Sana'a University Journal of Applied Sciences and Technology ISSN : 2958-9568
Vol. 1| No. 2 | Page 147 – 167 | 2023 | https://journals.su.edu.ye/index.php/jast

Breast Cancer-Risk Factors and Prediction Using Machine-Learning Algorithms and Data Source: A Review of Literature

Ayman Alsabry[1,*], Malek Algabri[1], Amin Mohamed Ahsan[2]
[1] Department of Computer Science, Faculty of Computer and Information Technology, Sana'a University, Sana'a, Yemen.
[2] Department of Computer Science, Faculty of Computer Science and Information Technology, International University of Technology Twintech, Sana'a, Yemen.
***Corresponding author**: *aymanalsabri@su.edu.ye*

Article Info.
Article history:
Received: April 11, 2023
Accepted: April 29, 2023
Published: April 30 2023

Keywords
1. Breast Cancer Risk Factors
2. Breast Cancer Data Source
3. Machine learning (ML)
4. Support Vector Machines (SVM)
5. Artificial Neural Networks (ANN)
6. Decision Trees (DT)
7. Random Forests (RF)
8. Clustering
9. Dimensionality Reduction

ABSTRACT: Breast cancer (BC) is a major health concern worldwide. It is a complex and multifactorial disease, and identifying its risk factors is crucial for early detection and effective treatment. This review article provides an overview of the literature on breast cancer risk factors, data sources, and machine learning algorithms for prediction. The paper discusses the various risk factors associated with breast cancer, including age, family history, lifestyle choices, and environmental factors. Additionally, the article explores the different data sources used in breast cancer research, including clinical data, genomic data, and lifestyle data. The paper then reviews the different machine-learning algorithms used for breast cancer prediction, including supervised and unsupervised learning. The performance of ML algorithms in predicting BC risk using different data sources was assessed. This review provides valuable insights into the current state of research on BC risk factors and prediction using ML algorithms and data sources, and the findings will be useful for healthcare professionals and researchers working in the field of BC.

Contents

1. Introduction

Breast tumors are unnatural growths of breast tissue that can be non-cancerous or cancerous. In 2020, 2.26 million new instances of BC were discovered, making up about 11.7% of all cancer cases in women worldwide [1]. An accurate and timely diagnosis of breast tumors is critical for effective treatment and improved patient outcomes.

Research has been conducted on various aspects of breast tumors, including risk factors, molecular subtypes, and treatment options. One area of active research is the development of new diagnostic methods, including those based on machine learning.

Traditionally, breast tumors have been diagnosed through physical examination, mammography, ultrasound, and biopsy. Mammography is a popular screening method, although it has drawbacks, especially for women with thick breast tissue. [2]. Ultrasound can be used to detect and evaluate breast tumors, but it is operator-dependent and can miss small tumors [3]. A biopsy involves the removal of a tissue sample for examination under a microscope and is the gold standard for BC detection [4].

ML-based methods have shown promise in improving the performance of BC diagnosis. By detecting patterns and correlations that may not be apparent to human observers, ML can help identify early-stage tumors and distinguish between benign and malignant tumors. One example of an ML-based method for breast tumor diagnosis is computer-aided detection (CAD). CAD systems use algorithms to analyze mammograms and identify areas that may be indicative of BC. According to research, CAD can increase mammography's sensitivity and specificity, especially in women with denser breast tissue [5].

The aims of this literature review are:

- To identify BC risk factors.
- To review the current state of research on BC prediction using ML algorithms.
- To evaluate ML's performance in predicting BC risk.
- To identify the data sources used in BC prediction using ML algorithms.
- To evaluate ML's performance with different data sources.

The following describes the structure of this paper: Section 2 identifies BC risk factors; Section 3 reviews the current state of research on BC prediction using ML models and evaluates its performance; Section 4 identifies the data sources used in BC prediction using ML models; and Section 5 concludes the paper.

2. BC Risk Factors

Uncertainty exists over the precise cause of cancer's beginning. Nonetheless, there are a number of well-known factors that raise the risk of developing BC. Gender and age are among the important factors, but other significant contributors to BC include hormonal factors, such as exposure to estrogens, as well as reproductive factors like the number of children a woman has, her age at the birth of her first child, and whether or not she has breastfed. The usage of hormone therapy and genetic genes, as well as unhealthy diets leading to obesity, can affect the development of BC. Moreover, there are additional important risk factors for BC that include the use of hormonal contraception, alcohol intake, and ionizing radiation exposure in adolescence [6-10].

2.1. Gender:

BC is one of the most frequent cancers that affect women worldwide. While BC can also occur in men, it is far less prevalent in men than in women [11]. The incidence of BC is strongly associated with sex and gender, with females being at a significantly higher risk than males [11].

The reason for this difference is due to the presence of estrogen and progesterone hormones in the female body. These hormones are known to stimulate the growth of breast cells, which can increase the risk of BC development [12]. Furthermore, women with a family history or genetic mutations such as those in the BRCA1 and BRCA2 genes have a higher risk of developing the disease [11].

Many racial and ethnic groups have varied incidence rates of BC, with African American women having a greater incidence and death rate than white women [11]. This disparity may be due to differences in socioeconomic status, access to healthcare, and lifestyle factors such as diet and physical activity.

2.2. Age:

Age-related increases in BC risk are observed, and they are more prevalent in women over 50. [11]. In fact, women aged 55 or older are diagnosed with invasive BC in two-thirds of cases [12]. It is believed that the accumulation of genetic mutations over time is what has caused this elevated risk.

As women age, the amount of estrogen and progesterone in their bodies decreases, which can cause cancerous changes in breast tissue [12]. Additionally, older women may have other health conditions or take medications that can affect their BC risks, such as hormone replacement therapy or certain types of birth control [11]. Although age poses a considerable risk for BC, it is not the sole risk factor. Women of all ages can develop breast cancer, and some women may have a higher risk due to factors such as family history or genetic mutations [7].

2.3. Hormonal Status:

Hormonal status is a significant factor in BC risk. BC is often associated with hormones, particularly estrogen and progesterone. Due to their ability to promote breast cell proliferation, these hormones can raise the chance of developing BC when they are present in high concentrations for an extended period of time [12].

Women who have a history of certain hormonal conditions, such as the early onset of menstruation, the later onset of menopause, or having their first child at a later age, may have a higher risk of developing BC [11]. Additionally, women who take hormone replacement therapy (HRT) for menopausal symptoms may be more likely to develop BC [12].

Brinton et al. demonstrated that the initial menstruation occurring at the age of 15 or later correlated with a 23% decrease in BC risk in comparison to early menarche, which happens before the age of 12 [13].

Furthermore, it was shown that, compared to delayed menopause, early menstruation offered a larger risk for BC development. The relative risk increased by 2.9% for each year that menopause came later, and those who experienced menopause after the age of 54 were at double the risk as those who did so before the age of 45 [9, 14].

The meta-analysis also revealed that premenopausal women had a greater chance of developing BC than postmenopausal women of the same age. Within this group, Body Mass Index's impact on disease risk was observed—obesity reduced the risk in premenopausal patients while increasing it in postmenopausal patients. The study also found that early menarche correlated with an increased likelihood of lobular BC and late menopause [9]. Moreover, those who experienced late menopause were more likely to have breast cancer with steroid expression [9].

Additional reproductive factors influencing BC risk, as supported by numerous studies, include breastfeeding, the age at which a woman gave birth to her first child, and the total number of children born [13].

According to research, transgender women experience BC at a higher rate than cisgender men, whereas transgender men experience BC at a lower rate than cisgender women [15]. Hormone therapy administered to transgender women over a brief period increases the risk of BC, which presents in a more feminine pattern [15].

2.4. Hormonal and Reproductive Factors Influencing BC:

Estrogens are crucial to the development of BC, as they are involved in its pathogenesis [10]. BC is known to be a hormone-dependent tumor, with increased estrogen levels and prolonged exposure being linked to an elevated risk of developing the disease [10].

Epidemiological studies have corroborated that heightened exposure to endogenous and

exogenous estrogens raises the likelihood of BC occurrence [16]. In postmenopausal women, elevated serum estrogen levels have been connected to an increased BC risk. Both hormonal and reproductive factors undeniably contribute to the growth of this risk [16]. Exposure duration to estrogen and pregnancy impact, determined by variables such as first menstruation age, first pregnancy age (particularly for mothers who had their first kid after turning 30), nulliparity, or menopause onset age, shape an individual's BC risk level [17].

Menstruation that starts early and ends late doubles the risk compared to menstruation that starts later and ends earlier [18].

Additionally, childlessness and having the first pregnancy later in life (over 30 years old) are factors linked to extended estrogen exposure [19].

Sterile women and those who get their first pregnancy after turning 30 are at a 2–5 times greater risk of getting BC. As opposed to complete pregnancies, which have the protective impact of progesterone throughout the second stage of pregnancy, spontaneous and artificial losses may increase the risk [20]. The impact of exogenous estrogen on BC remains a contentious issue and the subject of numerous studies.

