Radial Basis Function Neural Network with PCA and LDA for Breast Cancer Classification

Md Ismail Alam Mokul

Department of Electrical and Computer Engineering, University of Waterloo
200 University Ave W, Waterloo, ON, N2L 3G1, Canada
miamokul@uwaterloo.ca

***Abstract*— Breast Cancer is the most diagnosed disease among women. Early detection and diagnosis of breast cancer can significantly reduce the suffering and death count. In this work, the Radial Basis Function Neural Network is used as a classifier to predict the benign and malignant cases. RBFNN has a shallow architecture with a radial basis function that performs a nonlinear transformation, by mapping nonlinearity into a high-dimensional domain and making the data linearly separable. Wisconsin dataset on breast cancer diagnosis is used for the study. Principal Component Analysis and Linear Discriminant Analysis were used independently for dimension reduction of the dataset before feeding into two different classifiers with RFBNN. This work shows that without any data dimensionality reduction, RBFNN performs poor as a classifier (Accuracy 90%). Implementation of both PCA and LDA improves the performance of the classification model, but the LDA-RBFNN model (accuracy 97%) outperforms the PCA-RBFNN (accuracy 94%) for the classification task.**

***Keywords*— Classification, Radial Basis Function Neural Network, Principal Component Analysis, Linear Discriminant Analysis, Data dimension reduction**

I. Introduction

Cancer is ranked as one of the leading causes of death globally in recent decades [1]. According to a published report in 2018 by GLOBOCAN, there were an estimated 18.1 million new cases of cancers and 9.6 million deaths associated with cancer. Out of the 36 types of cancers, breast cancer is the most commonly diagnosed cancer among women, accounting for 11.6% of the global cancer death [1]. It is said that almost every woman on earth is susceptible to the risk of breast cancer regardless of family history and background [2]. One of the major impediments of cancer treatment is that it cannot be associated with any visible or physical symptoms the early stage [3]. Cancerous cell identification and discerning between benign and malignant tissues are complex tasks, and accurately predicting the risk of cancer is arduous even for the best professionals. Besides, the number of expert pathologists are also few, and the process is costly and slow [3]. Additionally, human error in the assessment and decision-making is a natural phenomenon, heavily dependent on the professional acumen and emotional state of an individual. Computer-aided systems, embarking on various artificial intelligence and machine learning techniques, have been demonstrated to be a faster and accurate means for cancer detection and prediction [4][5][6].

Cancer classification and prediction involve assessing and processing a large amount of complex high-dimensional diagnostic data. Data pre-processing and feature extraction are important steps for any classification problem. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are two widely known multivariate statistical techniques for data feature extraction and dimension reduction [7][8]. Feature extraction refers to picking the important features of the data that can be useful for decision-making and avoiding redundant information [9]. In PCA, feature extraction is carried out by transforming the high-dimensional data into a lower-dimensional space where the first few components reflect the majority information of the dataset. On the other hand, in LDA, data is projected on a lower-dimensional space to maximize the separability of classes within the dataset. Radial Basis Function Neural Network (RBFNN), a special kind of feedforward artificial neural network, is efficient in classification problems, for having a radial basis function for nonlinear transformation. Unlike multilayer perceptron or other multi-layered neural networks, the construction and training of RBFNN are comparatively easier.

In this work, the classification performance of RBFNN with PCA and LDA is studied for the diagnosis of breast cancer. In the subsequent sections, literature review and background on RBFFN, PCA and LDA are discussed along with a description of methodology and implementation of this study. Finally, the findings and results are discussed.

II. Background and Literature Review

Cancer classification with the aid of neural networks has been a growing interest among many researchers in recent years. Ahmed et al have used deep belief networks for breast cancer classification [10]. They initialized the weights with deep belief neural network and constructed a backpropagation network with Liebenberg Marquart learning. Their experiment resulted in 99% accuracy in classification. Kaymak and his colleagues have studied the Multi-layer Perceptron Neural Networks for the classification of breast cancer [4]. Ahmet et al explored the classification of breast cancer with ANN. After the prepossessing of the data, they reduced the features by independent component analysis. The showed 90% diagnostic accuracy for breast cancer [11]. Genetically Optimized Neural

