Journal of Current Science and Technology, September - December 2023 Vol. 13 No. 3, pp. 642-656
Copyright ©2018-2023, Rangsit University ISSN 2630-0656 (Online)

Cite this article: Pechprasarn, S., Wattanapermpool, O., Warunlawan, M., Homsud, P., & Akarajarasroj, T. (2023). Identification of Important Factors in the Diagnosis of Breast Cancer Cells Using Machine Learning Models and Principal Component Analysis. *Journal of Current Science and Technology*, *13*(3), 642-656.
DOI:

Journal of Current Science and Technology

Journal homepage: https://jcst.rsu.ac.th

Identification of Important Factors in the Diagnosis of Breast Cancer Cells Using Machine Learning Models and Principal Component Analysis

Suejit Pechprasarn[1*], Ohmthong Wattanapermpool[2], Maninya Warunlawan[2], Pornchaya Homsud[2], and Thumpussorn Akarajarasroj[2]

[1]College of Biomedical Engineering, Rangsit University, Pathum Thani, 12000, Thailand
[2]Satriwithaya School, Wat Bowon Niwet, Phra Nakhon, Bangkok, 10200, Thailand

*Corresponding author; E-mail: suejit.p@rsu.ac.th

Received 7 March, 2023; Revised 17 May, 2023; Accepted 14 June, 2023;
Published online 30 August 2023

Abstract

Breast cancer (BC) is now identified as a disease with a significant impact on morbidity and mortality that is growing and widespread worldwide. This study uses a publicly available clinical dataset of 699 patients from the University of Wisconsin with 9 variables: (1) clump thickness, (2) uniformity of cell size, (3) uniformity of cell shape, (4) marginal adhesion, (5) single epithelial cell size, (6) bare nuclei, (7) bland chromatin, (8) normal nucleoli, and (9) mitoses. This dataset has been used for many studies in the past to pinpoint critical factors in patient diagnosis. Here, we use this data to ensure its unbiasedness and accuracy. We then apply principal component analysis and machine learning models to identify factors in diagnosing a malignant or benign tumor. We investigate and compare the classification accuracy of different machine learning models, including tree, linear discriminant, quadratic discriminant, logistic regression, naive Bayes, support vector machine (SVM), K-nearest neighbor (KNN), ensemble, neural network, and kernel. The best models that can achieve the highest accuracy are medium Gaussian SVM, coarse Gaussian SVM, and cosine KNN, with an accuracy of 96.5%. The principal component analysis method is then performed to identify crucial components and build an accurate model with fewer parameters. The medium Gaussian SVM has the highest cross-validation classification accuracy of 96.98% and requires only three predictors: normal nucleoli, bare nuclei, and cell size uniformity.

Keywords: *breast cancer classification; classification of malignant and benign cells; machine learning; principal component analysis; complexity reduced model; intelligent diagnostic software*

1. Introduction

Breast Cancer (BC) is the most common life-threatening malignancy in women. Nowadays, breast cancer in the early stage can be roughly 80% curable due to advancements in treatment strategies (Sheikh et al., 2022). However, metastatic breast cancer (MBC) or stage IV is the stage in which cancer has spread beyond the breast to other areas of the body and is mainly considered incurable (McGuire et al., 2015). It is imperative to study this issue worldwide to maintain the quality of life and increase the life expectancy of those coping with breast cancer (Shimoi et al., 2020).

The chance of developing breast cancer in their lifetime for women in developing countries is approximately 13% or one in every 8 women (Yedjou et al., 2019), with an increasing number of cases of around 0.5% per year (Siegel et al., 2023). Primarily, breast cancer occurs in middle-aged to older women. However, the statistics show a significant decline of 43 percent in breast cancer mortality rates from 1989 to 2020 due to improved treatments and higher awareness (Giaquinto et al., 2022).

The invasions and malignancies of BC classify various types of BC. Non-invasive BC is the presence of atypical cells that do not invade other surrounding breast tissues and are confined to the ducts, such as ductal carcinoma in situ (DCIS) and also lobular carcinoma in situ (LCIS) (Dange et al., 2017). On the other hand, invasive breast cancer is the presence of spreading and invasive cells that break through the duct. Based on the National Breast Cancer Foundation Inc, DCIS is the most frequent type of non-invasive BC that has not spread out of the ducts, while LCIS presents a higher risk of developing cancer cells in either breast. Other common BC types are invasive, such as infiltrating lobular carcinoma (ILC) and infiltrating ductal carcinoma (IDC), and they usually metastasize to other fatty tissues of the breasts (Das et al., 2021).

Several studies from the Center for Disease Control and Prevention (CDC) clearly show that the risks or causes of BC come from many distinct factors. Female individuals older than 50 are at the highest risk of developing breast cancer (Richardson et al., 2022). Data from breast cancer patients indicate that the cause of BC may be attributed to environmental factors, stress level, heredity, or even lifestyle factors (Dumalaon-Canaria et al., 2014).

Individuals must observe their breasts for thickened tissues or lumps, the first symptoms usually present. Most breast lumps are not harmful, but women should be aware of the potential signs of BC and always get checked by specialists (Swathi et al., 2019). Changes in the size, shape, and appearance of itchy skin and swelling lumps on the breasts may also be disease symptoms (Oliveri et al., 2019).

Obesity, physical inactivity, and radiation exposure are more likely to be the primary risk factors for developing breast cancer among women. Nevertheless, breast cancer can also develop due to many other things, including starting menstruation early, late, or no pregnancy (Yau et al., 2022). Moreover, the consumption of alcohol and hormone replacement therapy can also escalate the risks of getting breast cancer (Liu et al., 2015).

Based on screening and doctor's recommendation, there are several treatments for BC patients with different diagnosed results and stages. The five standard treatments are surgery, radiotherapy, chemotherapy, targeted therapy, and hormone therapy. Patients with BC should strictly follow their doctors' guidance on medication, treatment(s), or a combination. Certain medicines such as Tamoxifen, Raloxifene, and Aromatase Inhibitors (Cornell et al., 2022) could be used to lessen the risk of breast cancer with a doctor's recommendation according to the individual's health risks and symptoms.

These days, the world has been developing with the breakthrough in computer science and many other things. There have been predicted statements by Artificial Intelligence (AI) about the scheme and achievement of screening and diagnosis of breast cancer in laboratories (Dileep & Gyani, 2022). Breast cancer claimed more than 600,000 lives globally in 2018. Health organizations worldwide suggest screening mammography for the early detection of malignant growth sites because it effectively lowers mortality from bosom disease by 20–40% (Iqbal et al., 2022).

AI has recently made massive progress in screening methods for breast cancer (Saelee et al., 2022), which is emerging and becoming helpful to the treatment in detecting breast cancer in the very early stage and more precise prognosis (Nassif et al., 2022; Shah et al., 2022). Advancement in digital pathology also helps create high-quality images of tissues specimen using computerized technology (Gupta et al., 2022). The AI lies in detecting variations of conditions in different stages to evaluate the disease's progress (Dileep & Gyani, 2022).