The relationship between hormonal contraceptive use and the risk of BC has been proved by the studies of Beral et al. [21] and Mørch et al. [22]. Both studies came to the same conclusion: prolonged hormonal contraceptive use increases the relative risk of BC. Furthermore, the risk increased as subjects continued taking hormonal contraceptives for longer durations [21, 22].

2.5. Genetic Factors and Family Incidence:

A mere 5–10% of BC cases are related to genetics. Notably, mutations in the BRCA1 and BRCA2 genes are the most recognized genetic links to this type of cancer [23].

On chromosome 17, the BRCA1 gene functions as a suppression gene that codes for a nuclear protein essential for maintaining genome integrity. This protein is a component of a complex that binds to RNA polymerase II, engages histone deacetylase, and affects activities like transcription, DNA repair, or recombination. This complex also contains other suppression genes and DNA damage-detecting products.

Homologous recombination is a unique mechanism used by BRCA2 genes, another suppression gene on chromosome 13, to help repair double-stranded damage to DNA [24].

Just 3–5% of people with BC have these gene mutations. However, because the BRCA1 and BRCA2 genes are so prevalent, these people need to be part of a proactive program. It is estimated that individuals carrying the BRCA1/BRCA2 mutation face a tenfold greater risk of developing BC [21]. Mutations in these genes correspond with a cumulative 60%+ risk of BC by age 70, with lifetime probabilities ranging from 41% to 90%. Additionally, BRCA1 gene mutations are linked to triple-negative cancer while BRCA2 gene mutations correlate with estrogen receptor-expressed BC [25-27].

High-penetration mutations in other suppressor genes, such as TP53 (Li-Fraumeni syndrome) and PTEN (Cowden syndrome), also predispose individuals to BC. Women with Li-Fraumeni syndrome have a 54% cumulative risk of developing BC by age 70. For those with Cowden's syndrome, the lifetime risk varies between 25% and 50%. However, both genetic conditions are quite uncommon [28, 29].

BC shows a moderate predisposition in individuals with mutations in the ATM, BRIP1, CHEK2, and PALB2 genes. Those carrying these mutations have a risk of developing malignant tumors that is 2-3 times greater than the general population [29].

It is estimated that less than 10% of breast cancers have genetic causes [30]. Conversely, more than 90% of breast cancers result from sporadic somatic mutations. The risk of developing BC doubles for women whose immediate family member (mother or sister) has been treated for this malignancy and increases three to six times if two of their closest relatives have undergone treatment. This risk decreases when the affected family members are older at the time of their cancer diagnosis [14].

2.6. Mild Breast Changes:

Benign changes in the breast tissue refer to non-cancerous alterations or abnormalities that can occur in the cells or structures of the breast. Some common types of benign breast conditions include fibrocystic changes, cysts, papillomata, and fibroadenomas. Although benign breast changes are non-cancerous and do not require treatment in most cases, they can increase the risk of developing BC. According to various studies, the chance of developing BC is considerably increased by the presence of benign breast disorders such as atypical ductal hyperplasia (ADH) or atypical lobular hyperplasia (ALH) [27, 31, 32]. Women with these types of benign breast conditions may require closer monitoring and more frequent BC screening.

According to a study by Hartmann et al. [27], women with ADH or ALH had up to five times greater chance of developing BC than those without any benign breast abnormalities.

Dupont et al. found similar findings in their meta-analysis, which discovered that women with proliferative breast disease who did not have ADH or ALH had a two-fold greater risk of BC [32].

Furthermore, Hartmann et al. discovered that women with a family history of BC and benign breast disorders like ADH or ALH are considerably more at risk of getting BC [31].

2.7. Ionizing Radiation:

Ionizing radiation is a recognized risk factor for BC. Ionizing radiation is a kind of energy that can penetrate the body and damage cells, including breast cells, which can lead to the development of cancer.

Boice Jr. and Monson [33] found that women who received repeated fluoroscopic examinations of the chest, such as for tuberculosis screening, had an increased risk of developing breast cancer.

Ronckers et al. [34] found that women who received multiple diagnostic X-rays for spine deformities during adolescence had an increased risk of developing breast cancer.

Gonzalez et al. [35] found, in their systematic review paper, that BC is one of the most common second cancers that develop after radiation therapy for other cancers, such as Hodgkin lymphoma.

Beir [36] reviewed the scientific evidence on the health risks of exposure to low levels of ionizing radiation and concluded that even low doses of ionizing radiation can increase the risk of cancer, including breast cancer.

Preston et al. [37] found, in their pooled analysis of eight cohort studies, that exposure to ionizing radiation increases the risk of developing breast cancer, with the risk increasing with increasing radiation dose.

Ronckers et al. [38] found that radiation exposure to the thyroid during childhood cancer treatment is associated with an increased risk of developing BC later in life.

Sigurdson et al. [39] found, in their study of radiologic technologists, that those who were exposed to ionizing radiation had an increased risk of developing breast cancer.

Cardis et al. [40] found, in their study of nuclear industry workers, that exposure to ionizing radiation increases the risk of BC development, with the risk increasing with increasing radiation dose. Shore et al. [41] found

that women who received X-ray therapy for acute postpartum mastitis had an increased risk of developing breast cancer.

Here are some ways to reduce the risk of developing BC from ionizing radiation exposure with references:

- ***Limit exposure:*** Avoid unnecessary exposure to ionizing radiation, such as medical imaging tests that use ionizing radiation, unless it is medically necessary [42].
- ***Shielding:*** When undergoing medical imaging that uses ionizing radiation, ask for a shield or apron to cover other parts of the body and limit radiation exposure [43].
- ***Alternative imaging:*** Ask your doctor if there are non-ionizing radiation imaging tests that can be used instead of those that use ionizing radiation [44].
- ***Follow radiation safety guidelines***: If working in an occupation that involves exposure to ionizing radiation, make sure to follow appropriate radiation safety guidelines and wear protective equipment [45].
- ***Breastfeeding:*** Studies have shown that BC risk can be decreased by breastfeeding, including in women who have been exposed to ionizing radiation [46].
- ***Healthy lifestyle:*** Overall BC risk can be decreased by leading a healthy lifestyle that includes frequent exercise, a balanced diet, and abstaining from smoking and excessive alcohol usage, including for those who may have been exposed to ionizing radiation [47].

2.8. Alcohol Consumption:

Alcohol use has been found to significantly increase the chance of developing breast cancer. The data consistently demonstrates that greater levels of alcohol intake are linked to an increased risk of breast cancer, according to several studies that have looked into the relationship between alcohol use and BC risk. The following is an overview of the connection between drinking alcohol and the risk of developing BC, with references to some of the important studies in this field:

Hamajima et al. [48] discovered that alcohol consumption was linked to an increased risk of breast cancer, with the risk increasing with increasing alcohol intake. The study involved a collaborative reanalysis of individual data from 53 epidemiological studies, including over 58,000 women with BC and nearly 95,000 women without the disease.

Similar results were obtained by Smith-Warner et al. [49], who discovered that women who drank more than one alcoholic beverage daily had a greater chance of acquiring BC than non-drinkers did. This connection was also noted in a study by Chen et al. [50], which discovered that even moderate alcohol use in adulthood was linked to an elevated risk of BC and that the risk rose with increasing alcohol use. Other populations have also shown indications of the relationship between drinking alcohol and the risk of BC. As an illustration, Romieu et al. [51] discovered that alcohol use was linked to a higher incidence of BC.

Alcohol's impact on the body's estrogen levels is assumed to be the cause of the association between drinking and the risk of BC. Estrogen is a hormone that is known to encourage the proliferation of BC cells, and it has been demonstrated that drinking alcohol raises estrogen levels [52]. Alcohol use being a risk factor for both estrogen receptor-positive and estrogen receptor-negative BC may be explained by this effect of alcohol on estrogen levels [53].

2.9. Diet:

Diet is a key modifiable risk factor for BC. A nutritious diet may lower the chance of getting BC, according to data from several studies that have looked into the connection between dietary components and BC risk [54-56].

BC risk has been linked to eating a diet high in fruits and vegetables. Aune et al.'s meta-analysis

studies [54] discovered a link between increased fruit and vegetable consumption and a lower risk of BC. In a similar vein, Zhang et al. [55] conducted meta-analyses of case-control studies and discovered a link between an increased intake of fruits and vegetables and a lower risk of breast cancer.

Brown rice, whole wheat, and oats are examples of whole grains that have been linked to a lower risk of breast cancer. According to a comprehensive review and meta-analysis of observational studies conducted by Xiao et al. [56], eating more whole grains was linked to a lower risk of breast cancer.

A higher risk of BC has been linked to high consumption of nutritional fat, particularly saturated fat. High consumption of total and saturated fat was linked to an elevated risk of CS, according to Willett et al. [57].

It is debatable if soy consumption increases the risk of BC. While some research has shown that consuming a lot of soy may raise your risk of breast cancer, some studies have suggested the opposite. Consuming soy was not linked to an increased risk of BC, according to Wu et al. [58].

