International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

A Comprehensive Survey on Different Machine Learning Approaches for Breast Cancer Prediction based on Medical Imaging Modalities and Microarray Gene Expression.

Md. Faisal Bin Abdul Aziz[1], Azree Nazri[2,*], Razali Yaakob[3], Fatematuz Zuhura Evamoni[4], Teh Noranis Mohd Aris[5], and Zamberi Sekawi[6]

[1]Department of Computer Science, Faculty of Computer Science & Information Technology
Universiti Putra Malaysia
Selangor, Malaysia
Department of Computer Science and Engineering
Comilla University
Cumilla, Bangladesh
faisal@cou.ac.bd

[2]Department of Computer Science, Faculty of Computer Science & Information Technology
Universiti Putra Malaysia
Selangor, Malaysia
azree@upm.edu.my

[3]Department of Computer Science, Faculty of Computer Science & Information Technology
Universiti Putra Malaysia
Selangor, Malaysia
razaliy@upm.edu.my

[4]Department of Biology, Faculty of Science
Universiti Putra Malaysia
Selangor, Malaysia

[5]Department of Computer Science, Faculty of Computer Science & Information Technology
Universiti Putra Malaysia
Selangor, Malaysia
nuranis@upm.edu.my

[6]Department of Medical Microbiology, Faculty of Medicine and Health Sciences
Universiti Putra Malaysia
Selangor, Malaysia
zamberi@upm.edu.my

Abstract—Cancer is a complex global health problem that causes a high death rate. Breast cancer (BC) is the second most common death-causing disease in women worldwide. BC develops in the cells of the ducts or lobules of the glandular tissue when breast cells become uncontrollably proliferative. It can be controlled if diagnosed early enough. There are many techniques used to diagnose or classify BC. Machine learning (ML) has a significant effect on BC classification. This article provides a comparative study of different ML approaches for BC prediction based on medical imaging and microarray gene expression (MGE) data. DT, KNN, RF, SVM, Naïve Bayes, ANN, etc. perform much better in their respective fields. Another method named ensemble, incorporates more than one single classifier to solve the same problem. The study shows how ML with supervised, unsupervised, and ensemble learning might help with BC prognosis. This paper observes ensemble methods provide better performance than a single classifier. Finally, a comprehensive review of various imaging modalities and microarray gene expression, different datasets, performance metrics and outcomes, challenges, and prospective research directions are provided for the new researchers in this fast-growing field.

Keywords- Breast cancer; Medical imaging modalities; Microarray gene expression; Classification; Machine learning; Ensemble methods.

I. Introduction

Breast cancer (BC) is a complicated disease with a variety of histological and biochemical characteristics [1]. It occurs when breast cells develop out of control. BC is considered the second largest disease responsible for women's deaths [2]. According to epidemiological data, BC affects 50% of women aged 50 to 69 years old [3]. Lobules, ducts, and connective tissues are the three main components of the breast, and the disease occurs mostly in ducts and lobules. The disease can also spread outside the breast through the blood and lymph vessels [4]. BC is categorized into two groups: benign and malignant. A benign is not harmful to the body and causes mortality in humans very infrequently. On the other hand, a malignant is more harmful and can lead to death in people because the cells grow out of control [5]. In the United States, it is expected to occur with 1,918,030 and 609,360 new cases and deaths, respectively, in 2022 [6]. As a result, early detection of BC is absolutely essential for a better prognosis. Although the symptoms are mild in the beginning, the chances of survival increase significantly if diagnosed early [7]. BC is identified using conventional methods such as breast self-examination, blood tests, mammograms, X-rays, and biopsy, but these techniques are time-consuming and prone to human error [8].

Medical imaging (MI) is the most efficient way to diagnose, treat, and detection of problems in the breast with the regular help of image processing, computer vision, and ML [9]. Microarrays can concurrently measure the expression levels of thousands of genes. MGE data can be helpful to support medical decision-making for a specific sick person, e.g. in oncology for classification purposes [10]. Clinical mistakes are the third-leading reason of death in the USA [11]. By 2040, the burden of breast cancer is expected to rise to more than 3M new cases and 1M deaths annually as a consequence of population growth and aging alone [12]. So, early BC diagnosis and treatment raise the chances of a cure while also decreasing mortality and the probability of recurrence [13].

ML is a subdivision of Artificial Intelligence (AI). ML is one of the best widely used models to quickly train machines and build prognostic models for effective decision-making. ML uses statistical approaches and algorithms to construct data models that can acquire and adjust knowledge to classify without human involvement [7]. Several ML classifiers are applied to microarray gene expression data and medical imaging to predict BC. Fig. 1 depicts the taxonomy that was taken into account in this review. The success of machine learning has increased over time due to the increase in computing power, despite the growth in data volume [9]. As a result, a reliable approach for predicting BC is needed, and ML algorithms are widely applied for BC classification. According to "pubmed.gov" database, we found a graphical representation that is shown in Fig. 2 by searching for BC and ML related research from the last ten years that showing the application of ML models is increasing in the research of BC day by day. We hope that the presented review provides a helpful resource for researchers who wish to conduct research on microarray gene expression and medical imaging pivoting to ML-based BC identification. We chose prominent studies from 2018 to 2023 according to popularity to conduct this review. The summary of the differences between existing studies and our review is displayed in Table I.

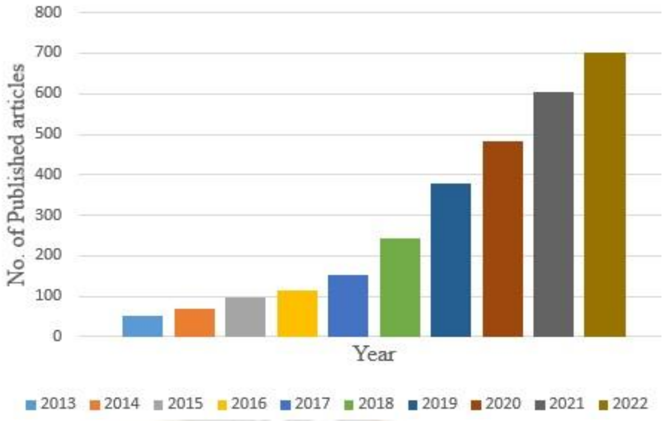


Figure 2. Published articles related to BC and ML.

The overall contributions of this study is summarized as follows:

1. We present a taxonomy of BC prediction based on ML.
2. In this study, we focus on different medical imaging datasets and their repositories for new researchers.
3. We describe MGE for BC diagnosis with some sources of public datasets. Some researchers used either medical images or gene expression data, but we reviewed both.
4. We explore a variety of popular ML techniques and represent a taxonomy by taking the distinctions in ML tasks into account. Here, we divide the techniques into 3 major types, such as (i) supervised learning, (ii) unsupervised learning, and (iii) reinforcement.
5. We discuss the dimensionality reduction technique for both linear and non-linear data.
6. We mention several important evaluation metrics that used to evaluate the performance of ML-based classification models.
7. We have summarized the results according to different ML classifiers to assist new researchers as well as academicians in the classification of BC.

