Early Detection of Breast Cancer Using Machine Learning Techniques

M. Tahmooresi[1], A. Afshar[2], B. Bashari Rad[1], K. B. Nowshath[1] and M. A. Bamiah[2]
[1]*Asia Pacific University of Technology and Innovation (APU), Malaysia.*
[2]*University of Malaya, Malaysia.*
maryam.tahmooresi@yahoo.com

***Abstract*—Cancer is the second cause of death in the world. 8.8 million patients died due to cancer in 2015. Breast cancer is the leading cause of death among women. Several types of research have been done on early detection of breast cancer to start treatment and increase the chance of survival. Most of the studies concentrated on mammogram images. However, mammogram images sometimes have a risk of false detection that may endanger the patient's health. It is vital to find alternative methods which are easier to implement and work with different data sets, cheaper and safer, that can produce a more reliable prediction. This paper proposes a hybrid model combined of several Machine Learning (ML) algorithms including Support Vector Machine (SVM), Artificial Neural Network (ANN), K-Nearest Neighbor (KNN), Decision Tree (DT) for effective breast cancer detection. This study also discusses the datasets used for breast cancer detection and diagnosis. The proposed model can be used with different data types such as image, blood, etc.**

***Index Terms*—Breast Cancer; Breast Cancer Detection; Medical Images; Machine Learning.**

I. Introduction

World Health Organization (WHO) reported the breast cancer is the most common cancer amongst women globally [1]. It is also the highest ranked type of cancer cause the death among women in the world [2, 3]. In Malaysia, Breast cancer has the highest rate of cancer deaths, around 25%, and it is the commonest cancer among women [4]. Around 5% of Malaysian women are at risk of breast cancer while Europe and the United States, it is around 12.5% [3]. It confirms that women with breast cancer in Malaysia present at a later stage of the disease compared to women from other countries [4]. Usually, breast cancer can be easily detected if specific symptoms appear. However, many women who are suffering from breast cancer have no symptoms. Hence, regular breast cancer screening is very important for early detection [3].

Early detection of breast cancer aids for early diagnosis and treatment, because the prognosis is very important for long-term survival [5]. Since early detection, diagnosis, and treatment of cancer can reduce the risk of death, it plays a significant role in saving the life of the patient. Any delay in detection of cancer in early stages leads to disease progression and complication of treatment [5], therefore long waiting time prior to diagnosis of breast cancer and starting the treatment process is of prognostic concern.

Previous studies on the investigation of the consequences of a late diagnosis of cancer confirm that it is strongly associated with progression of the disease to more advanced stages, consequently less chance to save the patient's life. In a systematic review conducted by Prof MA Richards et al. [6], an analysis of 87 studies strongly concluded that female patients with breast cancer who start their therapy less than 3 months after the appearance of symptoms significantly have a higher chance of survival compare to those who wait for more than 3 months.

Many previous studies confirm that detection of breast cancer in early stages significantly increase the chance of survival because it prevents the spreading of malignant cells throughout the entire body [6].

The main contribution of this paper is to review the role of machine learning techniques in early detection of the breast cancer.

Artificial Intelligence (AI) can be applied to improve breast cancer detection and diagnosis, as well as prevent overtreatment. Nevertheless, combining AI and Machine Learning (ML) methods enables the prediction and empower accurate decision making. For example, deciding on the biopsy results for detecting breast cancer if the patient needs surgery or not.

Currently, Mammograms are the most used test available, however, still, they have false positive (high-risk) results which shows abnormal cells that can lead to unnecessary biopsies and surgeries. Sometimes surgery is done to remove lesions reveals that it is benign which is not harmful. This means that the patient will go through unnecessary painful and expensive surgery.

ML Algorithms were introduced with many features such as effective performance on healthcare related dataset which involve images, x-rays, blood samples, etc. Some methods are appropriate for the small dataset whereby others are suitable for huge datasets. However, noise can be a problematic concern in some methods.

This paper is organized as follows, Section II introduces the breast cancer briefly, Section III explains the ML algorithms used for detecting breast cancer. A summary of previous related works is given in section IV. Finally, Section V concludes the paper.

II. Breast Cancer

Breast cancer is the most found disease in the women, worldwide, where abnormal growth of a mass of tissue, cause the expansion of malignant cells leads to acute breast cancer. These malignant cells are originally created from milk glands of the breast. These malignant cells which are the main reason for breast cancer can be classified into different groups according to their unusual progress and capability affecting other normal cells [7]. The capability of affecting means whether these malignant cells affect only the local cells or can spread throughout the full body. The effect of spreading these malignant cells throughout the whole body of the patient is called as metastasis [7]. It is very important to prevent this spreading effect by a diagnosis of cancer in the early stages using advanced techniques and equipment. In recent decades, there are many efforts to employ artificial intelligence and other related methods to assist in the detection of cancer in earlier stages.

Early detection of cancer boosts the increase of survival chance to 98% [8]. Figure 1. shows different types of cancers whereby breast cancer is leading with 24% as follows.

