Submitted: 31 October, 2023 Accepted: 23 November, 2023 Published: 15 August, 2024 DOI:10.22514/ejgo.2024.064

ORIGINAL RESEARCH

European Journal of Gynaecological Oncology

# Automated breast cancer mass diagnosis: leveraging artificial intelligence for detection and classification

Hiam Alquran[1,†], Mohammed Alsalatie[2,†], Wan Azani Mustafa[3,*], Abdullatif Hammad[1], Mohammad Tabbakha[1], Hassan Almasri[1], Reham Kaifi[4,5]

[1] Department of Biomedical Systems and Informatics Engineering, Yarmouk University, 21163 Irbid, Jordan
[2] The Institute of Biomedical Technology, King Hussein Medical Center, Royal Jordanian Medical Service, 11855 Amman, Jordan
[3] Faculty of Electrical & Engineering Technology, Campus Pauh Putra, Universiti Malaysia Perlis, 02000 Arau, Perlis, Malaysia
[4] College of Applied Medical Sciences, King Saud Bin Abdulaziz University for Health Sciences, 21423 Jeddah, Saudi Arabia
[5] King Abdullah International Medical Research Center, 22384 Jeddah, Saudi Arabia

***Correspondence**
wanazani@unimap.edu.my
(Wan Azani Mustafa)

† These authors contributed equally.

## Abstract

Breast cancer, a prevalent global concern affecting women, underscores the importance of early detection for improved treatment outcomes and reduced mortality rates. Mammogram image is widely employed as a tool for early detection of breast tumors. Incorrect diagnoses elevate the risk of cancer metastasis to vital organs like the lungs, stomach and lymph nodes. This study presents a software application categorizing mammogram images as benign or malignant. It relies on intrinsic features and employs twelve pre-trained deep-learning models. Additionally, ten feature selection algorithms are utilized to identify crucial attributes. Exploiting various feature selection techniques, pinpoint the most representative ones. The selected features from each algorithm contribute to building and testing the Gaussian Support Vector Machine (SVM) classifier. ReliefF selects the optimal features, reflecting the highest test accuracy in the SVM classifier. The recorded results demonstrate an accuracy, sensitivity, precision and specificity of 99.9%. These findings underscore the potential of combining diverse deep-learning structures with feature-reduction techniques to enhance diagnostic capabilities. The research highlights the technology's potential adoption in the healthcare sector, particularly considering the substantial volume of images involved and the heightened reliability it introduces to the mammogram image diagnosis process.

## Keywords

Deep learning; Breast cancer; Warper methods; PCA; ICA; Feature selection

## 1. Introduction

Breast cancer (BC) is one of the most common diseases around the world, especially for women, as around 50% of breast cancers develop in women who have no identifiable breast cancer risk factor other than gender (female) and age (over 40 years). In 2020, there were 2.3 million women diagnosed with breast cancer and 685,000 deaths globally. As of the end of 2020, there were 7.8 million women alive who were diagnosed with breast cancer in the past five years, making it the world's most prevalent cancer [1]. We could limit the severity of BC if we detected it early, and one of the methods that could be used for that purpose is the Mammogram, which is an imaging device that uses the X-ray principle but for the breast only at lower doses compared to the ordinary X-ray device [2]. It is worth mentioning that mammography is the best test for doctors to detect BC early, up to three years before it can be felt [3]. There are two types of tumors: benign and malignant. Benign cells are non-cancerous cells and are formed by abnormal cell growth. In many cases, they do not need treatment; they also have a smooth and regular shape and usually, when removed, do not grow back. Conversely, malignant tumors form when cancer cells multiply and develop into a mass, where they invade nearby tissues. They have an uneven shape and may also separate from tumors and spread throughout the body in a process called metastasis.

Developing an automated system for diagnosing breast tumors is an optimistic objective for researchers, with the goal of assisting doctors in reducing misdiagnoses, whether they are false positives or false negatives. The current research focuses on the field of computer-aided diagnosis (CAD) system development that revolves around identifying the most significant features for distinguishing between different types of breast masses. Consequently, this paper centers its attention on the most pertinent features for breast mass analysis, leveraging artificial intelligence techniques. One of the most extensive datasets utilized in various algorithms to construct a sensitive model for breast mass diagnosis is the Curated Breast Imaging Subset Digital Database for Screening Mammography (CBIS-DDSM) dataset [4–6].

Multiple experiments were carried out on the aforementioned dataset. Numerous research investigations have utilized the DDSM dataset, with examples including work by Hassan, S.A. and team [7] in 2020. They used the DDSM and INbreast datasets to train deep convolutional neural networks for classification, specifically AlexNet and GoogleNet. The models demonstrated impressive accuracy on both Convolution Neural Network (CNN) networks, with the AlexNet model achieving

100% accuracy on CBIS-DDSM and INbreast databases. The aim was to compare the performance of the two deep learning models. They just focused on using the AlexNet model for classification.

In 2021, Lou, M. and their research group [8] introduced a novel multi-level global-guided branch-attention network (MGBN) for mass classification, utilizing DDSM and INbreast datasets, yielding moderate results with an Area under the curve (AUC) of approximately 0.8375. Furthermore, A. Baccouche and team [9] employed the “You Only Look Once” model (YOLO) alongside three datasets for detection and classification. While their results were remarkable in detection, the classification did not exceed 74.4%. Niu J *et al.* [10] also utilized CNN with the Convolutional Block Attention Module (CBAM) to enhance feature extraction, resulting in improved performance in DDSM, not exceeding 96%. Meanwhile, Khaoula Belhaj Soulami and team [11] implemented an end-to-end UNet model for automatic detection, segmentation and classification tasks, achieving an F1-score of 0.99 with a focus on tumor detection and segmentation.

In 2022, S. R. Sannasi Chakravarthy and Harikumar Rajaguru [12] introduced an algorithm called the enhanced Crow-Search Optimized Extreme Learning Machine (ICS-ELM) for mass detection and classification. They utilized DDSM, Mammographic Image Analysis Society (MIAS) and INbreast datasets, achieving maximum overall classification accuracies of 97.193%, 98.137% and 98.266% for DDSM, MIAS and INbreast datasets, respectively.

Moreover, Kumar, I. and associates [13] explored dense tissue pattern characterization using conventional networks. They experimented with different activation functions and achieved the best accuracy on the DDSM dataset using ResNet-18, reaching 92.3%. The method focused on the impact of activation functions on the classification results of deep learning models.

Due to the influence of feature extraction on the classification accuracy of breast masses, Caiqing Liao and team [14] proposed the Feature Selection and Enhancement Network (FSE-Net) for classifying mammogram images. They introduced a novel feature selection and enhancement network FS and employed DDSM and INbreast datasets, achieving accuracy that did not exceed 80.6% and 95.6%, respectively.

All the preceding experiments have primarily concentrated on tasks like classification, segmentation or detection using deep learning with a single model. Since each deep learning model possesses its unique structure and behavior in feature extraction and image classification, previous studies have often revolved around specific sets of relevant features, potentially limiting the model’s generalizability to new test cases.

In contrast, this paper takes a different approach by employing a comprehensive range of techniques to extract pertinent features from various models. This approach provides additional insights into the nature of breast masses, and harnesses feature reduction methods to obtain the most relevant deep features for classification using machine learning classifiers. The novelty of this paper lies in the fusion of various feature engineering techniques, encompassing both feature extraction and reduction, in the classification of breast masses. The study then proceeds to compare the effectiveness of different scenarios. The corresponding sections illustrate the methods, results and discussion and end with the conclusion.

## 2. Materials and methods

The approach outlined in this paper is depicted in Fig. 1. It commences with the extraction of deep features using widely recognized convolutional neural networks. The most critical features are selected using 10 distinct feature selection techniques using a Gaussian support vector machine classifier to differentiate between benign and malignant breast masses.

