Turkish Journal of Computer and Mathematics Education *Vol.12 No.14 (2021), 1970- 1979*
Research Article

Breast Cancer Classification Using Machine Learning Techniques: A Review

Srwa Hasan Abdulla [1, a] Ali Makki Sagheer [2, b], Hadi Veisi [3,c]

[1] Basic Science Department, Sulaimani University, Iraq.
[2] Department of Computer Technical Engineering, Al-Qalam University College, Iraq.
[3] Department of Computer Engineering, University of Tehran, Iran
a)Corresponding author email: sirwa.abdulla@univsul.edu.iq
b) prof.ali@alqalam.edu.iq
c) h.veisi@ut.ac.ir

Article History: Received: 12 June 2021; Revised: 09 July 2021; Accepted: 14 August 2021; Published online: 19 August 2021

Abstract: Breast cancer remains one of the top diseases that lead to thousands of death in women every year. Artificial intelligence (AI) has been utilized for diagnosis early, rapidly, and accurately breast tumors. The objective of this paper is to review recent studies for classifying these tumors. Machine learning algorithms such as Support Vector Machine (SVM), K-Nearest Neighbour (K-NN), and Random Forest (RF) are used to classify medical images into malignant and benign. Moreover, deep learning has been employed recently for the same purpose, among them, Convolutional Neural Network (CNN) is one of the most popular techniques. The results showed that the SVM achieved high accuracy, about 97%, therefore, the researchers utilized various functions for this algorithm and added more features such as bagging and boosting to increase its efficacy. In addition, deep learning obtained high accuracy using CNN which is higher than 98%.

Keywords: Mamography, CAD, Classification, Machine learning. SVM**.**

1. Introduction

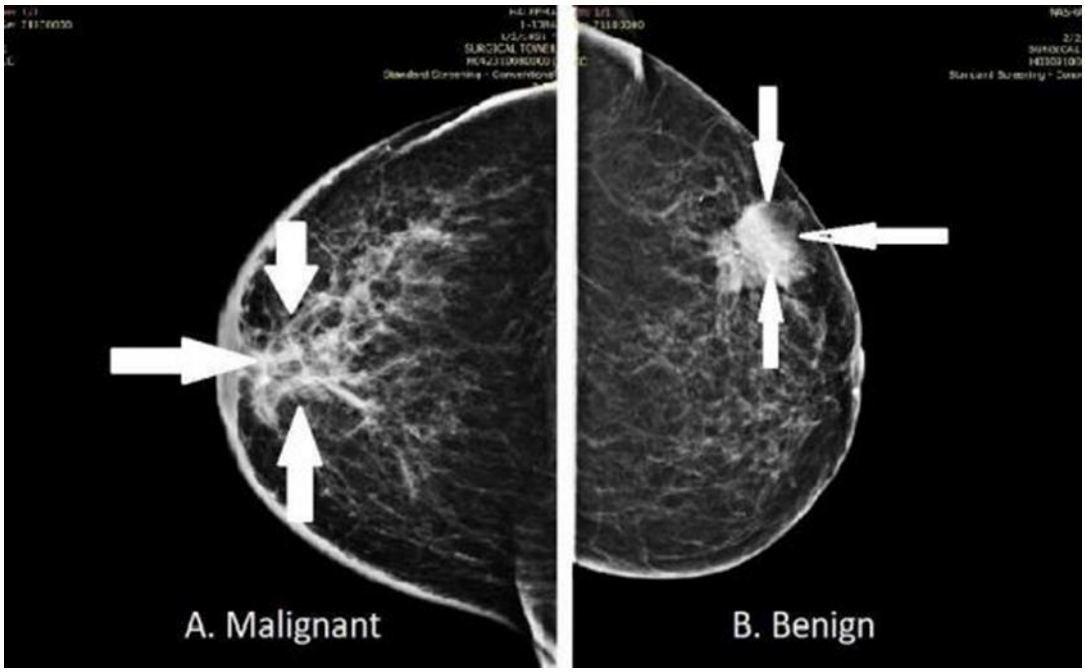
Recently, many scholars have mentioned that the mortality rate has raised in women due to breast cancer. According to the World Health Organization (WHO), the number of females that died in 2018 is about 627,000. Also, this organization predicts that the number may reach to 2.7 million in 2030 globally [1]. The late discovery of this disease and complex procedures are the main reasons for the low survival rate. Therefore, detection of breast cancer earlier is vital to decrease the risk of developing cancer in other tissue cells and carry out a proper treatment[2]. Cancer is a creation of abnormal cells that come from a modification in these cells genetically and spreads into the body, a late in diagnosis and treatment leads to death. There are two types of breast cancer, invasive and non-invasive. The former is harmful, malignant, ability to infect other organs, and classified as cancerous. The latter is non-invasive, not harmful, and not spread to other organs. This disease infects the women's chest and specifically glands and milk ducts, the spread of breast cancer to other organs is frequent and could be through the bloodstream [3]. Different techniques are used to capture breast cancer such as Ultrasound Sonography (ULS), Computerized Thermography (CT), Biopsy (Histological images), Magnetic-Resonance-Imaging (MRI), and Digital Mammography breast X-ray images (DMG). CT is a computerized x-ray imaging procedure that uses a narrow beam of x-ray focusing on a patient with rotation. This procedure produces signals that are dealing with computer to generate cross-sectional images. These images are called tomographic and include rich information from traditional x-ray.