Numerous optimization algorithms exist, such as the Binary Crocodiles Hunting Strategy, the proximal gradient method, and Dykstra's algorithm (Perrin & Roncalli, 2020). However, the Principal Component Analysis (PCA) approach is one of the simplest methods as it reduces the complexity of ML models by lessening the dimensions of models (Lever et al., 2017). Compared to other model reduction approaches, the indicated method results in a satisfactorily high speed and low computing demands.

Software screening plays a significant role in aiding doctors to diagnose patients, allowing them to prophesy and spot a clearer view of BC indications. With precise integration, AI helps reduce the healthcare staff's workload, computational burden, and resources (Gill et al., 2022). Furthermore, pathologists are rare in the medical field. According to the Thai Institute of Pathology, the Ministry of Public

Thailand's healthcare system has 18 surgical pathologists and approximately 102,000 worldwide (Colangelo et al., 2022). Moreover, the ratio of the number of pathologists to physicians in the world is 1 to 125. These statistics present a shortage of pathologists, leading to misdiagnoses of diseases, especially cancer (Troxel, 2006).

Dr. WIlliam H. Wolberg, University of Wisconsin Hospitals, Madison, Wisconsin, USA, has donated a dataset of 699 patients collected from January 1989 to November 1991 with benign and malignant cells as an open-source database, which can be downloaded from the University of California Irvin (UCI) machine learning repository website. Here, we have employed the dataset (Mangasarian & Wolberg, 1990; Wolberg & Mangasarian, 1990; Wolberg, Mangasarian & Setiono, 1989). The dataset is a biased dataset containing 458 patients having benign cells (B) and 241 patients having malignant cells (M). Here, the dataset was curated by randomly removing 217 patients from the B group, so the B and the G groups have the same size becoming an unbiased dataset. The curated dataset was then analyzed using built-in Machine Learning (ML) models under Matlab 2022b to see whether trained models correctly classified the cell type. We have also performed principal component analysis (PCA) (Panyamit et al., 2022) to reduce the model's size and identify crucial parameters to classify the type of breast cell.

2. Objectives

1. This research uses statistical machine learning models to determine the clinical variables that are most significant for clinicians to consider when diagnosing breast tumors. The study will evaluate the accuracy of the models in classifying breast tumors as malignant or benign and identify factors that affect prediction accuracy in breast cancer patients. The machine learning models will be trained using a dataset that includes clinical variables and tumor characteristics of breast cancer patients. The researchers will employ a 5-fold cross-validation method to assess the performance of the models, and they will calculate accuracy, recall, precision, and F_1 score for both the training dataset and the validation dataset.
2. Use machine learning (ML) to predict the accuracy rates of breast cancer diagnosis affected by each factor and compare them to the rates reported in the literature.
3. Apply PCA and ML models, comparing different statistical models' BC classification and predictive accuracy of different machine learning models: Tree, Linear Discriminant, Quadratic Discriminant, Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Ensemble, Neural Network, and Kernel.

3. Materials and methods

This section explains a detailed methodology, including data source, data curation, training dataset and testing dataset preparation, supervised classification training, and PCA computation. There is no universally accepted criterion or standard rule of thumb to ascertain that a particular model is more effective than another model by looking at the dataset in machine learning. The number of samples that we need is also incalculable. Even though regression usually necessitates a greater number of samples than classification does. From the theorem of no free lunch, a theoretical discovery proposes that every optimization algorithm performs equally well on average across all conceivable objective functions; it indicates no effortless solution in machine learning. As a result, these general guidelines are often more inaccurate than accurate. This is the reason why we have included all possible models that are available in MATLAB. The process flow of this research is depicted in Figure 1 and provided in the subsections below.

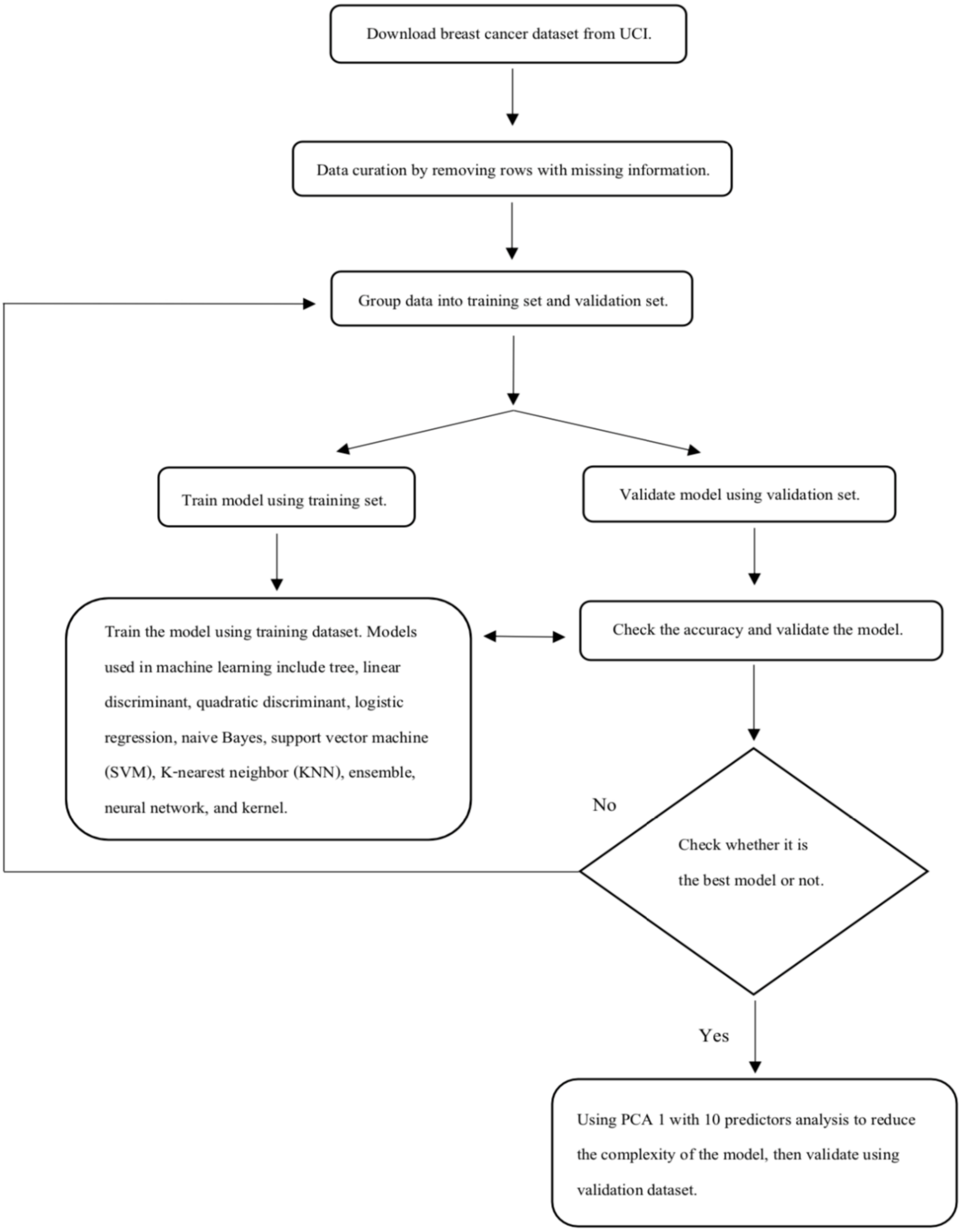
Download breast cancer dataset from UCI.
Data curation by removing rows with missing information.
Group data into training set and validation set.
Train model using training set.
Validate model using validation set.
Train the model using training dataset. Models used in machine learning include tree, linear discriminant, quadratic discriminant, logistic regression, naive Bayes, support vector machine (SVM), K-nearest neighbor (KNN), ensemble, neural network, and kernel.
Check the accuracy and validate the model.
No
Check whether it is the best model or not.
Yes
Using PCA 1 with 10 predictors analysis to reduce the complexity of the model, then validate using validation dataset.