### 2.1 Database

The study utilized the Digital Database for Screening Mammography (DDSM) [4], which can be accessed through the following link: `https://www.kaggle.com/datasets/tommyngx/breastcancermasses`. This dataset comprises both benign and malignant tumors found in mammograms. Initially, 2188 mass images were collected and incorporated into the dataset. Subsequently, the images underwent preprocessing, which included contrast-limited adaptive histogram equalization and data augmentation. Following these processes, the dataset expanded to contain 13,128 images, consisting of 5970 benign and 7158 malignant mammogram images. Each image was standardized to a size of 227 × 227 pixels. Fig. 2 shows sample of benign and malignant masses [4].

### 2.2 Deep learning models

This study employed 12 CNN structures for training and testing breast mammogram images. The transfer learning algorithm was utilized to extract features for training and testing from each network’s last fully connected layer. Two features were extracted for each of the two classes from every CNN model. The structure of each network is individually explained in the corresponding section. The dataset was divided into 70% for training and 30% for testing for each network, with a validation subset extracted from the training data, comprising 10% of the entire training set. Brief descriptions of each utilized network are provided in the corresponding sections.

#### 2.2.1 Efficient Net

In 2019, Mingxing Tan and Quoc V. Le introduced EfficientNet-B0. The aim of this architecture is to achieve a balance between model accuracy and computational efficiency. Using a compound scaling method based on a compound scalar, EfficientNet-B0 achieves state-of-the-art performance while being computationally efficient; this network is a part of the B0 family (B0 to B7). Mainly, EfficientNet-B0 is made of 237 layers, which can be summarized as an input layer, depth-wise convolution layer, batch normalization, activation function block, global average pooling layer and rescaling parameters. This combination succeeds in achieving a top-1 accuracy of 77.1% and a top-5 accuracy of 93.3%. EfficientNet-B0 demonstrates remarkable accuracy comparable to larger and more resource-intensive models in training and inference. However, highly specialized tasks with high precision may still benefit from more

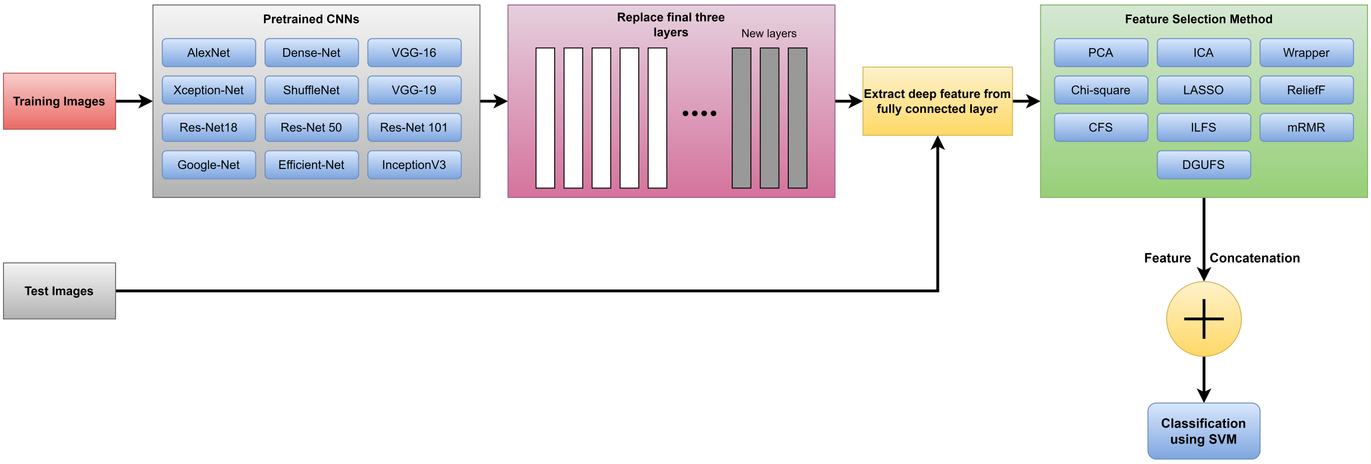

**FIGURE 1. The proposed approach in classification between benign and malignant breast masses.** CNN: Convolution Neural Network; VGG: Visual Geometry Group; ICA: Independent component analysis; DGUFS: Dependence Guided Unsupervised Feature Selection; lasso: Least Absolute Shrinkage and Selection Operator; ilfs: Infinite Latent Feature Selection; PCA: Principal Component Analysis; cfs: correlations based feature selection.

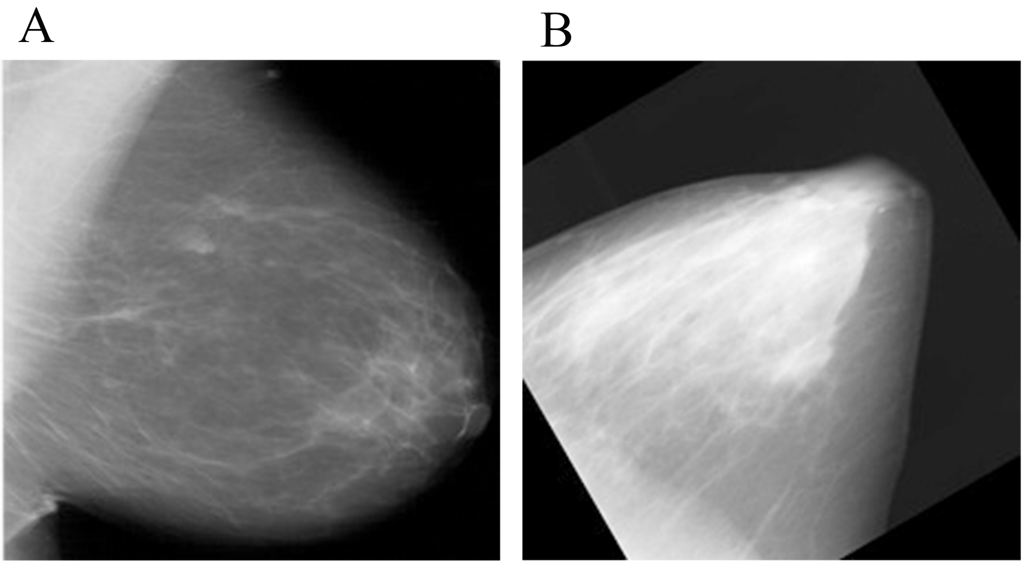

**FIGURE 2. Sample of images in database.** (A) benign, (B) malignant.

complex architectures, and, like any model, EfficientNet-B0's effectiveness can be affected by the specifics of the dataset [15].

### 2.2.2 Shuffle Net

Xiangyu Zhang *et al.* [16] introduced Shuffle Net in 2017. It has been designed to acquire high performance while significantly lowering computational complexity, making it ideal for resource-constrained environments. Essentially, Shuffle Networks by shuffling channels, allowing efficient group convolution operations. As part of Shuffle Net, channels are shuffled, which divides them into groups and alters their order within each group. By doing so, information can be exchanged between different groups while minimizing computational costs. "Shuffle Units" are the fundamental building blocks of the system, consisting of a pointwise convolution, a channel shuffling, and a depthwise convolution. It optimizes the flow of information and the computation of data; this network consists of a convolutional layer, max-pooling and a varying number of stages that contain varying numbers of shuffle units. ShuffleNet is known for its impressive accuracy given its low computational requirements, but if computational resources are not a constraint, more elaborate architectures may be able to achieve even higher accuracy. ShuffleNet was able to reach the top-1 error rate of 7.8% on the ImageNet classification task [16].

### 2.2.3 Dense Net201

Presented by Gao Huang, Zhuang Liu, Laurens van der Maaten and Kilian Q. Weinberger in 2017, The DenseNet201, consisting of 201 layers, obtains additional inputs from layers preceding them and passes their respective feature maps to subsequent layers. The method used is concatenation. Each layer receives "collective knowledge" from the previous layers. In general, DenseNet consists of the input layer, the initial convolutional layer with batch normalization and ReLU activation, dense blocks consisting of densely connected convolutional and concatenation layers, transition layers, global average pooling, and the fully connected layer. The benefits of this architecture are numerous. As a result of its remarkable accuracy, DenseNet201 rivaled more complex image classification models while remaining computationally efficient with a top-1 error rate of 22.58% and a top-5 error rate of 6.34%. As with any architecture, DenseNet201 has its limitations. The extensive connectivity of the system can cause high memory consumption during training. Additionally, some instances might cause redundant feature propagation due to the dense connections [17].