Network was explored by Arpit et al for the classification of breast cancer where they showed classification accuracy 100% for the training-testing partition of 70-30 [2]. An interesting study on the estimation of breast cancer risk was studied by Ayer and his colleagues. They used a 3-layered MLP with 1000 hidden layer nodes and trained the network with the 62219 mammography images. They showed an accuracy of 96.5% [12]. Bardou et al used Convolution Neural Networks for feature extraction from the breast biopsy images [13]. Ahmed and Shaimaa have studied Probabilistic Neural Networks for the classification of breast cancer where they showed a higher accuracy of 98% [14]. A good number of studies have focused on DNA microarray and gene subset for the classification of cancers. Cho and Won have studied on Leukemia, colon and B-cell Lymphoma cancer by extracting significant gene subsets with the aid of ANN [6]. Xu and his colleague have used support vector machine based recursive feature elimination for cancerous gene selection and prediction [15]. They showed that previously known 70-gene signatures for cancer classification can be reduced to 50-gene signatures for effectively predicting the prognoses. Chen et al have studied the gene expressions for lung cancer cells, collected from 4 hospitals, and used ANN for patient survival classification with an accuracy of 83% [16]. Garro and his colleagues have studied the nature-inspired Swarm Intelligence, Artificial Bee Colony algorithm, for features selection from DNA microarray that best describes the disease [17]. Dheeba and her colleague have used Particle Swarm Optimized Wavelet Neural Network for the detection of breast abnormalities in digital mammograms and showed 96.8% classification accuracy [18]. Fadzil et al worked with Genetic Algorithm for feature selection in breast cancer datasets and used MLP for classification [19].

The reason for this interest among the researcher is the efficacy of ANN and ML techniques in cancer classification. The major advantages of neural networks are that they can handle a large amount of complex non-linear data, are adaptive in learning and show self-organizing properties. However, the structure and complexity of neural network topology are major limitations of neural networks. Therefore, a problem-specific approach to neural network architecture is essential. For effective classification, a neural network has to be optimized in three areas: architecture, training, and data. Number of input neurons, number of hidden layers and hidden neurons, and number of output layers and their topology needs to be optimized for achieving high accuracy. Selection of the right training algorithm, weight optimization technique and number of training epochs need to be optimized for better outcome of the neural network. Besides, selection of the data, pre-processing, and quantity and quality of data are major parameters for neural network models [20].

RBFNN has the advantage of having only one hidden layer with radial basis functions. This makes the network architecture easy. However, choosing the correct spread of the hidden layer nodes and the algorithm for two steps training of RBFNN are important steps for efficient classification.

Another important element for successful classification is the structure of the data. PCA [7] and LDA [8] [21] are two widely used multivariate statistical techniques for data processing. They primarily reduce the data dimension, extract the important features and maximize the class gap. They result in better data subset that is useful for the classification task of the neural network. In the subsequent sections, RBFNN and different feature extraction and dimension reduction techniques are discussed.

III. Artificial neural network and radial basis function neural network

The artificial neural network is inspired by the biological construct of neurons in the human brain. It is a mathematically connected net of a large number of nodes, called neurons, and can deal with arbitrary non-linear function and 'learn'[22]. Two main classes of neural networks are feedforward network where the data signal flows in the forward direction, and feedback network (recurrent networks) where output layer nodes become the feedback to the previous layer nodes. Due to the different flow of signals in the traditional neural networks, two different types of learnings take place. One is supervised learning where the learning is obtained through examples. Another is unsupervised learning where the output characteristic of the network is achieved by the network itself through feedback in the absence of any target output.[22]. Based on the classes and learning paradigm, there exist several ANN architectures that are used for various applications: Multi-layer Perceptron (MLP), Radial Basis Function Neural Network (RBFNN), Kohonen's Self-organizing Network (KSON), and Hopfield Network. After the reformulation of Backpropagation learning in the 1980s, there was a breakthrough in the multi-layer architectures of neural networks[22]. Backpropagation learning is based on the gradient descent technique for minimizing cumulative error. Subsequently, several other optimization techniques, mainly based on evolutionary computation and nature-based swarm intelligence, have been used for the optimization of the network [23][20][19][2][18][24][25][26].

