Evaluating Deep and Statistical Machine Learning Models in the Classification of Breast Cancer from Digital Mammograms

Amel A. Alhussan[1], Nagwan M. Abdel Samee[2]*, Vidan F. Ghoneim[3], Yasser M. Kadah[4]
Computer Science Department, College of Computer & Information Sciences[1]
Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia[1]
Information Technology Department, College of Computer & Information Sciences[2]
Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia[2]
Computer Engineering Department, Misr University for Science and Technology, Giza, Egypt[2]
Biomedical Engineering Department, Helwan University, Cairo, Egypt[3]
Electrical and Computer Engineering Department, King Abdulaziz University, Jeddah, Saudi Arabia[4]
Biomedical Engineering Department, Cairo University, Giza, Egypt[4]

***Abstract*—The application of artificial intelligence techniques in computer aided detection and diagnosis problems has been among the most promising areas with interest from the scientific community and healthcare industry. Recently, deep learning has become the prime tool for such application with many studies focusing on developing variants that optimize diagnostic performance. Despite the widely accepted success of this class of techniques in this application by the scientific community, it is not prudent to consider it as the only tool available for such purpose. In particular, statistical machine learning offers a variety of techniques that can also be applied at a much lower computational cost. Unfortunately, the results from both strategies cannot be directly compared due to the differences in experimental setups and datasets used in available research studies. Therefore, we focus in this study on this direct comparison using the same dataset and similar data preprocessing as the input to both. We compare statistical machine learning to deep learning in the context of computer-aided detection of breast cancer from mammographic images. The results are compared using diagnostic performance metrics and suggest that simpler statistical machine learning techniques may provide better performance with simpler architectures that allow explanation of results.**

Keywords*—Computer-aided detection; computer-aided diagnosis; statistical machine learning; deep learning

I. Introduction

Breast cancer is the most frequently diagnosed cancer and accounts for a significant portion of the total cancer related deaths among women[1]. The early detection of cancer in general, and particularly in breast cancer, is crucial to patient survival. Therefore, periodic screening was recommended for women above a certain age or before that for those women with a family history of the disease. The primary imaging modality for such screening is x-ray mammography where two images in craniocaudal and mediolateral oblique directions are taken and examined carefully by a radiologist for early signs of abnormalities including microcalcifications [2]. The resultant images in their digital form have very high resolution and quantization level (for example, 5k resolution is common at 12-16 bits of grayscale). As a result, the process for reading such images is tiring, lengthy, costly, and prone to errors. Moreover, the shortage of radiologists compounded by increase in volume of image data due to better awareness and introduction of 3D techniques such as tomosynthesis poses a challenge for healthcare services in this area. Therefore, computer-aided diagnosis is now pursued as a possible solution to this problem. Even though many such systems were proposed early on as applications of the growing artificial intelligent systems, the digital transition of radiology departments made the utilization of such assisting tools more readily available in many applications including mammography.

Computer-aided diagnosis (CAD) is generally defined as a diagnosis made by a radiologist who uses the output of a computer analysis of the images when making his/her interpretation. CAD systems can play different roles in the diagnostic process. For example, it can be used for as pre-screening where the CAD system is utilized as the first reader then the radiologist verifies such reading and makes the final diagnosis. Alternatively, concurrent reading of images between the radiologist and the CAD system, which in this case serves as a second independent reader. Also, another approach is to make the diagnosis interactively using the CAD system where the radiologist marks suspicious areas on image and uses analysis from the CAD system to confirm the likely diagnosis. Therefore, this approach improves diagnostic performance, reduces performance intra- and inter-observer variability of radiologists, improves radiologist productivity and hence serves as a mitigation of global shortage of radiologists.

The early CAD systems relied on statistical machine learning techniques (e.g., [3]), while most recent scientific studies targeting this field were overwhelmingly using deep learning techniques (e.g., [4] [5] [6]). This is a natural consequence of the usual technology hype cycle (sometimes called Gartner hype curve) of deep learning technologies where this area is within the peak of inflated expectations. In order to speed up the process of reaching the plateau of productivity of

*Corresponding Author

such curve with its associated wide adoption of the technology, it is important to consider comparisons with earlier technologies in order to better assess potential and realize limitations.

In this study, we address the direct comparison of statistical machine learning and deep learning techniques in the context of computer-aided detection of breast cancer from mammographic images. The same dataset and similar data preprocessing are used as the input to several techniques that represent both categories. The details of all steps of implementation are provided to allow reproducibility of results and the performance is compared using statistical diagnostic performance metrics to allow objective comparison.

Performance evaluation rests at the heart of any machine learning model[7]. It is necessary in selecting the input features and it decides which model is appropriate for each data set. This is a fact that should be considered when we select an algorithm for cancer detection, classical or deep learning-based method, for a CAD system. Currently, there are a huge number of researches[5] [8] [9] that have utilized the deep learning approaches and recommending them as they have yielded higher accuracies, above 97%. However, a comparable accuracy, 96%, has been attained using the conventional machine learning paradigms as surveyed in [10]. Therefore, a comparison of deep learning techniques with the earlier artificial intelligence techniques based on statistical machine learning would provide a valuable insight in this regard. Unfortunately, there are no studies that targeted such direct comparison with common datasets and classification tasks and hence, addressing this comparison would be a useful addition to guide researchers in this field.

This paper is organized as follows. Section II includes the literature review. Section III gives detailed description about the data set employed in this work. The methodologies used in statistical machine learning models and deep learning techniques are presented in Section IV. Sections V and VI presents the results and discussion of both statistical and deep learning methods and the comparison between them. Finally, the conclusions in Section VII summarize this work and its significant results.