### 2.2.4 Xception Net

François Chollet introduced Xception in 2017. In principle, standing for “Extreme Inception”, Xception utilizes depthwise separable convolutions to enhance feature extraction and computational efficiency. The architecture is based on two main principles: Depthwise Separable Convolutions, which divide the convolution operation into two steps: depthwise convolution followed by pointwise convolutions, significantly reducing computational requirements. With a linear bottleneck structure and depthwise separable convolutions and residual connections, this design optimizes information flow while minimizing parameters and computations. In addition to a customizable depth that opens the door for variations like Xception-41, Xception-65 and more. Xception reached a top-1 accuracy of 79% and a top-5 accuracy of 94.5% on the ImageNet classification task [18].

### 2.2.5 Inception-v3

After the success of GoogLeNet, another evolutional version of it appeared called Inception-v3, brought by Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens and Zbigniew Wojna in 2016. With a depth of 48 layers, the primary advantage of Inception-v3 over its predecessor is its improved depth and complexity. Furthermore, Inception-v3 introduces auxiliary classifiers as regularizes to increase the stability and effectiveness of gradient propagation. As well as batch normalization, it also improves convergence during training and boosts performance overall. To sum it up, Inception V3’s major modifications include smaller convolutions, asymmetric convolution from spatial factorization, the use of auxiliary classifiers, and efficient reduction of grid size. Inception-v3 achieved a top-1 error rate of 17.2% and a top-5 error rate of 3.58% when compared to participants in (ImageNet Large Scale Visual Recognition Challenge) ILSVRC 2012. There are still some challenges associated with Inception-v3, despite its advantages. The deeper and more complex it is, the more computational resources it needs [19].

### 2.2.6 ResNets

In 2015, Kaiming He *et al.* [20] developed ResNet, a foundational architecture for deep learning. The ResNet consists of several residual blocks, mainly consisting of two 3 × 3 convolutional layers, each followed by batch normalization and a ReLU activation, each with multiple convolutional layers and a shortcut connection that skips some layers. As a result of these shortcut connections, the network can learn residual functions, which capture the differences between the desired output and the current prediction. The main advantage of ResNet is that it can train very deep networks without compromising performance. With residual connections, gradient flow becomes easier, allowing networks with hundreds of layers to be trained successfully. A variety of computer vision tasks can be solved using ResNet, due to its strong generalization capabilities. When applied to high-resolution images or devices with limited resources, ResNet may encounter memory consumption challenges. Moreover, residual connections might not benefit the architecture’s initial layers as much, potentially leading to redundant computations. ResNet-18 has 18 layers comprising a convolutional layer, max-pooling and four stacks; each stack contains multiple residual blocks. While ResNet-50 and ResNet-101 have 50 and 101 layers, respectively. The only difference is in-depth, design, shortcut connections and several blocks. Regarding performance, on ImageNet validation, ResNet-18, ResNet-50 and ResNet-101 achieved a top-1 error rate of 27.88%, 22.85% and 21.75%, respectively [20].

### 2.2.7 GoogLeNet

Consisting of 22 layers, GoogLeNet won the ILSVRC 2014 [9] and introduced the concept of inception by C. Szegedy *et al.* [21]. The main advantage of the inception lies in applying different convolutional filters with different sizes in parallel and stacking their output to generate the net output. In addition, GoogLeNet uses 1 × 1 convolution and global average pooling that allows it to produce much deeper architecture. These 22 layers mainly consist of convolution with ReLU activation, max pool layer, inception, average pool layer, dropout regularization, linear and softmax classifier. GoogLeNet stands in the field of classification with a top-5 error rate of 6.67% on ImageNet. It is important, however, to take into account some drawbacks. It is possible for the complexity of the architecture to result in an increase in memory consumption and computational requirements [21].

### 2.2.8 VGG nets

VGG16 and VGG19 were proposed by the Visual Geometry Group (VGG) at the University of Oxford in 2014. Simple yet efficient in design, VGG16 comprises 16 layers, 13 of which are convolutional layers, and three are fully connected layers. Alternatively, VGG19 consists of 19 layers, including 16 convolutional layers and three fully connected layers. The convolutional layers contain 3 × 3 filters and max-pooling layers of 2 × 2 pools distributed as alternating patterns, while the fully connected layers act as classifiers; all hidden layers in VGG architecture use ReLU, this unambiguous design allows a very high performance due to its ability to learn complicated features. However, this depth of architecture might introduce overfitting; another problem of this design is the long training time and large model size [22].

### 2.2.9 AlexNet

As the ILSVRC (ImageNet Large Scale Visual Recognition Competition) winner in 2012. AlexNet was marked as a great breakthrough in the realm of deep learning; this state-of-the-art was first introduced by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton in 2012 [23]. It consists of five convolutional layers and three fully connected layers. Using Rectified Linear Units (ReLUs) and Multiple Graphics Processing Unit (GPUs), AlexNet achieved a relatively short training time; the error rate was greatly reduced after using a local normalization scheme and overlapping pooling. Even though AlexNet has 60 million parameters, which may introduce overfitting in case of insufficient training examples, AlexNet overcomes this problem by data augmentation and using a dropout technique that involves setting each hidden neuron with probability 0.5 output to zero. In this way, "dropped out" neurons do not participate in the forward pass or the backpropagation. A top-1 error of 47.1% and a top-5 error of 28.2% were achieved by the best model in the 2010 version of the ImageNet competition. Compared to this, AlexNet had a 37.5% top-1 error rate and a 17.0% top-5 error rate [24].

## 2.3 Features reduction techniques

Recently, rapid development in the scientific field, especially the new measurement technologies and instruments, resulted in an exponential increase in data qualitatively and quantitatively, mainly increasing features will lead to the curse of dimensionality, in addition, more features will require more data points to represent this data that will consume a very large amount of time and computational resources, therefore the necessity of features reduction methods arise [25].

Dimensionality reduction can be executed through two distinct approaches: one involves retaining only the most pertinent variables from the initial dataset, referred to as feature selection, while the other involves capitalizing on input data redundancy to identify a smaller set of novel variables, each representing a combination of the input variables and essentially carrying the same information as the original variables. This latter technique is known as dimensionality reduction [26]. The subsequent section outlines the most widely recognized techniques for feature reduction, the corresponding section describes some of feature reduction and selection methods.

### 2.3.1 Principal component analysis

Principle component analysis (PCA): a statistical approach that can sum up the data content by converting a large set of variables into a small set while keeping most of the information of the large set [26].

We can express the mathematical process of computing PCA as follows: first, we remove the classes from the dataset, then we compute the mean of each attribute, and afterward, we compute the covariance matrix for the whole dataset using the formula below:

$$cov\left(X,Y\right)=\frac{1}{n}\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})$$

Where cov($X$,$Y$) represent the covariance between X and Y attributes.

$x$ and $y$ are the instances of X and Y respectively.

$\bar{x}$ and $\bar{y}$ the mean of X and Y respectively.

$n$ the number of instances.

The next step is to compute the eigenvectors and the corresponding eigenvalues from the covariance matrix by solving the following equation:

$$\det\left(A-\lambda I\right)=0$$

Where $A$ is the covariance matrix, $\lambda$ is the eigenvalue, $I$ is an identity matrix, and the eigenvalues of $A$ are roots of the characteristic equation.

The resulting eigenvectors are sorted by decreasing eigenvalues, and we choose a certain number of eigenvectors with the largest eigenvalues, these eigenvectors will form matrix $W$, which will be used to find our new subspace and transform the samples into it using the equation:

$$y=W^{'}\times x$$

Where $W'$ is the transpose of matrix $W$.