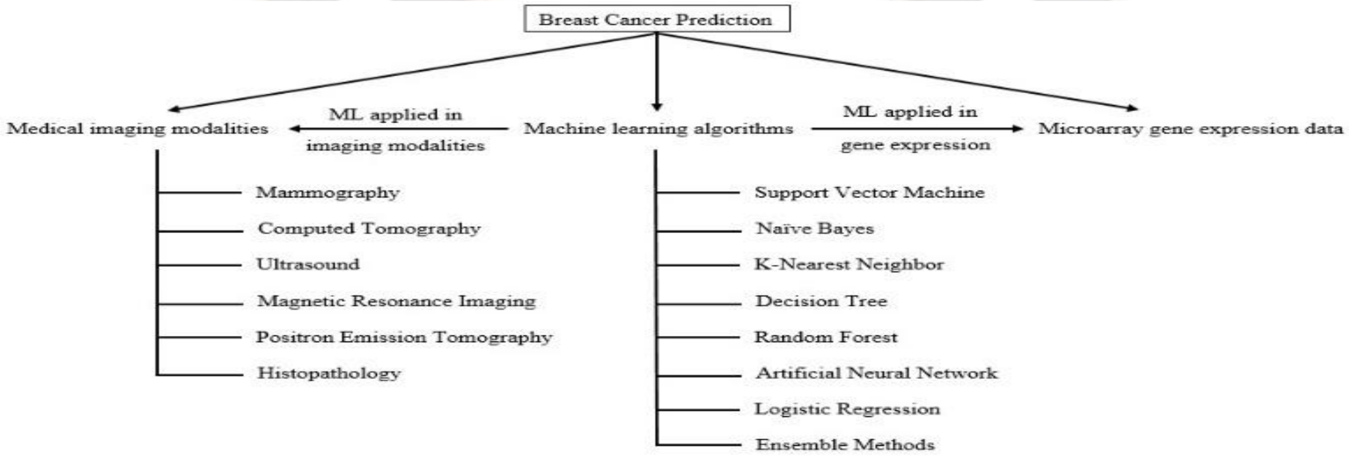


Figure 1. A comparative analysis of existing studies on the basis of the classification of breast cancer

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

TABLE I. A comparative analysis of existing studies on the basis of the classification of breast cancer

Study								ML Classifiers							
Ref.	**Year**	**Taxonomy**	**Imaging Modalities**	**Microarray gene expression**	**Datasets**	**Performance matrices**	**Challenges**	**SVM**	**RF**	**DT**	**ANN**	**Ensemble**	**kNN**	**NB**	**LR**
[14]	2023	×	√	×	√	√	√	√	√	×	√	×	×	×	×
[15]	2023	×	×	√	√	√	√	√	×	×	√	×	√	×	×
[16]	2022	×	×	√	√	√	×	√	√	×	√	×	√	√	√
[17]	2022	×	√	×	√	√	×	√	×	×	×	×	√	×	×
[18]	2022	×	√	×	√	×	√	√	×	√	√	×	×	×	√
[7]	2022	×	√	×	√	×	√	√	×	√	×	×	√	×	√
[19]	2022	×	√	×	√	√	√	√	×	√	√	×	√	√	√
[13]	2021	×	×	√	√	√	×	√	×	×	×	×	√	√	×
[20]	2021	×	×	√	√	√	×	√	√	×	√	×	×	×	√
[21]	2021	×	√	×	√	×	√	√	×	√	√	√	×	√	×
[22]	2021	×	√	×	√	×	√	√	√	√	×	×	√	√	√
[23]	2020	×	×	√	×	√	×	√	×	×	√	×	√	×	×
[24]	2019	×	×	√	×	√	√	×	×	×	√	×	×	×	×
[25]	2018	×	√	×	√	√	×	√	×	√	√	√	√	×	×
Ours	-	√	√	√	√	√	√	√	√	√	√	√	√	√	√

The remaining study is structured as follows: Section II overview of BC. Raw materials for breast cancer diagnosis and prediction in section III. In section IV, we explain ML approaches for BC classification in detail. We present a literature study in section V. Assessment Metrics for Medical Imaging modalities and Microarray Data Classification displayed in section VI, followed by discussion and challenges in section VII. Finally, this paper concludes in Section VIII with future prospects.

II. RAW MATERIALS FOR BREAST CANCER DIAGNOSIS AND PREDICTION

A. *Medical imaging (MI)*

Define MI refers to particular techniques that are used to examine different organs of the body in order to diagnose, observe, or treat diseases. Each technology delivers precise information regarding the part of the body being observed or treated, the sickness, injury, or the effectiveness of the medical assistance. MI plays a vital role in the diagnosis, treatment, and detection of problems in the breast with the help of image processing, computer vision, and ML [9]. MI is the most efficient way to diagnose primary-stage BC with the regular help of several imaging modalities such as mammography, ultrasonography, MRI, CT scan, PET, biopsy, and duplex ultrasound [26]. Some popular and publicly available BC datasets of MI modalities are shown in Table II.

1) Mammography

Mammography using low-energy X-rays is the best operative tool for detecting BC [14]. The screening tool helped diagnose lethal cancers earlier, improving prognoses and reducing death rates by up to 50%. Despite its good consequences, mammography has drawbacks. All US-screened females will have at least one false positive. An analysis of the advantages and risks of mammography found that 200 out of every 1000 women who get a mammogram every two years will get a false positive result. A false positive has negative effects on the mental health of women, causing unnecessary stress [18]. Mammography is inexpensive, easy, rapid, and commonly used as a screening diagnostic for breast cancer because it can detect even tiny changes in the breast that can't be seen with the naked eye [27].

2) Ultrasound (US)

US is used to diagnose BC due to its non-invasive, best-tolerated, and radiation-free nature. Mammography is typically ineffective at detecting BC in dense breast tissue, whereas US is a vastly effective analytic tool for detecting these cancers and also helps radiologists to analyze breast ultrasounds [28]. US imaging is cheaper and more portable than MRI and mammography [29].

3) Magnetic Resonance Imaging (MRI)

MRI generates images of the internal organs of the body, including the breasts, lungs, liver, and bones, using magnets and radio waves. The absence of radiation in an MRI makes it a superior investigation. Breast MRI images offer more detailed information on soft breast tissue than mammography, US, and CT scan. MRI datasets are not freely accessible, so few research have classified BC using MRI [30]. Xu et al. [31] implemented an ML-based model to evaluate preoperative clinical and MRI features of lymphovascular invasion (LVI) in invasive BC (IBC) to help make decisions about treatment and predict the outcome. Finally, XGBoost obtained an AUC of 0.832 in the training dataset and 0.838 in the validation dataset.