II. Related Work

Pretrained CNN models such as Alex net [11], VGG[12], and Googlenet[13] are the most popular pre-trained models for image classification. A survey of 83 research studies is presented by Abdelhafiz *et al.* [14] which demonstrate the significant results gained by CNN models in breast cancer detection and classification. They discuss the datasets used and all limitations and challenges that affect the results. They show the significant results in the latest research of breast cancer classification and emphasize on the significant effects of image preprocessing techniques. They also highlighted the effects of some important customized parameters such as validation techniques, activation function, and learning rates. They found that many studies depend on pretrained models, data augmentation, batch normalization, and dropout techniques to improve their results. Shen *et al.* [15] designed CAD system based on VGG and ResNet pretrained CNN networks. The datasets used to train their machine were DDSM and INbreast datasets. They designed batch classifier and whole image classifier to detect and classify breast cancer. The networks have been adapted by adjusting the number of layers, learning rate and number of epochs. Different techniques such as batch normalization, and data augmentation have been employed to improve the model. The achieved results surpassed the results of previous studies. Also, the CAD system that proposed by Al-antari et. al.[16] based on Deep belief network (DBN) to automatically detect and classify breast cancer. They used two techniques for mass diagnosis, the whole mass ROIs, and Randomly extracted ROI with size 32x32, then they classify the detected masses using their proposed DBN system. The results of their proposed system outperformed other conventional classifiers. Al-masni *et al*. [17] CAD system based on CNN model that is called You Only Look Once (YOLO) technique [18] for automatic detection and classification of breast cancer. In their system, DDSM dataset is used to train and test their system, in addition to the augmented data produced by different techniques such as rotation, translation, and scaling to avoid overfitting. They also utilized number of preprocessing techniques to eliminate irrelevant characteristics of the mammograms. Their CAD system achieved 99.7% for mass detection and 97% for classification. Several studies show that the pretrained CNN models such as Resnet [19], Alexnet [20], [21] and GoogleNet [30] demonstrate higher results using unaugmented patches and more enhanced results with augmented ones. Different CNN models [22][23] show different detection and classification accuracies and performance depending on the application, techniques and datasets used. On other hand, the state-of-the art methods in building CAD system have been compared to the deep learning-based methods in few studies such as the work done in [24] that have evaluated some of the classical methods against the CNN based system. However, the complexity of the incredible performance of the pretrained CNN networks has not been yet fairly assessed and compared to the simple conventional machine learning methods.

III. Data Preparation

In The data used in this work were obtained from the popular Mammography Image Analysis Society (MIAS) database [25]. This database was prepared from x-ray films carefully selected from the United Kingdom National Breast Screening Program and digitized with to a resolution of 50 microns using a device with a linear optical density mapping range from 0 to 3.2 and quantization of 8-bits per pixel. Then, the images were reduced to 200-micron resolution and clipped/padded to maintain size of all images at 1024×1024 pixels in the mini-MIAS version of this database, which was used in this study [26]. The database contains left and right breast images for 161 patients with a total of 322 images. The images represent samples from normal, benign and malignant cases with 208, 63 and 51 images respectively. The database provided the ground truth diagnoses for all images and exact locations of abnormalities that may be present within each image given as the center and radius of the surrounding circle for each lesion. A square region of interest (ROI) of size 32×32 was selected inside the lesion. The size of the ROI was selected this way to ensure adequate statistical representation of the lesion while keeping the size as small as possible to subsequently allow better lesion localization ability for the developed system [3]. A database of 144 ROIs was built with equal number of normal and abnormal regions (72 each). The abnormal regions were selected from the available lesions with 41 benign and 31 malignant samples such that each of them was obtained from a different case to avoid bias in testing. Also, the samples represented the various abnormality subclasses having lesion or cluster sizes that are large enough to contain the selected ROI size. The database of labelled ROIs was then used as the input to both statistical machine learning and deep learning techniques.

IV. Methods

A. Statistical Machine Learning Methods

In this study, several statistical machine learning techniques representing the spectrum of methods in this area are implemented and their parameters are optimized to allow proper performance comparison to be conducted. The general block diagram for all statistical machine learning systems is shown in Fig. 1.

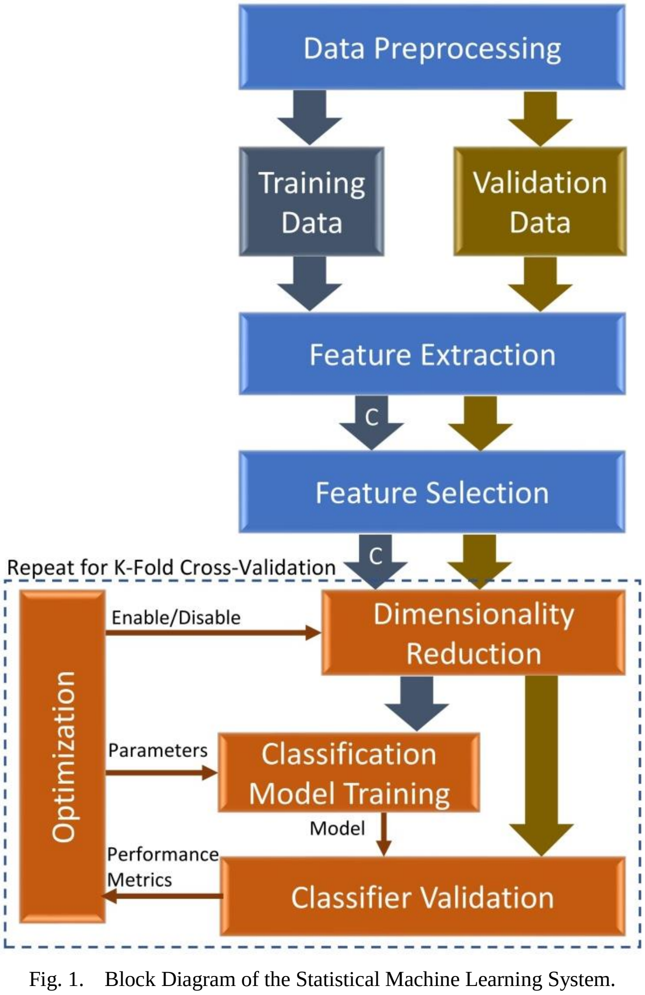

Fig. 1. Block Diagram of the Statistical Machine Learning System.

The learning system in this class of techniques is based on multiple-fold cross-validation to obtain reliable results and minimize the problem of overfitting. In particular, the data for normal and abnormal cases are randomly divided into training and validation sets where the validation set is equal to the number of cases divided by the number of folds and the rest are assigned to the training set. In our system, the number of folds was selected to be 5 such that in each fold, 14 images from each of the normal and abnormal data are used for validation while 58 images from each are used to train the classifier. This is repeated 5 times (same as number of folds) and the results from all folds are averaged together to provide the overall system performance.