### 2.3.2 Independent component analysis (ICA)

Independent component analysis (ICA) is a statistical technique designed to explore sources, from sets of random variables and check if these sources are independent [27]. ICA has several applications in unsupervised learning and classification studies, where it has superiority in extracting independent features from the original feature dataset, which reduces the dimensionality of the feature space and improves classification accuracy effectively. In mathematical terms, the original features can be expressed as ($X_1$, $X_2$, … $X_n$). These features come from different sources such as ($S_1$, $S_2$, $S_n$); the combination is x equals cap A. A represents the mixing matrix. This equation can be written as y = Wx, where W represents the demixing matrix and y stands for the independent component. The ICA's components are derived using the fastICA algorithm. Moreover, the set of extracted components (y = $y_1$, $y_2$, ..., $y_n$) is characterized by their non-Gaussian and maximally independent nature. One common method for measuring this independence is using the kurtosis measure adopted in this paper to rank the extracted independent components [28].

### 2.3.3 Wrapper methods

Wrapper methods select features based on a machine learning algorithm that is fitted to a particular dataset. A greedy search approach is used to evaluate all possible combinations of features against the evaluation criteria. Evaluation criteria are simply performance measures that vary according to the problem type, such as accuracy, precision for classification, or R-squared for regression [29]. This paper employed the most commonly used techniques under the Wrapper methods, which are based backward.

### 2.3.4 Chi-square test

The chi-square test is applied to determine the independence between two variables; in other words, the chi-square hypothesis that variables A and B are independent. Then, test this hypothesis by calculating the sum of differences between the observed count and the expected count of the variables based on the following equation.

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{\left(o_{ij}-e_{ij}\right)^2}{e_{ij}}$$

Supposing A has c distinct values and B has r distinct values. The data tuples described by A and B can be shown as a contingency table, with the c values of A making up the columns and the r values of B making up the rows.

$$e_{ij} = \frac{count\left(A = a_i\right) \times count\left(B = \ b_i\right)}{n}$$

Where n is the number of data tuples, count(A=a$_i$) is the number of tuples having values ai for A, and so on for B. If the calculated chi-square value ($\chi^2$) is less than the significance level for (r – 1) × (c – 1) degree of freedom, then the chi-square hypothesis is true; otherwise, the hypothesis is rejected, which means the null hypothesis is rejected and the features are dependent [30].

### 2.3.5 LASSO method

The Least Absolute Shrinkage and Selection Operator (LASSO) is a strong method that mainly consists of regularization and feature selection. First, the LASSO method calculates the sum of the absolute values of the model parameters. The sum is then constrained, but the sum must be less than a fixed value called the upper bound. To make this possible, a process called shrinking or regularization is applied. As a result, some of the coefficients of the regression variables will be penalized and shrunk to zero; afterward, the remaining non-zero coefficient will be selected to be part of the model [31].

For more explanation, assuming the LASSO regression problem, the generalized cost function in LASSO regression can be written as follows:

$$E\left(\beta\right) + \ \lambda R(\beta)$$

Here, cap E of beta is the error rate, cap R of beta is the regularization term, and lambda is the parameter that determines the power of the pena the shrinking increases [32].

### 2.3.6 ReliefF

ReliefF is capable of handling multiple class problems, and it is also able to deal with noisy and incomplete data. ReliefF depends on determining the most significant attributes by finding out how efficiently the attribute values can separate close instances to the same or different class. We can sum the algorithm as the following steps: first, the algorithm selects a random instance called *Ri* then search in a k number of the nearest neighbors, k neighbors from the same class of the neighbors' *Hi* called nearest hits, then another k nearest neighbors from each of the different classes called nearest misses *Mi*. Then the algorithm calculate a parameter called quality estimation *W[A]* for all the attributes *A* based on their instance, nearest Hits, nearest misses [33].

### 2.3.7 Minimum redundancy maximum relevance algorithm

Minimum redundancy maximum relevance algorithm (MRMR) is aimed at searching for the essential set of features by minimizing the redundancy of the features set while at the same time maximizing the relevance of a certain feature to predict the right output (class in case of classification), assume feature Xi, the importance of Xi based on MRMR algorithm can be expressed as:

$$f^{MRMR}(X_i) \ = \ I(Y, Xi) \ - \frac{1}{|S|}\sum_{Xs\in S} I(Xs, \ Xi)$$

Y in this expression represents the class label, the chosen features are expressed as S, and |S| is the count of features, while I(Y, Xi) is a function that describes the common or shared information, this method will search for the highest f$^{MRMR}$ score for each feature and rank the features based on that score to select the features with the highest score [34, 35].

### 2.3.8 Infinite latent feature selection

The Infinite Latent Feature Selection (ILFS) method can be put in 3 stages. First is the preprocessing stage starts with breaking the dataset into vectors. Each vector represents the distribution of features, and the possible unique values of that vector are huge therefore, the process of mapping the set of values into smaller and manageable groups commonly known as tokens is necessary, the process of assigning tokens to the features is called discriminative quantization, the second step is to weight the graph based on relevance and other conditional probabilities, lastly, the thirst step begins after obtaining the matrix from the previous procedure, its geometric series is calculated in order to expand its path into infinity [36, 37].

### 2.3.9 Pairwise correlations feature selection

The process of choosing a subset of features from a dataset *via* examination of the pairwise correlations between the features is known as pairwise correlation-based feature selection. In order to enhance model performance and lower dimensionality, it seeks to discover the most illuminating and least redundant features.

First, the pairwise correlation is computed for each pair of features in the dataset. Pearson's correlation coefficient, which assesses the linear relationship between two variables, is the often utilized correlation coefficient. It has a range of −1 to 1, with 1 denoting a high positive correlation, −1 a strong negative correlation, and 0 denoting no association. Once we've located them, we can evaluate the non-redundant features' significance or relevance to the target variable [38].

### 2.3.10 Dependence guided unsupervised feature selection

Dependence Guided Unsupervised Feature Selection (DGUFS) is one of the feature reduction methods that is used to determine the most important features by first taking the measurement of pairwise dependencies using correlation coefficients, mutual information or other statistical methods, then, a rank will be assigned for the features based on the dependence test; afterward the most important features will be selected from the ranking [39].

## 2.4 Support vector classifier

SVM is a supervised machine learning technique that can be used for classification as well as regression. SVMs locate a hyperplane in the data that divides the various classes or values. The hyperplane is selected in such a way that the difference between the two classes is maximized.

SVMs are useful for machine learning tasks since they are highly resistant to noise and outliers and can classify data in high-dimensional domains.

One popular option for non-linear classification and regression applications is to use SVM with a Gaussian kernel. The Gaussian kernel implicitly maps the entered data to a high-dimensional feature space, enabling SVMs to simulate complex decision boundaries effectively.

The model seeks to identify the ideal hyperplane that maximizes the margin among distinct classes in the feature space while employing SVM with a Gaussian kernel. Next, new, unobserved data points are classified using the decision function. The decision function predicts the class label of an input point by assigning it a positive or negative value.

The sigma parameter in the Gaussian kernel function controls how quickly the similarity or distance measure decreases as the separation between the two values increases. A decision boundary that is smoother and has a bigger value of sigma has a narrower kernel, whereas a decision boundary that is more complicated has a larger value of sigma.

The Gaussian kernel function using sigma ($\sigma$) can be defined mathematically as follows:

$$K(x, y) = e^{-\frac{|x-y|^2}{2\sigma^2}}$$

Where $\sigma$ is the sigma parameter, $|x - y|^2$ is the squared Euclidean distance of x and y, and the input data points are x and y are the label [40, 41].

# 3. Results & discussion

For every convolutional neural network, the final fully connected layer is substituted with a new one by adjusting the number of output features to match the desired classes, a well-established technique in deep learning known as transfer learning. The remaining layers remain unaltered. Subsequently, each network undergoes training using the Adam optimizer, iteratively updating weights until convergence is achieved. The training employs mini-batch sizes 32, a maximum of 10 epochs, and an initial learning rate of 0.001. Early stopping is implemented during training to prevent overfitting. This identical procedure is applied across all the mentioned networks.