The latter is a prominent technique that is utilized to detect the edges of the tumor from various angles [4]. Figure 1 shows two types of breast cancer (i.e., malignant and benign) through using digital mammography. These images support a radiologist with information about the tumor and how it spread in the area. The radiologist investigates and analyses them manually and then he/she decides the result after taking part with other experts [5]. This process takes time and the results depend strongly upon the knowledge and experiences of the staff, moreover, the experts are not available in each area in the world. Therefore, the research community proposed an automatic system called a CAD (Computer-Aided Diagnosis) for better classification of tumors, accurate results, and rapid executing without needing for radiologists or experts [6]. Machine learning algorithms (MLs) are suggested as an alternative to human vision and experience for analysing medical images and taking the final decisions with high accuracy[7]. The procedures for employing ML approaches include six main steps (i.e., loading images, pre-processing, segmentation, features extraction, features selection, and classification). The first three steps are responsible for removing any pixels that are unrelated to the tumor. Features extraction and selection are responsible for converting the images to statistical features and reducing their size by selecting the most relevant ones. Finally,

applying one of the algorithms of machine learning to classify datasets and obtain a final result. A new approach has been applied recently for classifying breast tumors which is deep learning. This approach skips the initial steps of pre-processing and feature extractions, a large dataset can increase the performance of this method. This study introduces a comprehensive study about the recent methods that are used in this field and presents a full description of the initial steps for the complete process of classification.

This paper contains the main steps of cleaning and preparing for classification that is shown in Section 2. Section 3 presents ML approaches, and a comprehensive work for the previous studies regarding breast cancer disease introduces in section 4. Finally, discussion and conclusions are put in Section 5.

Figure 1. Images of mammography for breast cancer types [4]



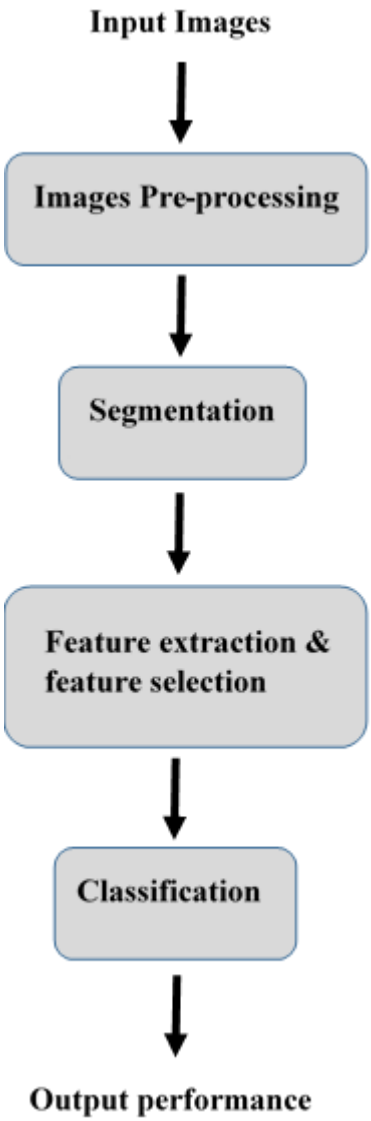
2. Classification stages

The researchers and companies have developed CAD systems to automate breast cancer classification to benign and malignant. These systems can improve a radiologist to find and discriminate tumors in the tissues. Selecting the appropriate algorithms in CAD system requires a better understanding of the contents of cancer images. Generally, a structure for high level of CAD system for cancer diagnosis is shown in Figure 2 . There are four main processing stages for classifying breast tumors into benign and malignant. Complete descriptions of these steps are introduced in detail in the next subsections.

2.1 Data acquisition

At this stage, a dataset is selected to train a model firstly and then evaluate it secondly. There are various datasets for breast cancer tumors that are available globally, but the most popular ones are Wisconsin Breast Cancer dataset (WBCD), Wisconsin Diagnostic Breast Cancer (WDBC), Digital Database for Screening Mammography (DDSM), and Mammographic Image Analysis Society (MAIS). Each dataset include necessary information for processing and modelling, for instance, the second dataset contains the ID for a patient, features, and diagnosis. The ID refers to the identification number for the patient, the number of features are 10 that are computed from a digital image of a breast mass, and the diagnosis for the patient (i.e., positive or negative) [8]. The number of patients are 569, 357 of them are diagnosed as benign and 212 are classified as malignant. Table 1 shows a detail of these features, other studies proposed more attributes based on these through computing mean and standard error for each feature

Figure 2. A general structure of CAD system for breast cancer diagnosis



2.2 Data pre-processing

The next stage is data pre-processing that contains removing redundant and irrelevant data that can improve hugely from the performance of ML algorithms. The following tasks are applied in this step [9]:

Table 1. Features of WDBC dataset [10]

No.	Feature
1	radius (mean of distances from the centre to points on the perimeter)
2	texture (standard deviation of gray-scale values)
3	Perimeter
4	Area
5	smoothness (local variation in radius lengths)
6	compactness (perimeter^2 / area - 1.0)
7	concavity (severity of concave portions of the contour)
8	concave points (number of concave portions of the contour)
9	Symmetry
10	fractal dimension ("coastline approximation" - 1)

•Removing any duplicate row in the dataset.