A critical part of all statistical machine learning systems is feature extraction. This is the main difference between statistical machine learning and deep learning techniques where features are implicitly learned from the data in the training process. Here, a set of 175 statistical textural features were calculated including 25 first-order features (e.g., mean, standard deviation, percentiles) [27] and 150 higher-order statistical textural features utilizing different textural analysis methods with different attributes including features from the gray-level co-occurrence matrix (GLCM) (also known as spatial gray level dependence method or SGLDM) [28], neighborhood gray tone difference matrix (NGTDM) [29] [30], spatial frequency-based method (SFM) [31], texture energy transform [32][33], fractal analysis [34], and Fourier power spectrum [35].

In order to select the best features that show statistically significant changes between normal and abnormal cases, two-sample t-test was performed between the normal and abnormal training samples in the first fold of the cross-validation process. This avoids any bias from including the validation samples in this process directly in the first fold or indirectly through their inclusion in other folds. The significance level was set at a p-value of 0.05 whereby features showing p-values lower than that are indicated as good features, while the others are deemed indiscriminate and discarded in subsequent steps. This resulted in a total of 43 selected features with 17 first-order and 26 higher-order features.

The next step in the processing pipeline is responsible for dimensionality reduction to minimize the feature space by combining features into major directions that are orthogonal to each other and spanning the directions of most variance in the data. This is done using principal component analysis (PCA) where the features are reduced to only 7 combined features or principal components that explain 95% of the variance of the data. This helps remove redundancies from multiple correlated features. Since the subsequent classification step may include techniques that rely on distance measurements, the dimensionality reduction may not always be needed. In fact, the small amount of variance that was not explained by the output of PCA may contribute to the accuracy of the classification. Therefore, the optimization of different classifiers was allowed to enable or disable such dimensionality reduction in its search for the best performance for each classifier.

The statistical machine learning system of this study included the implementation of six different families of traditional classifiers that included parametric and nonparametric classification methods. Such methods are decision trees, discriminant analysis, ensemble, k-nearest neighbor (KNN), naïve Bayes, and support vector machines (SVM) [36]. The different parameters and variants for each classifier were optimized to reach the best performance using 5-fold cross-validation to obtain reliable estimation of the performance. The performance was measured using the accuracy (Acc), sensitivity (Sens), specificity (Spec), positive predictive value (PPV), and negative predictive value (NPV) [36]. While the accuracy is an important performance metric since it gives the percentage of correct outcomes to the total number of cases, it does not differentiate between false-positive and false-negative errors. This is problematic in the context of a computer-aided detection system where a false-negative diagnosis could have much more severe consequences than a false-positive one. The sensitivity metric addresses that where it gives the percentage of abnormal cases that were correctly diagnosed. On the other hand, the specificity gives the percentage of normal cases that were correctly diagnosed. Together they give the complete picture that allows an observer to compare different systems. For example, if two systems have the same accuracy, a system with a better sensitivity is preferred in a computer-aided detection system. The two other metrics of PPV and NPV address the post examination questions of the reliability of the results. For example, given a particular positive diagnosis outcome from a classifier, a question can be posed as how reliable such result is. Such metrics depend on disease prevalence in the patient population, unlike the sensitivity and specificity metrics.

Given the higher risk of false-negative outcomes of classification, a custom cost function that makes the cost of such errors twice that of false-positives was included in the optimization process. That is, for each classifier, the optimization process that searches through different parameter values and classifier variants is also allowed to add such custom cost function to compare the outcomes of different selections and choose the technique that minimizes this new cost function rather than the one with uniform cost for all types of error.

The results from different classifiers are reported as the cross-validation performance metrics for the best variant of optimal parameters for each classier family. Also, to investigate the effects of dimensionality reduction and custom cost function, the best results are reported for three cases including using no dimensionality reduction or custom cost function, using custom cost function, and using both dimensionality reduction and custom cost function.

B. Deep Learning Methods

Three deep learning networks were considered in this study. These networks are AlexNet [11][37], GoogLeNet [38] , and VGG-16 [12] networks that represent different architectures with different levels of complexity as represented by the number of parameters (weights and biases). In particular, the numbers of parameters for these networks are approximately 61 million for AlexNet, 7 million for GoogLeNet, and 138 million for VGG-16. Even with the least complex of them, the huge number of parameters suggests that it is not possible to properly train such networks with the limited data set available in this study. In fact, it is difficult to collect sufficiently large data sets for such purpose for most medical diagnosis problems given the difficulty of such collection in a standardized manner, privacy issues that prevent access without consent, as well as the severe data imbalance usually encountered in medical data with much less abnormal cases than normal cases. Therefore, two strategies are commonly utilized to mitigate this problem. The first is to use data augmentation whereby each image in the original dataset is used to generate multiple images that include the same diagnostic characteristics as the original [39]. The idea behind this approach is that changing the orientation of the abnormality in the image does not affect the diagnosis by a doctor. Therefore, using rotated or flipped versions of the image would present the network with different images that still represent the diagnostic classification of the original one. In this work, 8-fold data augmentation is used whereby each ROI in the dataset is augmented using flipping (left-right and up-down), rotation (90 and 270 degrees), image matrix transposition, in addition to their combinations of flipping left-right of 90 degrees rotated images and flipping left-right of up-down flipped images. This augmentation results in increasing the size of the dataset to 1152 samples representing 576 normal and 576 abnormal ROIs. An illustration of such augmentation is shown in Fig. 2.

Although the size of the dataset is significantly larger after augmentation, this size is still much smaller than the number of network parameters, which means that it cannot be used as is for training the network without compromising the performance due to the certain overfitting problem. Therefore, the second strategy relies on transfer learning to start from pre-trained networks and fine tune such networks to address the classification task at hand. The idea behind this strategy is that the early stages of deep learning networks are trained to extract low-level image features, which is common and similar between different image classification problems, while the rest of the network are intended to learn the specific classes for each application. Therefore, keeping such early stages intact and replacing only the application-specific final stages would make the training requirements much less demanding without sacrificing the overall performance. This is the essence of network-based transfer learning [40]. In this study, this strategy is applied using the selected networks pre-trained using ImageNet database with more than 14 million images [41]. A block diagram of the transfer learning process is shown in Fig. 3.