The training features are derived from each network's final fully connected layer, and model evaluation is conducted using an unseen test dataset. Test features are also extracted. Various feature reduction and selection algorithms are employed to select the most representative features. The resulting output is fed into a Gaussian support vector machine classifier (SVM). The subsequent results elucidate the performance of the SVM when combining deep features and reduction techniques.

The provided table outlines the top 10 relevant features obtained through the feature above reduction and ranking techniques.

Table 1 showcases ten feature reduction techniques, outlining the sequence of features for each method from the most representative to the least. The primary goal in singling out these top 10 attributes is to ensure consistent accuracy, even when the number of features is augmented. Fig. 3, on the other hand, delineates the utilization of deep features from each mentioned CNN across all reduction methods by emphasizing the most crucial features. The X-axis represents the CNN names, while the y-axis illustrates the frequency of incorporating the extracted network in all feature reduction methods.

Fig. 3 illustrates the efficacy of DenseNet in extracting the most pertinent features for breast masses, having been employed 11 times across all reduction methods. Following closely are ResNet18 and Inceptionv3, each utilized ten times, and subsequently, GoogleNet used nine times. Conversely, both Shuffle and AlexNet demonstrate a limitation in revealing the most representative features of breast tumors. This underscores the advantage of the proposed approach in attaining optimal representative features by leveraging various CNNs with diverse structures and approaches to extracting deep features from mammogram images.

The synergy between deep features and feature reduction techniques distinguishes pertinent and redundant features. This process diminishes the intricacy of the machine learning model and minimizes the risk of overfitting. The Gaussian Support Vector Machine (SVM) was chosen to evaluate the effectiveness of each method in creating a robust model for distinguishing between benign and malignant breast masses. The accompanying figure illustrates the test confusion matrix for each of these methods.

Fig. 4 demonstrates the superiority of our approach across all methods, with the best results achieved using the ReliefF method. The selected features from this method are detailed in Table 1, starting with the most significant feature, the AlexNet deep attribute, and concluding with the least significant, represented by the ResNet18 features. This method excluded deep features from VGG16, Xception and ShuffleNets. The highest accuracy achieved was an impressive 99.9%, with a sensitivity of 99.9%, specificity of 100%, precision of 99.9%, an F1-score of 1, and a negative predictive value of 99.9%.

The corresponding Table 2 and Fig. 4 summarize the results obtained for all reduction methods, including accuracy, sensitivity, specificity and precision.

**TABLE 1. This is a sample table caption.**

| Method | Ranked Features | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Wrapper | ResNet-50 | VGG16 | *Inception-v3* | Dense Net201 | Efficient Net | Xception Net | GoogLeNet | Xception Net | ResNet-101 | VGG19 |
| lasso | Dense Net201 | Shuffle Net | ResNet-18 | ResNet-101 | Dense Net201 | Shuffle Net | *Inception-v3* | ResNet-101 | GoogLeNet | *Inception-v3* |
| ReliefF | AlexNet | Efficient Net | ResNet-50 | VGG19 | *Inception-v3* | GoogLeNet | Dense Net201 | Dense Net201 | ResNet-101 | ResNet-18 |
| cfs | ResNet-50 | VGG16 | Xception Net | VGG19 | Efficient Net | GoogLeNet | ResNet-18 | ResNet-18 | *Inception-v3* | *Inception-v3* |
| ilfs | ResNet-18 | ResNet-50 | ResNet-50 | Dense Net201 | VGG16 | VGG16 | VGG19 | AlexNet | VGG19 | Dense Net201 |
| PCA | *Inception-v3* | ResNet-18 | GoogLeNet | Xception Net | ResNet-50 | VGG16 | *Inception-v3* | Dense Net201 | Dense Net201 | AlexNet |
| mRMR | ResNet-18 | Efficient Net | ResNet-18 | Xception Net | AlexNet | VGG16 | Efficient Net | VGG19 | Dense Net201 | GoogLeNet |
| chi-Square | Efficient Net | Efficient Net | GoogLeNet | GoogLeNet | *Inception-v3* | *Inception-v3* | ResNet-101 | ResNet-101 | ResNet-18 | ResNet-18 |
| ICA | GoogLeNet | ResNet-50 | ResNet-101 | GoogLeNet | Dense Net201 | VGG19 | AlexNet | VGG16 | ResNet-50 | Shuffle Net |
| DGUFS | Dense Net201 | Dense Net201 | Xception Net | VGG19 | ResNet-50 | ResNet-50 | ResNet-18 | Inception-v3 | ResNet-101 | ResNet-18 |

*VGG: Visual Geometry Group; ICA: Independent component analysis; DGUFS: Dependence Guided Unsupervised Feature Selection; lasso: Least Absolute Shrinkage and Selection Operator; ilfs: Infinite Latent Feature Selection; PCA: Principal Component Analysis; cfs: correlations based feature selection.*

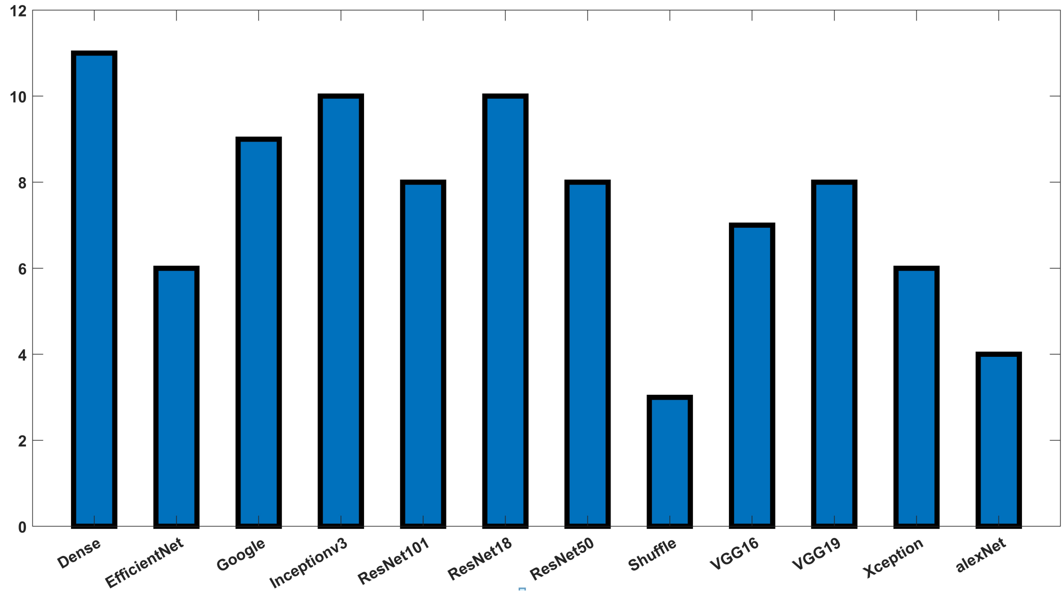

**FIGURE 3. The frequency of using deep features from each network in all reduction techniques.** VGG: Visual Geometry Group.

As indicated in Table 2, the mRMR method yielded the highest precision, followed by ReliefF. Conversely, ReliefF and Warper exhibited the most heightened sensitivity. Moreover, mRMR, along with ReliefF, achieved the best specificity. Ultimately, ReliefF attained the highest accuracy at 99.9%. Across all methods, ReliefF consistently demonstrated superior accuracy, top specificity, noteworthy sensitivity and precision.

Fig. 5 displays the outcomes achieved through the suggested methodology. The most favorable results are evident when utilizing features selected by the ReliefF method. The rationale behind this lies in its capability to prioritize optimistic features, choosing the top ten from diverse sources, including two features from the DenseNet and the remaining from 8 distinct networks. This amalgamation of varied features contributes to a more precise characterization of the tumor's nature, as demonstrated by the Gaussian support vector machine classifier.

For further analysis, the Receiver Operating Characteristic curve (ROC) is depicted in Fig. 6. The ROC curves illustrate the relationship between all feature selection methods' true and false positive rates.