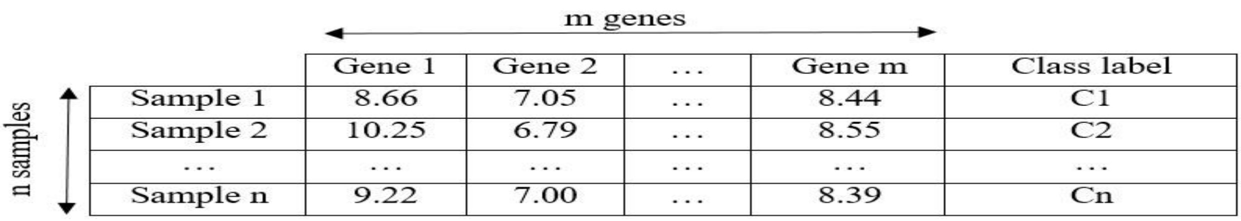
B. *Microarray gene expression (MGE)*

Microarrays (MAs) are a tool that archives gene expression from DNA or RNA. This system displays interesting features, such as producing high-dimensional data from a small sample size [15]. Fig. 3 describes a symbolic n × (m+1) matrix representation of MGE data, with n rows representing samples and m columns representing genes. It can be implemented in different kinds of research including epigenetics, genotyping, translation profiling, and gene expression profiling. Different steps are involved to analyze gene expression data. Feature extraction is the procedure of transforming the scanned microarray image into computable values that are stored in binary (.CEL) or text format and annotating it with the basic gene information. A package ‘limma’ that is implemented to find differentially expressed genes integrates a technique to accurate for various testing.

Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

TABLE II. Publicly available dataset of MI modalities for BC

Datasets	**Types for dataset**	**URL**
BCDR [32], [33]	Mammography and ultrasound images	https://bcdr.ceta-ciemat.es/
BUSI [34]	Ultrasound	https://scholar.cu.edu.eg/?q=afahmy/pages/dataset
DCE-MRI [35]	MRI	https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=70226903
Private [36]	CT scan	NA
WBCD [37]	Digitized image of a fine needle aspirate (FNA) of a breast mass.	https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29
WDBC [37]	Digitized image of a fine needle aspirate (FNA) of a breast mass.	https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic) or https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data
BreakHis [38]	Histopathological Image	https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/
MIAS [39]	Mammographic image	https://www.repository.cam.ac.uk/handle/1810/250394
DDSM [40]	Mammographic image	https://www.mammoimage.org/databases/


m genes
n samples

	Gene 1	Gene 2	...	Gene m	Class label
Sample 1	8.66	7.05	...	8.44	C1
Sample 2	10.25	6.79	...	8.55	C2
...	...	...	...	...	...
Sample n	9.22	7.00	...	8.39	Cn

Figure 3. A typical MGE matrix where rows denote the samples (different condition like several cells, growing stages and treatments), and the columns represent genes (usually genes of the whole genome). The final column is the class label i.e. the information on the sample going to which group.

This technique forms a log2 fold change ratio between the test and control state and an ‘adjusted’ p-value that evaluates the significance of the difference [41]. We obtained the most significant genes from microarray dataset using a threshold value like p-value < 0.05 or 0.01. Each gene or transcript is denoted on the GeneChip by 11 probe sets.

1) *Datasets of Microarray Gene Expression*

There are a lot of biological data on BC is available in online repositories but relatively fewer datasets of MAs are related to ML. There are 2 most popular databases named NCBI (https://www.ncbi.nlm.nih.gov/) and ArrayExpress (https://www.ebi.ac.uk/biostudies/arrayexpress). Another famous online repository named Kent Ridge Biomedical Data Set Repository for different diseases datasets such as BC, central nervous system (CNS), colon cancer, leukemia, etc. based on gene expression (https://leo.ugr.es/elvira/DBCRepository/). Publicly available datasets of microarray gene expression stored in Table III. The Curated Microarray Database (CuMiDa) [42] comprises 78 cancer microarray datasets extracted for machine learning from 30,000 studies of Gene Expression Omnibus (GEO) using several filtering processes.

C. Feature selection and extraction (FSE)

Feature selection (FS) preserves a subset of the features, whereas feature extraction (FE) techniques transfer the data to a new feature set. FE is an important technique for extracting appropriate features from input data. FSE has a significant effect on image processing, data mining, ML, and bioinformatics [43]. FS is distributed into three parts: filters, wrappers, and hybrid methods. Filters that retrieve features from the input data without learning, and wrappers [44] that are based on learning methods to estimate which features are beneficial.

The hybrid strategy is a combination of filter and wrapper techniques. First, the filter technique is used to reduce the feature space, then the wrapper technique is used to select feature subsets. Fig. 4, displayed the overall process of BC classification from input to outcome. The most used dimensionality reduction approaches are principal component analysis (PCA) [45] for linear data and t-Distributed Stochastic Neighbor Embedding (t-SNE) for non-linear data. t-SNE is an unsupervised, non-linear approach mostly applied to data mining and the visualization of high-dimensional data into 2D or 3D. PCA is extensively used in bioinformatics, image processing, speech processing, and NLP. The t-SNE technique produces better outcomes than PCA and other linear dimensionality reduction approaches [46]. Extracted features (EF) are obtained from specific datasets using PCA, t-SNE, and other algorithms [47]. EF plays an important role in classification and clustering.

III. MACHINE LEARNING (ML) APPROACHES FOR BC CLASSIFICATION

Artificial intelligence (AI) has a branch called ML that is used to develop and test algorithms for prediction, pattern recognition, and classification [48]. Three main steps of ML are preprocessing, feature selection or extraction, and prediction, which can be used to provide a prognosis for BC. ML techniques are classified into 3 major groups: supervised, unsupervised, and reinforcement learning. The most common ML algorithms to predict BC are as follows:

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

A. Logistic regression (LR)

A SL model that is extensively implemented for fraud detection and clinical tests. It has more dependent variables. It has a popular choice for modeling and major advantage of accepting binary responses [49].