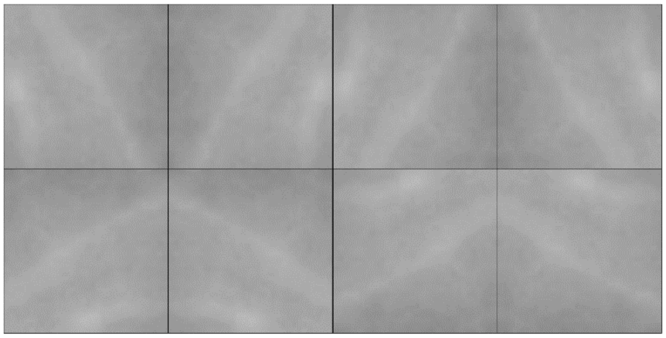

Fig. 2. Illustration of Augmentation Applied to a Sample Abnormal ROI. The Original Image is shown at the Top Left Corner with its Seven Orthogonal Transformations.

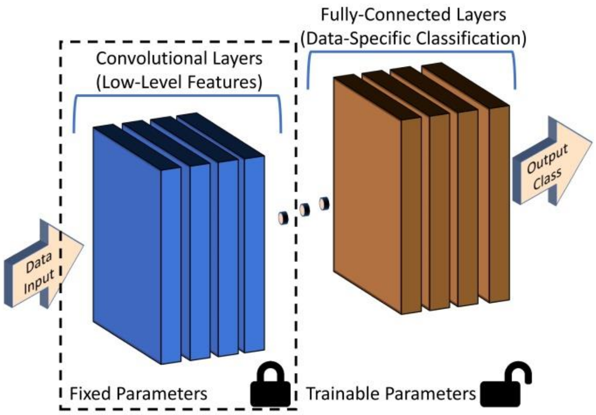

Fig. 3. Illustration of Network-based Transfer Learning.

Given that the input layer must not be altered as a part of the transfer learning strategy, the s ROIs in the dataset were resized to the respective size of each network (227×227 for AlexNet and 224×224 for the others) using bilinear interpolation with an anti-aliasing filter to meet network requirements and maintain the quality of the images and keep them free of aliasing artifacts. All networks require color images rather than grayscale images. This was dealt with by using the same grayscale image for each of the three color components of the network input. Then, the available resized dataset was divided into 3 independent sets at the beginning of the process with 60% designated as training, 20% for validation and 20% for testing. The training data are used to train the parameters of the trainable part of the network to minimize the error in the diagnostic task classification outcome using a suitable optimization technique. On the other hand, the validation data can be used to optimize the hyperparameters of the network including the optimization technique selection and parameters to optimize the performance metrics. Therefore, given that they are both utilized, even in different ways, in the optimization process, it is important to keep an independent set for testing to avoid bias and to be able to assess the presence of overfitting.

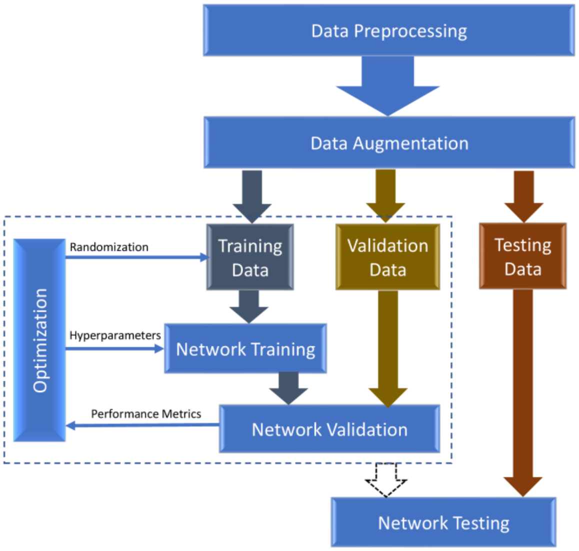

Fig. 4. Block Diagram of the Deep Learning System.

The process of estimating the optimal parameters for each network to give the best performance is a challenging optimization problem because of its high dimensionality and non-convex objective function. Therefore, stochastic optimization techniques are conventionally utilized. In this study, stochastic gradient descent with momentum (SGDM) optimizer is used with a learning rate: 0.0001, momentum term factor of 0.9, L_2-Regularization: 0.0005, and gradient threshold method of L_2-norm. The selected mini-batch size was 16 images, and the maximum number of training epochs was set to 100. Such settings were selected by observing validation results through experimentation and were used for all networks in order to allow direct comparison of their results and computational costs. A diagram representing the deep learning systems used is shown in Fig. 4.

V. Results

The statistical machine learning and deep learning systems were implemented on an academic license of Matlab 2020b (Mathworks, Inc.) with Statistics and Machine Learning and Deep Learning toolboxes. The computer system uses a quad-core Intel® Core™ i7-6700HQ CPU running at 2.60GHz and 16 GB of RAM and NVIDIA GeForce GTX 950M graphics card with 4 GB of memory and CUDA-supported graphics processing units (GPUs). The operating system is Windows 10 Home Edition (version 20H2). Due to the differences in machine configurations and software development environments, the reported computational time measurements of the conducted experiments may be specific to the machine and environment used but the findings derived from their relative values can still be useful to compare different techniques.

The results for the statistical machine learning techniques are presented in Table I. As can be observed, the best accuracy of 99.3% was obtained using a support vector machine classifier with a linear kernel with no feature standardization, uniform cost function, and no PCA feature dimensionality reduction. The second-best accuracy of 98.6% was obtained from a KNN classifier with a Minkowski distance metric, an exponent of 3, a number of neighbors (K) of 5, an inverse distance weight, standardized features, custom cost function that penalizes false negatives double that of false positives, and no PCA feature dimensionality reduction. From the point of view of computer-aided detection, the most important performance metric is the sensitivity. The best sensitivity of 100% was obtained from multiple classifiers including both classifiers with best accuracy, in addition to the other variants of the support vector machine classifier with custom cost function and PCA feature dimensionality reduction. The best specificity of 98.6%, best positive predictive value of 98.6% and best negative predictive value of 100% were also obtained by the same support vector machine classifier variant that provided the best accuracy and sensitivity results. This indicates that this particular classifier has the best overall performance among statistical classification techniques.