2.10. Obesity:

BC risk is generally known to be increased by obesity. Many studies have examined the link between obesity and the risk of BC, and the data consistently support the notion that a greater body mass index (BMI) is linked to a higher risk of BC [59-61].

Higher BMI was linked to an increased incidence of postmenopausal breast cancer, according to research by Renehan et al. [59]. With every 5 kg/m2 rise in BMI, the chance of BC rose by 12%. Similar to this, Huang et al. [60] observed that greater BMI was linked to an increased incidence of BC in both premenopausal and postmenopausal women in a meta-analysis of case-control studies.

There might be a number of causes for the link between obesity and the risk of BC. Estrogen, a hormone that is known to encourage the proliferation of BC cells, is produced by adipose tissue. As a result, having more adipose tissue may result in having more estrogen, which might raise your chance of developing BC. Chronic inflammation is another aspect of obesity that has been linked to breast cancer risk [61].

The link between obesity and the risk of BC may depend on factors other than BMI, such as body fat distribution. A greater waist-to-hip ratio (WHR), a measure of central obesity, was linked to a higher risk of BC in postmenopausal women, according to Chan et al. meta-analyses of cohort studies [62].

2.11. Nicotine Consumption:

Consuming nicotine, mostly through cigarette smoking, is known to increase the chance of developing numerous cancers, including lung, bladder, and pancreatic cancer. In addition, there is research that suggests nicotine use may be linked to a higher risk of BC.

According to Kang et al.'s research [63], both postmenopausal and premenopausal women who smoke have a higher chance of developing BC.

Similar results were obtained by Luo et al. [64], who discovered that current smokers had a considerably higher risk of BC than never smokers in a sizable prospective cohort of American women. Women who had smoked for more than 50 years were most at risk for BC.

There might be a number of reasons why smoking increases the risk of BC. Smoke from cigarettes has been proven to include compounds, including nicotine, that damage DNA, raising the possibility of cancer-causing mutations [65]. Furthermore, polycyclic aromatic hydrocarbons (PAHs), which are well-known carcinogens, are present in cigarette smoke [66].
Table 1 summarizes the BC risk factors.

Table 1: BC risk factors [67]

Hormonal	Beginning menstruation at a young age Experiencing menopause later in life Having a first pregnancy at an age over 30 No pregnancies Being in a postmenopausal state, taking oral contraceptives Receiving hormone replacement therapy
Physiological and health status factors	Older age (over 35 years) Familial BC history Familial cancer history Benign changes in the breasts Ionizing radiation An oncogenic viral infection (e.g., Epstein-Barr)
Nutritional	Consumption of fats, red and fried meat, and iron intake Obesity A small intake of fresh produce Limited consumption of plant-based compounds with estrogen-like effects
lifestyle	Frequent moderate/high alcohol use Insufficient regular exercise Working at night.

3. BC Prediction Using ML

BC detection models are computational models that use various ML algorithms to process medical data and predict the likelihood of a patient having BC. The models are developed using features extracted from medical images, such as mammograms, ultrasound, and magnetic resonance imaging (MRI), as well as clinical and pathological factors, such as age, family history, hormonal status, and tumor size.

The ultimate goal of BC detection models is to improve the accuracy and efficiency of BC diagnosis, which can be challenging due to the high variability in BC symptoms and the potential for false positives and negatives. By analyzing large amounts of data and identifying patterns, these models can help physicians make more informed decisions about patient care and potentially reduce the number of unnecessary biopsies and other invasive procedures.

The most often employed ML methods in BC prediction are as follows:

3.1. Supervised Learning

Supervised learning techniques are widely used in BC prediction, with numerous studies employing SVM, ANN, DT, and RF for classification tasks [68-71]. These models are trained using labeled datasets with the objective of learning to predict the outcomes (e.g., benign or malignant) based on the input features (e.g., tumor characteristics).

3.1.1. SVM

SVMs have been extensively applied in BC prediction, with many studies reporting high classification accuracy [72]. For instance, Liu et al. [68] employed SVMs to classify BC cells based on WDBC dataset features. The result showed an accuracy of 99.83%. Nurlaily et al. [73] used SMOTE technology to solve the imbalance problem with the BC Microarray dataset. The SVM with SMOTE showed higher accuracy of 92.5%. Akinnuwesi et al. [74] utilized principal component analysis (PCA) to preprocess two datasets related to BC and, subsequently, SVM algorithm to categorize BC in

180 instances. These instances were selected from the Oncology Department of LASUTH. Resmini et al. [75] applied a genetic algorithm to select optimal features to improve the accuracy of SVM with the DMR-IR dataset. The model achieved 96.15% accuracy. In their study, Abdullah et al. [76] achieved an accuracy of 98% by using a multi-class SVM. The performance of the SVM model throughout the numerous studies is shown in Table 2.

Table 2: The performance review of SVM in the various studies

Year	Ref.	Method	Dataset	Accuracy
2017	[72]	SVM and KNN	WDBC	98.57% with SVM
2018	[68]	SVM	WDBC	98.83%
2019	[73]	SVM with SMOTE	Microarray Dataset	92.5%
2020	[74]	SVM PCA	180 instances from the Oncology Department of LASUTH	97.62%
2021	[75]	SVM with a genetic algorithm	DMR-IR	96.15%
2022	[77]	RBF-SVM	WDBC	100%
2023	[76]	SVM	Mammography images collected from the University of Salford	98%

3.1.2. ANN

ANN is a popular ML technique that has been widely investigated for BC prediction. ANNs are capable of handling high-dimensional data and nonlinear relationships, making them a promising tool for BC prediction. Kaymak et al. [69] used digital images from the Near East University Hospital and classified the images using a back-propagation neural network (BPPN). The accuracy achieved is 70.4%. Charan, Khan, and Khurshid [78] applied CNN to the Mammograms-MIAS dataset, which achieved 65% accuracy. Ragab et al. [79] compared the performance of two different models, including SVM and the deep convolutional neural network (DCNN). The result showed an accuracy of 73.6% for DCNN, while the use of the CNN with the Bayes algorithm achieved a superior accuracy result of 98.95% [80]. Desai and Shah [81] compared the performances of two different models, including CNN and MLP, from the literature. The result showed that CNN outperformed the MLP. Bourouis et al. [82] proposed the wavelet neural network (WNN) and the grey wolf optimization (GWO) algorithm. The proposed model achieved 98% accuracy. Abunasser et al. [83] trained a deep learning algorithm on the imageNet database, collected from the Kaggle depository. The study achieved 98.30% accuracy. The performance of the ANN model throughout the numerous studies is shown in Table 3.

Table 3: The performance review of ANN in the various studies

Year	Ref.	Method	Dataset	Accuracy
2017	[69]	BPPN	Digital images from the Near East University Hospital	70.4%
2018	[78]	CNN	Mammograms-MIAS dataset	65%
2019	[79]	DCNN	DDSM	73.6%

2020	[80]	CNN with the Bayes algorithm	Thermal image dataset (140 instances)	98.95%
2021	[81]	CNN, MLP	Review paper	CNN outperformed MLP
2022	[82]	WNN with GWO	Breast ultrasound (US) images	98%
2023	[83]	deep neural networks	ImageNet database (Kaggle depository)	98.30%

3.1.3. DT

DT is a popular ML algorithm that can be used for both classification and regression tasks. In a decision tree, the data is split into smaller subsets based on the values of features, and a tree-like structure is created, with branches representing the decisions made at each split. At the end of the tree, the leaves represent the final decision or prediction.

Al-Salihy and Ibrikci [84] proposed a BC detection model using decision trees. They evaluated the model using the WDBC dataset. The model achieved 88.0% accuracy.

Higa [70] conducted a comparative study of DT and artificial neural network algorithms for BC detection. The result showed 94% accuracy for both models.

Sathiyanarayanan et al. [85] conducted a comparative study of DT, SVM, KNN, and MLP for BC detection. The result showed 99% accuracy for both models.

Musa and Aliyu [86] applied a DT classifier to data collected from the Department of Radiotherapy and Oncology at Usmanu Danfodiyo University Teaching Hospital, Sokoto State, Nigeria. The accuracy achieved was 87%.

Basker et al. [87] applied the DT classifier to the WBC dataset. The accuracy achieved was 98.20%.

Nasser and Behadili [88] conducted a comparative study of DT and KNN algorithms for BC detection. The DT with feature selection achieved 87.83% accuracy.

Nemade and Fegade [89] conducted a comparative study of Naïve Bayes (NB), Logistic Regression (LR), SVM, KNN, DT, RF, AdaBoost, and XGBoost algorithms for BC detection. The DT achieved a higher accuracy of 97%.

Table 4 summarizes the performance of the DT model in the various studies.