•Filling a missing value in a cell of a dataset with an appropriate value.

•Converting any string attribute to numerical value as MLs cannot deal with string.

•Normalizing values in ranges either from 0 to 1 or from -1 to 1 as MLs deal better with small values.

•Splitting a dataset into two parts (i.e., training and testing), the splitting could be (80% - 20%) or (50% - 50%) for training and testing respectively.

2.3 Segmentation

As mentioned in the introduction, different imaging modalities are used by radiologists. However, the most techniques that are used for classifying breast cancer are ultrasound imaging and mammograms. The former imaging technique is low contrast and blurry boundaries that impact automatic segmentation. In the contrast, the later imaging technique is high resolution, low energy x-ray and it aids in discovering abnormalities in tissue cells [11]. For mammography images, segmentation is applied to extract the interest region from the image and remove any noise such as lesions of the breast tumor, pectoral muscles, and any region that does not belong to the breast [12]. There are various techniques that are used to apply for segmentation on the selected image such as region-based, threshold-based, and edge-based techniques. For instance, the authors in the study [13] proposed a novel approach to locate a boundary of the pectoral muscle. The study utilized a differentiation operator for edge detection of the boundaries, also, to estimate the value of intensity function. A convex image is produced by determining the endpoint of the breast body edges. Finally, a topographic map is generated by developing a convex hull function.

2.4 Features extraction

Features extraction refers to a technique that minimizes the number of features by generating a new set of attributes with the same information as the old ones. Working with a huge dataset with hundreds or thousands of features without extracting the most ones that represent the actual observations about given variables could lead to overfitting in a model of ML. Therefore, applying this technique reduces the risk of overfitting and increases the performance of the machine learning model. In other words, the purpose of features extraction is to discard the original features by proposing new ones that summarize most of the information of the old characteristics in the dataset. In addition, this technique increases the speed of training, accuracy, and enables visualization [14]. Examples of feature extraction technique are Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA), and t-Distributed Stochastic Neighbour Embedding (t-SNE). The former (PCA) is a statistical technique that reduces a data with losing its properties. Large set of data is identified by this technique to produce a smaller set of uncorrelated features called principal components. PCA is an efficient tool that is used in different fields such as image processing, computer graphics, and face recognition [15]. The second technique is LDA, which is very similar to the PCA regarding a reduction in data and looking to the axes that maximize the variance in the data. LDA finds out another axes that maximize the separation between classes. Therefore, PCA is unsupervised as it ignores the class label and computes the direction of variance. On the other hand the LDA is considered as supervised as it relies on classes to compute a separation between them [16]. In the contrast, (t-SNE) is a technique that was invented recently in 2008 compared with PCA which was developed in 1933. The former gives better presenting for the data than the later as the t-SNE deals with nonlinearity in the data and preserves a small pairwise distance compared with large distance using PCA [17].

After segmenting a region of interest for a breast tumor, features are measured based on the architecture of the tumor. During observation, radiologists noticed that the benign tumor can be characterized as smooth, round, and its boundary is clear. Whilst the malignant tumor is commonly rough, blurred, and speculated [18]. The authors in [19] used a co-occurrence matrix and run-length matrix to extract features related to a texture that contain two types (i.e., structural and statistical). A work in [20] extracted statistical features such as median, mean, variance, and standard deviation and utilized Gray Level Co-Occurrence Matrix (GLCM) for analysis. The GLCM technique is a second-order method that provides statistical information about texture features of the image. This technique calculates a relationship between two pixels, which are reference pixel and neighbouring pixel. This tool is helpful to identify various regions in the image through extracting texture features [21].

However, these techniques have various limitations depending on the technique that is used in extracting features. Therefore, deep learning approach is suggested as an alternative solution, which skips initial steps of preparation and obtains high accuracy [22]. The advantages of applying this approach over traditional MLs are dealing directly with raw images, less requiring to expert knowledge, less effort to tune important features, and

reducing in time consuming. The level of accuracy by using this technique increases with a big data, which is one of the most limitations of using deep learning technique [23].