Figure 1 Process flow of the ML analysis in this study.

3.1 Dataset

The BC dataset analyzed here was obtained from the open-source UCI Machine Learning Repository (Mangasarian & Wolberg, 1990) (assessed January 12, 2023). It contains data from 699 patients' breast biopsied tissue to identify cytological characteristics of being or malignant using the Fine-Needle Aspirates (FNAs) method (Frable, 1983; Mouriquand & Pasquier, 1980), including the following 10 features based on a 1 to

10 scale graded at the time when the samples were collected. Note that 1 is closest to benign, whereas 10 is most relative to the anaplastic case (Wolberg & Mangasarian, 1990). It is crucial to point out that the original paper from Wolberg & Mangasarain (1990) did not elaborate on how these values were determined and evaluated (Ohno-Machado & Bialek, 1998). The nine features with the corresponding diagnosis are shown in Table 1 below. There are only nine features since the first column in the dataset is the sample numbers, which are used instead of patients' names for anonymity. This feature has been excluded from this study.

Table 1 Predictors and labels obtained from the UCI breast cancer dataset.

Predictors	**Details**	**Values**	**Mean ($\bar{x}$)**	**Standard deviation (S)**
Clump Thickness	If the cell is grouped, forming monolayers, it is likely to be benign cells since cancerous cells tend to be grouped in multilayers.	1-10	4.42	2.82
Uniformity of Cell Size	Cancer cells tend to vary in size, which is used to estimate the consistency in the cell size.	1-10	3.13	3.05
Uniformity of Cell Shape	Cell uniformity is used to verify the shape of the types of cells (cancer and non-cancer) because cancer cells tend to vary in shape.	1-10	3.21	2.97
Marginal Adhesion	Cancer cells tend to lose the ability of cells sticking together, so this method can be used to identify carcinogenic and non-carcinogenic cells.	1-10	2.81	2.86
Single Epithelial Cell Size	Most cancer usually occurs inside epithelial tissues, which is a tissue covering the internal and external surfaces of our body.	1-10	3.22	2.21
Bare Nuclei	This feature is for nuclei in a cytologic preparation not in or surrounded by cytoplasm.	1-10	3.46	3.64
Bland Chromatin	It describes a uniform texture of the nucleus.	1-10	3.44	2.44
Normal Nucleoli	It is a term for the largest structure in the nucleus, where they synthesize and assemble the cell's ribosomes.	1-10	2.87	3.05
Mitoses	This type of cell division is where replicated or daughter cells have the same kind and number of chromosomes as their original or parent cells.	1-10	1.59	1.72
Label	**Details**	**Values**	**Number in each class**	
Class:	There are two class types: benign and malignant	2 = benign 4 = malignant	458 cases (65.52%) 241 cases (34.48%)	

Table 2 Classification Model Predictors and Labels

Variables	**Type**	**Variables**	**Type**
Clump Thickness	Predictor	Bare Nuclei	Predictor
Uniformity of Cell Size	Predictor	Bland Chromatin	Predictor
Uniformity of Cell Shape	Predictor	Normal Nucleoli	Predictor
Marginal Adhesion	Predictor	Mitoses	Predictor
Single Epithelial Cell Size	Predictor	Malignant/Benign	Label

3.2 Data curation

The biased 699 data rows were then reduced to 683 rows by removing 16 rows that contained missing information. The 683 rows were 444 for the B cases and 239 for the M cases; this dataset was biased toward the B case, which is inappropriate for ML training. The dataset size was further reduced to 478 rows containing 239 for B and 239 for M, respectively. This curation ensured that the dataset was unbiased by randomly removing 205 rows from the B cases. The 10 variables were first treated as predictors and labels for supervised ML training and classification tasks. The 10 predictor variables are summarized in Table 2.

3.3 Dataset for training and validation

The 478 rows were then divided into 2 datasets: the training dataset (430 rows) and the validation dataset (48 rows) at a ratio of 90% to 10%. The training and validation datasets were selected randomly by ensuring that both datasets contained the same number of B cases and M cases, in other words, unbiased training dataset and unbiased validation dataset.

3.4 Machine Learning Training

We then used a built-in Classification Learner application in MATLAB 2022b to train the supervised ML models listed in Table 3 using the training dataset. The numerous models were trained to compare their performance for the given dataset.

Classification performance was evaluated using a 5-fold cross-validation accuracy, precision, recall, and F_1 score, as shown in Equation (1)-(4).

$$\text{Accuracy}= \frac{T_p+T_n}{T_p+T_n+F_p+F_n} \quad (1)$$

$$\text{Precision}= \frac{T_p}{T_p+F_p} \quad (2)$$

$$\text{Recall}= \frac{T_p}{T_p+F_n} \quad (3)$$

$$F_1= 2\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \quad (4)$$

Where T_p and T_n are defined as true positive and true negative cases, respectively. F_p and F_n are defined as false positive and false negative cases, respectively.

The trained models were then validated using the separate validation dataset to determine whether the trained models were generalized and could provide a correct response for the unseen dataset. The same performance parameters, including the 5-fold cross-validation accuracy, precision, recall, and F_1 score for the validation dataset, were also calculated to validate the model performance against the training dataset.

3.5 Principal component analysis

A PCA model (Vidal et al., 2016) with a statistical confidence of 95% was then applied to identify essential predictors contributing to the model classification accuracy using the built-in PCA analysis tool in MATLAB 2022b. After identifying critical predictors, a less complex classification model was trained to show that with fewer predictors, the classification model can still perform similarly to the models trained using all the available predictors. We will also discuss in the result section that the PCA provides similar results to other feature section methods, including probability density functions of $\chi2$-test (Chi-square), ANOVA, and the p-value of the Kruskal-Wallis method (Ostertagova et al., 2014).

4. Results

4.1 ML Classification accuracy based on all 9 Predictors.

The ML models in Table 3 were first trained using all 9 predictors and malignant/benign labels, as listed in Table 1. The classification accuracy of each model is shown in Table 4. The top three ML models were the Medium Gaussian SVM (SVM), Coarse Gaussian SVM (SVM), and Cosine KNN (KNN), with the same classification accuracy of 96.5%.