Table 4: The performance review of DT in the various studies

Year	Ref.	Method	Dataset	Accuracy
2017	[84]	DT	WDBC	88.0%
2018	[70]	DT ANN	WDBC	94% for DT
2019	[85]	DT SVM KNN MLP	WDBC	99%
2020	[86]	DT	Department of Radiotherapy and Oncology of Usmanu Danfodiyo University Teaching Hospital, Sokoto State, Nigeria	87% for DT
2021	[87]	DT	WBC	98.20 %
2022	[88]	DT	Oncology Teaching Hospital, Medical City, Baghdad, Iraq, in 2014-2016	87.83%
2023	[89]	NB	WDBC	97% for DT

		LR SVM KNN DT RF AdaBoost XGBoost		

3.1.4. RF

RF is a popular ML technique that has shown promise for BC prediction. RF is an ensemble learning method that combines multiple DTs to improve the accuracy and robustness of the model. In RF, multiple DTs are created by randomly selecting a subset of features and a subset of instances from the dataset. Each decision tree is trained on a different subset of the data, and the predictions from all the trees are combined to get the final prediction. 14 GA-based features were proposed in the Forest model by Aličković and Subasi [90]. The models achieved 99.48% accuracy. Al-Quraishi et al. [71] developed a predictive model based on the RF algorithm to predict the recurrence of breast cancer. The model shows an accuracy of 98.63%. 98.6% accuracy was obtained by training RF on the WDBC dataset [91]. Aroef et al. [92] conducted a comparative study of RF and SVM algorithms for BC detection. With feature selection using Boruta, the RF achieved an accuracy of 90%, while the SVM achieved the highest accuracy of 95%. Gopal et al. [93] conducted a comparative study of MLP, LR, and RF algorithms for BC detection. With feature selection. The accuracy achieved with MLP, LR, and RF was 98%, 79%, and 95%, respectively. Sakib et al. [94] performed a comparison among DL, SVM, LR, RF, and KNN to predict breast cancer. The RF model outperforms others with 96.66% accuracy. Hossin et al. [95] performed a comparison among RF, KNN, DT, LR, AdaBoost, SVM, and Gaussian Naive Bayes to predict breast cancer. The LR model outperforms other models with 96.66% accuracy. Table 5 summarizes the performance of the RF model in the various studies.

Table 5: The performance review of RF in the various studies

Year	Ref.	Method	Dataset	Accuracy
2017	[90]	RF with GA-based 14 features	WDBC	99.48%
2018	[71]	RF	Wisconsin prognostic BC (35/198)	98.63%
2019	[91]	RF	WDBC	98.6%
2020	[92]	RF SVM	BC Coimbra Data Set	RF = 90 SVM = 95
2021	[93]	MLP LR RF	WBCD	MLP = 98% LR = 79% RF = 95%
2022	[94]	DL SVM RF KNN	WDBC	RF = 96.66%
2023	[95]	RF	WBCD	LR = 96.66%

		KNN DT LR AdaBoost SVM Gaussian Naive Bayes		RF = 95.61%

3.2. Unsupervised Learning:

Unsupervised learning techniques such as clustering and dimensionality reduction have been widely studied for their potential to predict breast cancer. The accuracy of supervised learning approaches can be increased by using these methods, which can find patterns and correlations in the data [96, 97]. Clustering is a technique that involves grouping similar data points together based on their features. In the context of BC prediction, clustering can be used to identify subgroups of patients with similar clinical or pathological characteristics. Sehhati et al. [98] addressed the fact that unsupervised clustering of BC patient data using K-means clustering reveals gene expression subgroups with prognostic value. The study conducted by N. Singh et al. [99] used the clustering methods of K-means and fuzzy C-means (FCM) to detect BC in mammograms. The results showed the capability of this combination to detect BC at an early stage. The experimental results of the study by Moftah et al. [100] showed the superior accuracy of adaptive K-means over the traditional K-means clustering technique in detecting breast cancer. Dubey et al. [101] found that the k-means algorithm was better than FCM in terms of computational time, while FCM showed accurate clustering on the BCW dataset.

Dimensionality reduction is a technique that involves reducing the number of features in the data while retaining as much of the original information as possible. This can improve the efficiency of ML algorithms and reduce the risk of overfitting. In the context of BC prediction, dimensionality reduction can be used to identify the most important features for prediction [102]. For example, a study by Sharma et al. [103] used principal component analysis (PCA) to identify the most important features for BC diagnosis and achieved an accuracy of 98.99%. (Sharma et al. [104] used the proposed t-SNE dimensionality reduction techniques and applied a predictive model based on snapshot assembly. The proposed model was examined on the WBCD dataset and achieved 86.6% accuracy. Niţică et al. [105] conducted a comparative study of t-SNE, autoencoders, and a self-organizing map unsupervised algorithm for BC detection. The autoencoders achieved a higher AUC of 93.5%. Pan et al. [106] integrated a genetic algorithm with t-SNE to improve the performance of neural network models. The experimental result confirmed that the proposed model can predict patient survival rates accurately.

4. Data Sources

4.1. Imaging Data:

Imaging data, such as mammograms, MRIs, and ultrasounds, are commonly used as input for ML models in BC prediction [107-109].

In order to predict BC risk using mammography data, Yala et al. [107] applied deep learning. With a 95% confidence interval [CI] of 0.66 to 0.75, the study's AUC was 0.70. El Rahman [110] explored the use of a mammographic dataset with deep learning and a genetic algorithm feature selection technique to detect breast cancer. The research found a number of ideal features and had an accuracy rate of 83.74%. Pramanik et al. [111] explored the use of deep features extracted from breast mammograms to predict breast mass. The study identified 25% of the predictive features and achieved an accuracy of 96.07%. Samee et al. [112] used the

CNN algorithm with features extracted from breast mammograms by AlexNet, VGG, and GoogleNet to predict breast mass. The study achieved an accuracy of 98.50%. Bhukya and Sadanandam [113] explored the use of a roughSet-based feature selector with an RF mammographic dataset to detect breast cancer. The study achieved an accuracy of 95.23%.

Witowski et al. [114] used a deep learning algorithm to improve the performance of detecting BC with MRI images. The study achieved an AUC of 0.92 (95% CI: 0.92-0.93). Nasir et al. [108] used a fine-tuning method to extract features from MRI images and then used a neural network to detect breast cancer. The study achieved 98.1% accuracy. Lu et al. [115] explored the use of the Relief algorithm's feature selection method on breast MRI images and applied ensemble techniques of the ML models, including ensemble under-sampling, AdaBoost, and the subspace method. The study achieved an AUC of 0.9617.

Pourasad et al. [109] explored the use of the fractal feature selection method on breast ultrasound images. The result showed that CNN achieved a higher accuracy of 99.8%. Xu et al. [116] developed an ML model to predict HER2 BC status from ultrasound images. The result showed that deep learning achieved an accuracy of 85.79%. T. Wu et al. [117] developed an ML model to predict triple-negative BC status from ultrasound images. The result showed that LR achieved 85%. Masud et al. [118] used a pre-trained CNN model to predict BC using ultrasound images. The result showed that using Adam and RMSprop optimizers with CNN improved accuracy to 100%.

4.2. Clinical Data:

Clinical data, such as patient demographics, tumor characteristics, and treatment histories, are also used as input features in ML models for BC prediction. For instance, Telsang and Hegde [119] used the SVM model to predict BC using the WDBC. The result showed that the model achieved 96.25% accuracy and an AUC of 99.4. Alshayeji et al. [120] used the ANN model to predict BC using the WDBC. The result showed that the model achieved 99.85% accuracy. S. Singh et al. [121] used the ANN model to predict BC using the WDBC. The result showed that the model achieved 99.41% accuracy. Prahmana and Sitepu [122] used the KNN model to predict BC using the BCCD. The result showed that the model achieved 97% accuracy. Rasool et al. [123] used the Polynomial SVM model to predict BC using the BCCD. The result showed that the model achieved 76.83% accuracy. Elsadig et al. [124] used the SVM model to predict BC using the WBDC. The result showed that the model achieved 97.7% accuracy.

5. Conclusion

BC is a significant health concern worldwide, and early detection through risk factor identification and prediction is crucial for improving patient outcomes. ML algorithms have emerged as a promising tool for predicting BC risk and identifying relevant risk factors. The selection of appropriate data sources is essential for improving the accuracy and reliability of predictive models. This literature review has provided valuable insights into the current state of research on BC risk factors, prediction using ML algorithms, and data sources. The review identified several risk factors associated with the development of breast cancer, including age, family history, and hormone levels. ML algorithms such as SVMs, decision trees, artificial neural networks, RFs, clustering, and dimensionality reduction have been used to predict BC risk with varying degrees of success. The review also identified several data sources used in BC prediction, including mammography, magnetic resonance imaging, and electronic health records. The performance of ML algorithms in predicting BC risk using different data sources varied depending on the study's design and the features used. The review identified gaps in current research, including the need for more studies that use a combination of multiple data sources and the need for more

studies that validate the performance of ML algorithms on independent datasets.