TABLE III. Datasets of microarray gene expression for BC classification

Microarray datasets	**GEO accession**	**No. of Genes**	**No. of Samples**	**URL**
Breast & Colon cancer	GSE3726	22283	52	https://file.biolab.si/biolab/supp/bi-cancer/projections/info/BC_CCGSE3726_frozen.html
Breast cancer	GSE349_350	12625	24	https://file.biolab.si/biolab/supp/bi-cancer/projections/info/BCGSE349_350.html
Breast cancer	GSE33447	36623	16	https://sbcb.inf.ufrgs.br/cumida [42]
Breast cancer	-	24481	78	https://leo.ugr.es/elvira/DBCRepository/BreastCancer/BreastCancer.html [50]
NKI Breast Cancer Data	-	1570	272	https://data.world/datasets/breast-cancer
Breast cancer	-	24481	97	https://csse.szu.edu.cn/staff/zhuzx/Datasets.html
^{2}BC-TCGA [51]	-	17814	590	https:// data.mendeley.com/datasets/v3cc2p38hb/1
Breast cancer	GSE2034	12634	286	https:// data.mendeley.com/datasets/v3cc2p38hb/1

In paper [52], author used modified LR to analyze microarray gene expression for the classification of BC. Hypothesis function for Logistic Regression:

$$h_\theta(x) = g(\theta^T x) \quad (1)$$

where $g(z) = \frac{1}{1+e^{-z}}$ *and* $z \in \mathbb{R}$

B. Decision tree (DT)

DT is a very famous SL that employs a tree-structured classification system with nodes representing input factors as well as leaves indicating decision outcomes. DTs are prominent and extensively used ML techniques for regression and classification of BC [53]. The basic learning strategy of DT is constructed on the divide and conquer method. The decisions are grounded on some circumstances and are easy to infer with the highest accuracy [5]. The DT technique includes classification and regression tree (CART), C4.5, C5.0 and conditional tree [54].

C. Random forest (RF)

A collection of tree-type classifiers known as the RF algorithm was first presented in a study by Breiman [55]. The RF algorithm is only capable of identifying significant features, not redundant ones [56]. RF was used to reveal that genetic expression of the androgen receptor pathway may be used to molecularly identify the two kinds of BC [57].

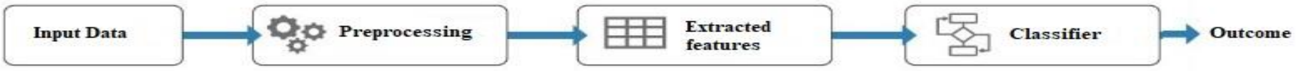


Figure 4. Diagram for applying feature extraction to input data for a ML classifier.

D. Support Vector Machine (SVM)

SVM is an SL approach that can handle regression and classification related tasks[58]. A special kernel function, like a polynomial or radial basis function (RBF), must be specified while developing the SVM classifier because it is a vital learning element [59]. It has the highest accuracy rate when it comes to large dataset prediction. Fig. 5, SVM used the highest number among different ML classifiers from our survey of BC research.

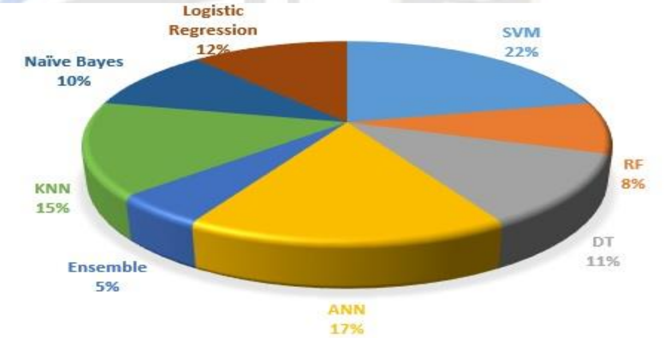


Figure 5. ML algorithms aare re used in different studied research.

E. Naïve Bayes (NB)

NB classifier is prominent for its simplicity as well efficiency. It is a probabilistic classifier and based upon the desired class, it learns the probabilities of features [5]. It achieves the highest accuracy while evaluating the probability of large datasets as well as noisy data as an input [60]. The author analyzed RNA-seq data that contained 110 triple-negative and 992 non-triple-negative samples to categorize BC types using the NB classifier [61].

F. Artificial Neural Network (ANN)

ANN is a popular data mining technique. A neural network is made up of 3 layers: input, a hidden, and an output layer. An ANN algorithm is utilized to extract the important patterns that are so complicated. ANNs operate for a variety of classification or pattern recognition purposes [53]. Researchers using ANN and DT classifiers on the WBCD dataset obtained an accuracy of 98.55% for BC classification [62].

G. K-Nearest Neighbor (KNN)

In KNNs, k is applied to designate the no. of nearest neighbors that must be incorporated into the voting processes. In KNN, parameter tuning is completed by selecting a suitable value of k to improve performance. The similarity of two points is evaluated by, e.g., the Euclidean distance [5]. It is used in pattern recognition and well-known approach for BC classification. In paper [61], the author tested RNA-seq data of BC samples to classify BC types using the KNN classifier.

H. K-means clustering

K-means is the simplest clustering technique and commonly used unsupervised ML methods. It is used to decide the degree of similarity between two or more data points. At least one cluster makes up each data point, making it ideal for analyzing big datasets [63]. According to Dey and Mukhopadhyay, they implemented a model using a Particle Swarm Optimization (PSO)-based K-means clustering technique for clustering MGE data.

I. Ensemble Learning

Ensemble learning is an approach that combines several base models to get better outcomes. Ensemble methods typically produce outcomes that are more accurate than those of a single model [64]. The accuracy and variety of the base learners have an important role in the success of ensemble methods [65]. Three major types of ensemble methods are i) bagging, ii) boosting, and iii) stacking.

IV. LITERATURE STUDY

Rajaguru and Chakravarthy [66] analyzed the performance of the classification of BC. The authors used DT and KNN classifiers to identify either benign or malignant using WBCD datasets after feature selection using PCA. Then finally, results indicated that the KNN provided better than the DT for the classification of BC.

Based on the digital mammogram images, the authors implemented a model to classify BC using an ML classifier with the MIAS dataset. The researchers split the dataset by 70% and 30% for training and testing purposes. Finally, SVM classified the normal and abnormal mammogram images with 100% accuracy [67].

Loey et al. [68] proposed an intelligent decision support system (IDSS) using gene expression profiles obtained from DNA microarrays for early BC detection. Information gain (IG) was used to identify the informative genes from the dataset. They used the grey wolf optimization (GWO) technique to reduce the number of particular features. Lastly, the authors employed an SVM model for cancer prediction.

Yu et al. [69] explained an RNA-seq-based BC prediction using ML. The authors obtained the significant genes from differentially expressed genes (DEGs) using gene ontology (GO) enrichment. They got better result using weighted DEGs for different performance metrics.

A deep learning method was suggested by Danaee et al. [70] to detect critical genes for the prediction of BC. They extracted functional genes from high-dimensional gene expression profiles by the Stacked Denoising Autoencoder (SDAE). The performance of the extracted features was assessed through 5-fold CV (cross-validation) and supervised classification models such as a single-layer ANN and both a linear kernel SVM and an SVM with a radial basis function kernel (SVM-RBF) to categorize benign or malignant. Lastly, when SDAE features were utilized, the SVM-RBF classifier yielded the highest accuracy.

LR-based model was developed by Morais-Rodrigues et al. [52] to predict BC using MGE data like series GSE65194, GSE20711, and GSE25055 from the Gene Expression Omnibus (GEO). A minimum of 80% performance was achieved in classification (sensitivity and specificity).