2.5 Features selection

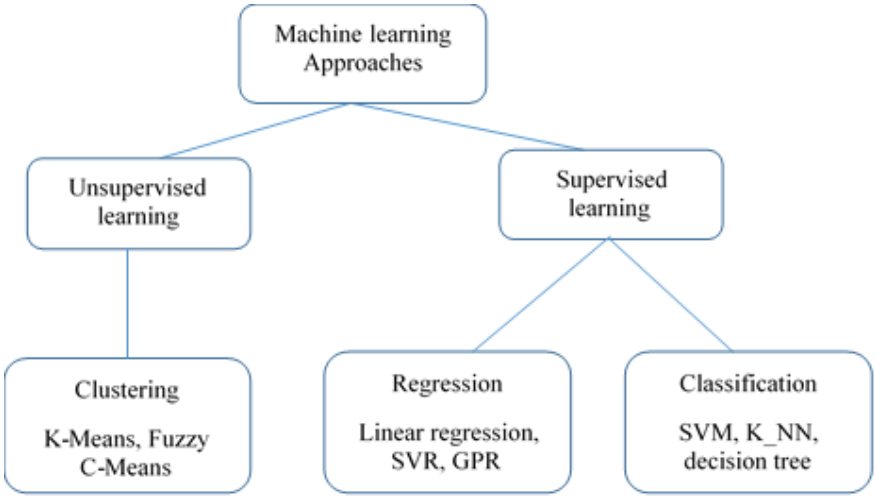
Effective classification scheme also depends highly on selecting features technique, which reduces in a number of features and also ranks them from most important one to least important ones. This reduction possibly gives many benefits through statistical analysis (e.g., improving accuracy, reducing the risk of overfitting, increasing training speed, the possibility of data visualization) [24]. There are several methods that can be utilized to apply feature selection (i.e., filter method, wrapper method, and embedded method). The former includes selecting a subset from the dataset that contains only relevant features by using filtering methods such as Pearson Correlation. The second method is more accurate than the previous one as it uses machine learning algorithms for evaluating features, but it needs more processing time. The technique involves adding and removing features based on the performance of the model [25].

3. Classification using machine learning algorithms

Machine learning algorithms can be classified mainly into two types, Supervised Learning (SL) and Unsupervised Learning (USL). The first type requires training through a labeled dataset that contains inputs and output as a target. In this type, there are two phases, the training phase, and the testing phase. In the training phase, the model in SL is first built through data that are labelled manually by human intervention, and then in the testing phase, new data that are not seen by the model tests the model [26]. The second type of machine learning algorithm which is USL, there is no need to train a model. The data samples are classified based on common features of these samples; this type is suitable when there is no labeled data. There is another type of machine learning that is placed between the two aforementioned types. This type called Semi-Supervised Learning (SSL) that needs a few samples of labeled data that are used to label unlabelled samples. The SL algorithms can be used to solve problems of classification and regression. Classification is a learning process that deals with discrete data and classifies them into classes. In the contrast, the regression deals with continuous data or real data variables such as temperature or time [27]. In the USL, clustering is used to solve the problem of identification of samples based on common characteristics. There are many algorithms that are utilized by researchers or developers to solve problems of classification and identification of samples automatically and rapidly. Support Vector Machine (SVM), K-NN, Naïve Bayes (NB), C-means are the most famous algorithms that are used to identify breast cancer tumors [28]. Figure 3 shows the main types of machine learning approaches and their algorithms.

Support vector machine (SVM): This algorithm has been utilized by researchers in solving various problems in regression and classification, the latter is commonly used. According to the number of features, n-spaces are formed where each coordinate is created for each feature. This algorithm tries to draw different new lines, which are called hyperplanes, among the n-spaces to find out a best line that has maximum margin. The maximum margin can be defined as s margin that segregates between different classes, which represented by data points [29, 30]. Various studies have used this algorithms such as [31, 32] to classify breast cancer tumors that achieved promising results. These studies utilized different algorithms (i.e., SVM, K-NN, C4.5, NB, K-means, EM, PAM, and fuzzy c-means). They found that the SVM algorithm achieved higher accuracy than other algorithms.

Figure 3. Machine learning approaches [28]



K-NN (Nearest Neighbour): This algorithm has been used in different applications such as healthcare, finance, image and video recognition and also in handwriting. The algorithm firstly trains a model with labelled data with

different classes and then tests the model using new points. The algorithm calculates the nearest known neighbour points to the new data points using one of the approaches such as Manhattan distance, Hamming distance, Minkowski distance, and Euclidean distance. The new point is classified to a known class that is nearest to this point, the algorithm repeats the same procedure for all new points [33]. The authors in studies [34, 35] claimed that combing K-NN with SVM can improve the efficient of the scheme.

Random Forest (RF): Random forest is a technique that used widely in large datasets efficiently and quickly. Many researchers have used this method in their projects and also utilized in different real applications [36]. the technique relies on principle of ensemble learning that creates various classifiers and merges their results [37]. The performance of a single classifier is less than of multiple weak classifiers that both using the same dataset. There are many ensembles methods such as boosting, bagging and lately Random Forest. The boosting method [38] firstly initializes all instances with the same weights and then sequentially gives more weights for the instances that are misclassified and less weights for the instances that are classified correctly. In bagging method [39], the dataset is divided into different training subsets that are fed in parallel to classifier using a majority vote. In contrast, the Random Forest is a type of ensemble approaches that classifies new instances through constructing many decision trees and majority voting. In this algorithm, an entire set of features is divided into many subsets and each one represents a decision tree that selected randomly. Random Forest is faster than bagging and boosting, and more robust regarding noise from boosting [40].