All the preceding ROC curves clearly illustrate that the deep features consistently achieve a high Area Under the Curve (AUC) score of 1. This underscores the effectiveness of deep learning in extracting pertinent features for breast masses, particularly when combined with feature selection to identify the most representative ones. This suggests the potential of considering the model for deployment as software in healthcare units.

A comprehensive comparison with the most recent literature on breast mass classification was conducted to evaluate the proposed method's effectiveness further. Table 3 compares our work with existing literature, focusing on accuracy and AUC scores.

The table shows that our proposed approach achieves the highest accuracy among the compared methods. This indicates the potential for developing the model into a software application for implementation in healthcare sectors, which aims to reduce false positive cases. Such an advancement would significantly enhance the breast cancer diagnosis process and subsequent treatment. Many studies employing feature engineering techniques to improve the diagnosis approach [28, 40, 41].

## 4. Conclusions

Breast cancer stands as one of the most prevalent and potentially life-threatening conditions affecting women worldwide, emphasizing the critical importance of early detection and appropriate treatment. Mammogram images serve as a primary tool for the timely identification of breast tumors. An inaccurate diagnosis can significantly increase the risk of cancer metastasizing to other organs, such as the lungs. Extracting relevant features is vital in improving detection through deep learning techniques that extract attributes representing malignant and benign tumors.

Integrating deep learning with feature selection methods is pivotal in identifying essential features while reducing dimensionality to retain the most impactful ones. Coupled with a Support Vector Machine (SVM) classifier, this approach yields a robust model. This research introduces a software application that employs deep learning techniques to classify mammogram images as benign or malignant based on inherent features. Furthermore, feature selection algorithms are employed to isolate the most critical attributes. Moreover, utilizing the entire image for diagnosis purposes enhances accuracy by incorporating the surrounding region into the diagnostic process and capturing the most relevant features, considering the impact of tumors on the neighboring areas [42, 43].

Various feature selection approaches are exploited, with ReliefF emerging as the most suitable choice. The results

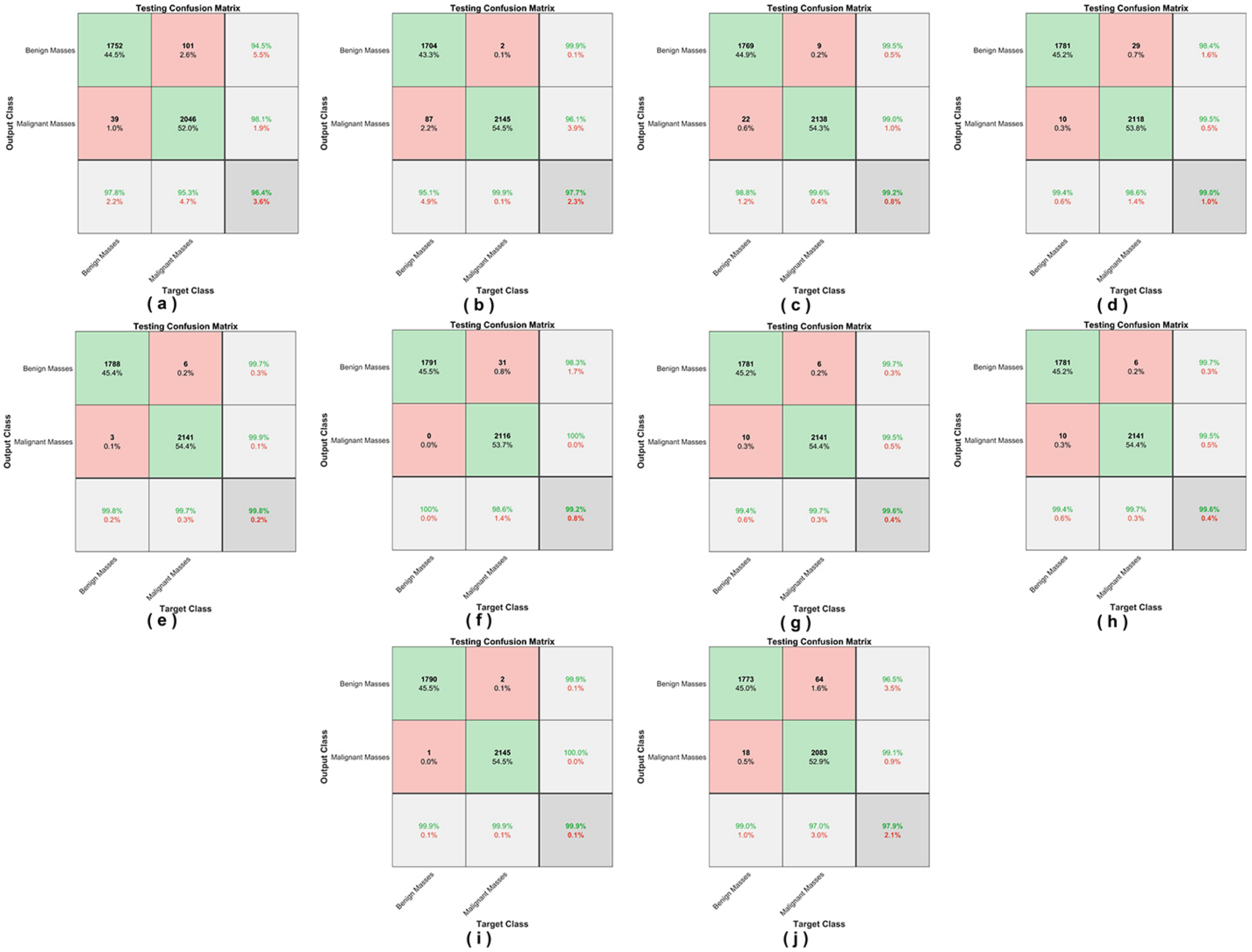

**FIGURE 4. The test confusion matrices of SVM using (a) cfs, (b) CHI, (c) DGUFS, (d) ICA, (e) ilfs, (f) lasso (g) mRMR, (h) PCA, (i) ReliefF, (j) warpper.**

**TABLE 2. The obtained results for all feature reduction methods.**

| Reduction_Method | Precision | sensitivity | specificity | Accuracy |
|---|---|---|---|---|
| CFS | 97.80% | 94.50% | 98.10% | 96.40% |
| CHI | 99.00% | 96.50% | 99.10% | 97.70% |
| DGUFS | 98.80% | 99.50% | 99.00% | 99.20% |
| ICA | 99.40% | 98.40% | 99.50% | 99.00% |
| ILFS | 99.80% | 99.70% | 99.90% | 99.80% |
| mRMR | 100.00% | 98.30% | 100.00% | 99.20% |
| Lasso | 99.40% | 99.70% | 99.50% | 99.60% |
| PCA | 99.40% | 99.70% | 99.50% | 99.60% |
| ReliefF | 99.90% | 99.90% | 100.00% | 99.90% |
| Warpper | 95.10% | 99.90% | 96.10% | 97.70% |

*DGUFS: Dependence Guided Unsupervised Feature Selection; ICA: Independent component analysis; ILFS: Infinite Latent Feature Selection; Lasso: Least Absolute Shrinkage and Selection Operator; PCA: Principle component analysis; CFS: Correlation-based Feature Selection; CHI: Chi-square Test.*