To classify BC, Ragab et al. [71] suggested a deep convolutional neural network (DCNN). They achieved it by developing four separate experiments and evaluating results with two datasets: CBIS-DDSM and MIAS. The experiment deployed pre-trained, fine-tuned DCNN to improve the classification accuracy.

The paper [72] developed an ensemble method using AdaBoost (PCA-AE-Ada) to better classify BC using five gene expression datasets. Finally, they compared the proposed PCA-AE-Ada classifier to the created classifier PCA-Ada and found that PCA-AE-Ada performed better.

An ensemble based system established by Mahesh and Mohan Kumar [73] for BC classification on Coimbra dataset. The system extracted features, then created an ensemble ML model using Naïve Bayes, RBF Neural Network, and LDA. Finally, they are labelled as benign or malignant. The used method recorded a 75.86% accuracy where NB, RBFNN, and LDA obtained 62.06%, 69%, and 58.17%, respectively.

The CWV-BANN-SVM model was established by Abdar and Makarenkov [74] and incorporates the SVM and boosting artificial neural network (BANN) methods with ensemble confidence-weighted voting (CWV). The authors discovered that this model was effective in detecting BC and that the model had served better performance by equally splitting the data into train and test sets.

Bhardwaj et al. [75] analyzed WDCB data for BC classification. The authors applied different classifiers, mentioning multilayer perceptron, KNN, genetic programming, and RF to the dataset. Among the classifiers, RF showed the best result that was obtained; the accuracy was 96.24%.

V. ASSESSMENT METRICS FOR MEDICAL IMAGES AND MICROARRAY DATA CLASSIFICATION

There are many performance matrices involved in BC classification as follows:

Accuracy: The proportion of accurate predictions to all predictions is calculated. The accuracy can be expressed by equation (2). Sensitivity: The true positive rate. It is evaluated by equation (3). Specificity: The false positive rate. The specificity can be defined by equation (4). Precision: It refers to the no. of true positives divided by the overall no. of positive predictions. It can be computed by equation (5). Recall: a probabilistic portion to decide if an actual positive case is appropriately categorized with the positive class and defined by equation (6). F-measure / F score: Evaluated as the geometric mean of precision as well as recall. It can be expressed by equation (7). Table IV represents a confusion matrix that can be used to visualize a classifier's performance.

TABLE IV. Confusion matrix

		Predicted Class	
		P	N
Actual Class	P	True Positive (TP)	False Negative (FN)
	N	False Positive (FP)	True Negative (TN)

Accuracy = (TP+TN) / (TP+FN+FP+TN) (2)

Sensitivity = TP / (TP+FN) (3)

Specificity = FP / (FP+TN) (4)

Precision = TP / (TP+FP) (5)

Recall = TP / (TP+FN) (6)

F-measure = 2 × Precision × Recall / (Precision + Recall) (7)

Matthews correlation coefficient (MCC) = ((TP × TN) – (FP × FN)) / √((TP+FP)(TP+FN)(TN+FP)(TN+FN)) (8)

ROC curve = Sensitivity *vs* 1 – Specificity.

VI. Discussion and Challenges

We discuss different critical analyses regarding MI modalities and MGE data from BC research. Table V compares ML approaches for the prediction of BC in different contexts based on data sources, classification methods, and results of each algorithm. The results are evaluated using various datasets with several outputs such as accuracy, precision, recall, sensitivity, specificity, F-measure, and so on. One challenge of microarray data contains huge number of genes (more than 50K), but their sample size is small. So, the curse of dimensionality occurs in MAs data and needs to handle carefully to obtain the most significant features for a better outcome. Another challenge of MA data is imbalance which may occur poor accuracy. In paper [74], an ensemble based CWV-BANN-SVM achieved 100% accuracy, whereas ANN and SVM gave 97.365% and 99.70%, respectively. In a systematic review [87], the highest level of accuracy using medical imaging data was 99.3%, while the most accurate performance using gene expression data was 99.8%.

TABLE V: Summary of recent research on BC classification using ML.

Ref.	Datasets	Image modalities/ Gene expression	Classification methods	Results (%)
[76]	Ultrasound Images [77]	US	SVM	93.08 (Accuracy) 0.9712 (AUC)
[78]	GSE45255 GSE15852	Gene expression	SVM ET RF	97.78 (Accuracy) 93.33 (Accuracy) 93.33 (Accuracy)
[79]	WBCD [80]	Blood analysis	RF and ET based ensemble	100 (Accuracy) 100 (Sensitivity) 100 (Specificity)
[68]	BC [81]	Gene expression	SVM	94.87 (Accuracy) 0.95 (Recall) 0.90 (Precision) 0.92 (F1)
[82]	BRCA	Gene expression	SVM ANN KNN DT RF NB DISCR	74.9094 ± 0.48 (Accuracy) 74.9094 ± 0.37 (Accuracy) 67.1014 ± 0.35 (Accuracy) 64.4565 ± 0.46 (Accuracy) 76.5761 ± 0.33 (Accuracy) 70.5978 ± 0.35 (Accuracy) 73.1884 ± 0.35 (Accuracy)
[52]	GSE65194, GSE20711, GSE25055 from NCBI	Gene expression	Modified logistic regression	80 (Sensitivity) 80 (Specificity)
[71]	CBIS-DDSM^{5} [83] -------------------- MIAS [39]	Mammography	SVM	97.90 (Accuracy) ………………… 97.40 (Accuracy)
[84]	BRCA RNA-seq [85]	Gene expression	Ensemble	98.41 (Accuracy) 97.77 (Precision) 97.20 (Recall)
[73]	Coimbra dataset [86]	Blood analysis	NB RBFNN LDA Ensemble	62.06 (Accuracy) 69 (Accuracy) 58.17 (Accuracy) 75.86 (Accuracy)
[74]	WBCD [80]	Blood analysis	ANN SVM Ensemble (CWV-BANN-SVM)	97.365 (Accuracy) 99.707 (Accuracy) 100 (Accuracy)

VII. CONCLUSIONS

BC is one of the world's most severe and lethal cancers, ranking in the top two. ML classifiers have recently become more and more prominent in BC prediction due to their high accuracy. We have summarized the state of the art of ML in BC diagnosis and prognosis for both MI modalities and gene expression data. Among all the classifiers studied, researchers used SVM at its maximum to classify BC. Early diagnosis of BC is the most significant issue for a healthy life, and early BC treatment increases the chances of a cure while also lowering the mortality rate. Early detection of BC has the potential to save more lives. According to prior studies, ML techniques provide better results in their respective fields. We reviewed studies that use gene expression and imaging data to detect or categorize BC. Both strategies were effective. Numerous imaging features were easily extracted, although not all were efficient but gene expression data may be more effective despite having fewer features or samples. In ensemble methods, every classifier showed their best performance and finally combined their outcomes to deliver the best result. This review concluded that the ensemble method deals with better outcomes than other single-classification algorithms. Due to harmful ionizing radiation in medical imaging such as CT scan, mammography, and X-rays, microarray gene expression data are superior for classifying BC. In this study, we presented an approach to systematic reviews that will assist the next generation of researchers in determining the general framework of ML-based BC classification. It will be most helpful to new and young researchers regarding BC prediction using ML using medical imaging modalities and microarray gene expression data.