Table 4 also shows that the top 3 highest F_1 performance models are the Medium Gaussian SVM model with the F1 score of 96.57%, followed by the Coarse Gaussian SVM model and the Cosine KNN model with the same F1 score of 96.52%. The confusion matrices and the receiver operating characteristic (ROC) plots for these three models are shown in Figure 2 and Figure 3, respectively.

Table 3 The ML models employed in this study.

Model	Details
Tree	Fine Tree
	Medium Tree
	Coarse Tree
Linear Discriminant	Linear Discriminant
Quadratic Discriminant	Quadratic Discriminant
Logistic Regression	Logistic Regression
Naïve Bayes	Gaussian Naïve Bayes
	Kernel Naïve Bayes
SVM	Linear SVM
	Quadratic SVM
	Cubic SVM
	Fine Gaussian SVM
	Medium Gaussian SVM
	Coarse Gaussian SVM
Kernel	SVM Kernel
	Logistic Regression Kernel
	Fine KNN
	Medium KNN
K-Nearest Neighbor (KNN)	Coarse KNN
	Cosine KNN
	Cubic KNN
	Weighted KNN
Ensemble	Boosted Trees
	Bagged Trees
	Subspace Discriminant
	Subspace KNN
	RUSBoosted Trees
Neural Network	Narrow Neural Network
	Medium Neural Network
	Wide Neural Network
	Bilayered Neural Network
	Trilayered Neural Network

Figures 2 and 3 show that although the Medium Gaussian SVM model has the highest F_1 score, it does have higher false positive cases. The other two models provide similar false positive performance and are better than the Medium Gaussian SVM model. The promising candidate model for malignant screening is the Medium Gaussian SVM model since having higher false positive cases is better than having higher false negative cases.

4.2 Validation of the trained model using the validation dataset

The separated validation dataset was then employed to predict the classification output compared to their known label. For the 48 validation cases, the trained Medium Gaussian SVM model can predict 23 true positive results, 22 true positive results, and only 1 false negative result and 2 false positive results, respectively, as shown in Figure 4.

Table 4 Classification Accuracy of ML Models Trained with all 9 Predictors.

Model	Details	Accuracy	Precision	Recall	F_1
Tree	Fine Tree	93.26%	92.66%	93.95%	93.30%
	Medium Tree	93.3%	92.66%	93.95%	93.30%
	Coarse Tree	92.8%	90.00%	96.28%	93.03%
Linear Discriminant	Linear Discriminant	95.1%	97.55%	92.56%	94.99%
Quadratic Discriminant	Quadratic Discriminant	94.9%	92.14%	98.14%	95.05%
Logistic Discriminant	Logistic Regression	95.8%	96.68%	94.88%	95.77%
Naïve Bayes	Gaussian Naïve Bayes	96.3%	94.22%	98.60%	96.36%
	Kernel Naïve Bayes	95.6%	96.23%	94.88%	95.55%
SVM	Linear SVM	96.0%	95.83%	96.28%	96.05%
	Quadratic SVM	96.0%	95.41%	96.74%	96.07%
	Cubic SVM	94.9%	94.88%	94.88%	94.88%
	Fine Gaussian SVM	95.1%	91.10%	100.00%	95.34%
	Medium Gaussian SVM	96.5%	95.05%	98.14%	96.57%*
	Coarse Gaussian SVM	96.5%	96.30%	96.74%	96.52%**
KNN	Fine KNN	94.0%	95.65%	92.09%	93.84%
	Medium KNN	96.0%	92.26%	95.81%	96.03%
	Coarse KNN	93.3%	97.45%	88.84%	92.95%
	Cosine KNN	96.5%	96.30%	96.74%	96.52%**
	Cubic KNN	95.6%	95.79%	95.35%	95.57%
	Weighted KNN	96.0%	96.26%	95.81%	96.03%
Ensemble	Boosted Trees	58.8%	93.18%	19.07%	31.66%
	Bagged Trees	94.9%	94.47%	95.35%	94.91%
	Subspace Discriminant	95.1%	97.09%	93.02%	95.01%
	Subspace KNN	95.1%	94.91%	95.35%	95.13%
	RUSBoosted Trees	95.1%	92.05%	37.67%	53.46%
Neural Network	Narrow Neural Network	94.0%	93.55%	94.42%	93.98%
	Medium Neural Network	93.5%	93.90%	93.02%	93.46%
	Wide Neural Network	93.5%	93.90%	93.02%	93.46%
	Bilayered Neural Network	93.5%	93.09%	93.95%	93.52%
	Trilayered Neural Network	94.7%	94.04%	95.35%	94.69%
Kernel	SVM Kernel	96.3%	94.62%	98.14%	96.35%
	Logistic Regression Kernel	96.3%	95.43%	97.21%	96.31%

These 4 values give the following performance parameters: the accuracy of 93.75%, the precision of 92.00%, the recall of 95.83%, and the F_1 score of 93.88%. The performance parameters for the validation cases were slightly less than the performance parameters of the training dataset, with discrepancies within 3%. For the ROC performance, the trained Medium Gaussian SVM model had a similar performance when validated using the validation dataset compared to the ROC plots of the training dataset, as shown in Figure 4.

4.3 Number of predictors reduction using PCA

In the previous sections, we have demonstrated that accurate classification models can be trained using all the available 9 predictors, which are (1) Clump Thickness, (2) Bare Nuclei, (3) Uniformity of Cell Size, (4) Bland Chromatin, (5) Uniformity of Cell Shape, (6) Normal Nucleoli, (7) Marginal Adhesion, (8) Mitoses, and (9) Single Epithelial Cell Size to predict whether the given cell image is benign or malignant. Here, we perform PCA analysis using the built-in Matlab PCA feature, and the coefficients are as shown in Table 5, calculated at 95% confidence.

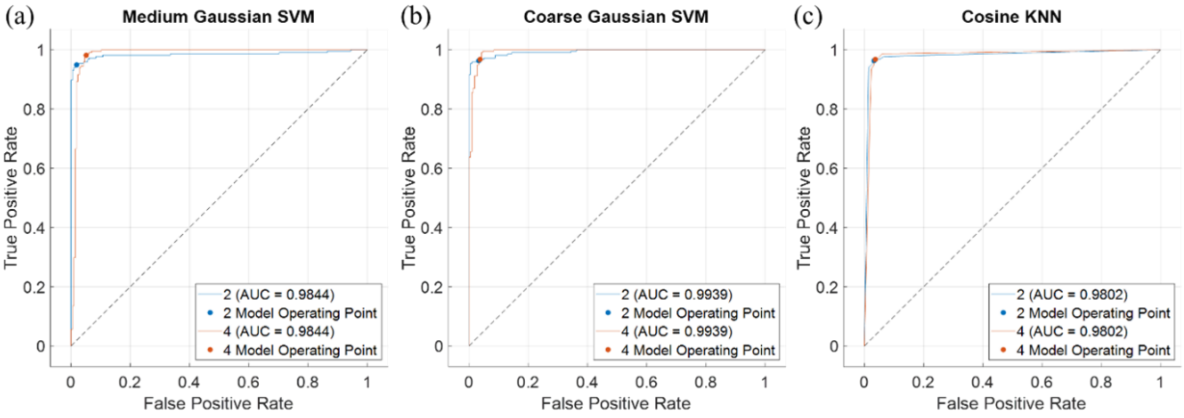

Figure 3 ROC plots of 3 best performance models (a) Medium Gaussian SVM, (b) Coarse Gaussian SVM, and (c) Cosine KNN