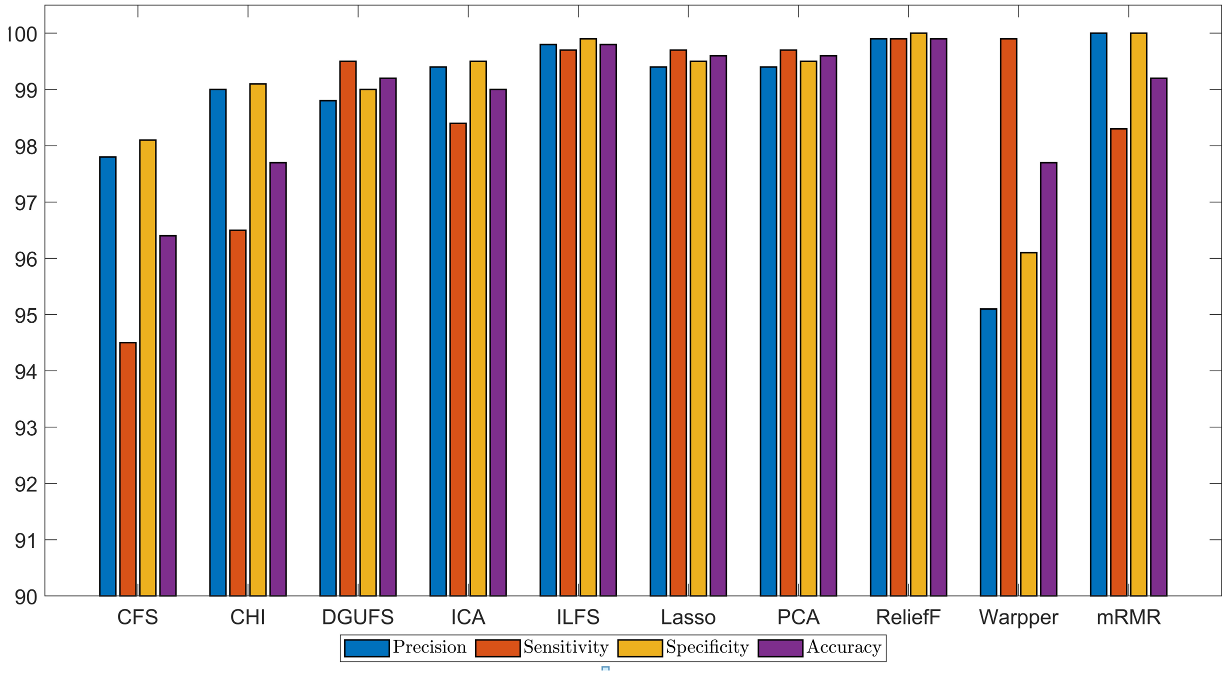

**FIGURE 5. The results using various reduction methods.** DGUFS: Dependence Guided Unsupervised Feature Selection; ICA: Independent component analysis; ILFS: Infinite Latent Feature Selection; Lasso: Least Absolute Shrinkage and Selection Operator; PCA: Principle component analysis; CFS: Correlation-based Feature Selection; CHI: Chi-square Test.

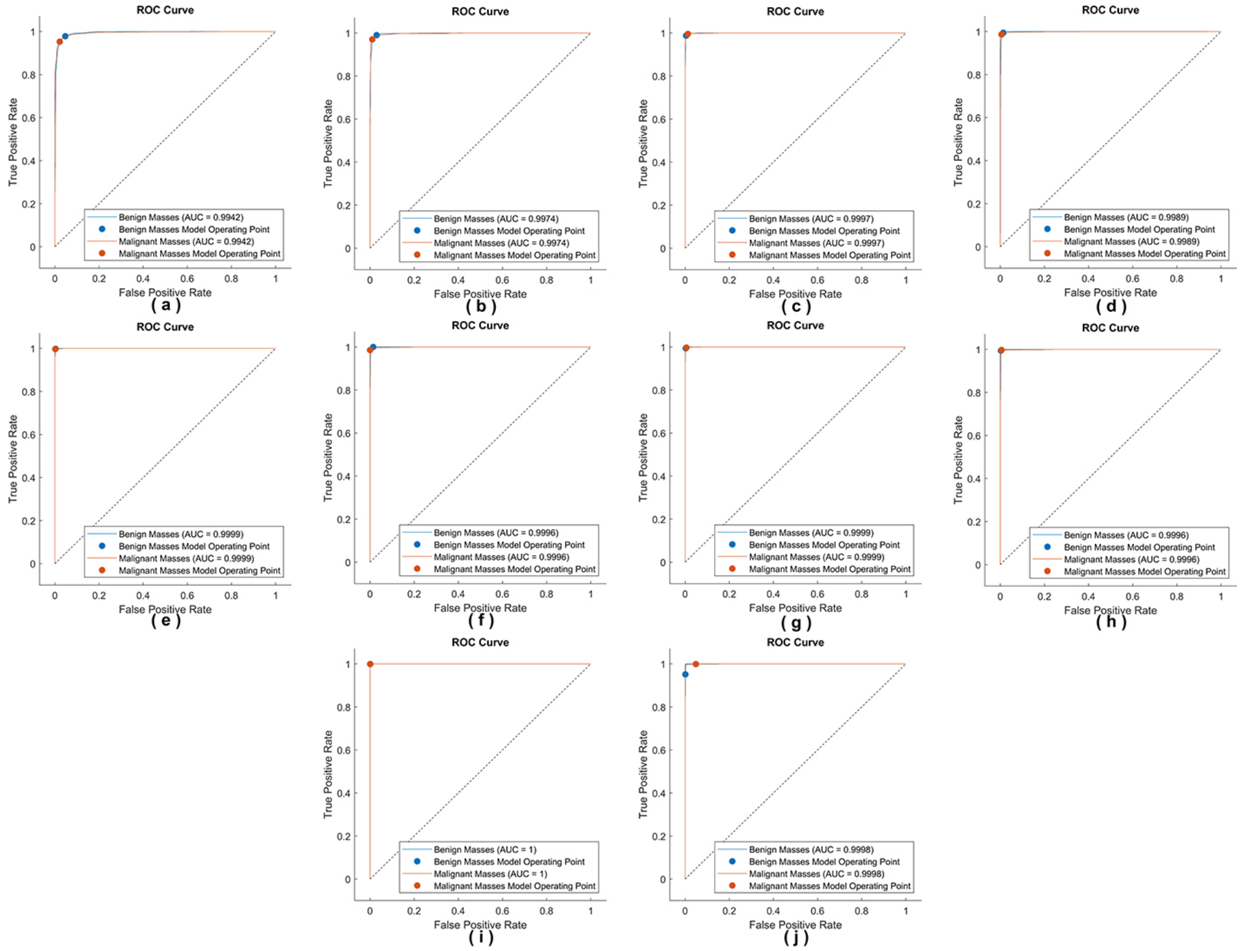

**FIGURE 6. The test ROC of SVM using (A) cfs, (B) CHI, (C) DGUFS, (D) ICA, (E) ilfs, (F) lasso, (G) mRMR, (H) PCA, (I) ReliefF, (J) warpper.** ROC: Receiver Operating Characteristic; AUC: Area under the curve.

**TABLE 3. Comparison with literature.**

| Ref# | Method | Results |
|---|---|---|
| [7] 2020 | Deep CNNs (Alex and Google Nets) | AlexNet (Accuracy 100%) |
| [8] 2021 | A novel multi-level global-guided branch-attention network (MGBN) for mass classification | AUC (0.8375) |
| [9] 2021 | YOLO based method | Detection accuyracy 95.7%, classification accuracy 74.4% |
| [10] 2021 | Deep learning for feature extraction | Accuracy 96.26% |
| [12] 2022 | Improved Extreme Learning Machine with Deep Learning | accuracy is 97.193% |
| [13] 2022 | Dense Tissue Pattern Characterization Using Deep Neural Network | accuracy 92.3% |
| [14] 2023 | FSE-Net: feature selection and enhancement network | accuracy for CBIS-DDSM 80.6% |
| Proposed Method | Using 12 deep learning structures, with 10 feature selection methods and Machine learning Classifier | Accuracy 99.9%, AUC 1 |

*AUC: Area under the curve.*

underscore the model's remarkable ability to effectively handle unseen data (test data), achieving an accuracy rate of nearly 100%. The findings of this research suggest the potential adoption of this technology within the healthcare sector, particularly considering the substantial number of images involved and the reliability it offers in the mammogram image diagnosis process.

## AVAILABILITY OF DATA AND MATERIALS

The dataset utilized in this manuscript can be accessed on the following website: https://www.kaggle.com/datasets/tommyngx/breastcancermasses.

## AUTHOR CONTRIBUTIONS

HiA, MA—designed the research study. HiA, MS, WAM, AH, HaA, MT and RK—performed the research; wrote the manuscript. HiA—analyzed the data. All authors contributed to editorial changes in the manuscript. All authors read and approved the final manuscript.

## ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

## ACKNOWLEDGMENT

Not applicable.

## FUNDING

This research received no external funding.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## REFERENCES

[1] World Health Organization. Cancer. 2023. Available at: https://www.who.int/health-topics/cancer#tab=tab_1 (Accessed: 24 October 2023).

[2] Vainio H, Bianchini F. Evaluation of cancer-preventive agents and strategies a new program at the international agency for research on cancer. Annals of the New York Academy of Sciences. 2001; 952: 177–180.

[3] Dong G, Liu H. Feature engineering for machine learning and data analytics. 4th edn. CRC Press: Boca Raton, FL. 2018.

[4] Wyer-Lee R, Gimenez F, Hoogi A, Rubin D. Curated breast imaging subset of digital database for screening mammography (CBIS-DDSM). 2016. Available at: https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=22516629 (Accessed: 05 December 2023).

[5] Heath M, Bowyer K, Kopans D, Kegelmeyer P, Moore R, Chang K, *et al*. Current status of the digital database for screening mammography. In Karssemeijer N, Thijssen M, Hendriks J, van Erning L (eds.) Computational Imaging and Vision (pp. 457–460). Springer: Dordrecht. 1998.

[6] Agarwal R, Diaz O, Lladó X, Yap MH, Martí R. Automatic mass detection in mammograms using deep convolutional neural networks. Journal of medical imaging. 2019; 6: 031409.

[7] Hassan SA, Sayed MS, Abdalla MI, Rashwan MA. Breast cancer masses classification using deep convolutional neural networks and transfer learning. Multimedia Tools and Applications. 2020; 79: 30735–30768.

[8] Lou M, Wang R, Qi Y, Zhao W, Xu C, Meng J, *et al*. MGBN: convolutional neural networks for automated benign and malignant breast masses classification. Multimedia Tools and Applications. 2021; 80: 26731–26750.

[9] Baccouche A, Garcia-Zapirain B, Castillo Olea C, S Elmaghraby A. Breast lesions detection and classification *via* YOLO-based fusion models. Computers Materials & Continua. 2021; 69: 1407–1425.

[10] Niu J, Li H, Zhang C, Li D. Multi-scale attention-based convolutional neural network for classification of breast masses in mammograms. Medical Physics. 2021; 48: 3878–3892.

[11] Soulami KB, Kaabouch N, Saidi MN, Tamtaoui A. Breast cancer: one-stage automated detection, segmentation, and classification of digital mammograms using UNet model based-semantic segmentation. Biomedical Signal Processing and Control. 2021; 66: 102481.

[12] Sannasi Chakravarthy SR, Rajaguru H. Automatic detection and classifi-

cation of mammograms using improved extreme learning machine with deep learning. IRBM. 2022; 43: 49–61.

[13] Kumar I, Kumar A, Kumar VDA, Kannan R, Vimal V, Singh KU, *et al*. Dense tissue pattern characterization using deep neural network. Cognitive Computation. 2022; 14: 1728–1751.

[14] Liao C, Wen X, Qi S, Liu Y, Cao R. FSE-Net: feature selection and enhancement network for mammogram classification. Physics in Medicine & Biology. 2023; 68: 195001.

[15] Tan M, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105–6114). 36th International Conference on Machine Learning. Long Beach, USA. 2019.

[16] Zhang X, Zhou X, Lin M, Sun J. ShuffleNet: an extremely efficient convolutional neural network for mobile devices. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018; 6848–6856.

[17] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017; 4700–4708.

[18] Chollet F. Xception: deep learning with depthwise separable convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017; 1251–1258.

[19] Szegedy C, Wei Liu, Yangqing Jia, Sermanet P, Reed S, Anguelov D, *et al*. Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015; 1–9.

[20] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016; 770–778.

[21] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016; 2818–2826.

[22] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014; 1–14.

[23] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, *et al*. Imagenet large scale visual recognition challenge. International Journal of Computer Vision. 2015; 115: 252–211.

[24] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems. 2012; 1097–1105.

[25] Destrero A, Mosci S, De Mol C, Verri A, Odone F. Feature selection for high-dimensional data. Computational Management Science. 2009; 6: 25–40.

[26] Sorzano C, Vargas J, Pascual Montano A. A survey of dimensionality reduction techniques. arXiv preprint arXiv:1403.2877. 2014; 1–35.

[27] Bell AJ, Sejnowski TJ. The "independent components" of natural scenes are edge filters. Vision Research. 1997; 37: 3327–3338.

[28] Tawalbeh S, Alquran H, Alsalatie M. Deep feature engineering in colposcopy image recognition: a comparative study. Bioengineering. 2023; 10: 105.

[29] Verma V. A comprehensive guide to feature selection using Wrapper methods in python. 2022. Available at: https://www.analyticsvidhya.com/blog/2020/10/a-comprehensive-guide-to-feature-selection-using-wrapper-methods-in-python/ (Accessed: 21 September 2023).

[30] Han J, Kamber M, Pei J. Data mining concepts and techniques third edition. The Morgan Kaufmann Series in Data Management Systems. 2011; 5: 83–124.

[31] Fonti V, Belitser E. Feature selection using LASSO. 2017. Available at: https://vu-business-analytics.github.io/internship-office/papers/paper-fonti.pdf (Accessed: 05 December 2023).

[32] Doosa G. The mathematical background of lasso and Ridge regression. 2021. Available at: https://medium.com/codex/mathematical-background-of-lasso-and-ridge-regression-23b74737c817 (Accessed: 18 October 2023).

[33] Robnik-Šikonja M, Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learning. 2003; 53: 23–69.

[34] Zhao Z, Anand R, Wang M. Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform. 2019 IEEE International Conference on Data Science and Advanced Analytics. 2019; 442–452.

[35] fscmrmr. Rank features for classification using minimum redundancy maximum relevance (MRMR) algorithm—MATLAB. 2023. Available at: https://www.mathworks.com/help/stats/fscmrmr.html (Accessed: 22 October 2023).

[36] Roffo G, Melzi S, Castellani U, Vinciarelli A. Infinite latent feature selection: a probabilistic latent graph-based ranking approach. 2017 IEEE International Conference on Computer Vision. 2017; 1407–1415.

[37] Miftahushudur T, Ali Wael CB, Praludi T. Infinite latent feature selection technique for hyperspectral image classification. Jurnal Elektronika dan Telekomunikasi. 2019; 19: 32–37.

[38] Li S, Oh S. Improving feature selection performance using pairwise pre-evaluation. BMC Bioinformatics. 2016; 17: 1–13.

[39] Guo J, Zhu W. Dependence guided unsupervised feature selection. Proceedings of the AAAI Conference on Artificial Intelligence. 2018; 32: 2232–2239.

[40] Ustuner M, Sanli FB, Dixon B. Application of support vector machines for land use classification using high-resolution rapideye images: a sensitivity analysis. European Journal of Remote Sensing. 2015; 48: 403–422.

[41] Alquran H, Al-Issa Y, Alslatie M, Abu-Qasmieh I, Alqudah A, Azani Mustafa W, *et al*. Liver tumor decision support system on human magnetic resonance images: a comparative study. Computer Systems Science and Engineering. 2023; 46: 1653–1671.

[42] Alsalatie M, Alquran H, Mustafa WA, Zyout A, Alqudah AM, Kaifi R, *et al*. A new weighted deep learning features using particle swarm and ant lion optimization for cervical cancer diagnosis on pap smear images. Diagnostics. 2023; 13: 2762.

[43] Khader A, Alquran H. Automated prediction of osteoarthritis level in human osteochondral tissue using histopathological images. Bioengineering. 2023; 10: 764.

**How to cite this article:** Hiam Alquran, Mohammed Alsalatie, Wan Azani Mustafa, Abdullatif Hammad, Mohammad Tabbakha, Hassan Almasri, *et al*. Automated breast cancer mass diagnosis: leveraging artificial intelligence for detection and classification. European Journal of Gynaecological Oncology. 2024; 45(4): 24-36. doi: 10.22514/ejgo.2024.064.