ACKNOWLEDGMENT

The authors are thankful to the Universiti Putra Malaysia, Malaysia; Bangabandhu Science and Technology Fellowship Trust under the Ministry of Science and Technology, Bangladesh and Comilla University, Bangladesh for their support in conducting this research.

REFERENCES

[1] S. Wang *et al.*, “Studying the pathological and biochemical features in breast cancer progression by confocal Raman microspectral imaging of excised tissue samples,” *J. Photochem. Photobiol. B Biol.*, vol. 222, p. 112280, 2021, doi: 10.1016/j.jphotobiol.2021.112280.

[2] G. L. Wong, S. G. Manore, D. L. Doheny, and H. W. Lo, “STAT family of transcription factors in breast cancer: Pathogenesis and therapeutic opportunities and challenges,” *Semin. Cancer Biol.*, Aug. 2022, doi: 10.1016/j.semcancer.2022.08.003.

[3] M. Kamińska, T. Ciszewski, K. Łopacka-Szatan, P. Miotła, and E. Starosławska, “Breast cancer risk factors,” *Prz. Menopauzalny*, vol. 14, no. 3, pp. 196–202, 2015.

[4] CDC, “Breast Cancer.” https://www.cdc.gov/cancer/breast/basic_info/what-is-breast-cancer.htm (accessed Jan. 10, 2023).

[5] A. S. Assiri, S. Nazir, and S. A. Velastin, “Breast Tumor Classification Using an Ensemble Machine Learning Method,” *J. Imaging*, vol. 6, pp. 1–13, 2020.

[6] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, “Cancer statistics, 2022,” *CA. Cancer J. Clin.*, vol. 72, no. 1, pp. 7–33, 2022.

[7] Y. Amethiya, P. Pipariya, S. Patel, and M. Shah, “Comparative analysis of breast cancer detection using machine learning and biosensors,” *Intell. Med.*, vol. 2, no. 2, pp. 69–81, May 2022.

[8] M. D. Purbolaksono, K. C. Widiastuti, M. S. Mubarok, Adiwijaya, and F. A. Ma’ruf, “Implementation of mutual information and bayes theorem for classification microarray data,” in *Journal of Physics: Conference Series*, 2018, vol. 971, no. 1, pp. 1–8.

[9] E. H. Houssein, M. M. Emam, A. A. Ali, and P. N. Suganthan, “Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review,” *Expert Syst. Appl.*, vol. 167, no. October 2020, pp. 1–20, 2021.

[10] N. Pochet, F. De Smet, J. A. K. Suykens, and B. L. R. De Moor, “Systematic benchmarking of microarray data classification: Assessing the role of non-linearity and dimensionality reduction,” *Bioinformatics*, vol. 20, no. 17, pp. 3185–3195, 2004.

[11] M. A. Makary and M. Daniel, “Medical error-the third leading cause of death in the US,” *BMJ*, vol. 353, 2016, doi: 10.1136/bmj.i2139.

[12] M. Arnold *et al.*, “Current and future burden of breast cancer: Global statistics for 2020 and 2040,” *The Breast*, vol. 66, pp. 15–23, Dec. 2022.

[13] M. Abd-Elnaby, M. Alfonse, and M. Roushdy, “Classification of Breast Cancer Using Microarray Gene Expression Data: A Survey,” *J. Biomed. Inform.*, vol. 117, p. 103764, 2021.

[14] K. Loizidou, R. Elia, and C. Pitris, “Computer-aided breast cancer detection and classification in mammography : A comprehensive review,” *Comput. Biol. Med.*, vol. 153, no. January, p. 106554, 2023.

[15] S. Osama, H. Shaban, and A. A. Ali, “Gene reduction and machine learning algorithms for cancer classification based on microarray gene expression data: A comprehensive review,” *Expert Syst. Appl.*, vol. 213, p. 118946, 2023.

[16] N. Mohd Ali, R. Besar, and N. A. Nor, “Hybrid Feature Selection of Breast Cancer Gene Expression Microarray Data Based on Metaheuristic Methods: A Comprehensive Review,” *Symmetry (Basel).*, vol. 14, no. 10, p. 1955, 2022, doi: 10.3390/sym14101955.

[17] D. Painuli, S. Bhardwaj, and U. köse, “Recent advancement in cancer diagnosis using machine learning and deep learning techniques: A comprehensive review,” *Comput. Biol. Med.*, vol. 146, p. 105580, Jul. 2022,.

[18] J. Mendes, J. Domingues, H. Aidos, N. Garcia, and N. Matela, “AI in Breast Cancer Imaging: A Survey of Different Applications,” *J. Imaging*, vol. 8, p. 228, 2022, doi: 10.3390/jimaging8090228.

[19] A. Mashekova, Y. Zhao, E. Y. K. Ng, V. Zarikas, S. C. Fok, and O. Mukhmetov, “Early detection of the breast cancer using infrared technology – A comprehensive

review," *Therm. Sci. Eng. Prog.*, vol. 27, p. 101142, 2022, doi: 10.1016/j.tsep.2021.101142.

[20] S. Albaradei *et al.*, "Machine learning and deep learning methods that use omics data for metastasis prediction," *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 5008–5018, 2021.

[21] S. R. Sannasi Chakravarthy and H. Rajaguru, "A Systematic Review on Screening, Examining and Classification of Breast Cancer," 2021, doi: 10.1109/STCR51658.2021.9588828.

[22] M. Tariq, S. Iqbal, H. Ayesha, I. Abbas, K. T. Ahmad, and M. F. K. Niazi, "Medical image based breast cancer diagnosis: State of the art and future directions," *Expert Syst. Appl.*, vol. 167, p. 114095, 2021.

[23] M. A. Hambali, T. O. Oladele, and K. S. Adewole, "Microarray cancer feature selection: Review, challenges and research directions," *Int. J. Cogn. Comput. Eng.*, vol. 1, no. October, pp. 78–97, 2020.

[24] M. Daoud and M. Mayo, "A survey of neural network-based cancer prediction models from microarray data," *Artif. Intell. Med.*, vol. 97, no. January, pp. 204–214, 2019.

[25] N. I. R. Yassin, S. Omran, E. M. F. El Houby, and H. Allam, "Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review," *Comput. Methods Programs Biomed.*, vol. 156, pp. 25–45, 2018.

[26] D. Simos *et al.*, "Imaging for distant metastases in women with early-stage breast cancer: A population-based cohort study," *Cmaj*, vol. 187, no. 12, pp. E387–E397, 2015.

