer Prev.* **2018**, *19*, 1747–1752. [[CrossRef](#)] [[PubMed](#)]
68. Malvezzi, M.; Carioli, G.; Bertuccio, P.; Boffetta, P.; Levi, F.; La Vecchia, C.; Negri, E. European cancer mortality predictions for the year 2019 with focus on breast cancer. *Ann. Oncol.* **2019**, *30*, 781–787. [[CrossRef](#)] [[PubMed](#)]
69. Tahmooresi, M.; Afshar, A.; Bashari Rad, B.; Nowshath, K.B.; Bamiah, M.A. Early detection of breast cancer using machine learning techniques. *J. Telecommun. Electron. Comput. Eng.* **2018**, *10*, 21–27.
70. Khan, T.M.; Leong, J.P.Y.; Ming, L.C.; Khan, A.H. Association of knowledge and cultural perceptions of Malaysian women with delay in diagnosis and treatment of breast cancer: A systematic review. *Asian Pac. J. Cancer Prev.* **2015**, *16*, 5349–5357. [[CrossRef](#)]
71. Mujar, N.M.M.; Dahlui, M.; Emran, N.A.; Hadi, I.A.; Wai, Y.Y.; Arulanantham, S.; Hooi, C.C.; Taib, N.A.M. Complementary and alternative medicine (CAM) use and delays in presentation and diagnosis of breast cancer patients in public hospitals in Malaysia. *PLoS ONE* **2017**, *12*, e0176394. [[CrossRef](#)]
72. Caplan, L. Delay in breast cancer: Implications for stage at diagnosis and survival. *Front. Public Health* **2014**, *2*, 87. [[CrossRef](#)]
73. Freitas, A.G.Q.; Weller, M. Patient delays and system delays in breast cancer treatment in developed and developing countries. *Cien. Saude Colet.* **2015**, *20*, 3177–3189. [[CrossRef](#)]
74. Innos, K.; Padrik, P.; Valvere, V.; Eelma, E.; Kütner, R.; Lehtsaar, J.; Tekkel, M. Identifying women at risk for delayed presentation of breast cancer: A cross-sectional study in Estonia. *BMC Public Health* **2013**, *13*, 947. [[CrossRef](#)]
75. Norsa'adah, B.; Rampal, K.G.; Rahmah, M.A.; Naing, N.N.; Biswal, B.M. Diagnosis delay of breast cancer and its associated factors in Malaysian women. *BMC Cancer* **2011**, *11*, 141. [[CrossRef](#)]
76. Sun, Y.; Wong, A.K.C.; Kamel, M.S. Classification of imbalanced data: A review. *Int. J. Pattern Recognit. Artif. Intell.* **2009**, *23*, 687–719. [[CrossRef](#)]
77. Johnson, R.H.; Anders, C.K.; Litton, J.K.; Ruddy, K.J.; Bleyer, A. Breast cancer in adolescents and young adults. *Pediatr. Blood Cancer* **2018**, *65*, e27397. [[CrossRef](#)] [[PubMed](#)]
78. Rajaram, N.; Mariapun, S.; Eriksson, M.; Tapia, J.; Kwan, P.Y.; Ho, W.K.; Harun, F.; Rahmat, K.; Czene, K.; Taib, N.A.M.; et al. Differences in mammographic density between Asian and Caucasian populations: A comparative analysis. *Breast Cancer Res. Treat.* **2017**, *161*, 353–362. [[CrossRef](#)] [[PubMed](#)]
79. Yap, Y.S.; Lu, Y.S.; Tamura, K.; Lee, J.E.; Ko, E.Y.; Park, Y.H.; Cao, A.Y.; Lin, C.H.; Toi, M.; Wu, J.; et al. Insights into Breast Cancer in the East vs the West: A Review. *JAMA Oncol.* **2019**, *5*, 1489–1496. [[CrossRef](#)] [[PubMed](#)]
80. Kumari Chelliah, K.; Shatirah Mohd Fandi Voon, N.; Ahamad, H. Breast Density: Does It Vary among the Main Ethnic Groups in Malaysia? *Open J. Med. Imaging* **2013**, *03*, 105–109. [[CrossRef](#)]
81. Mariapun, S.; Li, J.; Yip, C.H.; Taib, N.A.M.; Teo, S.H. Ethnic differences in mammographic densities: An Asian cross-sectional study. *PLoS ONE* **2015**, *10*, e0117568. [[CrossRef](#)]
82. Hanis, T.M.; Arifin, W.N.; Haron, J.; Wan Abdul Rahman, W.F.; Ruhaiyem, N.I.R.; Abdullah, R.; Musa, K.I. Factors Influencing Mammographic Density in Asian Women: A Retrospective Cohort Study in the Northeast Region of Peninsular Malaysia. *Diagnostics* **2022**, *12*, 860. [[CrossRef](#)]
83. Rahayu, A.; Zaharuddin, B.; Le, T.Q.; Rifhana, I.; Muhamad, B.; Mahmud, R.; Hamid, S.A. Relation of Breast Density with Age and Ethnicity in Malaysia. *Front. Health Inform.* **2013**, *2*, 1–4.
84. McGuire, A.; Brown, J.A.L.; Malone, C.; McLaughlin, R.; Kerin, M.J. Effects of age on the detection and management of breast cancer. *Cancers* **2015**, *7*, 908–929. [[CrossRef](#)]
85. Murphy, B.L.; Day, C.N.; Hoskin, T.L.; Habermann, E.B.; Boughey, J.C. Adolescents and Young Adults with Breast Cancer have More Aggressive Disease and Treatment Than Patients in Their Forties. *Ann. Surg. Oncol.* **2019**, *26*, 3920–3930. [[CrossRef](#)]
86. Tao, Z.Q.; Shi, A.; Lu, C.; Song, T.; Zhang, Z.; Zhao, J. Breast Cancer: Epidemiology and Etiology. *Cell Biochem. Biophys.* **2015**, *72*, 333–338. [[CrossRef](#)]
87. Chan, D.S.M.; Abar, L.; Cariolou, M.; Nanu, N.; Greenwood, D.C.; Bandera, E.V.; McTiernan, A.; Norat, T. World Cancer Research Fund International: Continuous Update Project—Systematic literature review and meta-analysis of observational cohort studies on physical activity, sedentary behavior, adiposity, and weight change and breast cancer risk. *Cancer Causes Control* **2019**, *30*, 1183–1200. [[CrossRef](#)] [[PubMed](#)]
88. Ligibel, J.A.; Basen-Engquist, K.; Bea, J.W. Weight Management and Physical Activity for Breast Cancer Prevention and Control. *Am. Soc. Clin. Oncol. Educ. B.* **2019**, *39*, e22–e33. [[CrossRef](#)] [[PubMed](#)]
89. Ahmed, Z.; Mohamed, K.; Zeeshan, S.; Dong, X.Q. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database* **2020**, *2020*, baaa010. [[CrossRef](#)] [[PubMed](#)]