The results for the deep learning techniques from pretrained networks using transfer learning are shown in Table II. The results from AlexNet and GoogLeNet networks showed similar results that are consistent in all performance metrics with a value of 98.3% each. On the other hand, the best results were obtained from the VGG-16 network with an accuracy of 98.7%, sensitivity of 99.1%, specificity of 98.3%, positive predictive value of 98.3%, and a negative predictive value of 99.1%. The training hyperparameters for all networks were the same where the optimization algorithm was stochastic gradient descent with momentum with a rate of 0.0001 and L2-regularrization of 0.0005, and 8x data augmentation (total of 1152 images) divided as 60% for training, 20% for validation and 20 for testing. The training was done only to the fully-connected layers of all networks whereby the convolutional layers were kept the same as a part of the transfer learning strategy used. The training was performed on GPU with 100 training epochs and a mini-batch size of 16 with a total time of 25 minutes for AlexNet, 53 minutes for GoogLeNet, and 279 minutes for VGG-16 networks.

TABLE I. PERFORMANCE METRICS OF STATISTICAL LEARNING TECHNIQUES

Method	**PCA**	**Cost**	**Accuracy**	**Sensitivity**	**Specificity**	**PPV**	**NPV**
Decision Tree[1]	-	Custom	97.20%	97.22%	97.22%	97.22%	97.22%
Decision Tree[2]	-	Equal	94.40%	97.22%	91.67%	92.11%	97.06%
Decision Tree[3]	95%	Custom	94.40%	95.83%	93.06%	93.24%	95.71%
Discriminant Analysis[4]	-	Custom	96.50%	95.83%	97.22%	97.18%	95.89%
Discriminant Analysis[4]	-	Equal	95.10%	93.06%	97.22%	97.10%	93.33%
Discriminant Analysis[4]	95%	Custom	97.90%	98.61%	97.22%	97.26%	98.59%
Ensemble[5]	-	Custom	96.50%	97.22%	95.83%	95.89%	97.18%
Ensemble[6]	-	Equal	95.10%	97.22%	93.06%	93.33%	97.10%
Ensemble[7]	95%	Custom	95.10%	97.22%	93.06%	93.33%	97.10%
KNN[8]	-	Custom	98.60%	100.00%	97.22%	97.30%	100.00%
KNN[9]	-	Equal	97.90%	97.22%	98.61%	98.59%	97.26%
KNN[10]	95%	Custom	97.20%	98.61%	95.83%	95.95%	98.57%
Naïve Bayes[11]	-	Custom	92.40%	90.28%	94.44%	94.20%	90.67%
Naïve Bayes[11]	-	Equal	93.10%	88.89%	97.22%	96.97%	89.74%
Naïve Bayes[12]	95%	Custom	96.50%	98.61%	94.44%	94.67%	98.55%
SVM[13]	-	Custom	97.20%	100.00%	94.44%	94.74%	100.00%
SVM[14]	-	Equal	99.30%	100.00%	98.61%	98.63%	100.00%
SVM[14]	95%	Custom	97.90%	100.00%	95.83%	96.00%	100.00%

[1] Maximum number of splits: 4

[2] Maximum number of splits: 13

[3] Maximum number of splits: 12

[4] Linear discriminant

[5] Method: Bag, Number of learning cycles: 13

[6] Method: Bag, Number of learning cycles:119

[7] Method: LogitBoost, Number of learning cycles: 11

[8] Minkowski distance, Number of neighbors: 5, Distance weight: Inverse, Standardized data

[9] Cosine distance, Number of neighbors: 3, Distance weight: Inverse, Standardized data

[10] City block distance, Number of neighbors: 5, Distance weight: Squared inverse

[11] Normal kernel

[12] Triangle kernel

[13] Linear kernel, Standardized data

[14] Linear kernel

TABLE II. PERFORMANCE METRICS OF DEEP LEARNING TECHNIQUES*

Network	**Training Time (minutes)**	**Accuracy**	**Sensitivity**	**Specificity**	**PPV**	**NPV**
AlexNet	25	98.26%	98.26%	98.26%	98.26%	98.26%
GoogLeNet	53	98.26%	98.26%	98.26%	98.26%	98.26%
VGG-16	279	98.70%	99.13%	98.26%	98.28%	99.12%

*Training options: Stochastic Gradient Descent with Momentum (SGDM) optimizer, Mini-Batch Size: 16, Maximum number of training epochs: 100, Learning rate: 0.0001, Data shuffling: every epoch', Validation frequency: 10 steps, Validation Patience: infinity, L_2-Regularization: 0.0005, Execution environment: GPU.

In order to better visualize the results and allow direct comparison between all techniques from both statistical and deep learning approaches, the results from all methods are shown in a graphical form in Fig. 5 and Fig. 6. In Fig. 5, the accuracy for all methods is shown as a square marker that varies in color for different variants. On the other hand, to distinguish the sensitivity and specificity values on the plot, the sensitivity is marked with an upward-pointing triangle, while the specificity is marked as a downward-pointing triangle. This provides an easy visual comparison of the accuracy, sensitivity and specificity values from all techniques. This also allows direct visualization of those techniques where specificity values are higher than those of the sensitivity. In Fig. 6, a similar strategy was used to mark the results of positive predictive values as circles, while those of negative predictive values as asterisks. It is clear that deep learning techniques provide better results than several statistical learning techniques, but they are comparable to several other techniques. Furthermore, the results indicate that deep learning techniques are outperformed by a support vector machine classifier variant in all performance metrics and by several classifiers when the sensitivity metric is emphasized.