[27] H. Li, S. Zhang, Q. Wang, and R. Zhu, "Clinical value of mammography in diagnosis and identification of breast mass," *Pakistan J. Med. Sci.*, vol. 32, no. 4, 2016, doi: 10.12669/pjms.324.9384.

[28] K. Jabeen *et al.*, "Breast Cancer Classification from Ultrasound Images Using Probability-Based Optimal Deep Learning Feature Fusion," *Sensors*, vol. 22, no. 3, p. 807, 2022, doi: 10.3390/s22030807.

[29] M. Byra, "Breast mass classification with transfer learning based on scaling of deep representations," *Biomed. Signal Process. Control*, vol. 69, 2021, doi: 10.1016/j.bspc.2021.102828.

[30] M. F. Mridha *et al.*, "A comprehensive survey on deep-learning-based breast cancer diagnosis," *Cancers (Basel).*, vol. 13, no. 23, p. 6116, 2021, doi: 10.3390/cancers13236116.

[31] Z. Xu *et al.*, "Using Machine Learning Methods to Assess Lymphovascular Invasion and Survival in Breast Cancer: Performance of Combining Preoperative Clinical and MRI Characteristics.," *J. Magn. Reson. Imaging*, Feb. 2023, doi: 10.1002/jmri.28647.

[32] G. Murtaza *et al.*, "Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges," *Artif. Intell. Rev.*, vol. 53, no. 3, pp. 1655–1720, 2020.

[33] D. C. Moura and M. A. Guevara López, "An evaluation of image descriptors combined with clinical data for breast cancer diagnosis," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 8, no. 4, pp. 561–574, 2013, doi: 10.1007/s11548-013-0838-2.

[34] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data Br.*, vol. 28, 2020, doi: 10.1016/j.dib.2019.104863.

[35] A. Saha *et al.*, "A machine learning approach to radiogenomics of breast cancer: A study of 922 subjects and 529 dce-mri features," *Br. J. Cancer*, vol. 119, no. 4, pp. 508–516, 2018.

[36] J. Koh, Y. Yoon, S. Kim, K. Han, and E. K. Kim, "Deep Learning for the Detection of Breast Cancers on Chest Computed Tomography," *Clin. Breast Cancer*, vol. 22, no. 1, pp. 26–31, 2022.

[37] W. H. Wolberg and O. L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 87, no. 23, pp. 9193–9196, 1990.

[38] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A Dataset for Breast Cancer Histopathological Image Classification," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1455–1462, 2016.

[39] J. Suckling *et al.*, "Mammographic Image Analysis Society (MIAS) database v1.21," 2015. https://www.repository.cam.ac.uk/handle/1810/250394 (accessed Dec. 24, 2022).

[40] J. E. E. Oliveira, M. O. Gueld, A. de A. Araújo, B. Ott, and T. M. Deserno, "Toward a standard reference database for computer-aided mammography," *Med. Imaging 2008 Comput. Diagnosis*, vol. 6915, p. 69151Y, 2008, doi: 10.1117/12.770325.

[41] "Analysis of microarray data." https://www.ebi.ac.uk/training/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/microarrays/analysis-of-microarray-data/ (accessed Dec. 21, 2022).

[42] B. C. Feltes, E. B. Chandelier, B. I. Grisci, and M. Dorn, "CuMiDa: An Extensively Curated Microarray Database for Benchmarking and Testing of Machine Learning Approaches in Cancer Research.," *J. Comput. Biol. a J. Comput. Mol. cell Biol.*, vol. 26, no. 4, pp. 376–386, Apr. 2019.

[43] A. O. Salau and S. Jain, "Feature Extraction: A Survey of the Types, Techniques, Applications," in *2019 International Conference on Signal Processing and Communication, ICSC 2019*, 2019, pp. 158–164.

[44] M. S. Iqbal, W. Ahmad, R. Alizadehsani, S. Hussain, and R. Rehman, "Breast Cancer Dataset, Classification and Detection Using Deep Learning," *Healthc.*, vol. 10, no. 12, p. 2395, 2022.

[45] M. A. Jawad and F. Khurshid, "Deep and Dense Convolutional Neural Network (D^2 CN^2 for Multi Category Classification of Magnification Specific and Magnification Independent Breast Cancer Histopathological Images," *SSRN Electron. J.*, 2022, doi: 10.2139/ssrn.4017036.

[46] Saurabh, "Comprehensive Guide on t-SNE algorithm with implementation in R & Python.," 2022. https://www.analyticsvidhya.com/blog/2017/01/t-sne-

implementation-r-python/ (accessed Apr. 05, 2023).

[47] L. Yang and Z. Xu, "Feature extraction by PCA and diagnosis of breast tumors using SVM with DE-based parameter tuning," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 3, pp. 591–601, 2019.

[48] S. H. Giordano, "Breast cancer in men," *N. Engl. J. Med.*, vol. 378, no. 24, pp. 2311–2320, 2018.

[49] H. Tran, "A Survey of Machine Learning and Data Mining Techniques used in Multimedia System," *A Prepr.*, pp. 1–30, 2019.

[50] L. J. Van't Veer *et al.*, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[51] H. Xie, J. Li, T. Jatkoe, and C. Hatzis, "Gene Expression Profiles of Breast Cancer," *Mendeley Data*, vol. V1, 2017.

[52] F. Morais-Rodrigues *et al.*, "Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression," *Gene*, vol. 726, no. September 2019, pp. 1–8, 2020.

[53] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015.

[54] E. Budiman, Haviluddin, N. Dengan, A. H. Kridalaksana, M. Wati, and Purnawansyah, "Performance of Decision Tree C4.5 Algorithm in Student Academic Evaluation," in *Lecture Notes in Electrical Engineering*, 2018, vol. 488, pp. 380–389.

[55] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[56] K. Moorthy and M. S. Mohamad, "Random forest for gene selection and microarray data classification," *Bioinformation*, vol. 7, no. 3, pp. 142–146, 2011.

[57] T. Hu, G. Zhao, Y. Liu, and M. Long, "A Machine Learning Approach to Differentiate Two Specific Breast Cancer Subtypes Using Androgen Receptor Pathway Genes," *Technol. Cancer Res. Treat.*, vol. 20, 2021, doi: 10.1177/15330338211027900.

[58] T. Evgeniou and M. Pontil, "Support vector machines: Theory and applications," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2001, vol. 2049 LNAI, pp. 249–257.

[59] M. W. Huang, C. W. Chen, W. C. Lin, S. W. Ke, and C. F. Tsai, "SVM and SVM ensembles in breast cancer prediction," *PLoS One*, vol. 12, no. 1, pp. 1–14, 2017.

[60] A. A. Ibrahim, A. I. Hashad, and N. E. M. Shawky, "A Comparison of Open Source Data Mining Tools for Breast Cancer Classification," in *Handbook of Research on Machine Learning Innovations and Trends*, A. E. Hassanien and T. Gaber, Eds. Hershey PA, USA: IGI Global, 2017, pp. 636–651.