4. Literature review

Artificial intelligence (AI) has played an important role in the healthcare field for providing safety and improved quality of care. Machine learning and deep learning are a branch of AI that are used widely in this field and especially in identifying and classifying tumors in the breast and brain[41]. The author in the study [31] used four SVM, C4.5, NB, and k-NN for classifying breast cancer based on Wisconsin Breast Cancer (original) dataset that contains 11 attributes and 699 instances. The results showed that the SVM achieved higher accuracy from other classifiers reached to 97.13%. Based on these results, other studies such as [29] have begun to investigate other kernel functions (i.e., linear, polynomial, and RBF) for SVM and advanced features such as bagging and boosting. The study evaluated these parameters by using two datasets, the first dataset has 11 attributes and 699 instances and the second dataset has 117 attributes and 102294 instances. The study found that the linear kernel based on SVM that uses bagging feature and RBF kernel based on SVM with boosting feature are suitable for a small dataset. Also, the latter achieved better results than other classifiers for large datasets. Similarly, in 2018, authors Y. Khourdifi and M. Bahaj in two different works [34, 35] applied four machine learning algorithms that are Random Forest, Naive Bayes, SVM, KNN using WEKA tool. The studies evaluated the algorithms using a dataset that consisted of 699 instances with 30 attributes. They found also that the SVM model obtained high accuracy among the others with accuracy reached to 97.9%. Other study [32] compared clustering algorithms K-means, Expectation Maximization, Partitioning Around Medoids (PAM) and Fuzzy c-means with classification algorithms SVM and C5.0. The study showed that SVM and C5.0 surpassed clustering models with 81% accuracy.

In the contrast, a new study [10] showed different results, the study compared and evaluated 9 machine learning algorithms. These algorithms are Logistic regression, Gaussian Naive Bayes, Linear Support vector machine, RBF Support vector machine, Decision Tree, Random Forest, Xgboost, Gradient Boosting, and KNN. The study utilized Wisconsin Diagnostic Breast Cancer (WDBC) dataset to evaluate these models. The study compared SL with semi-supervised SSL; the results showed that k-NN and logistics regression algorithms achieved higher accuracies. The accuracies for both algorithm were (SL = 98% & SSL = 97%) and (SL = 97% & SSL = 98%) respectively. Moreover, the study showed high accuracy for linear SVM with 97%.

Other studies [41, 42] proposed using ensemble learning to classify breast cancer tumors using WBCD dataset. The study [42] combined Boosting Artificial Neural Network (BANN) with two SVMs. The authors claimed that they obtained very high accuracy reached 100%. The study [41] proposed combing three classifiers SVM learning with stochastic gradient descent optimization, simple logistic regression learning, and multilayer perceptron network. These classifiers are utilized as ensemble classification and using a voting scheme. The study achieved high accuracy of about 99.42%. Similarly, the authors in the study [43] proposed an approach of an ensemble learning method by combining Multi-Verse Optimizer (MVO) and Gradient Boosting Decision Tree (GBDT). The former is responsible for tuning the parameters of the latter and also optimizing the selection of the features. The study used two datasets Wisconsin Diagnostic Breast Cancer and Wisconsin Breast Cancer to evaluate the proposed method. The proposed method showed more accuracy and has low variance from other models that are suggested from other studies. The authors in work [44] presented an observation related an ensemble learning that this scheme increases a base learner, but it reduces the bias or variance. While the authors in the study [45] claimed that the accuracy is improved in ensemble learning when a boosting feature is used.

On the other hand, deep learning has taken the attention of schoolers in recent years. On the other hand, deep learning has taken the attention of schoolers in recent years. This approach does not need to apply feature preparation. Instead, it can extract the features automatically from the medical images without the need for human intervention. The study [46] utilized a deep learning approach to classify breast cancer images, convolution neural network algorithm was employed CNN. The study evaluated the method by using three datasets DDSM, IN breast, and BCDR with accuracies 97.35%, 95.50%, and 96.67% respectively. Another study [47] achieved higher accuracy using more instances about 5699 instances and also applied the CNN algorithm, the accuracy was 98.62%. Other work [48] obtained lower accuracy reached to 87% from dataset was collected at two medical institutions, the Sun Yat-Sen University Cancer Centre and Nanhai Affiliated Hospital of Southern Medical University. Two studies [49, 50] collected very huge datasets, the first study was collected from (2010-2016) at five imaging sites affiliated with the New York University School of Medicine with around a million images and more than 140,000 patients. The second study collected datasets with 12,000 cases and both studies applied CNN, the accuracies in both works calculated under the curve (AUC). The authors in the study [51] collected 67, 520 images privately and achieved high accuracy compared to the enormous dataset about 95%. Table 2 summarizes these studies.