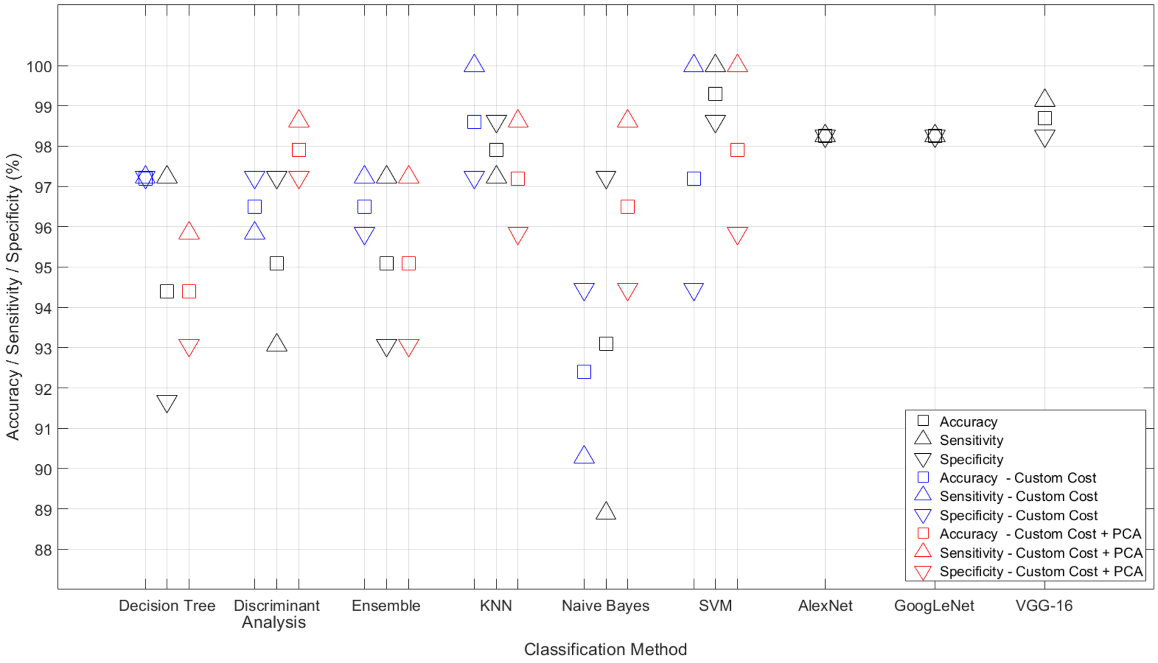

Fig. 5. Comparison of the Results of all Techniques with respect to Accuracy, Sensitivity, and Specificity (or Pretest) Performance Metrics.

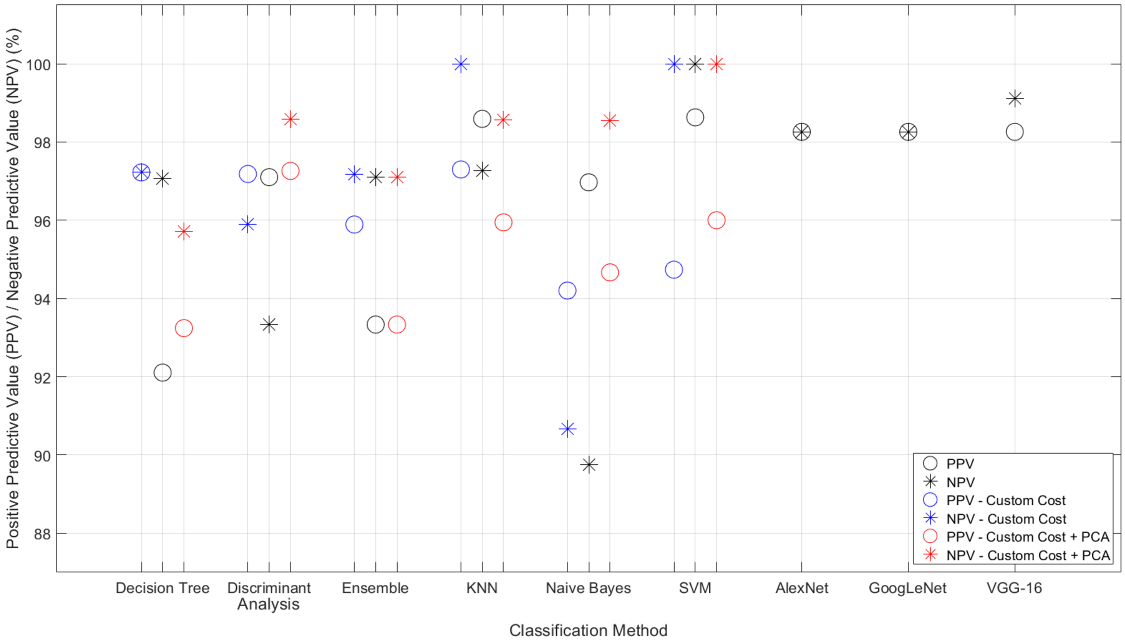

Fig. 6. Comparison of the Results of all Techniques with respect to Positive Predictive Value and Negative Predictive Value (or Post-Test) Performance Metrics.

VI. Discussion

The results suggest that the performance of statistical classification and deep learning methods are both generally acceptable with performance metrics mostly well above 90%. The best result was obtained with a statistical classification technique, which is clearly a simpler, faster alternative to the deep learning methods. While all statistical learning techniques took less than a minute to train, the computational requirements of training deep learning networks were much more demanding with training times up to 279 minutes even though transfer learning was used to keep the convolutional weights the same. Even though deep learning has been heavily emphasized in most recent research studies in the field of computer-aided diagnosis, the results of this work provide an objective comparison that suggests that simpler traditional approaches may yield comparable if not better results.

The results of all techniques showed sensitivity and specificity values that are generally close as they should be in clinical practice. Given that the data used were balanced with equal numbers of normal and abnormal cases, the accuracy values are the average of those of the sensitivity and specificity where the accuracy is exactly in the middle. It should be noted that the results for several classifiers showed higher values for specificity than sensitivity such as all variants of discriminant analysis, one variant of KNN, and two variants of Naïve Bayes classifiers. Even though some of these classifiers provide good overall accuracy, they are not suitable for computer-aided detection especially in screening studies.

The results of statistical classifier variants indicate that effect of using a custom cost function varied across different variants, but the sensitivity was improved or remained the same in all variants after applying such customization. This was particularly evident in the KNN classifier where the application of such customization provided significant improvement becoming the second-best statistical classifier and showing better accuracy than two deep learning networks with a sensitivity that is better than all of them. Therefore, it is suggested that this customization is considered in experimental evaluation of different statistical classification techniques. On the other hand, the results of using PCA for dimensionality reduction along with custom cost function showed a desirable effect of making the sensitivity values go above those of the specificity for the same classifier. This is evident in discriminant analysis and Naïve Bayes classifiers in particular. This allows the use of such classifier variants in computer-aided detection rather than discarding them as suggested above. Therefore, it is suggested that such dimensionality reduction is considered in such cases when sensitivity is lower than specificity. It should be noted that the explicability of the results of the system given that the features used and their weights are explicitly defined in the eigenvectors (principal components) of PCA outcomes.