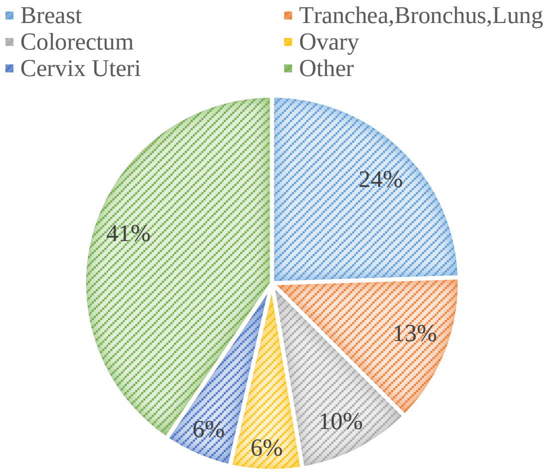

Figure 1: Types of cancer

III. Machine Learning Methods

Machine Learning is a process that machines (computers) are trained with data to make the decision for similar cases [9]. ML is employed in various applications, such as object recognition, network, security, and healthcare. There are two ML types i.e. single and hybrid methods like ANN, SVM, Gaussian Mixture Model (GMM), K-Nearest Neighbor (KNN), Linear Regressive Classification (LRC), Weighted Hierarchical Adaptive Voting Ensemble (WHAVE), etc. Following are the used ML algorithms:

A. Artificial Neural Network (ANN)

ANN is a model like human brains nerve system that has a large number of nodes connected to each other. Each node has two states: 0 means active and 1 means active. Also, each node has a positive or negative weight that adjusts the strength of the node and can activate or deactivate it. ANN provides samples of data to train the machine. The trained machine is used to detect the pattern of hidden date. It can search for patterns among patients' healthcare and personal records to identify high-risk lesions [10].

B. Support Vector Machine (SVM)

SVM is a supervised pattern classification model which is used as a training algorithm for learning classification and regression rule from gathered data [11]. The purpose of this method is to separate data until a hyperplane with high minimum distance is found. SVM is used to classify two or more data types. SVM include single or hybrid models such as Standard SVM (St-SVM), Proximal Support Vector Machine (PSVM), Newton Support Vector Machine (NSVM), Lagrangian Support Vector Machines (LSVM), Linear Programming Support Vector Machines (LPSVM), and Smooth Support Vector Machine (SSVM).

C. K-Nearest Neighbors (KNN)

KNN is a supervised learning method which is used for diagnosing and classifying cancer [12]. In this method, the computer is trained in a specific field and new data is given to it. Additionally, similar data is used by the machine for detecting (K) hence, the machine starts finding KNN for the unknown data. It is recommended to choose a large dataset for training also K value must be an odd number.

D. Decision Tree (DT)

DT is a data mining technique used for early detection of breast cancer. It is a model that presents classifications or regressions as a tree. In this model, the data set is broken to small sub-data, then to smaller ones. As a result, the tree is developed and at the last level, the result is revealed. In a tree structure, the leaves characterize the class labels whereby the branches characterize conjunctions of feature leading to the class labels Hence, DT is not sensitive to noise [13].

E. Random Forest (RF) Algorithm

RF algorithm is used at the regularization point where the model quality is highest, variance and bias problems are compromised [14]. RF builds numerous numbers of DTs using random samples with a replacement to overcome the problem of DTs. Each tree classifies its observations, and majority votes decision is chosen. RF is used in the unsupervised mode for assessing proximities among data points.

F. AdaBoost Classifier

This algorithm is used for classification and regression to predict breast cancer existence. It converts weak learners to strong ones by combining all weak learners to form a single strong rule. It gets the weight of the node and changes it continuously until an accurate result is found. However, it is sensitive to noise and quality of features [15].

G. Naïve Bayes (NB) Classifier

Naïve Bayes refers to a probabilistic classifier that applies Bayes' theorem with robust independence assumptions [16]. In this model, all properties are considered separately to detect any existing relationship between them. It assumes that predictive attributes are conditionally independent given a class. Moreover, the values of the numeric attributes are distributed within each class. NB is fast and performs well even with a small dataset. However, it is difficult to find independent properties in real life. [16]. have deployed NB classifier for breast cancer detection and it gave the maximum accuracy with only five dominant.

IV. Previous Related Works

Several studies have been conducted on the implementation of ML on Breast Cancer detection and diagnosis using different methods or combination of several algorithms to increase the accuracy. S. Gc *et al.* [17] worked on extracting features including variance, range, and compactness. They used SVM classification to evaluate the performance. Their findings showed the highest variance of 95%, range 94%, compactness 86%. According to their results, SVM can be considered as an appropriate method for Breast Cancer Detection.

Chunqiu Wang *et al.* [18] chose Microwave Tomography Imaging (MTI) to extract features and classify the images using ANN. Two different techniques were compared in this study, GMM and KNN. Their results showed that the sensitivity obtained by KNN is 87%, while for GMM is 67%. The accuracy was 85% for KNN and 75% for GMM. The result for Matthews Correlation Coefficient (MCC) was 67% and 48% for KNN and GMM, respectively. Finally, the specificity was 84% for KNN and 86% for GMM. According to their findings, Sensitivity, Accuracy, and MCC for KNN were better than GMM, but GMM was better in Specificity and Precision.

Chowdhary and Acharjya [19] focused on mammogram images as they are cheaper and more efficient in detection. However, since selecting and extracting features are important for improving performance, Fuzzy Histogram Hyperbolization (FHH) was chosen to increase the quality of images, Fuzzy C-mean for segmenting, and Gray level dependence model for extracting the features. Their method showed 94% accuracy for detecting malignant breast lesions.

In a study conducted by Aminikhanghahi *et al.* [20], wireless cyber mammography images were explored. After selecting features and extracting them, the researcher has chosen two different ML techniques, SVM and GMM to check their accuracy. Their findings showed that SVM is more accurate if there is no noise or error, else GMM is better and safer.

Durai *et al.* [21] Have selected Data Mining technique for detecting diseases including breast cancer. They used LRC and compared it with four other techniques including BFI, ID3, J48, and SVM. The result shows that LRC is the most accurate one with 99.25% accuracy.

Wang and Yoon [22] chose four methods of Data Mining to measure their effectiveness in detection. These models were: SVM, ANN, Naïve Bayes Classification and Adaboost tree. In addition, PCs and PCi were used for making hybrid models. After checking the accuracy, they have found out that Principal Component Analysis (PCA) can be a critical factor to improve performance.