Table 2. Overview of recent methods based on features and classifiers

Ref.	Dataset	Features	Method	Accuracy
[31]	Wisconsin Breast Cancer (original) dataset (WBCD)	11 attributes 699 instances	SVM, C4.5, NB, k-NN	Best accuracy for SVM 97.13%
[29]	Two datasets	First dataset: 11 attributes with 699 instances Second dataset: 117 attributes with 102294 instances	SVM with three functions and two features: bagging and boosting	96.85%, 95%
[34, 35]	Wisconsin Breast Cancer dataset	30 attributes with 699 instances	Random Forest, Naive Bayes, SVM, KNN	97.9%
[32]	Wisconsin Prognostic Breast Cancer dataset	32 attributes with 194 instances	Compare (K-means, EM, PAM, and fuzzy c-means) with SVM and C5.0	Better results for SVM with accuracy 97%
[10]	Wisconsin Diagnostic Breast Cancer (WDBC) dataset	30 attributes of 569 patients with 569 instances	Compare supervised learning (SL) with semi-supervised learning (SSL) for 9 algorithms	K-NN (SL = 98% & SSL = 97%) and logistics regression (SL = 97% & SSL = 98%)
[42]	Wisconsin Breast Cancer (original) dataset (WBCD)	11 attributes with 699 instances	Boosting Artificial Neural Network (BANN) with two SVMs	100%
[41]	Wisconsin Breast Cancer dataset (WBCD)	32 attributes with 569 instances	SVM with stochastic gradient descent optimization, simple logistic regression learning, and multilayer perceptron network	99.44%
[43]	Wisconsin Diagnostic Breast	11 attributes 699 instances	Multi-Verse Optimizer (MVO) and Gradient	This model is more accurate and has low

	Cancer and Wisconsin Breast Cancer	32 attributes and 569 instances	Boosting Decision Tree (GBDT)	variance from other models
[46]	DDSM, IN breast, and BCDR	DDSM: 5316 images, 641 cases of patients IN breast: 200 images for 50 cases BCDR: 600 images from 300 patients	CNN	97.35%, 95.50%, 96.67% for three datasets respectively
[47]	WBCD	9 attributes with 5699 instances	Deep neural network (DNN)	98.62%
[48]	The Datasets were collected at two medical institutions	990 images, 540 Malignant masses, and 450 benign lesions	CNN	87.68%
[49]	Data was collected from (2010-2016) at five imaging sites affiliated with New York University School of Medicine	1,001,093 images from 141,473 patients	CNN	The accuracy was calculated based on area under a curve (AUC)
[50]	Data was collected independently	12,000 cases, including 4000 samples proven cancers	CNN	The accuracy was calculated based on area under a curve (AUC)
[51]	Private dataset	67,520 mammographic images from 16,968 women	CNN	95%

5. Discussion and Conclusions

Two approaches are used to classify breast cancer tumors, traditional machine learning algorithms, and deep learning. The former applied various algorithms, but the most promising algorithm is the support vector machine (SVM). For instance, the studies [31, 32] compared SVM with other algorithms such as K-NN, C4.5, NB, K-means, EM, PAM, fuzzy c-means, they found that the SVM surpassed these methods with accuracy reached to 97%. Therefore, researchers tried to investigate more in these algorithms such as studies [29] [34, 35] that applied different functions for SVM and added more features such as bagging and boosting as in the study [29]. The study achieved high accuracy of about 95% with a huge dataset of about 102294 records. The other studies [34, 35] combined other methods with SVM (i.e., random forest, Naïve Bayes, and KNN), the accuracy reached a highest level of about 98%. Using traditional machine learning requires initial pre-processing and feature selection, which take time a computational consumption. As a result, recent studies have used the deep learning approach as it is able to extract suitable features automatically. Different studies employed convolution neural network (CNN) to classify breast tumors, a study [46] evaluates three datasets with different sizes and achieved very high accuracy reached to 97%. Another study used a popular dataset that was used in traditional MLs and obtained high accuracy of more than 98%. The deep learning approach obtained very high accuracy but it needs very lager datasets and requires more resources.

In conclusion, this study reviewed recent studies regarding breast cancer tumors. The study observed two lanes, the first lane used traditional MLs that applied various algorithms, but the most accurate one is SVM with an

accuracy of about 97%. The authors showed from the results that this method achieves higher accuracy if it combines with other methods such as random forest, Naïve Bayes, and K-NN. Deep learning also achieved higher accuracy reached 98% using a Convolution Neural Network (CNN). For future work, the error rate can be minimized by investigating more in deep learning algorithms and increasing datasets for medical breast images.

References

[1] "WHO | World Health Organization." [Online]. Available: https://www.who.int/. [Accessed: 28-Jun-2021].

[2] D. Bardou, K. Zhang, and S. M. Ahmad, "Classification of Breast Cancer Based on Histology Images Using Convolutional Neural Networks," IEEE Access, vol. 6, pp. 24680–24693, 2018.