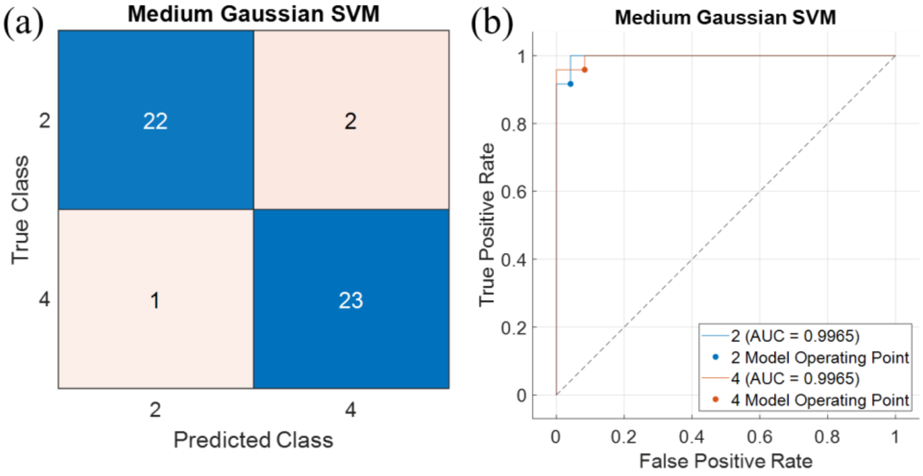

Figure 4 (a) Confusion matrix, and (b) ROC plots for the Medium Gaussian SVM model when validated using the 48 cases of the validation dataset.

Table 5 shows that the top 3 highest PCA coefficients are Normal Nucleoli, Bare Nuclei, and Uniformity of Cell Size. We also performed other statistical methods to identify the crucial factors from these 9 predictors by computing probability density functions of χ2- test, ANOVA, and the p-value of the Kruskal-Wallis method, as shown in Table 5. Although the results from the different approaches show slightly different predictor rankings, the top three highest scores remained the same for all the methods.

The importance of the 3 predictors was then justified by training the Medium Gaussian SVM models using a different number of predictors from 1 predictor to 4 predictors based on the PCA coefficients in Table 5. The confusion matrices for the models trained using a different number of predictors are shown in Figure 5, and the summarized performance parameters are shown in Table 6. The optimum number of predictors is three, and it can be seen in Table 5 that if the fourth predictor is added to the model, the model’s performance is degraded. It is well-established that having more information does not always result in a better model, for example, if the additional information contains more noise than helpful information for machine learning. Figure 6 shows ROC performance for 4 trained models using a different number of predictors. It can be seen that the ROC performance does not change or improve much after 2 predictors and provides a similar ROC response when trained using all 9 predictors.

4.4 Validation for the trained Medium Gaussian SVM models using 3 predictors.

The trained Medium Gaussian SVM models using 3 predictors were validated using the validation dataset. The confusion matrix and the ROC plots for the validation dataset are shown in Figure 7. From the confusion matrix, the performance parameters can be computed, and it found that the accuracy, precision, recall, and F_1 scores are 93.75%, 95.65%, 91.67%, and 93.62%, respectively. The performance of these parameters agrees with the model trained using all 9 predictors. Therefore, we are confident that the dimension-reduced model can be employed to classify breast cancer cells, reducing the resources and time for data collection and analysis.

Table 5 PCA coefficients, probability density functions of χ2- test, ANOVA, the p-value of Kruskal-Wallis test of the 9 predictors

Predictors	ANOVA	Predictors	Kruskal-Wallis	Predictors	χ2- test	Predictors	PCA coefficients
Normal Nucleoli	224.16	Bare Nuclei	160.57	Uniformity of Cell Size	160.59	Normal Nucleoli	0.43
Uniformity of Cell Size	219.11	Uniformity of Cell Size	158.51	Bare Nuclei	160.29	Bare Nuclei	0.49
Bare Nuclei	213.00	Normal Nucleoli	152.42	Normal Nucleoli	149.45	Uniformity of Cell Size	0.39
Marginal Adhesion	172.45	Marginal Adhesion	132.32	Marginal Adhesion	129.97	Bland Chromatin	0.38
Clump Thickness	147.31	Uniformity of Cell Shape	126.40	Uniformity of Cell Shape	128.85	Mitoses	0.37
Mitoses	136.70	Bland Chromatin	122.31	Mitoses	115.84	Marginal Adhesion	0.30
Bland Chromatin	136.44	Mitoses	120.51	Bland Chromatin	115.31	Clump Thickness	0.29
Uniformity of Cell Shape	112.78	Clump Thickness	109.14	Clump Thickness	98.53	Uniformity of Cell Shape	0.23
Single Epithelial Cell Size	36.03	Single Epithelial Cell Size	52.78	Single Epithelial Cell Size	48.56	Single Epithelial Cell Size	0.13

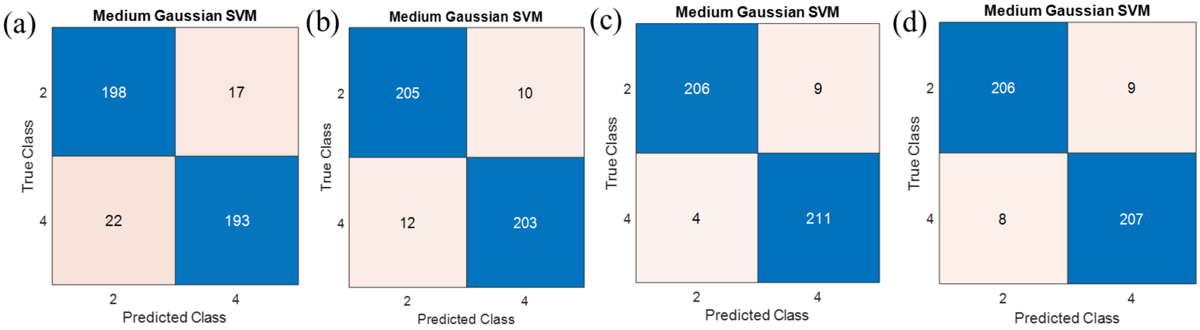

Figure 5 Confusion matrices for the models trained using a different number of predictors (a) 1 predictor: Normal Nucleoli, (b) 2 predictors: Normal Nucleoli & Bare Nuclei, (c) 3 predictors: Normal Nucleoli & Bare Nuclei & Uniformity of Cell Size, and (d) 4 predictors: Normal Nucleoli & Bare Nuclei & Uniformity of Cell Size & Bland Chromatin

Table 6 Accuracy, Precision, Recall, and F$_1$ scores of the Medium Gaussian SVM trained using a different number of predictors from 1 predictor to 4 predictors based on the PCA coefficient ranking in Table 5.