6. References

[1] B. S. Chhikara and K. Parang, "Global Cancer Statistics 2022: the trends projection analysis," *Chemical Biology Letters*, vol. 10, pp. 451-451, 2023.

[2] N. F. Boyd, L. J. Martin, M. J. Yaffe, and S. Minkin, "Mammographic density and breast cancer risk: current understanding and future prospects," *Breast cancer research*, vol. 13, pp. 1-12, 2011.

[3] W. A. Berg, Z. Zhang, D. Lehrer, R. A. Jong, E. D. Pisano, R. G. Barr, *et al.*, "Detection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography in women with elevated breast cancer risk," *Jama*, vol. 307, pp. 1394-1404, 2012. doi:10.1001/jama.2012.388

[4] R. Chen, Y. Qi, Y. Huang, W. Liu, R. Yang, X. Zhao, *et al.*, "Diagnostic value of core needle biopsy for determining HER2 status in breast cancer, especially in the HER2-low population," *Breast Cancer Research and Treatment*, vol. 197, pp. 189-200, 2023. https://doi.org/10.1007/s10549-022-06781-3

[5] A. Al-Faris, "Some Segmentation Approaches of Breast Magnetic Resonance Imaging Tumor in Computer Aided Detection Systems: A Review," *WSEAS Transactions on Computer Research*, vol. 11, pp. 1-9, 2023. DOI: 10.37394/232018.2023.11.1

[6] American Cancer Society. Breast Cancer Facts & Figures 2021-2022. Available at: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/2022-2024-breast-cancer-fact-figures-acs.pdf. Accessed on March 31, 2023.

[7] National Cancer Institute: Breast Cancer Treatment. Available at: https://www.cancer.gov/types/breast/hp/breast-treatment-pdq. Accessed on March 31.

[8] American Cancer Society. (2022). Breast Cancer Early Detection. Available at: https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/american-cancer-society-recommendations-for-the-early-detection-of-breast-cancer.html. Accessed on March 31.

[9] C. G. o. H. F. i. B. Cancer, "Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies," *The lancet oncology*, vol. 13, pp. 1141-1151, 2012. https://doi.org/10.1016/S1470-2045(12)70425-4

[10] W. Yue, J. P. Wang, Y. Li, P. Fan, G. Liu, N. Zhang, *et al.*, "Effects of estrogen on breast cancer development: Role of estrogen receptor independent mechanisms," *International journal of cancer*, vol. 127, pp. 1748-1757, 2010. https://doi.org/10.1002/ijc.25207

[11] American Cancer Society:Cancer Facts & Figures 2022. Available at: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2022/2022-cancer-facts-and-figures.pdf.

[12] National Cancer Institute:Risk Factors. Available at: https://www.nationalbreastcancer.org/breast-cancer-risk-factors.

[13] L. A. Brinton, C. Schairer, R. N. Hoover, and J. F. Fraumeni, Jr., "Menstrual factors and risk of breast cancer," *Cancer Invest*, vol. 6, pp. 245-54, 1988. https://doi.org/10.3109/07357908809080645

[14] S. Nardin, E. Mora, F. M. Varughese, F. D'Avanzo, A. R. Vachanaram, V. Rossi, *et al.*, "Breast cancer survivorship, quality of life, and late toxicities," *Frontiers in oncology*, vol. 10, p. 864, 2020. https://doi.org/10.3389/fonc.2020.00864

[15] C. J. de Blok, C. M. Wiepjes, N. M. Nota, K. van Engelen, M. A. Adank, K. M. Dreijerink, *et al.*, "Breast cancer risk in transgender people receiving hormone treatment: nationwide cohort study in the Netherlands," *Bmj*, vol. 365, 2019. doi: https://doi.org/10.1136/bmj.l1652

[16] G. V. Dall and K. L. Britt, "Estrogen effects on the mammary gland in early and late life and breast cancer risk," *Frontiers in oncology*, vol. 7, p. 110, 2017. https://doi.org/10.3389/fonc.2017.00110

[17] S. E. Singletary, "Rating the risk factors for breast cancer," *Annals of surgery*, vol. 237, p. 474, 2003. doi: 10.1097/01.SLA.0000059969.64262.87

[18] S. D. Harlow and P. Paramsothy, "Menstruation and the menopausal transition," *Obstetrics and Gynecology Clinics*, vol. 38, pp. 595-607, 2011. https://doi.org/10.1016/j.ogc.2011.05.010

[19] J. L. Kelsey, M. D. Gammon, and E. M. John, "Reproductive factors and breast cancer," *Epidemiologic reviews*, vol. 15, pp. 36-47, 1993. DOI: 10.1093/oxfordjournals.epirev.a036115

[20] J. S. Cohain, R. E. Buxbaum, and D. Mankuta, "Spontaneous first trimester miscarriage rates per woman among parous women with 1 or more pregnancies of 24 weeks or more," *BMC pregnancy and childbirth*, vol. 17, pp. 1-7, 2017. DOI 10.1186/s12884-017-1620-1

[21] V. Beral, D. Bull, R. Doll, R. Peto, G. Reeves, P. van den Brandt, *et al.*, "Collaborative group on hormonal factors in breast cancer: breast cancer and abortion: collaborative reanalysis of data from 53 epidemiological studies, including 83000 women with breast cancer from 16 countries," *Lancet*, vol. 363, pp. 1007-1016, 2004. https://doi.org/10.1016/S0140-6736(04)15835-2

[22] L. S. Mørch, C. W. Skovlund, P. C. Hannaford, L. Iversen, S. Fielding, and Ø. Lidegaard, "Contemporary hormonal contraception and the risk of breast cancer," *New England Journal of Medicine*, vol. 377, pp. 2228-2239, 2017. DOI: 10.1056/NEJMoa1700732

[23] A. Mehrgou and M. Akouchekian, "The importance of BRCA1 and BRCA2 genes mutations in breast cancer development," *Medical journal of the Islamic Republic of Iran*, vol. 30, p. 369, 2016. PMCID: PMC4972064

[24] R. Roy, J. Chun, and S. N. Powell, "BRCA1 and BRCA2: different roles in a common pathway of genome protection," *Nature Reviews Cancer*, vol. 12, pp. 68-78, 2012.

[25] H. Chen, J. Wu, Z. Zhang, Y. Tang, X. Li, S. Liu, *et al.*, "Association between BRCA status and triple-negative breast cancer: a meta-analysis," *Frontiers in pharmacology*, vol. 9, p. 909, 2018. https://doi.org/10.3389/fphar.2018.00909

[26] S. De Talhouet, J. Peron, A. Vuilleumier, A. Friedlaender, V. Viassolo, A. Ayme, *et al.*, "Clinical outcome of breast cancer in carriers of BRCA1 and BRCA2 mutations according to molecular subtypes," *Scientific reports*, vol. 10, p. 7073, 2020. https://doi.org/10.1038/s41598-020-63759-1

[27] H. Abdel-Razeq, F. Tamimi, L. Abujamous, S. Edaily, M. Abunasser, R. Bater, *et al.*, "Patterns and Prevalence of BRCA1 and BRCA2 Germline Mutations Among Patients with Triple-Negative Breast Cancer: Regional Perspectives," *Cancer Management and Research*, pp. 4597-4604, 2021. https://doi.org/10.2147/CMAR.S316470

[28] D. Angeli, S. Salvi, and G. Tedaldi, "Genetic predisposition to breast and ovarian cancers: how many and which genes to test?," *International journal of molecular sciences*, vol. 21, p. 1128, 2020. https://doi.org/10.3390/ijms21031128

[29] R. S. Chamseddine, C. Wang, K. Yin, J. Wang, P. Singh, J. Zhou, *et al.*, "Penetrance of male breast cancer susceptibility genes: A systematic review," *Breast Cancer Research and Treatment*, pp. 1-8, 2022.

[30] S. Armaou, M. Pertesi, F. Fostira, G. Thodi, P. Athanasopoulos, S. Kamakari, *et al.*, "Contribution of BRCA1 germ-line mutations to breast cancer in Greece: a hospital-based study of 987 unselected breast cancer cases," *British journal of cancer*, vol. 101, pp. 32-37, 2009. doi:10.1038/sj.bjc.6605115

[31] L. C. Hartmann, D. C. Radisky, M. H. Frost, R. J. Santen, R. A. Vierkant, L. L. Benetti, *et al.*, "Understanding the Premalignant Potential of Atypical Hyperplasia through Its Natural History: A Longitudinal Cohort StudyBreast Cancers after Atypical Hyperplasia," *Cancer Prevention Research*, vol. 7, pp. 211-217, 2014. https://doi.org/10.1158/1940-6207.CAPR-13-0222

[32] W. D. Dupont and D. L. Page, "Risk factors for breast cancer in women with proliferative breast disease," *New England Journal of Medicine*, vol. 312, pp. 146-151, 1985. DOI: 10.1056/NEJM198501173120303

[33] J. D. Boice Jr and R. R. Monson, "Breast cancer in women after repeated fluoroscopic examinations of the chest," *Journal of the National Cancer Institute*, vol. 59, pp. 823-832, 1977. https://doi.org/10.1093/jnci/59.3.823

[34] C. M. Ronckers, M. M. Doody, J. E. Lonstein, M. Stovall, and C. E. Land,

"Multiple diagnostic X-rays for spine deformities and risk of breast cancer," *Cancer Epidemiology Biomarkers & Prevention*, vol. 17, pp. 605-613, 2008. https://doi.org/10.1158/1055-9965.EPI-07-2628

[35] A. B. de Gonzalez, E. Gilbert, R. Curtis, P. Inskip, R. Kleinerman, L. Morton, *et al.*, "Second solid cancers after radiation therapy: a systematic review of the epidemiologic studies of the radiation dose-response relationship," *International Journal of Radiation Oncology* Biology* Physics*, vol. 86, pp. 224-233, 2013. https://doi.org/10.1016/j.ijrobp.2012.09.001

[36] V. Beir, "Health risks from exposure to low levels of ionizing radiation," *The National Academies report in brief*, 2005.