[3] Priyanka and K. Sanjeev, "A review paper on breast cancer detection using deep learning," IOP Conf. Ser. Mater. Sci. Eng., vol. 1022, no. 1, 2021.

[4] T. Mahmood, J. Li, Y. Pei, F. Akhtar, A. Imran, and K. Ur Rehman, "A brief survey on breast cancer diagnostic with deep learning schemes using multi-image modalities," IEEE Access, vol. 8, pp. 165779–165809, 2020.

[5] V. Lahoura, H. Singh, A. Aggarwal, B. Sharma, M. A. Mohammed, R. Damaševičius, S. Kadry, and K. Cengiz, "Cloud computing-based framework for breast cancer diagnosis using extreme learning machine," Diagnostics, vol. 11, no. 2, pp. 1–19, 2021.

[6] G. Battineni, N. Chintalapudi, and F. Amenta, "Performance analysis of different machine learning algorithms in breast cancer predictions," EAI Endorsed Trans. Pervasive Heal. Technol., vol. 6, no. 23, pp. 1–7, 2020.

[7] Op. P. Nave and M. Elbaz, "Artificial immune system features added to breast cancer clinical data for machine learning (ML) applications," BioSystems, vol. 202, no. April, 2021.

[8] R. Sarmento, "Breast Cancer Wisconsin (Diagnostic) Data Set," 2019.

[9] M. M. Jalal, Z. Tasnim, and M. N. Islam, "Exploring the Machine Learning Algorithms to Find the Best Features for Predicting the Risk of Cardiovascular Diseases," no. April, pp. 559–569, 2021.

[10] N. Al-Azzam and I. Shatnawi, "Comparing supervised and semi-supervised Machine Learning Models on Diagnosing Breast Cancer," Ann. Med. Surg., vol. 62, no. December 2020, pp. 53–64, 2021.

[11] R. V. and M. H. J. Dabass, S. Arora, "Segmentation Techniques for Breast Cancer Imaging Modalities-A Review," in 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2019, pp. 658–663.

[12] S. F. Khorshid and A. M. Abdulazeez, "Breast Cancer Diagnosis Based on K-Nearest Neighbors: a Review," PalArch's J. Archaeol. Egypt/Egyptology, vol. 18, no. 4, pp. 1927–1951, 2021.

[13] B. Mughal, N. Muhammad, M. Sharif, A. Rehman, and T. Saba, "Removal of pectoral muscle based on topographic map and shape-shifting silhouette," BMC Cancer, vol. 18, no. 1, pp. 1–14, 2018.

[14] Pier Ippolito, "Feature Extraction Techniques," 2019. [Online]. Available: https://towardsdatascience.com/feature-extraction-techniques-d619b56e31be. [Accessed: 20-Jun-2021].

[15] M. Mateen, J. Wen, Nasrullah, S. Song, and Z. Huang, "Fundus image classification using VGG-19 architecture with PCA and SVD," Symmetry (Basel)., vol. 11, no. 1, 2019.

[16] S. Raschka, "Linear Discriminant Analysis," 2014. [Online]. Available: https://sebastianraschka.com/Articles/2014_python_lda.html. [Accessed: 23-Jan-2021].

[17] A. Violante, "An Introduction to t-SNE with Python Example," 2018. [Online]. Available: https://towardsdatascience.com/an-introduction-to-t-sne-with-python-example-5a3a293108d1. [Accessed: 23-Jul-2021].

[18] R. R. and J. E. L. D. N. R. Mudigonda, "Gradient and texture analysis for the classification of mammographic masses," in EEE Transactions on Medical Imaging, 2000, pp. 1032–1043.

[19] U. R. Acharya, E. Y. K. Ng, J. H. Tan, and S. V. Sree, "Thermography based breast cancer detection using texture features and support vector machine," J. Med. Syst., vol. 36, no. 3, pp. 1503–1510, 2012.

[20] S. Francis, S.V., Sasikala, M., Saranya, "Detection of breast abnormality from thermograms using curvelet transform based feature extraction," J. Med. Syst., vol. 38, 2014.

[21] D.BhargavaS.VyasAyushiBansal, "Comparative analysis of classification techniques for brain magnetic resonance imaging images," Adv. Comput. Tech. Biomed. Image Anal., pp. 133–144, 2020.

[22] M. Bakator and D. Radosav, "Deep learning and medical diagnosis: A review of literature," Multimodal Technol. Interact., vol. 2, no. 3, 2018.

[23] H. A. Khan, W. Jue, M. Mushtaq, and M. U. Mushtaq, "Brain tumor classification in MRI image using convolutional neural network," Math. Biosci. Eng., vol. 17, no. 5, pp. 6203–6216, 2020.

[24] H. Zhang, G. Lu, M. T. Qassrawi, Y. Zhang, and X. Yu, "Feature selection for optimizing traffic classification," Comput. Commun., vol. 35, no. 12, pp. 1457–1471, 2012.