As a part of the ongoing efforts to develop regulatory standards to govern the artificial intelligence solutions, an emphasis on explicability, or the ability to explain the outcomes, is a requirement that clinical systems must be able to meet. This is clearly an advantage for such traditional methods where simpler equations can be used to do that. On the other hand, this is largely not possible with deep learning methods due to the complexity of the networks structure and its huge number of parameters that make it difficult to understand the decision-making process inside the network and also render such networks prone to such issues as data gaps and overfitting. This is particularly evident in medical systems because of the much less data sizes available and also the data imbalance where several abnormal classes can be significantly underrepresented in the training and testing processes.

The yielded results in this research using the AlexNet have surpassed the recent results achieved in the literature. The overall accuracy of the AlexNet, GoogLeNet on the MIAS Dataset achieved in [42], and [9] was 95.70%, 91.58% respectively. And as depicted in Table I, the conventional machine learning approached has yielded an extraordinary result that is greater than the results achieved the pretrained CNN networks. It also surpasses the current results achieved in the literature. The retrieved accuracy in the work done in [43] using SVM, and KNN on the MIAS database was 87.69%, and 88.54% respectively. The higher performance of the state-of-art classification methods retrieved here in this work has been achieved by the employed image augmentation paradigm. The data augmentation has helped in increasing the size of the training and testing data. The augmented images have been shuffled before being submitted to the classification models. In addition, the data has been split into three totally independent subsets for training, validation, and testing subset to minimize the problem of overfitting. For statistical machine learning methods, the use of K-fold cross-validation explained in the Methods section also addresses the overfitting problem. The variations across different experiments were found to be less than 1% indicating that the proposed framework does not suffer from overfitting.

VII. Conclusion

In this study, direct comparison between the performances of statistical machine learning to deep learning in the context of computer-aided detection of breast cancer from mammographic images was performed. The results are compared using diagnostic performance metrics and suggest that simpler statistical machine learning techniques may provide better performance with simpler architectures that require much fewer demanding computations while allowing explanation of results. In particular, a support vector machine classifier variant provided a better performance overall, while other statistical machine learning techniques such as KNN classifier variants provided comparable results to those of three widely used deep learning networks. The obtained accuracies above 98% using both classical and deep learning models surpassed reported results in the literature. Furthermore, the present study suggests that statistical machine learning based methods might be closer to meeting regulatory approval requirements for clinical use. This also emphasizes the importance of addressing such issues as data gaps and explicability of outcomes in deep learning techniques to boost their transition to clinical use.

ACKNOWLEDGMENT

This research was funded by the Deanship of Scientific Research at Princess Nourah bint Abdulrahman University through the Research Funding Program (Grant No# FRP-1440-30).

REFERENCES

[1] U. Bick and F. Diekmann, “Mammographic Signs of Malignancy,” in Digital Mammography, Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 175–185.

[2] M. P. Sampat, M. K. Markey, and A. C. Bovik, “Computer-Aided Detection and Diagnosis in Mammography,” in Handbook of Image and Video Processing, Elsevier Inc., 2005, pp. 1195–1217.

[3] Y. M. Kadah, A. A. Farag, J. M. Zurada, A. M. Badawi, and A. B. M. Youssef, “Classification algorithms for quantitative tissue characterization of diffuse liver disease from ultrasound images,” IEEE Trans. on Med. Imaging., vol. 15, no. 4, pp. 466–478, 1996.

[4] M. A. Al-Antari, M. A. Al-Masni, S. U. Park, J. Park, M. K. Metwally, Y. M. Kadah, S. M. Han, and T. S. Kim, “An Automatic Computer-Aided Diagnosis System for Breast Cancer in Digital Mammograms via Deep Belief Network,” J. Med. Biol. Eng., vol. 38, no. 3, pp. 443–456, Jun. 2018.

[5] M. A. Al-antari, S. M. Han, and T. S. Kim, “Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions in digital X-ray mammograms,” Comput. Methods Programs Biomed., vol. 196, p. 105584, Nov. 2020.

[6] W. E. Fathy and A. S. Ghoneim, “A deep learning approach for breast cancer mass detection,” Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 1, pp. 175–182, 2019.

[7] D. J. Hand, “Evaluating Statistical and Machine Learning Supervised Classification Methods,” in Statistical Data Science, pp. 37–53, Jul. 2018.

[8] F. Mohanty, S. Rup, B. Dash, B. Majhi, and M. N. S. Swamy, “An improved scheme for digital mammogram classification using weighted chaotic salp swarm algorithm-based kernel extreme learning machine,” Appl. Soft Comput., vol. 91, p. 106266, Jun. 2020.

[9] S. A. Hassan, M. S. Sayed, M. I. Abdalla, and M. A. Rashwan, “Breast cancer masses classification using deep convolutional neural networks and transfer learning,” Multimed. Tools. and Appl., vol. 79, no. 41, pp. 30735–30768, Aug. 2020.

[10] G. Meenalochini and S. Ramkumar, “Survey of machine learning algorithms for breast cancer detection using mammogram images,” Proc. Mater. Today, vol. 37, no. Part 2, pp. 2738–2743, Jan. 2021.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” Commun. ACM., vol. 60, no. 6, pp. 84–90, Jun. 2017.

[12] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” [Online]. Available: http:// arXiv:1409.1556.

[13] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning,” [Online]. Available: http:// arXiv:1602.07261.

[14] D. Abdelhafiz, C. Yang, R. Ammar, and S. Nabavi, “Deep convolutional neural networks for mammography: advances, challenges and applications,” BMC bioinform., vol. 20, no. Suppl 11, Jun. 2019.

[15] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, “Deep Learning to Improve Breast Cancer Detection on Screening Mammography,” Sci. Rep., vol. 9, no. 1, pp. 1–12, 2019.