[37] D. L. Preston, A. Mattsson, E. Holmberg, R. Shore, N. G. Hildreth, and J. D. Boice Jr, "Radiation effects on breast cancer risk: a pooled analysis of eight cohorts," *Radiation research*, vol. 158, pp. 220-235, 2002. https://doi.org/10.1667/0033-7587(2002)158[0220:REOBCR]2.0.CO;2

[38] C. M. Ronckers, A. J. Sigurdson, M. Stovall, S. A. Smith, A. C. Mertens, Y. Liu, *et al.*, "Thyroid cancer in childhood cancer survivors: a detailed evaluation of radiation dose response and its modifiers," *Radiation research*, vol. 166, pp. 618-628, 2006. https://doi.org/10.1667/RR3605.1

[39] A. J. Sigurdson, M. M. Doody, R. S. Rao, D. M. Freedman, B. H. Alexander, M. Hauptmann, *et al.*, "Cancer incidence in the US radiologic technologists health study, 1983–1998," *Cancer*, vol. 97, pp. 3080-3089, 2003. https://doi.org/10.1002/cncr.11444

[40] E. Cardis, M. Vrijheid, M. Blettner, E. Gilbert, M. Hakama, C. Hill, *et al.*, "The 15-country collaborative study of cancer risk among radiation workers in the nuclear industry: estimates of radiation-related cancer risks," *Radiation research*, vol. 167, pp. 396-416, 2007. https://doi.org/10.1667/RR0553.1

[41] R. E. Shore, N. Hildreth, E. Woodard, P. Dvoretsky, L. Hempelmann, and B. Pasternack, "Breast cancer among women given X-ray therapy for acute postpartum mastitis," *JNCI: Journal of the National Cancer Institute*, vol. 77, pp. 689-696, 1986. https://doi.org/10.1093/jnci/77.3.689

[42] National Cancer institute: Breast Cancer *Prevention*. Available at:https://www.cancer.gov/types/breast/patient/breast-prevention-pdq#_12. Accessed March 5, 2023.

[43] Radiological Society of North America. Radiation Safety in Adult Medical Imaging. available at: https://www.radiologyinfo.org/en/info.cfm?pg=safety-xray. Accessed March 5.

[44] American College of Radiology. Image Wisely. https://www.imagewisely.org/. Accessed March 5.

[45] Occupational Safety and Health Administration. Radiation. https://www.osha.gov/radiation. Accessed March 5.

[46] A. M. Stuebe, W. C. Willett, F. Xue, and K. B. Michels, "Lactation and incidence of premenopausal breast cancer: a longitudinal study," *Archives of internal medicine*, vol. 169, pp. 1364-1371, 2009. doi:10.1001/archinternmed.2009.231

[47] S. K. Clinton, E. L. Giovannucci, and S. D. Hursting, "The world cancer research fund/American institute for cancer research third expert report on diet, nutrition, physical activity, and cancer: impact and future directions," *The Journal of nutrition*, vol. 150, pp. 663-671, 2020. https://doi.org/10.1093/jn/nxz268

[48] N. Hamajima, K. Hirose, K. Tajima, T. Rohan, E. Calle, C. Heath, *et al.*, "Alcohol, tobacco and breast cancer--collaborative reanalysis of individual data from 53 epidemiological studies, including 58,515 women with breast cancer and 95,067 women without the disease," *British journal of cancer*, vol. 87, pp. 1234-1245, 2002. https://dx.doi.org/10.1038/sj.bjc.6600596

[49] S. A. Smith-Warner, D. Spiegelman, S.-S. Yaun, P. A. Van Den Brandt, A. R. Folsom, R. A. Goldbohm, *et al.*, "Alcohol and breast cancer in women: a pooled analysis of cohort studies," *Jama*, vol. 279, pp. 535-540, 1998. doi:10.1001/jama.279.7.535

[50] W. Y. Chen, B. Rosner, S. E. Hankinson, G. A. Colditz, and W. C. Willett, "Moderate alcohol consumption during adult life, *drinking* patterns, and breast cancer risk," *Jama*, vol. 306, pp. 1884-1890, 2011. doi:10.1001/jama.2011.1590

[51] I. Romieu, C. Scoccianti, V. Chajes, J. D. Batlle, and S. Rinaldi, "Alcohol intake

and breast cancer in the European Prospective *investigation* into Cancer and Nutrition (EPIC) study," *Cancer Research*, vol. 74, pp. 1271-1271, 2014. https://doi.org/10.1158/1538-7445.AM2014-1271

[52] K. W. Singletary and S. M. Gapstur, "Alcohol and breast *cancer*: review of epidemiologic and experimental evidence and potential mechanisms," *Jama*, vol. 286, pp. 2143-2151, 2001. doi:10.1001/jama.286.17.2143

[53] G. A. Colditz, B. A. Rosner, W. Y. Chen, M. D. *Holmes*, and S. E. Hankinson, "Risk factors for breast cancer according to estrogen and progesterone receptor status," *Journal of the national cancer institute*, vol. 96, pp. 218-228, 2004. https://doi.org/10.1093/jnci/djh025

[54] D. Aune, D. Chan, A. Vieira, D. N. *Rosenblatt*, R. Vieira, D. Greenwood, *et al.*, "Fruits, vegetables and breast cancer risk: a systematic review and meta-analysis of prospective studies," *Breast cancer research and treatment*, vol. 134, pp. 479-493, 2012. http://dx.doi.org/10.1007/s10549-012-2118-1

[55] C. X. Zhang, S. C. Ho, Y. M. Chen, J. H. Fu, S. Z. Cheng, and F. Y. Lin, "Greater vegetable and fruit intake is associated with a lower risk of breast cancer among Chinese women," *International journal of cancer*, vol. 125, pp. 181-188, 2009. https://doi.org/10.1002/ijc.24358

[56] Y. Xiao, Y. Ke, S. Wu, S. Huang, S. Li, Z. Lv, *et al.*, "Association between whole grain intake and breast cancer risk: a systematic review and meta-analysis of observational studies," *Nutrition Journal*, vol. 17, pp. 1-12, 2018.

[57] W. WILLETT and M. STAMPFER, "Dietary fat intake and breast cancer risk," *JNCI: Journal of the National Cancer Institute*, vol. 81, pp. 1422-1422, 1989. https://doi.org/10.1093/jnci/81.18.1422

[58] A. Wu, W. Koh, R. Wang, H. Lee, and M. Yu, "Soy intake and breast cancer risk in Singapore Chinese Health Study," *British journal of cancer*, vol. 99, pp. 196-200, 2008.

[59] A. G. Renehan, M. Tyson, M. Egger, R. F. Heller, and M. Zwahlen, "Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies," *The lancet*, vol. 371, pp. 569-578, 2008. https://doi.org/10.1016/S0140-6736(08)60269-X

[60] Z. Huang, S. E. Hankinson, G. A. Colditz, M. J. Stampfer, D. J. Hunter, J. E. Manson, *et al.*, "Dual effects of weight and weight gain on breast cancer risk," *jama*, vol. 278, pp. 1407-1411, 1997. doi:10.1001/jama.1997.03550170037029

[61] H. Rodríguez-Hernández, L. E. Simental-Mendía, G. Rodríguez-Ramírez, and M. A. Reyes-*Romero*, "Obesity and inflammation: epidemiology, risk factors, and markers of inflammation," *International journal of endocrinology*, vol. 2013, 2013. https://doi.org/10.1155/2013/678159

[62] D. S. Chan, A. Vieira, D. Aune, E. V. Bandera, D. *Greenwood*, A. McTiernan, *et al.*, "Body mass index and survival in women with breast cancer—systematic literature review and meta-analysis of 82 follow-up studies," *Annals of Oncology*, vol. 25, pp. 1901-1914, 2014. https://doi.org/10.1093/annonc/mdu042