Hafizah *et al.* [23] compared SVM and ANN using four different datasets of breast and liver cancer including WBCD, BUPA JNC, Data, Ovarian. The researchers have demonstrated that both methods are having high performance but still, SVM was better than ANN.

Azar and El-Said [24] worked on six different methods of SVM. They have compared ST-SVM with LPSVM, LSVM, SSVM, PSVM, and NSVM to find out which method performs the best in accuracy, sensitivity, specificity, and ROC. LPSVM proved to be the best with accuracy 97.1429%, sensitivity 98.2456%, specificity 95.082%, and ROC 99.38%. Therefore, LPSVM has the highest performance.

Deng and Perkowski [25] used a new method called Weighted Hierarchical Adaptive Voting Ensemble (WHAVE). They compared the accuracy of WHAVE with seven other methods that had the highest accuracies in previous researchers. WHAVE proved to achieve the highest performance value of 99.8%.

Rehman *et al.* [26] extracted different features including Phylogenetic trees, Statistical Features and Local Binary Patterns from mammography images. They used a hybrid model combined with SVM and RBF for classification. They checked the accuracy of each feature separately. In this step the best accuracy value was 76% for 90 features that were chosen based on Taxonomic Indices based Feature (TIF) Vector, 68% for Statistical and LBP based Feature Vector, then the features were combined (Taxonomic Indices, Statistical and LBP based Feature Vector) and again checked for accuracy. The evaluation results were the best after 4 times testing. The researchers claimed that to increase performance and efficiency of detecting breast cancer is performed by using different features.

Mejia *et al.* [27] have chosen Thermogram images for detecting breast cancer as it is cheaper and safer than other methods. It can detect cancer in the earlier stage compared to other images or tests, and it doesn’t have any limitation such as pregnancy, size or density of breast. Also, it doesn’t need any complex features for extracting. They selected 18 cases with 9 abnormal and 9 normal cases. KNN classifier was used to improve the accuracy. The results were 88.88% for abnormal and 94, 44% for normal cases.

Ayeldeen *et al.* [28] used AI and its techniques for breast cancer detection. They used 5 different methods for performance comparison. RF algorithm showed the highest result with 99% performance.

Avramov and Si [29] worked on feature extraction and the impact of the selection on performance. They applied 4 ways of correlation selection (PCA, T-Test Significance and Random feature selection) and 5 models of classification (LR, DT, KNN, LSVM, and CSVM). Best result was achieved by stacking the logistic, SVM and CSVM improve accuracy to 98.56%.

Ngadi *et al.* [30] used NSVC algorithm to test different classification methods including RBF, Poly, and Linear. Then they compared the results with other classification methods such as Naïve Bayes, DT, K-NN, SVM, RF, and Adaboost. RF has the best performance result with 93% accuracy. This proves that NSVC was better than the other methods.

Jiang and Xu [31] used Diffusion-Weighted Magnetic Resonance Image (DWI) for breast cancer detection. They used two types of features; one based on ROI and another one based on ADC- on 61 patient’s data. Moreover, they implemented RF-RFE and RF algorithm was used. The study findings show that the accuracy of RF-RFE and RF and Histogram + GLCM is 77.05% which indicates that feature-based texture has a critical role in improving performance and detection.

Salma [32] selected two different data sets from WBCD and KDD also they used FM-ANN for both of them. They compared the results with other techniques (RBF, FNN, and MNN). After training and testing KDD achieved better accuracy of 99.96% due to the number of features were more. Comparing the results FM- ANN proved to be more accurate.

Bevilacqua *et al.* [33] selected MR images for training and testing. After extracting data and processing, they used ANN for classification and detecting breast cancer. However, when Genetic Algorithm was used to optimize ANN, the observed specificity was 90.46%, sensitivity was 89.08% and the average accuracy was improved to 89.77% and high accuracy changed to 100%.

Table 1 represents all the related work ML method used in this study [17-33]. It contains the references, type of extracted features, data sets and measured performances. Performance is the most significant feature in choosing the proper method.

Table 1
Related work on different types of methodology, features, dataset, and references for breast cancer detection