[16] Al-Antari M.A., M. A. Al-Masni, and T. S. Kim, “Deep Learning Computer-Aided Diagnosis for Breast Lesion in Digital Mammogram,” Adv Exp Med Biol., vol 1213. Springer, 2020.

[17] M. A. Al-Masni M. A. Al-Antari, J. M. Park, G. Gi, T. Y. Kim, P. Rivera, E. Valarezo, M.T. Choi, S.M. Han, and T. S. Kim, “Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system,” Comput. Methods. Programs. Biomed., vol. 157, pp. 85–94, Apr. 2018.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), vol. 2016-December, pp. 779–788, 2016.

[19] N. Ismail and C. Sovuthy, “Breast Cancer Detection Based on Deep Learning Technique,” Proc. 2019 Int. UNIMAS STEM 12th Eng. Conf. (EnCon), pp. 89–92, 2019.

[20] E. L. Omonigho, M. David, A. Adejo, and S. Aliyu, “Breast Cancer:Tumor Detection in Mammogram Images Using Modified AlexNet Deep Convolution Neural Network,” Proc. 2020 Int. Conf. Math. Comput. Eng. Comput. Sci. (ICMCECS), pp. 1–6, 2020.

[21] A. Marchesi et al., “The Effect of Mammogram Preprocessing on Microcalcification Detection with Convolutional Neural Networks,” Proc. 2017 IEEE 30th Int. Symp. Comput.-Based Med. Sys. (CBMS), vol. 2017-June, pp. 207–212, Jun. 2017.

[22] S. A. Agnes, J. Anitha, S. I. A. Pandian, and J. D. Peter, “Classification of Mammogram Images Using Multiscale all Convolutional Neural Network (MA-CNN),” J. Med. Sys., vol. 44, no. 1, pp. 1–9, Dec. 2019.

[23] R. S. Patil and N. Biradar, “Automated mammogram breast cancer detection using the optimized combination of convolutional and recurrent neural network,” Evol. Intell., pp. 1–16, Apr. 2020.

[24] T. Kooi, G. Litjens, B. Van Ginneken, A. Gubern-Mérida, C. I. Sánchez, R. Mann, A. den Heeten, and N. Karssemeijer, “Large scale deep learning for computer aided detection of mammographic lesions,” Med. Image Anal., vol. 35, pp. 303–312, Jan. 2017.

[25] J. Suckling, J. Parker, D. Dance, S. Astley, I. Hutt, and et al., « Mammographic Image Analysis Society (MIAS) database v1.21”, [Online]. Available: https://www.repository.cam.ac.uk/handle/ 1810/250394.

[26] J. Suckling, S. Astley, D. Betal, N. Cerneaz, D. Dance, et al, “The Mammographic Image Analysis Society Digital Mammogram Database, Int. Congr. Ser. - Excerpta Med., Vol. 1069, pp375-378. [Online]. Available: http://peipa.essex.ac.uk/info/mias.html

[27] Z. Abduh, M. A. Wahed, and Y. M. Kadah, “Robust computer-Aided detection of pulmonary nodules from chest computed tomography,” J. Med. Imaging. Health. Inform., vol. 6, no. 3, pp. 693–699, Jun. 2016.

[28] R. M. Haralick, I. Dinstein, and and K. Shanmugam, “Textural Features for Image Classification,” IEEE Trans. Syst. Man Cybern.: Syst., vol. SMC-3, no. 6, pp. 610–621, 1973.

[29] J. S. Weszka, C. R. Dyer, and A. Rosenfeld, “A Comparative Study of Texture Measures for Terrain Classification,” IEEE Trans. Syst. Man Cybern.: Syst., vol. SMC-6, no. 4, pp. 269–285, 1976.

[30] M. Amadasun and R. King, “Texural Features Corresponding to Texural Properties,” IEEE Trans. Syst. Man Cybern.: Syst., vol. 19, no. 5, pp. 1264–1274, 1989.

[31] C. M. Wu and Y. C. Chen, “Statistical feature matrix for texture analysis,” CVGIP-GRAPH. MODEL. IM., vol. 54, no. 5, pp. 407–419, Sep. 1992.

[32] K. I. Laws, “Rapid Texture Identification,” Proc. SPIE 0238, Image Proc. Missile Guid., Dec. 1980, vol. 0238, pp. 376–381.

[33] R. M. Haralick, “Image texture survey,” in Handbook of Statistics, vol. 2. Elsevier, pp. 399–415, 1982.

[34] P. Shanmugavadivu and V. Sivakumar, “Fractal dimension based texture analysis of digital images,” Procedia Eng., vol. 38, pp. 2981–2986, 2012.

[35] M. I. Owis, A. B. M. Youssef, and Y. M. Kadah, “Characterisation of electrocardiogram signals based on blind source separation,” Med. Biol. Eng. Comput., vol. 40, no. 5, pp. 557–564, 2002.

[36] M. Sugiyama, Introduction to Statistical Machine Learning, Morgan Kaufmann, Elsevier, 2016.

[37] O. Russakovsky, J. Deng, H. Su, J. Krause S. Satheesh S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, and A.C. Berg, “ImageNet Large Scale Visual Recognition Challenge,” Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, Dec. 2015.

[38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Oct. 2015, pp. 1–9.

[39] C. Shorten and T. M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning,” J. Big Data, vol. 6, no. 1, pp. 1-48, Dec. 2019.

[40] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, “A survey on deep transfer learning,” Lect. Notes Comput. Sci., Oct. 2018, vol. 11141 LNCS, pp. 270–279.

[41] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, “ImageNet: A large-scale hierarchical image database,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 248–255, 2010.

[42] E. L. Omonigho, M. David, A. Adejo, and S. Aliyu, “Breast Cancer:Tumor Detection in Mammogram Images Using Modified AlexNet Deep Convolution Neural Network,” Proc. Int. Conf. Math. Comput. Eng. Comput. Sci. (ICMCECS 2020), pp. 1-6, 2020.

[43] V. Devakumari, “A Hybrid Algorithm with Modified SVM and KNN for Classification of Mammogram Images using Medical Image Processing with Data Mining Techniques,” Eur. J. Mol. Clin. Med., vol. 7, no. 10, pp. 2956–2964, 2021.