[25] P. Ippolito, "Feature Selection Techniques," 2019. [Online]. Available: https://towardsdatascience.com/feature-extraction-techniques-d619b56e31be. [Accessed: 28-Jun-2021].

[26] M. Fatima and M. Pasha, "Survey of Machine Learning Algorithms for Disease Diagnostic," J. Intell. Learn. Syst. Appl., vol. 09, no. 01, pp. 1–16, 2017.
[27] 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, no. May 2015, pp. 8–17, 2015.
[28] D. M. Abdulqader, A. M. Abdulazeez, and D. Q. Zeebaree, "Machine learning supervised algorithms of gene selection: A review," Technol. Reports Kansai Univ., vol. 62, no. 3, pp. 233–244, 2020.
[29] 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.
[30] Z. Li, R. Yuan, and X. Guan, "Accurate classification of the internet traffic based on the SVM method," ICC, June 2007, pp. 1373–1378.
[31] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," Procedia Comput. Sci., vol. 83, no. Fams, pp. 1064–1069, 2016.
[32] R. Rawal, "BREAST CANCER PREDICTION USING MACHINE LEARNING," J. Emerg. Technol. Innov. Res., vol. 7, no. 5, 2020.
[33] W. Cherif, "Optimization of K-NN algorithm by clustering and reliability coefficients: Application to breast-cancer diagnosis," Procedia Comput. Sci., vol. 127, pp. 293–299, 2018.
[34] Y. K. and M. Bahaj, "Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification," in International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), 2018, pp. 1–5.
[35] Y. K. and M. Bahaj, "Feature Selection with Fast Correlation-Based Filter for Breast Cancer Prediction and Classification Using Machine Learning Algorithms," in 2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT), 2018, pp. 1–6.
[36] C. Nguyen, Y. Wang, and H. N. Nguyen, "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic," J. Biomed. Sci. Eng., vol. 06, no. 05, pp. 551–560, 2013.
[37] N. Sirikulviriya and S. Sinthupinyo, "Integration of rules from a random forest," Int. Conf. Inf. …, vol. 6, pp. 194–198, 2011.
[38] Y. Freund, R. E. Schapire, and M. Hill, "Experiments with a New Boosting Algorithm Rooms f 2B-428 , 2A-424 g," 1996.
[39] R. Richman and M. V. Wüthrich, "bagging predictors," Risks, vol. 8, no. 3, pp. 1–26, 2020.
[40] Y. L. Pavlov, "Random forests," Random For., pp. 1–122, 2019.
[41] A. S. Assiri, S. Nazir, and S. A. Velastin, "Breast Tumor Classification Using an Ensemble Machine Learning Method," J. Imaging, vol. 6, no. 6, p. 39, May 2020.
[42] 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, no. May, pp. 557–570, 2019.
[43] H. Tabrizchi, M. Tabrizchi, and H. Tabrizchi, "Breast cancer diagnosis using a multi-verse optimizer-based gradient boosting decision tree," SN Appl. Sci., vol. 2, no. 4, pp. 1–19, 2020.
[44] D. G. and L. C. S. Lee, M. Amgad, M. Masoud, R. Subramanian, "An Ensemble-based Active Learning for Breast Cancer Classification," in IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2019, pp. 2549–2553.
[45] A. H. Osman and H. M. A. Aljahdali, "An Effective of Ensemble Boosting Learning Method for Breast Cancer Virtual Screening Using Neural Network Model," IEEE Access, vol. 8, pp. 39165–39174, 2020.
[46] A. O. Chougrad H, Zouaki H, "Deep convolutional neural networks for breast cancer screening," Comput Methods Prog Biomed, vol. 157, pp. 19–30, 2018.
[47] and P. V. S. S. R. C. M. S. Karthik, R. Srinivasa Perumal, "Breast cancer classification using deep neural networks," Knowl Comput Its Appl Knowl Manip Process Tech, vol. 1, pp. 227–241, 2018.
[48] H. Cai, Q. Huang, W. Rong, Y. Song, J. Li, J. Wang, J. Chen, and L. Li, "Breast Microcalcification Diagnosis Using Deep Convolutional Neural Network from Digital Mammograms," Comput. Math. Methods Med., vol. 2019, no. January 2020, 2019.
[49] N. W. et Al, "Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening," IEEE Trans. Med. Imaging, vol. 39, no. 4, pp. 1184–1194, 2020.
[50] E. F. Conant, A. Y. Toledano, S. Periaswamy, S. V. Fotin, J. Go, J. E. Boatsman, and J. W. Hoffmeister, "Improving Accuracy and Efficiency with Concurrent Use of Artificial Intelligence for Digital Breast Tomosynthesis," Radiol. Artif. Intell., vol. 1, no. 4, p. e180096, 2019.
[51] G. V. Ionescu, M. Fergie, M. Berks, E. F. Harkness, J. Hulleman, A. R. Brentnall, J. Cuzick, D. G. Evans, and S. M. Astley, "Prediction of reader estimates of mammographic density using convolutional neural networks," J. Med. Imaging, vol. 6, no. 03, p. 1, 2019.