<table>
<tr><th>R</th><th>Methodology</th><th>Features</th><th>Data Base</th><th>Performance</th><th>Dataset</th></tr>
<tr><td>[17]</td><td>SVM</td><td>Variance, Range, Compactness</td><td>Mammogram</td><td>
<table>
<tr><th></th><th>MCC</th><th>Sensitivity</th><th>Specificity</th><th>Accuracy</th></tr>
<tr><td>Variance</td><td>83.2%,</td><td>95%,</td><td>88%</td><td>91.5%</td></tr>
<tr><td>Range</td><td>82.1%</td><td>94%</td><td>88%</td><td>90.5%</td></tr>
<tr><td>Compactness</td><td>70%</td><td>86%</td><td>84%</td><td>85%</td></tr>
</table>
</td><td>Digital Database for Screening Mammography (DDSM)</td></tr>
<tr><td>[18]</td><td>GMM
KNN</td><td>Tissue</td><td>Microwave Tomography Image</td><td>
<table>
<tr><th></th><th>MCC%</th><th>Sensitivity</th><th>Specificity</th><th>Precision</th><th>Accuracy</th></tr>
<tr><td>KNN</td><td>**67%**</td><td>**87%,**</td><td>84%</td><td>70%</td><td>80-90%</td></tr>
<tr><td>GMM</td><td>48%</td><td>67%</td><td>**86%**</td><td>**70.8%**</td><td>70-80%</td></tr>
</table>
</td><td>ETRI</td></tr>
<tr><td>[19]</td><td>SVM, KNN, RSDA</td><td>Fuzzy Histogram Hyperonization, Fuzzy C-mean, and Gray level dependence model</td><td>Mammogram</td><td>
<table>
<tr><th></th><th>Training set</th><th>Accuracy %</th></tr>
<tr><td>Normal</td><td>70</td><td>100</td></tr>
<tr><td>Benign</td><td>60</td><td>96.67</td></tr>
<tr><td>**Malignant**</td><td>50</td><td>**94**</td></tr>
</table>
</td><td>Mammographic Image Analysis Society (MIAS)</td></tr>
<tr><td>[20]</td><td>SVM, GMM</td><td>Contrast, Homogeneity, Mean, Correlation, Energy, Maximum</td><td>Mammography</td><td>
<table>
<tr><th></th><th>MCC</th><th>Sensitivity</th><th>Specificity</th></tr>
<tr><td>SVM</td><td>78.78%</td><td>82%</td><td>96%</td></tr>
<tr><td>GMM</td><td>72.06%</td><td>84%</td><td>86%</td></tr>
</table>
</td><td>DDSM University of South Florida</td></tr>
<tr><td>[21]</td><td>LRC</td><td>Mitoses, Marginal-Adhesion, Normal Nucleoli, Clump Thickness, Bland Chromatin, Uniformity of cell shape, Single Epithelial cell size, Uniformity of cell size, Bare Nuclei</td><td>Standard Data</td><td>
<table>
<tr><th></th><th>Accuracy percentage</th></tr>
<tr><td>**LRC**</td><td>**99.25**</td></tr>
<tr><td>BFI</td><td>95.46</td></tr>
<tr><td>ID3</td><td>92.99</td></tr>
<tr><td>J48</td><td>98.14</td></tr>
<tr><td>SVM</td><td>96.40</td></tr>
</table>
</td><td>UCI</td></tr>
<tr><td>[22]</td><td>SVM, ANN, NB, Adaboost tree, PCA</td><td>WBC: Mitoses, Marginal-Adhesion, Normal Nucleoli, Clump Thickness, Bland Chromatin, Uniformity of cell shape, Single Epithelial cell size, Uniformity of cell size, Bare Nuclei
WDBC, Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave Points Symmetry, Fractal Dimension</td><td>Standard Data</td><td>
<table>
<tr><th></th><th colspan="2">Accuracy percentage</th></tr>
<tr><th></th><th>WBC</th><th>WDBC</th></tr>
<tr><td>**SVM**</td><td>97.10</td><td>97.99</td></tr>
<tr><td>PCs-SVM</td><td>**97.47**</td><td>98.12</td></tr>
<tr><td>PCi-SVM</td><td>96.73</td><td>97.90</td></tr>
<tr><td>**ANN**</td><td>89.88</td><td>99.60</td></tr>
<tr><td>PCs-ANN</td><td>95.52</td><td>99.61</td></tr>
<tr><td>PCi-ANN</td><td>94.33</td><td>**99.63**</td></tr>
<tr><td>**Naïve**</td><td>96.21</td><td>93.32</td></tr>
<tr><td>PCs-Naïve</td><td>96.50</td><td>91.79</td></tr>
<tr><td>PCi-Naïve</td><td>96.16</td><td>91.72</td></tr>
<tr><td>**Adaboost**</td><td>95.84</td><td>97.19</td></tr>
<tr><td>PCs-Adaboost</td><td>96.24</td><td>96.73</td></tr>
<tr><td>PCi-AdaBoost</td><td>96.32</td><td>96.83</td></tr>
</table>