Predictors included in the model training	Accuracy	Precision	Recall	F$_1$
1 predictor: Normal Nucleoli	90.93%	91.90%	89.77%	90.82%
2 predictors: Normal Nucleoli & Bare Nuclei	94.88%	95.31%	94.42%	94.86%
3 predictors: Normal Nucleoli & Bare Nuclei & Uniformity of Cell Size	96.98%	95.91%	98.14%	97.01%
4 predictors: Normal Nucleoli & Bare Nuclei & Uniformity of Cell Size & Bland Chromatin	96.05%	95.83%	96.28%	96.05%

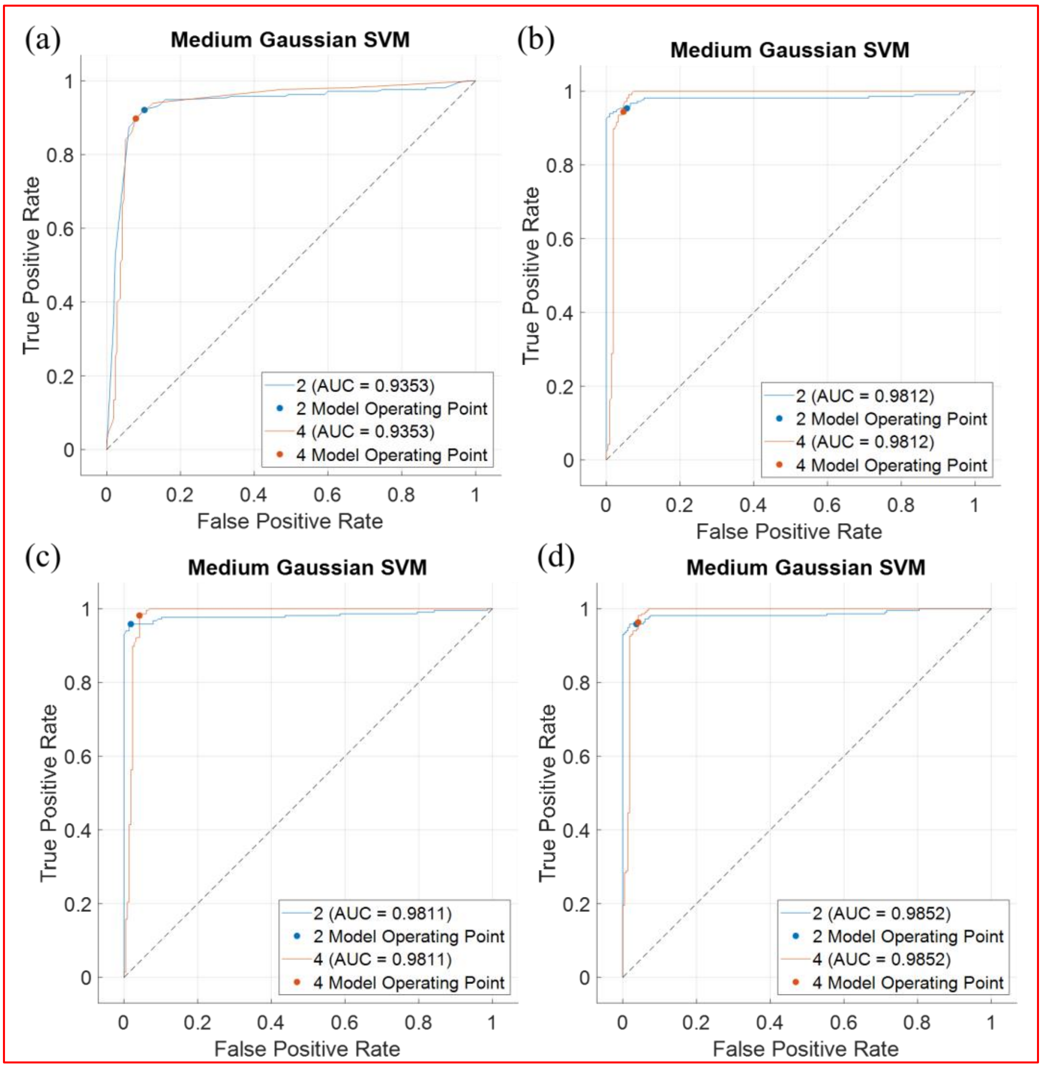

Figure 6 ROC plots for the models trained using a different number of predictors (a) 1 predictor: Normal Nucleoli, (b) 2 predictors: Normal Nucleoli & Bare Nuclei, (c) 3 predictors: Normal Nucleoli & Bare Nuclei & Uniformity of Cell Size, and (d) 4 predictors: Normal Nucleoli & Bare Nuclei & Uniformity of Cell Size & Bland Chromatin

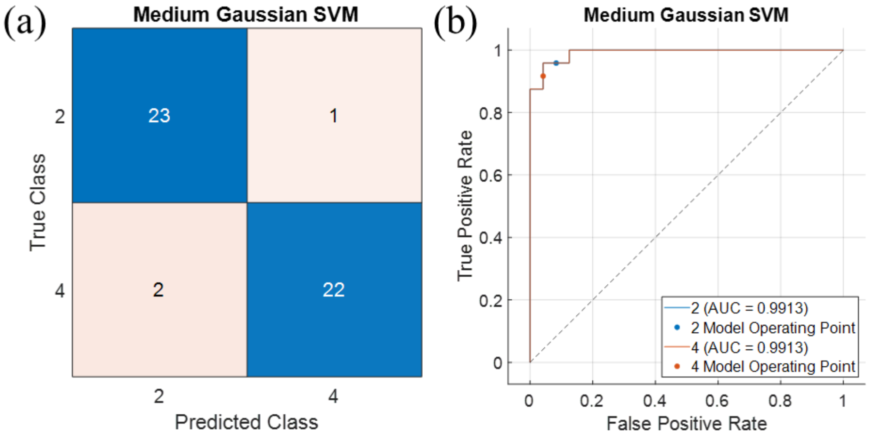

Figure 7 (a) Confusion matrix, and (b) ROC plots for the Medium Gaussian SVM model trained using 3 predictors when validated using the 48 cases of the validation dataset.

5. Discussion

Nowadays, as people around the globe are suffering an epidemic of breast cancer, techniques, and technologies to detect this painful disease still require complicated and time-consuming procedures such as biopsy images of suspected clumped cells (Versaggi & De Leucio, 2020). Based on an open-source dataset of 10 variables, we identified three predictor variables contributing most to diagnosing breast cancer. Alternatively, our developed and size-reduced AI can achieve a high classification accuracy of 96.97%, and it only requires 3 extracted features from FNA biopsied breast epithelial cell image. This finding allows physicians to focus more on these three predictors and reduce the healthcare staff's workload, computational burden, and resources. Furthermore, pathologists are rare in the medical field. There are 3 surgical pathologists in Thailand and approximately 102,000 worldwide (Colangelo et al., 2022). Moreover, the ratio of the number of pathologists to physicians in the world is 1 to 125. These statistics clearly show a shortage of pathologists, leading to misdiagnoses of diseases, especially cancer.