[63] K. Wang, F. Li, X. Zhang, Z. Li, and H. Li, "Smoking increases risks of all-cause and breast cancer specific mortality in breast cancer *individuals*: a dose-response meta-analysis of prospective cohort studies involving 39725 breast cancer cases," *Oncotarget*, vol. 7, p. 83134, 2016. doi: 10.18632/oncotarget.13366

[64] J. Luo, K. L. Margolis, J. Wactawski-Wende, K. Horn, C. Messina, M. L. Stefanick, *et al.*, "Association of active and passive smoking with *risk* of breast cancer among postmenopausal women: a prospective cohort study," *Bmj*, vol. 342, 2011. https://doi.org/10.1136/bmj.d1016

[65] S. S. Hecht, "Lung carcinogenesis by tobacco smoke," *International journal of cancer*, vol. 131, pp. 2724-2732, 2012. https://doi.org/10.1002/ijc.27816

[66] J. Korsh, A. Shen, K. Aliano, and T. Davenport, "Polycyclic aromatic *hydrocarbons* and breast cancer: a review of the literature," *Breast Care*, vol. 10, pp. 316-318, 2015. https://doi.org/10.1159/000436956

[67] S. R. Lakhani, I. O. Ellis, S. Schnitt, P. H. Tan, and M. van de Vijver, "WHO Classification of Tumours of the Breast," 2012. ISBN: 9789283224334

[68] N. Liu, J. Shen, M. Xu, D. Gan, E.-S. Qi, and B. Gao, "Improved cost-sensitive support vector machine classifier for breast cancer diagnosis," *Mathematical Problems in Engineering*, vol. 2018, pp.

1-13, 2018. https://doi.org/10.1155/2018/3875082

[69] S. Kaymak, A. Helwan, and D. Uzun, "Breast cancer image classification using artificial neural networks," *Procedia computer science*, vol. 120, pp. 126-131, 2017. https://doi.org/10.1016/j.procs.2017.11.219

[70] A. Higa, "Diagnosis of breast cancer using decision tree and artificial neural network algorithms," *cell*, vol. 1, pp. 23-27, 2018.

[71] T. Al-Quraishi, J. H. Abawajy, M. U. Chowdhury, S. Rajasegarar, and A. S. Abdalrada, "Breast cancer recurrence prediction using random forest model," in *Recent Advances on Soft Computing and Data Mining: Proceedings of the Third International Conference on Soft Computing and Data Mining (SCDM 2018), Johor, Malaysia, February 06-07, 2018*, 2018, pp. 318-329. DOI: 10.1007/978-3-319-72550-5_31

[72] M. M. Islam, H. Iqbal, M. R. Haque, and M. K. Hasan, "Prediction of breast cancer using support vector machine and K-Nearest neighbors," in *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, 2017, pp. 226-229. DOI:10.1109/R10-HTC.2017.8288944

[73] D. Nurlaily, Irhamah, S. W. Purnami, and H. Kuswanto, "Support vector machine for imbalanced microarray dataset classification using ant colony optimization and genetic algorithm," in *AIP Conference Proceedings*, 2019, p. 020076. https://doi.org/10.1063/1.5139808

[74] B. A. Akinnuwesi, B. O. Macaulay, and B. S. Aribisala, "Breast cancer risk assessment and early diagnosis using Principal Component Analysis and support vector machine techniques," *Informatics in medicine unlocked*, vol. 21, p. 100459, 2020. https://doi.org/10.1016/j.imu.2020.100459

[75] R. Resmini, L. Silva, A. S. Araujo, P. Medeiros, D. Muchaluat-Saade, and A. Conci, "Combining genetic algorithms and SVM for breast cancer diagnosis using infrared thermography," *Sensors*, vol. 21, p. 4802, 2021. https://doi.org/10.3390/s21144802

[76] A. K. Abdullah, R. M. Azawi, I. T. Ibrahim, and A. A. Ajwad, "Mammography images classification system based texture analysis and multi class support vector machine," in *AIP Conference Proceedings*, 2023, p. 070003. https://doi.org/10.1063/5.0110733

[77] A. S. Elkorany, M. Marey, K. M. Almustafa, and Z. F. Elsharkawy, "Breast cancer diagnosis using support vector machines optimized by whale optimization and dragonfly algorithms," *IEEE Access*, vol. 10, pp. 69688-69699, 2022. DOI: 10.1109/ACCESS.2022.3186021

[78] S. Charan, M. J. Khan, and K. Khurshid, "Breast cancer detection in mammograms using convolutional neural network," in *2018 International* Conference *on Computing, Mathematics and Engineering Technologies (iCoMET)*, 2018, pp. 1-5. DOI: 10.1109/ICOMET.2018.8346384

[79] D. A. Ragab, M. Sharkas, S. Marshall, and J. Ren, "Breast cancer detection using deep convolutional neural networks and support vector machines," *PeerJ*, vol. 7, p. e6201, 2019. https://doi.org/10.7717/peerj.6201

[80] S. Ekici and H. Jawzal, "Breast cancer diagnosis using thermography and convolutional neural networks," *Medical hypotheses*, vol. 137, p. 109542, 2020. https://doi.org/10.1016/j.mehy.2019.109542

[81] M. Desai and M. Shah, "An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN)," *Clinical eHealth*, vol. 4, pp. 1-11, 2021. https://doi.org/10.1016/j.ceh.2020.11.002

[82] S. Bourouis, S. S. Band, A. Mosavi, S. Agrawal, and M. Hamdi, "Meta-heuristic algorithm-tuned neural network for breast cancer diagnosis using ultrasound images," *Frontiers in Oncology*, vol. 12, p. 834028, 2022. doi: 10.3389/fonc.2022.834028

[83] B. S. ABUNASSER, S. M. DAUD, I. S. ZAQOUT, and S. S. ABU-NASER, "CONVOLUTION NEURAL NETWORK FOR BREAST CANCER DETECTION AND CLASSIFICATION–FINAL RESULTS," *Journal of Theoretical and Applied Information Technology*, vol. 101, 2023.

[84] N. K. Al-Salihy and T. Ibrikci, "Classifying breast cancer by using

decision tree algorithms," in *Proceedings of the 6th International Conference on Software and Computer Applications*, 2017, pp. 144-148. https://doi.org/10.1145/3056662.3056716

[85] P. Sathiyanarayanan, S. Pavithra, M. S. Saranya, and M. Makeswari, "Identification of breast cancer using the decision tree algorithm," in *2019 IEEE International conference on system, computation, automation and networking (ICSCAN)*, 2019, pp. 1-6. DOI: 10.1109/ICSCAN.2019.8878757

[86] A. Musa and U. Aliyu, "Application of machine learning techniques in predicting of breast Cancer metastases using decision tree algorithm," *Sokoto Northwestern Nigeria. J Data Mining Genomics Proteomics*, vol. 11, pp. 2153-0602.20, 2020. DOI: 10.35248/2153-0602.20.11.220.

[87] N. Basker, S. Theetchenya, D. Vidyabharathi, J. Dhaynithi, G. Mohanraj, M. Marimuthu, *et al.*, "Breast cancer detection using machine learning algorithms," *Annals of the Romanian Society for Cell Biology*, pp. 2551-2562, 2021.

[88] F. K. Nasser and S. F. Behadili, "Breast Cancer Detection using Decision Tree and K-Nearest Neighbour Classifiers," *Iraqi Journal of Science*, pp. 4987-5003, 2022. https://doi.org/10.24996/ijs.2022.63.11.34

[89] V. Nemade and V. Fegade, "Machine Learning Techniques for Breast Cancer Prediction," *Procedia Computer Science*, vol. 218, pp. 1314-1320, 2023. https://doi.org/10.1016/j.procs.2023.01.110

[90] E. Aličković and A. Subasi, "Breast cancer diagnosis using GA feature selection and Rotation Forest," *Neural Computing and applications*, vol. 28, pp. 753-763, 2017.