</td><td>Wisconsin Breast Cancer Database Original (WBC)
Wisconsin Diagnostic Breast Cancer Database (WDBC)</td></tr>
<tr><td>[23]</td><td>ANN,
SVM</td><td>Mitoses, Marginal-Adhesion, Normal Nucleoli, Clump Thickness, Bland Chromatin, Uniformity of cell shape and size, Single Epithelial cell size, Bare Nuclei</td><td>Standard Data</td><td>
<table>
<tr><th></th><th>Accuracy</th><th>Sensitivity</th><th>Specificity</th><th>AUC</th></tr>
<tr><td>SVM</td><td>99.51%</td><td>99.25%</td><td>100%</td><td>99.63%</td></tr>
<tr><td>ANN</td><td>98.54%</td><td>99.25%</td><td>97.22%</td><td>98.24%</td></tr>
</table>
</td><td>Wisconsin Breast Cancer Database (WBCD)</td></tr>
<tr><td>[24]</td><td>St-SVM,
PSVM,
LSVM,
NSVM,
LPSVM,
SSVM</td><td>Mitoses, Marginal-Adhesion, Normal Nucleoli, Clump Thickness, Bland Chromatin, Uniformity of cell shape and size, Single Epithelial cell size, Bare Nuclei</td><td>Mammography</td><td>
<table>
<tr><th></th><th>Accuracy</th><th>Sensitivity</th><th>Specificity</th><th>ROC</th></tr>
<tr><td>**LPSVM**</td><td>**97.1429**</td><td>**98.2456**</td><td>95.082</td><td>**99.38**</td></tr>
<tr><td>LSVM</td><td>95.4286</td><td>96.5217</td><td>93.3333</td><td>97.18</td></tr>
<tr><td>SSVM</td><td>96.5714</td><td>96.5812</td><td>**96.**5517</td><td>98.35</td></tr>
<tr><td>PSVM</td><td>96</td><td>97.3684</td><td>93.4426</td><td>97.75</td></tr>
<tr><td>NSVM</td><td>96.5714</td><td>96.5812</td><td>**96.**5517</td><td>98.35</td></tr>
<tr><td>ST-SVM</td><td>94.86</td><td>95.65</td><td>93.33</td><td>96.61</td></tr>
</table>
</td><td>WBCD</td></tr>
<tr><td>[25]</td><td>Weighted Hierarchical Adaptive Voting Ensemble (WHAVE) Disjunctive Normal Form (DNF) rule-based method,
DT, NB, SVM</td><td>Mitoses, Marginal-Adhesion, Normal Nucleoli, Clump Thickness, Bland Chromatin, Uniformity of cell shape and size, Single Epithelial cell size, Bare Nuclei</td><td></td><td>
<table>
<tr><th>Method</th><th>Accuracy Percentage</th></tr>
<tr><td>DNF</td><td>65. 72</td></tr>
<tr><td>DT</td><td>94.74</td></tr>
<tr><td>NB</td><td>84.5</td></tr>
<tr><td>SVM</td><td>99.54</td></tr>
<tr><td>Hybrid</td><td>99.54</td></tr>
<tr><td>KNN</td><td>97.14</td></tr>
<tr><td>Quadratic Classifier</td><td>97.14</td></tr>
<tr><td>**WHAVE**</td><td>**99.8**</td></tr>
</table>
</td><td>WBCD</td></tr>
<tr><td>[26]</td><td>SVM
RBF kernel</td><td>Phylogenetic trees, Statistical Features, and Local Binary Patterns</td><td>DDSM</td><td>
<table>
<tr><th rowspan="3">Training</th><th rowspan="3">Testing (%)</th><th colspan="2">Model I</th><th colspan="2">Model II</th><th colspan="2">Model III</th></tr>
<tr><th colspan="2">TIF %</th><th colspan="2">(LBP) %</th><th colspan="2">TIF and LBP %</th></tr>
<tr><th>Accuracy</th><th>Specificity</th><th>Accuracy</th><th>Specificity</th><th>Accuracy</th><th>Specificity</th></tr>
<tr><td>80</td><td>20</td><td>64</td><td>58</td><td>54</td><td>51</td><td>66</td><td>60</td></tr>
<tr><td>70</td><td>30</td><td>71</td><td>66</td><td>52</td><td>49</td><td>65</td><td>61</td></tr>
<tr><td>**60**</td><td>**40**</td><td>**76**</td><td>73</td><td>68</td><td>**64**</td><td>**80**</td><td>**76**</td></tr>
<tr><td>50</td><td>50</td><td>70</td><td>76</td><td>64</td><td>60</td><td>72</td><td>67</td></tr>
</table>
</td><td>MIAS</td></tr>
<tr><td>[27]</td><td>KNN</td><td>Mean, Standard Deviation</td><td>Thermogram</td><td>
<table>
<tr><th></th><th colspan="2">Accuracy</th></tr>
<tr><td rowspan="2">KNN</td><td>Normal</td><td>Abnormal</td></tr>
<tr><td>94.44%</td><td>88.88%</td></tr>
</table>
</td><td>Federal Fluminense University Hospital</td></tr>
</table>