We have utilized patient information from Fine-Needle Aspiration (FNA) pictures to use them for different purposes separately; to train and validate the models. Those datasets for training were put and calculated in MATLAB for classification accuracy, precision, recall, and F1, while the dataset for validation was calculated in the confusion matrix. The tested results from the abovementioned dataset were created, collected, and given as an open-source database that excluded patients' IDs.

Previous studies were conducted on using machine learning models to diagnose breast cancer, such as the research conducted in 2010 by Osareh & Shadgar (2010). The results achieved from this study were equal to 98.80% and 96.33% by applying support vector machines and classifier models. Other studies were also implemented using machine learning to predict the survival rates of BC patients. The research concludes that the Trees Random Forest model had the most satisfying performance, compared to other techniques due to the accuracy, sensitivity, and the area under the ROC curve of this approach being 96%, 96%, and 93%, respectively (Montazeri et al., 2016).

We are fully aware that there are also some other model reduction methods and optimization tools, such as Genetic Algorithm (GA), Patricle-Swarm-Optimization (PSO), and Fast Forward Quantum Optimization Algorithm (FFQOA). However, this study's main objective is to reduce the model complexity; when the PCA models were calculated, it took approximately 2 minutes. Therefore, employing more sophisticated algorithms (Kaur et al., 2022) to perform the task was not essential in our study.

In this research, our current limitation is that the database's 1-10 scoring system was not clearly defined from the database source. For future studies, our research team strongly advises referring back to the standard scoring system for each predictor, such as using the TNM system (Cserni et al., 2018), which groups tumors' clump thickness in millimeters or centimeters. Regarding standard scoring systems, we plan to reverse engineer the original dataset to traditional medical measurement frameworks to convert the 1-10 scale to appropriate methods. We would also like to implement an automated system to interpret FNA biopsied images of suspected breast tumor cells and apply this new approach for high-accuracy breast cancer classification and screening.

6. Conclusion

We identified three crucial factors in classifying breast tumors, leading to breast cancer diagnosis from an open-source dataset in the UCI Machine Learning Repository containing 10 predictor variables in 699 patients. Firstly, we curated the data to create an unbiased training environment. Next, we trained ML models and found that Medium Gaussian SVM is the most suitable. We then applied PCA and supervised ML models to diagnose breast cancer in MATLAB 2022b, comparing them by their cross-validation accuracy. PCA models were developed using 1-4 variables, excluding the first predictor and the label. The model with 3 predictors is adequate, with an accuracy of 96.98%. We found the critical predictive variables to be normal nuclei, bare nuclei, and uniformity of cell size, which physicians should pay special attention to when diagnosing breast cancer.

7. Acknowledgements

We acknowledge the dataset used in this study, obtained from the breast cancer databases from the University of Wisconsin Hospitals, Madison, by Dr. William H. Wolberg, downloaded from the UCI machine learning open-source database on January 12, 2023.

8. References

Colangelo, T., Carbone, A., Mazzarelli, F., Cuttano, R., Dama, E., Nittoli, T., . . . Palumbo, O. (2022). Loss of circadian gene Timeless induces EMT and tumor progression in colorectal cancer via Zeb1-dependent mechanism. *Cell Death & Differentiation*, *29*(8), 1552-1568. https://doi.org/10.1038/s41418-022-00935-y

Cornell, L., Sahni, S., Couch, F., & Clune, C. (2022). Clinical Implications and Utility of Polygenic Risk Scores in Women at Elevated Risk for Breast Cancer. *Journal of Precision Medicine*, *8*(3), 408-413. https://doi.org/10.7326/M20-5874

Cserni, G., Chmielik, E., Cserni, B., & Tot, T. (2018). The new TNM-based staging of breast cancer. *Virchows Archiv*, *472*, 697-703. https://doi.org/10.1007/s00428-018-2301-9

Dange, V., Shid, S., Magdum, C., & Mohite, S. (2017). A review on breast cancer: An overview. *Asian Journal of Pharmaceutical Research*, *7*(1), 49-51. https://doi.org/10.5958/2231-5691.2017.00008.9

Das, A. K., Biswas, S. K., Bhattacharya, A., & Alam, E. (2021, March 19-20). *Introduction to Breast Cancer and Awareness* [Conference presentation]. 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India. https://doi.org/10.1109/ICACCS51430.2021.9441686

Dileep, G., & Gyani, S. G. G. (2022). Artificial intelligence in breast cancer screening and diagnosis. *Cureus*, *14*(10) Article e30318. https://doi.org/10.7759/cureus.30318

Dumalaon-Canaria, J. A., Hutchinson, A. D., Prichard, I., & Wilson, C. (2014). What causes breast cancer? A systematic review of causal attributions among breast cancer survivors and how these compare to expert-endorsed risk factors. *Cancer Causes & Control*, *25*, 771-785. https://doi.org/10.1007/s10552-014-0377-3

Frable, W. J. (1983). Fine-needle aspiration biopsy: a review. *Human pathology*, *14*(1), 9-28. https://doi.org/10.1007/s10552-014-0377-3

Giaquinto, A. N., Sung, H., Miller, K. D., Kramer, J. L., Newman, L. A., Minihan, A., . . . Siegel, R. L. (2022). Breast cancer statistics, 2022. *CA: a cancer journal for clinicians*, *72*(6), 524-541. https://doi.org/10.3322/caac.21754

Gill, S. S., Xu, M., Ottaviani, C., Patros, P., Bahsoon, R., Shaghaghi, A., . . . Abraham, A. (2022). AI for next generation computing: Emerging trends and future directions. *Internet of Things*, *19*, Article 100514. https://doi.org/10.1016/j.iot.2022.100514

Gupta, R., Kurc, T., & Saltz, J. H. (2022). Introduction to Digital Pathology from Historical Perspectives to Emerging Pathomics. *Whole Slide Imaging: Current*

Applications and Future Directions, 1-22. https://doi.org/10.1007/978-3-030-83332-9_1

Iqbal, M. S., Ahmad, W., Alizadehsani, R., Hussain, S., & Rehman, R. (2022). Breast Cancer Dataset, Classification and Detection Using Deep Learning. *Healthcare, 10*(12), Article 2395. https://doi.org/10.3390/healthcare10122395

Kaur, K., Sagar, A. K., & Chakraborty, S. (2022). Accelerating the performance of sequence alignment using machine learning with RAPIDS enabled GPU. *Journal of Current Science and Technology*, *12*(3), 462-481.