[91] T. I. Rohan, A. B. Siddik, M. Islam, and M. S. U. Yusuf, "A precise breast cancer detection approach using ensemble of random forest with AdaBoost," in *2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)*, 2019, pp. 1-4. DOI: 10.1109/IC4ME247184.2019.9036697

[92] C. Aroef, Y. Rivan, and Z. Rustam, "Comparing random forest and support vector machines for breast cancer classification," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 18, pp. 815-821, 2020. http://doi.org/10.12928/telkomnika.v18i2.14785

[93] V. N. Gopal, F. Al-Turjman, R. Kumar, L. Anand, and M. Rajesh, "Feature selection and classification in breast cancer prediction using IoT and machine learning," *Measurement*, vol. 178, p. 109442, 2021. https://doi.org/10.1016/j.measurement.2021.109442

[94] S. Sakib, N. Yasmin, A. K. Tanzeem, F. Shorna, K. Md. Hasib, and S. B. Alam, "Breast cancer detection and classification: A comparative analysis using machine learning algorithms," in *Proceedings of Third International* Conference *on Communication, Computing and Electronics Systems: ICCCES 2021*, 2022, pp. 703-717. DOI: 10.1007/978-981-16-8862-1_46

[95] M. M. Hossin, F. J. M. Shamrat, M. R. Bhuiyan, R. A. Hira, T. Khan, and S. Molla, "Breast cancer detection: an effective comparison of different machine learning algorithms on the Wisconsin dataset," *Bulletin of Electrical Engineering and Informatics*, vol. 12, pp. 2446-2456, 2023. DOI: https://doi.org/10.11591/eei.v12i4.4448

[96] M. E. Celebi and K. Aydin, *Unsupervised learning* algorithms vol. 9: Springer, 2016. ISBN: 978-3-319-24211-8

[97] M. Dash, H. Liu, and J. Yao, "Dimensionality reduction of unsupervised data," in *Proceedings ninth ieee international conference on tools with artificial intelligence*, 1997, pp. 532-539. DOI: 10.1109/TAI.1997.632300

[98] M. Sehhati, M. A. Tabatabaiefar, A. H. Gholami, and M. Sattari, "Using classification and K-means methods to predict breast cancer recurrence in gene expression data," *Journal of Medical Signals and Sensors*, vol. 12, p. 122, 2022. doi: 10.4103/jmss.jmss_117_21

[99] N. Singh, A. G. Mohapatra, and G. Kanungo, "Breast cancer mass detection in mammograms using K-means and fuzzy C-means clustering," *International Journal of Computer Applications*, vol.

22, pp. 15-21, 2011.

[100] H. M. Moftah, A. T. Azar, E. T. Al-Shammari, N. I. Ghali, A. E. Hassanien, and M. Shoman, "Adaptive k-means clustering algorithm for MR breast image segmentation," *Neural Computing and Applications*, vol. 24, pp. 1917-1928, 2014.

[101] A. K. Dubey, U. Gupta, and S. Jain, "Comparative study of K-means and fuzzy C-means algorithms on the breast cancer data," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 8, pp. 18-29, 2018.

[102] S. E. Henouda, F. Z. Laallam, O. Kazar, S. Harous, and D. Houfani, "On the Effectiveness of Dimensionality Reduction Techniques on High Dimensionality Datasets," in *12th International Conference on Information Systems and Advanced Technologies "ICISAT 2022" Intelligent Information, Data Science and Decision Support System*, 2023, pp. 156-166. DOI: 10.1007/978-3-031-25344-7_15

[103] N. Sharma, M. Mangla, S. N. Mohanty, and S. Satpaty, "A stochastic neighbor embedding approach for cancer prediction," in *2021 international conference on emerging smart computing and informatics (ESCI)*, 2021, pp. 599-603. DOI: 10.1109/ESCI50559.2021.9396902

[104] N. Sharma, K. Sharma, M. Mangla, and R. Rani, "Breast cancer classification using snapshot ensemble deep learning model and t-distributed stochastic neighbor embedding," *Multimedia Tools and Applications*, vol. 82, pp. 4011-4029, 2023.

[105] Ş. Niţică, G. Czibula, and V.-I. Tomescu, "A comparative study on using unsupervised learning based data analysis techniques for breast cancer detection," in *2020 IEEE 14th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, 2020, pp. 000099-000104. DOI: 10.1109/SACI49304.2020.9118783

[106] X. Pan, T. Zhang, Q. Yang, D. Yang, J.-C. Rwigema, and X. S. Qi, "Survival prediction for oral tongue cancer patients via probabilistic genetic algorithm optimized neural network models," *The British journal of radiology*, vol. 93, p. 20190825, 2020. https://doi.org/10.1259/bjr.20190825

[107] A. Yala, C. Lehman, T. Schuster, T. Portnoi, and R. Barzilay, "A deep learning mammography-based model for improved breast cancer risk prediction," *Radiology*, vol. 292, pp. 60-66, 2019. https://doi.org/10.1148/radiol.2019182716

[108] M. U. Nasir, T. M. Ghazal, M. A. Khan, M. Zubair, A.-u. Rahman, R. Ahmed, *et al.*, "Breast cancer prediction empowered with fine-tuning," *Computational Intelligence and Neuroscience*, vol. 2022, 2022. https://doi.org/10.1155/2022/5918686

[109] Y. Pourasad, E. Zarouri, M. Salemizadeh Parizi, and A. Salih Mohammed, "Presentation of novel architecture for diagnosis and identifying breast cancer location based on ultrasound images using machine learning," *Diagnostics*, vol. 11, p. 1870, 2021. https://doi.org/10.3390/diagnostics11101870

[110] S. A. El_Rahman, "Predicting breast cancer survivability based on machine learning and features selection algorithms: a comparative study," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 8585-8623, 2021. https://doi.org/10.1007/s12652-020-02590-y

[111] P. Pramanik, S. Mukhopadhyay, S. Mirjalili, and R. Sarkar, "Deep feature selection using local search embedded social ski-driver optimization algorithm for breast cancer detection in mammograms," *Neural Computing and Applications*, vol. 35, pp. 5479-5499, 2023. https://doi.org/10.1007/s00521-022-07895-x(0123456789().,-volV)(0123456789().,-volV)

[112] N. A. Samee, G. Atteia, S. Meshoul, M. A. Al-antari, and Y. M. Kadah, "Deep Learning Cascaded Feature Selection Framework for Breast Cancer Classification: Hybrid CNN with Univariate-Based Approach," *Mathematics*, vol. 10, p. 3631, 2022. https://doi.org/10.3390/math10193631

[113] H. Bhukya and M. Sadanandam, "RoughSet based Feature Selection for Prediction of Breast Cancer," *Wireless Personal Communications*, pp. 1-18, 2023. https://doi.org/10.21203/rs.3.rs-1542645/v1

[114] J. Witowski, L. Heacock, B. Reig, S. K. Kang, A. Lewin, K.

Pysarenko, *et al.*, "Improving breast cancer diagnostics with deep learning for mri," *Science Translational Medicine*, vol. 14, p. eabo4802, 2022. DOI: 10.1126/scitranslmed.abo4802

[115] W. Lu, Z. Li, and J. Chu, "A novel computer-aided diagnosis system for breast MRI based on feature selection and ensemble learning," *Computers in biology and medicine*, vol. 83, pp. 157-165, 2017. http://dx.doi.org/10.1016/j.compbiomed.2017.03.002

[116] Z. Xu, Q. Yang, M. Li, J. Gu, C. Du, Y. Chen, *et al.*, "Predicting HER2 status in breast cancer on ultrasound images using deep learning method," *Frontiers in oncology*, vol. 12, p. 829041, 2022. doi: 10.3389/fonc.2022.829041

[117] T. Wu, L. R. Sultan, J. Tian, T. W. Cary, and C. M. Sehgal, "Machine learning for diagnostic ultrasound of triple-negative breast cancer," *Breast cancer research and treatment*, vol. 173, pp. 365-373, 2019. https://doi.org/10.1007/s10549-018-4984-7

[118] M. Masud, M. S. Hossain, H. Alhumyani, S. S. Alshamrani, O. Cheikhrouhou, S. Ibrahim, *et al.*, "Pre-trained convolutional neural networks for breast cancer detection using ultrasound images," *ACM Transactions on Internet Technology (TOIT)*, vol. 21, pp. 1-17, 2021. https://doi.org/10.1145/3418355

[119] V. A. Telsang and K. Hegde, "Breast cancer prediction analysis using machine learning algorithms," in *2020 International Conference on Communication, Computing and Industry 4.0 (C2I4)*, 2020, pp. 1-5. DOI: 10.1109/C2I451079.2020.9368911

[120] M. H. Alshayeji, H. Ellethy, and R. Gupta, "Computer-aided detection of breast cancer on the Wisconsin dataset: An artificial neural networks approach," *Biomedical Signal Processing and Control*, vol. 71, p. 103141, 2022. https://doi.org/10.1016/j.bspc.2021.103141

[121] S. Singh, S. K. Jangir, M. Kumar, M. Verma, S. Kumar, T. S. Walia, *et al.*, "Feature Importance Score-Based Functional Link Artificial Neural Networks for Breast Cancer Classification," *BioMed Research International*, vol. 2022, 2022. https://doi.org/10.1155/2022/2696916

[122] I. G. Prahmana and K. A. B. Sitepu, "Knearst Algorithm Analysis–Neighbor Breast Cancer Prediction Coimbra," *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, vol. 1, pp. 226-230, 2022. DOI: https://doi.org/10.53842/jaiea.v1i3.97

[123] A. Rasool, C. Bunterngchit, L. Tiejian, M. R. Islam, Q. Qu, and Q. Jiang, "Improved machine learning-based predictive models for breast cancer diagnosis," *International journal of environmental research and public health*, vol. 19, p. 3211, 2022. https://doi.org/10.3390/ijerph19063211

[124] M. A. Elsadig, A. Altigani, and H. T. Elshoush, "Breast cancer detection using machine learning approaches: a comparative study," *International Journal of Electrical & Computer Engineering (2088-8708)*, vol. 13, 2023. DOI: 10.11591/ijece.v13i1.pp736-745