<table>
<tr><th>R</th><th>Methodology</th><th>Features</th><th>Data Base</th><th>Performance</th><th>Dataset</th></tr>
<tr><td>[28]</td><td>Bayes Net (BN), Multi-Class Classifier, DT, Radial Basis Function, RF</td><td>TP Rate, FP Rate, Precision, Recall, F-measure, ROC area</td><td>Blood Serum</td><td>
<table>
<tr><th></th><th>RF on TP rate</th><th>FP Rate</th><th>Precision</th><th>Recall</th><th>F</th><th>ROC</th></tr>
<tr><td>BN</td><td>0.947</td><td>0.035</td><td>0.949</td><td>0.947</td><td>0.945</td><td>0.995</td></tr>
<tr><td>Multi CC</td><td>0.933</td><td>0.043</td><td>0.933</td><td>0.933</td><td>0.93</td><td>0.987</td></tr>
<tr><td>DT</td><td>0.87</td><td>0.084</td><td>0.878</td><td>0.87</td><td>0.868</td><td>0.966</td></tr>
<tr><td>RBF</td><td>0.774</td><td>0.128</td><td>0.722</td><td>0.774</td><td>0.739</td><td>0.908</td></tr>
<tr><td>RF</td><td><b>0.99</b></td><td><b>0.007</b></td><td><b>0.99</b></td><td><b>0.99</b></td><td><b>0.99</b></td><td><b>1</b></td></tr>
</table>
</td><td>Department of Biochemistry and Molecular Biology of Kasr Alainy</td></tr>
<tr><td>[29]</td><td>Logistic Regression (LR), DT. KNN, Cubic SVM (CSVM)</td><td>Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave Points, Symmetry, Fractal, Dimension</td><td>Microscope Digital Image</td><td>
<table>
<tr><th></th><th>Accuracy percentage</th></tr>
<tr><td><b>DT with 30 features</b></td><td>92.51</td></tr>
<tr><td><b>KNN with 30 features</b></td><td>91.56</td></tr>
<tr><td>LR with 3 features</td><td>96.27</td></tr>
<tr><td>LR with 6 features</td><td><b>97.77</b></td></tr>
<tr><td><b>LR with 30 features</b></td><td>95.65</td></tr>
<tr><td>LSVM with 3 features</td><td>97.47</td></tr>
<tr><td>LSVM with 10 features</td><td><b>97.87</b></td></tr>
<tr><td><b>LSVM with 30 features</b></td><td>97.30</td></tr>
<tr><td>CSVM with 11 features</td><td>97.98</td></tr>
<tr><td>SVM and CSVM</td><td><b>98.56</b></td></tr>
<tr><td><b>CSVM with 30 features</b></td><td>98</td></tr>
<tr><td>Stacking the Logistic, LSVM, and CSVM</td><td>98.56</td></tr>
</table>
</td><td>UCI</td></tr>
<tr><td>[30]</td><td>NSVC</td><td>BI-RADS, Age, Shape, Margin, Density, Severity</td><td>Mammography</td><td>Accuracy: 99%</td><td>UCI</td></tr>
<tr><td>[31]</td><td>RF-Recursive Feature Elimination (RF-RFE) method</td><td>ROI: Mean, Variance, Skewness, Kurtosis, Energy, Entropy
ADC: Contrast, Entropy, ASM, Correlation</td><td>Diffusion-Weighted Magnetic Resonance Image (DW (Convert to ADC)-MRI)</td><td>
<table>
<tr><th></th><th>Accuracy</th><th>Sensitivity</th><th>Specificity</th><th>AUC</th></tr>
<tr><td>RF-RFE and RF</td><td><b>77.05%</b></td><td><b>84.21%</b></td><td><b>65.21%</b></td><td><b>0.76</b></td></tr>
<tr><td>Histogram</td><td>68.85%</td><td>76.32%</td><td>56.52%</td><td>0.73</td></tr>
<tr><td>GLCM</td><td>65.57%</td><td>71.05%</td><td>56.52%</td><td>0.63</td></tr>
<tr><td>Histogram + GLCM</td><td><b>77.05%</b></td><td><b>84.21%</b></td><td><b>65.21%</b></td><td><b>0.76</b></td></tr>
</table>
</td><td>Zhejiang Cancer Hospital</td></tr>
<tr><td>[32]</td><td>Fast Modular Artificial Neural Network (FM-ANN)</td><td>WBCD: f4, f8, f12, f14, f24, f27, f28
KDD: f22, f29, f47, f50, f60, f61, f62, f63, f64, f65, f71, f97f80, f98, f108,</td><td>X-Ray</td><td>
<table>
<tr><th></th><th>Feedforward %</th><th>MLP %</th><th>RBF %</th><th>MNN %</th><th>FM-ANN</th></tr>
<tr><td>WBCD 70:30</td><td>98.45</td><td>91.50</td><td>93.75</td><td>99.22</td><td>99.80</td></tr>
<tr><td>WBCD 50:50</td><td>94.91</td><td>89.5</td><td>90.65</td><td>93.57</td><td>95.71</td></tr>
<tr><td colspan="2"><b>WBCD after training Accuracy</b></td><td colspan="2"><b>99.8</b></td><td></td><td></td></tr>
<tr><td>KDD 70:30</td><td>94.91</td><td>93.95</td><td>98.45</td><td>99.22</td><td>99.96</td></tr>
<tr><td>KDD 50:50</td><td>93.21</td><td>92.95</td><td>97.98</td><td>98.22</td><td>98.96</td></tr>
<tr><td colspan="2"><b>KDD cup 2008 after training Accuracy</b></td><td colspan="2"><b>99.96</b></td><td></td><td></td></tr>
</table>
</td><td>WBCD, KDD Cup 2008</td></tr>
<tr><td>[33]</td><td>Optimized ANN</td><td>Size, Convexity, Solidity, Eccentricity, Aspect ratio, Circularity, the standard deviation value of the gray levels of images with and without MC in ROIs;</td><td>MRI</td><td>
<table>
<tr><th></th><th>High Accuracy</th><th>Average Accuracy</th><th>Sensitivity</th><th>Specificity</th></tr>
<tr><td>Optimized ANN</td><td>100%</td><td>89.77%</td><td>89.08%</td><td>90.46%</td></tr>
</table>
</td><td>Radiologists of the University of Bari Aldo Moro</td></tr>
</table>