Lever, J., Krzywinski, M., & Altman, N. (2017). Points of significance: Principal component analysis. *Nature methods*, *14*(7), 641-643. https://doi.org/10.1038/nmeth.4346

Liu, Y., Nguyen, N., & Colditz, G. A. (2015). Links between alcohol consumption and breast cancer: a look at the evidence. *Women's health*, *11*(1), 65-77. https://doi.org/10.2217/WHE.14.62

Mangasarian, O. L., & Wolberg, W. H. (1990). *Cancer diagnosis via linear programming*. University of Wisconsin-Madison Department of Computer Sciences. Retrieved from https://minds.wisconsin.edu/bitstream/handle/1793/59346/TR958.pdf

McGuire, A., Brown, J. A., & Kerin, M. J. (2015). Metastatic breast cancer: the potential of miRNA for diagnosis and treatment monitoring. *Cancer and metastasis reviews, 34*, 145-155. https://doi.org/10.1007/s10555-015-9551-7

Montazeri, M., Montazeri, M., Montazeri, M., & Beigzadeh, A. (2016). Machine learning models in breast cancer survival prediction. *Technology and Health Care*, *24*(1), 31-42. https://doi.org/10.3233/THC-151071

Mouriquand, J., & Pasquier, D. (1980). Fine needle aspiration of breast carcinoma: a preliminary cytoprognostic study. *Acta cytologica*, *24*(2), 153-159. https://pubmed.ncbi.nlm.nih.gov/6245554/

Nassif, A. B., Talib, M. A., Nasir, Q., Afadar, Y., & Elgendy, O. (2022). Breast cancer detection using artificial intelligence techniques: A systematic literature review. *Artificial Intelligence in Medicine*, *127*, Article 102276. https://doi.org/10.1016/j.artmed.2022.102276

Ohno-Machado, L., & Bialek, D. (1998). Diagnosing breast cancer from FNAs: variable relevance in neural network and logistic regression models. *MEDINFO'98* (pp. 537-540). IOS Press Ebooks. https://doi.org/10.3233/978-1-60750-896-0-537

Oliveri, S., Faccio, F., Pizzoli, S., Monzani, D., Redaelli, C., Indino, M., & Pravettoni, G. (2019). A pilot study on aesthetic treatments performed by qualified aesthetic practitioners: efficacy on health-related quality of life in breast cancer patients. *Quality of Life Research*, *28*, 1543-1553. https://doi.org/10.1007/s11136-019-02133-9

Osareh, A., & Shadgar, B. (2010, April 20-22). *Machine learning techniques to diagnose breast cancer* [Conference presentation]. 2010 5th international symposium on health informatics and bioinformatics. https://doi.org/10.1109/HIBIT.2010.5478895

Ostertagova, E., Ostertag, O., & Kováč, J. (2014). *Methodology and application of the Kruskal-Wallis test.* Paper presented at the Applied mechanics and materials, Ankara, Turkey. https://doi.org/10.4028/www.scientific.net/AMM.611.115

Panyamit, T., Sukvivatn, P., Chanma, P., Kim, Y., Premratanachai, P., & Pechprasarn, S. (2022). Identification of factors in the survival rate of heart failure patients using machine learning models and principal component analysis. *Journal of Current Science and Technology*, *12*(2), 336-348. https://ph04.tci-thaijo.org/index.php/JCST/article/view/299

Perrin, S., & Roncalli, T. (2020). Machine learning optimization algorithms & portfolio allocation. *Machine Learning for Asset Management: New Developments and Financial Applications*, 261-328. https://doi.org/10.1002/9781119751182.ch8

Richardson, L. C., King, J. B., Thomas, C. C., Richards, T. B., Dowling, N. F., & King, S. C. (2022). Peer Reviewed: Adults Who Have Never Been Screened for Colorectal Cancer, Behavioral Risk Factor Surveillance System, 2012 and 2020. *Preventing Chronic Disease, 19*, Article E21. https://doi.org/10.5888/pcd19.220001

Saelee, P., Pongtheerat, T., Sophonnithiprasert, T., & Jinda, W. (2022). Clinicopathological significance of FANCA mRNA expression in Thai patients with breast cancer. *Journal of*

Current Science and Technology, *12*(3), 408-416. https://ph04.tci-thaijo.org/index.php/JCST/article/view/254

Shah, S. M., Khan, R. A., Arif, S., & Sajid, U. (2022). Artificial intelligence for breast cancer analysis: Trends & directions. *Computers in Biology and Medicine*, *142*, Article 105221. https://doi.org/10.1016/j.compbiomed.2022.105221

Sheikh, A., Md, S., & Kesharwani, P. (2022). Aptamer grafted nanoparticle as targeted therapeutic tool for the treatment of breast cancer. *Biomedicine & Pharmacotherapy*, *146*, Article 112530. https://doi.org/10.1016/j.biopha.2021.112530

Shimoi, T., Nagai, S. E., Yoshinami, T., Takahashi, M., Arioka, H., Ishihara, M., . . . Sagara, Y. (2020). The Japanese breast cancer society clinical practice guidelines for systemic treatment of breast cancer, 2018 edition. *Breast Cancer*, *27*, 322-331. https://doi.org/10.1007/s12282-020-01085-0

Siegel, R. L., Miller, K. D., Wagle, N. S., & Jemal, A. (2023). Cancer statistics, 2023. *CA: a cancer journal for clinicians*, *73*(1), 17-48. https://doi.org/10.3322/caac.21763

Swathi, T., Krishna, S., & Ramesh, M. V. (2019, March 21-23). *A survey on breast cancer diagnosis methods and modalities* [Conference presentation]. 2019 international conference on wireless communications signal processing and networking (WiSPNET), Chennai, India. https://doi.org/10.1109/WiSPNET45539.2019.9032799

Troxel, D. B. (2006). Medicolegal aspects of error in pathology. *Archives of Pathology & Laboratory Medicine*, *130*(5), 617-619. https://doi.org/10.5858/2006-130-617-MAOEIP

Versaggi, S. L., & De Leucio, A. (2020). *Breast Biopsy*. Europe PMC. Retrieved from https://europepmc.org/article/nbk/nbk559147

Vidal, R., Ma, Y., Sastry, S. S., Vidal, R., Ma, Y., & Sastry, S. S. (2016). *Principal component analysis (pp. 25-62).* Springer New York. https://doi.org/10.1007/978-0-387-87811-9_2

Wolberg, W. H., & Mangasarian, O. L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *AppliedMathematics:Wolberg and Mangasarian*, *87*(23), 9193-9196. https://doi.org/10.1073/pnas.87.23.9193

Wolberg, W. H., Mangasarian, O. L., & Setiono, R. (1989). *Pattern Recognition Via Linear Programming: Theory and Application to Medical Diagnosis*. University of Wisconsin-Madison Department of Computer Sciences. Retrieved from http://digital.library.wisc.edu/1793/59186

Yau, C., Osdoit, M., van der Noordaa, M., Shad, S., Wei, J., de Croze, D., . . . Sonke, G. S. (2022). Residual cancer burden after neoadjuvant chemotherapy and long-term survival outcomes in breast cancer: a multicentre pooled analysis of 5161 patients. *The Lancet Oncology*, *23*(1), 149-160. https://doi.org/10.1016/S1470-2045(21)00589-1

Yedjou, C. G., Sims, J. N., Miele, L., Noubissi, F., Lowe, L., Fonseca, D. D., . . . Tchounwou, P. B. (2019). Health and racial disparity in breast cancer. *Breast cancer metastasis and drug resistance: Challenges and progress*, 31-49. https://doi.org/10.1007/978-3-030-20301-6_3