[61] J. Wu and C. Hicks, "Breast cancer type classification using machine learning," *J. Pers. Med.*, vol. 11, no. 2, pp. 1–12, 2021.

[62] R. Hazra, M. Banerjee, and L. Badia, "Machine Learning for Breast Cancer Classification with ANN and Decision Tree," in *11th Annual IEEE Information Technology, Electronics and Mobile Communication Conference, IEMCON 2020*, 2020, pp. 522–527.

[63] Y. Li and H. Wu, "A Clustering Method Based on K-Means Algorithm," *Phys. Procedia*, vol. 25, pp. 1104–1109, 2012.

[64] N. Demir, "Ensemble Methods: Elegant Techniques to Produce Improved Machine Learning Results." https://www.toptal.com/machine-learning/ensemble-methods-machine-learning (accessed Jul. 10, 2022).

[65] A. C. Tan and D. Gilbert, "Ensemble machine learning on gene expression data for cancer classification.," *Appl. Bioinformatics*, vol. 2, no. 3 Suppl, pp. 1–10, 2003.

[66] H. Rajaguru and S. R. Sannasi Chakravarthy, "Analysis of decision tree and k-nearest neighbor algorithm in the classification of breast cancer," *Asian Pacific J. Cancer Prev.*, vol. 20, no. 12, pp. 3777–3781, 2019.

[67] F. A. K. Al-Fahaidy, B. Al-Fuhaidi, I. AL-Darouby, F. AL-Abady, M. AL-Qadry, and A. AL-Gamal, "A Diagnostic Model of Breast Cancer Based on Digital Mammogram Images Using Machine Learning Techniques," *Appl. Comput. Intell. Soft Comput.*, vol. 2022, pp. 1–17, 2022.

[68] M. Loey, M. W. Jasim, H. M. EL-Bakry, M. H. N. Taha, and N. E. M. Khalifa, "Breast and colon cancer classification from gene expression profiles using data mining techniques," *Symmetry (Basel).*, vol. 12, no. 3, pp. 1–16, 2020.

[69] Z. Yu, Z. Wang, X. Yu, and Z. Zhang, "RNA-Seq-Based Breast Cancer Subtypes Classification Using Machine Learning Approaches," *Comput. Intell. Neurosci.*, vol. 2020, 2020, doi: 10.1155/2020/4737969.

[70] P. Danaee, R. Ghaeini, and D. A. Hendrix, "A deep learning approach for cancer detection and relevant gene identification," *Pacific Symp. Biocomput.*, vol. 0, no. 212679, pp. 219–229, 2017.

[71] D. A. Ragab, O. Attallah, M. Sharkas, J. Ren, and S. Marshall, "A framework for breast cancer classification using Multi-DCNNs," *Comput. Biol. Med.*, vol. 131, no. December 2020, 2021, doi: 10.1016/j.compbiomed.2021.104245.

[72] D. Zhang, L. Zou, X. Zhou, and F. He, "Integrating Feature Selection and Feature Extraction Methods with Deep Learning to Predict Clinical Outcome of Breast Cancer," *IEEE Access*, vol. 6, pp. 28936–28944, 2018.

[73] V. G. V. Mahesh and M. Mohan Kumar, "An ensemble classification based approach for breast cancer prediction," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1065, no. 1, pp. 1–11, 2021.

[74] M. Abdar and V. Makarenkov, "CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer," *Meas. J. Int. Meas. Confed.*, vol. 146, pp. 557–570, 2019.

[75] A. Bhardwaj, H. Bhardwaj, A. Sakalle, Z. Uddin, M. Sakalle, and W. Ibrahim, "Tree-Based and Machine Learning Algorithm Analysis for Breast Cancer Classification," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–6, 2022, doi: 10.1155/2022/6715406.

[76] A. Chowdhury *et al.*, “Ultrasound classification of breast masses using a comprehensive Nakagami imaging and machine learning framework,” *Ultrasonics*, vol. 124, p. 106744, 2022, doi: https://doi.org/10.1016/j.ultras.2022.106744.

[77] S. Kaisar Alam, E. J. Feleppa, M. Rondeau, A. Kalisz, and B. S. Garra, “Ultrasonic multi-feature analysis procedure for computer-aided diagnosis of solid breast lesions,” *Ultrason. Imaging*, vol. 33, no. 1, pp. 17–38, 2011.

[78] N. M. Ali, R. Besar, and N. A. A. Aziz, “A case study of microarray breast cancer classification using machine learning algorithms with grid search cross validation,” *Bull. Electr. Eng. Informatics*, vol. 12, no. 2, pp. 1047–1054, 2023.

[79] M. M. Ghiasi and S. Zendehboudi, “Application of decision tree-based ensemble learning in the classification of breast cancer,” *Comput. Biol. Med.*, vol. 128, pp. 1–11, 2021.

[80] K. Bache and M. Lichman, “UCI Machine Learning Repository [http://archive. ics. uci. edu/ml]. University of California, School of Information and Computer Science,” *Irvine, CA*, 2013. http://archive.ics.uci.edu/ml.

[81] C. J. Alonso-González, Q. I. Moro-Sancho, A. Simon-Hurtado, and R. Varela-Arrabal, “Microarray gene expression classification with few genes: Criteria to combine attribute selection and classification methods,” *Expert Syst. Appl.*, vol. 39, no. 8, pp. 7270–7280, 2012.

[82] J. P. Sarkar, I. Saha, A. Sarkar, and U. Maulik, “Machine learning integrated ensemble of feature selection methods followed by survival analysis for predicting breast cancer subtype specific miRNA biomarkers,” *Comput. Biol. Med.*, vol. 131, no. October 2020, pp. 1–13, 2021.

[83] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, “Data Descriptor: A curated mammography data set for use in computer-aided detection and diagnosis research,” *Sci. Data*, 2017, doi: 10.1038/sdata.2017.177.

[84] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, “A deep learning-based multi-model ensemble method for cancer prediction,” *Comput. Methods Programs Biomed.*, vol. 153, pp. 1–9, 2018.

[85] J. N. Weinstein *et al.*, “The cancer genome atlas pan-cancer analysis project,” *Nat. Genet.*, vol. 45, no. 10, pp. 1113–1120, 2013.

[86] M. Patrício *et al.*, “Using Resistin, glucose, age and BMI to predict the presence of breast cancer,” *BMC Cancer*, vol. 18, no. 1, pp. 1–8, 2018.

[87] A. B. Nassif, M. A. Talib, Q. Nasir, Y. Afadar, and O. Elgendy, “Breast cancer detection using artificial intelligence techniques: A systematic literature review,” *Artif. Intell. Med.*, vol. 127, p. 102276, 2022, doi: 10.1016/j.artmed.2022.102276.