According to Figure 2, most researchers have worked on mammogram images as its quicker than other types of breast cancer detection and it is safe and more effective [34].

Figure 3 presents a comparison of using ML methods and algorithms methodologies employed for breast cancer detection in the reviewed literature listed in Table 1. It is observed that SVM is the most frequently used method. Whereby, Figure 4 presents the results of breast cancer detection using ML methods.

V. Conclusion

In the present paper, breast cancer and ML were introduced as well as an in-depth literature review was performed on existing ML methods used for breast cancer detection. The findings of these researchers suggest that SVM is the most popular method used for cancer detection applications. SVM was used either alone or combined with another method to improve the performance. The maximum achieved accuracy of SVM (single or hybrid) was 99.8% that can be improved to 100%. It was observed from the work of [33] who used optional ANN on MRI resulted in 100% accuracy in detecting breast cancer. This method can be applied and tested on another dataset like mammogram and ultrasound to check the performance of different data types. The mammogram was the most frequent data set used compared to other types of data such as ultrasound images, thermal images or blood features.

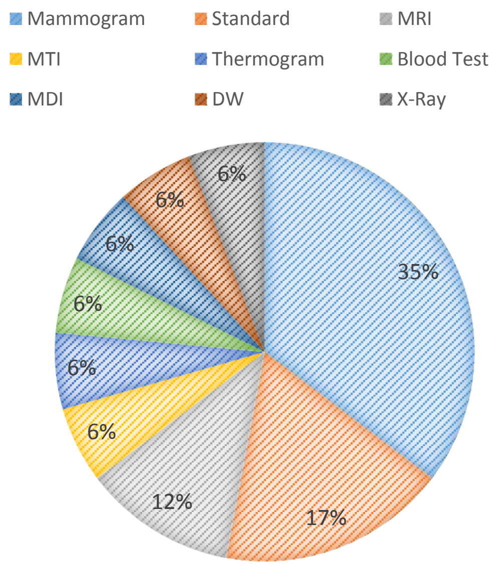

Figure 2: Different breast cancer detection methods

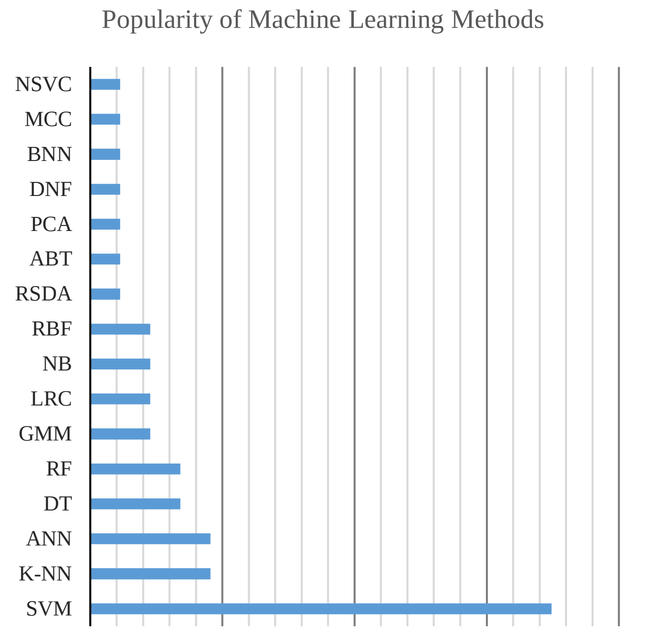

Figure 3: Using machine learning methods in cancer detection

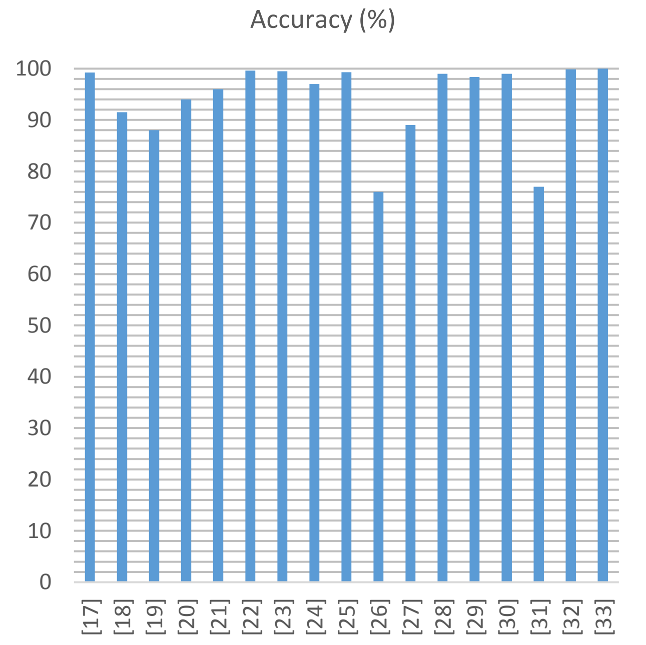

Figure 4: Accuracy percentages in different literatures

REFERENCES

[1] World Health Organization, “Cancer country profiles 2014,” WHO, http://www.who.int/cancer/country-profiles/en/

[2] M. Stalin, and R. Kalaimagal, “Breast cancer diagnosis from low-intensity asymmetry thermogram breast images using fast support vector machine,” *i-manager's Journal on Image Processing*, vol. 3, no. 3, pp. 17–26, 2016.

[3] R. Kirubakaran, T. C. Jia, and N. M. Aris, “Awareness of Breast Cancer among Surgical Patients in a Tertiary Hospital in Malaysia,” *Asian Pacific Journal of Cancer Prevention*, 2017, vol. 18, no. 1, pp. 115–120.

[4] T. M. Khan, and S. A. Jacob, “Brief review of complementary and alternative medicine use among Malaysian women with breast cancer,” *Journal of Pharmacy Practice and Research*, 2017, vol. 47, no. 2, pp. 147–152.

[5] L. Caplan, “Delay in breast cancer: implications for the stage at diagnosis and survival,” *Frontiers in Public Health*, 2014, vol. 2, Article 87, pp. 1–6.

[6] M.A. Richards, A.M. Westcombe, S.B. Love, P. Littlejohns, and A.J. Ramirez, “Influence of delay on survival in patients with breast cancer: a systematic review,” *The Lancet*, 1999, vol. 353, no. 9159, pp. 1119-1126.

[7] B. Stewart and C.P. Wild, *World Cancer Report 2014*, International Agency for Research on Cancer, WHO, 2014.

[8] S. A. Korkmaz, and M. Poyraz, “A New Method Based for Diagnosis of Breast Cancer Cells from Microscopic Images: DWEE—JHT,” *J. Med. Syst.*, vol. 38, no. 9, p. 92, 2014.

[9] P. Louridas, and C. Ebert, “Machine Learning,” *IEEE Softw*., vol. 33, no. 5, pp. 110–115, 2016.

[10] A. Simons, “Using artificial intelligence to improve early breast cancer detection, “2017. Retrieved on April 10, 2018, from https://www.csail.mit.edu/news/using-artificial-intelligence-improve-early-breast-cancer-detection

[11] E. Ali, and W. Feng, “Breast Cancer classification using Support Vector Machine and Neural Network,” *International Journal of Science and Research*, pp. 2013, 2319-7064.

[12] S. Medjahed, T. Saadi, and A. Benyettou, “Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules,” *International Journal of Computer Applications*, 2013, vol. 62, no. 1, pp. 0975 – 8887.

[13] R. Sumbaly, N. Vishnusri, and S. Jeyalatha, “Diagnosis of Breast Cancer using Decision Tree Data Mining Technique,” *International Journal of Computer Applications*, 2014, vol. 98, no. 10, pp. 0975 – 8887.

[14] M. Elgedawy, “Prediction of Breast Cancer using Random Forest, Support Vector Machines and Naïve Bayes,” *International Journal of Engineering and Computer Science*, 2017, vol. 6, no. 1, pp. 19884-19889.

[15] R. Senkamalavalli, and T. Bhuvaneswari,” Improved classification of breast cancer data using hybrid techniques, “*International Journal of Advanced Research in Computer Science.* 2017, vol. 8, no. 8, pp. 454-457.

[16] A. Hazra, S. Mandal, and A. Gupta” Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms,” *International Journal of Computer Applications.* 2016, vol. 145, no.2, pp. 0975 – 8887.

[17] S. Gc, R. Kasaudhan, T. K. Heo, and H.D. Choi, “Variability Measurement for Breast Cancer Classification of Mammographic Masses,” in *Proceedings of the 2015 Conference on research in adaptive and convergent systems (RACS)*, Prague, Czech Republic, 2015, pp. 177–182.

[18] C. Wang, W. Wang, S. Shin, and S. I. Jeon, “Comparative Study of Microwave Tomography Segmentation Techniques Based on GMM and KNN in Breast Cancer Detection,” in *Proceedings of the 2014 Conference on Research in Adaptive and Convergent Systems (RACS '14)*, Towson, Maryland, 2014, pp. 303–308.

[19] C. L. Chowdhary, and D. P. Acharjya, “Breast Cancer Detection using Intuitionistic Fuzzy Histogram Hyperbolization and Possibilitic Fuzzy c-mean Clustering algorithms with texture feature-based Classification on Mammography Images,” in *Proceedings of the International Conference on Advances in Information Communication Technology & Computing*, Bikaner, India, 2016, pp. 1–6.

[20] S. Aminikhanghahi, S. Shin, W. Wang, S. I. Jeon, S. H. Son, and C. Pack, “Study of wireless mammography image transmission impacts on robust cyber-aided diagnosis systems,” *Proc. 30th Annu. ACM Symp. Appl. Comput. - SAC ’15*, pp. 2252–2256, 2015.

[21] S. G. Durai, S. H. Ganesh, and A. J. Christy, “Novel Linear Regressive Classifier for the Diagnosis of Breast Cancer,” *In Computing and Communication Technologies (WCCCT), 2017 World Congress on* 2017.

[22] H. Wang, and S. W. Yoon, “Breast cancer prediction using data mining method,” *IIE Annu. Conf. Expo 2015*, pp. 818–828, 2015.

[23] S. Hafizah, S. Ahmad, R. Sallehuddin, and N. Azizah, “Cancer Detection Using Artificial Neural Network and Support Vector Machine: A Comparative Study,” *J. Teknol*, vol. 65, pp. 73–81, 2013.

[24] A. T. Azar, and S. A. El-Said, “Performance analysis of support vector machines classifiers in breast cancer mammography recognition,” *Neural Comput. Appl*., vol. 24, no. 5, pp. 1163–1177, 2014.

[25] C. Deng, and M. Perkowski, “A Novel Weighted Hierarchical Adaptive Voting Ensemble Machine Learning Method for Breast Cancer Detection,” *Proc. Int. Symp. Mult. Log*., vol. 2015–Septe, pp. 115–120, 2015.

[26] A. U. Rehman, N. Chouhan, and A. Khan, “Diverse and Discriminative Features Based Breast Cancer Detection Using Digital Mammography,” *2015 13th Int. Conf. Front. Inf. Technol*., pp. 234–239, 2015.

[27] T. M. Mejia, M. G. Perez, V. H. Andaluz, and A. Conci, “Automatic Segmentation and Analysis of Thermograms Using Texture

Descriptors for Breast Cancer Detection," 2015 *Asia-Pacific Conf. Comput. Aided Syst. Eng*., pp. 24–29, 2015.

[28] H. Ayeldeen, M. A. Elfattah, O. Shaker, A. E. Hassanien, and T.-H. Kim, "Case-Based Retrieval Approach of Clinical Breast Cancer Patients," *2015 3rd Int. Conf. Comput. Inf. Appl*., pp. 38–41, 2015.

[29] T. K. Avramov and D. Si, "Comparison of Feature Reduction Methods and Machine Learning Models for Breast Cancer Diagnosis," *Proc. Int. Conf. Comput. Data Anal. - ICCDA '17,* pp. 69–74, 2017.

[30] M. Ngadi, A. Amine, and B. Nassih, "A Robust Approach for Mammographic Image Classification Using NSVC Algorithm," *Proc. Mediterr. Conf. Pattern Recognit. Artif. Intell. - MedPRAI-2016*, pp. 44–49, 2016.

[31] Z. Jiang, and W. Xu, "Classification of benign and malignant breast cancer based on DWI texture features," *ICBCI 2017 Proceedings of the International Conference on Bioinformatics and Computational Intelligence* 2017.

[32] M. U. Salma, "Fast Modular Artificial Neural Network for the Classification of Breast Cancer Data," *Proc. Third Int. Symp. Women Comput. Informatics - WCI '15*, pp. 66–72, 2015.

[33] V. Bevilacqua, A. Brunetti, M. Triggiani, D. Magaletti, M. Telegrafo, and M. Moschetta, "An Optimized Feed-forward Artificial Neural Network Topology to Support Radiologists in Breast Lesions Classification," P*roc. 2016 Genet. Evol. Comput. Conf. Companion - GECCO '16 Companion*, pp. 1385–1392, 2016.

[34] M. Rmili, and A. El, "A Combined Approach for Breast Cancer Detection in Mammogram," *2016 13th International Conference on Computer Graphics, Imaging and Visualization*, pp. 350–353, 2016.