For the classification problem, ANN should have a robust learning method that is quick and efficient [27]. This learning method can be applied to an appropriate network for the given problem. The robustness in avoiding local optima, oscillation, or catastrophic forgetting is essential. Efficiency can be estimated by focusing on reducing the number of teaching examples [27].

Radial Basis Function (RBF) is a class of feedforward networks. It was first proposed by Broomhead and Lowe in 1989 [28]. It was inspired by the receptive fields of the cerebral cortex in the human brain that has a powerful function approximation capability [22]. The traditional RBF neural network is a class of nets with kernel functions in its hidden layer. A simple kernel function is distributed in a neighborhood and its response is essentially local to the input vector [27]. The

architecture of RBF consists of three neural layers: an input layer neurons for feeding input feature vectors, a hidden layer of RBF neurons that return the outcome of radial function, and an output layer neurons that can calculate linear combination [29] (refer to Figure 1). The hidden layer neurons are represented by a radial basis function with a center where each neuron executes a nonlinear transformation according to the radial basis activation function[30]. A linear transformation is performed between the hidden layer nodes and output nodes. Generally, LMS gradient descent, or pseudo-inverse solution is used for the determination of the weights between the kernel and output layer[29].

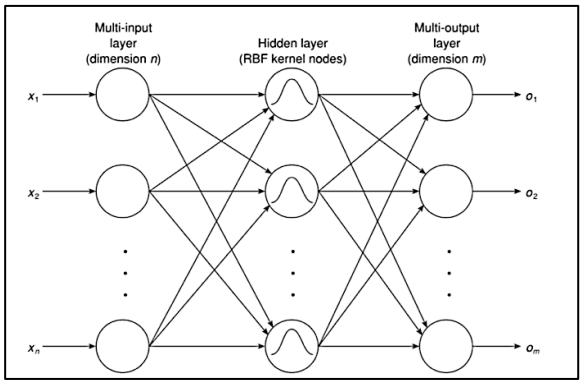

Figure 1 Schematic presentation of an RBFNN. Taken from [22]

The RBF function is generally given by the eq 1 [22]

$$g_i(x) = \gamma_i\left(\frac{\|x - v_i\|}{\sigma_i}\right) \quad (1)$$

where x is the input vector. v_i is the center of the receptive field of g_i, and σ_i is the width.

For RBFNN, Gaussian function is used widely [22]

$$g_i(x) = exp\left(\frac{-\|x - v_i\|^2}{2\sigma_i^2}\right) \quad (2)$$

Thus the output of the RBFNN is given by the eq 3 [22]

$$o_i(x) = \sum_{i=1}^{n} w_{ij} g_i(x) \qquad j = 1, 2, \ldots r \quad (3)$$

where n is the number of hidden layer kernel nodes and r is the number of output nodes. w_{ij} is the connection wights between the hidden nodes and output nodes.

The important parameters for designing and training of RBF are threefold [27][29] and the training is done in a hybrid approach as in the last two stages of the training design [22]:

- Determining the number of kernel functions (hidden layer neurons,) to use for the problem
- Adjusting and determining the center and the width of the kernel functions.
- Calculating the connecting weights between the hidden layers and the output nodes.

The first part of RBF training is the adaptation of the kernel function centers and scaling parameters [22] (Refer to Figure 2). The center and the width of the kernel function can be determined by clustering methods such as K-means clustering method [27] and P-nearest neighbor heuristics [27], supervised vector quantization and supervised classification tree algorithm [29]. In the absence of target signals, clustering and vector quantization techniques can be used to group the data points. The objective is to minimize the error by determining a representative set of the kernel function centers from a large dataset. In the case of supervised training, learning vector quantization or k-means clustering can be used to determine the centers [29]. The LVQ algorithm was proposed by Kohonen in 1990 as a competitive network where each output represents a class. It can support both binary and multi-class classification problems. On the other hand, the decision tree algorithm breaks down the feature space into nodes. Each node represents a region and a terminal node represents a class where all data points within the region belong to the same class [29]. Figure 3 shows the distribution of 12 Gaussian nodes with centers and the spread of three data points.

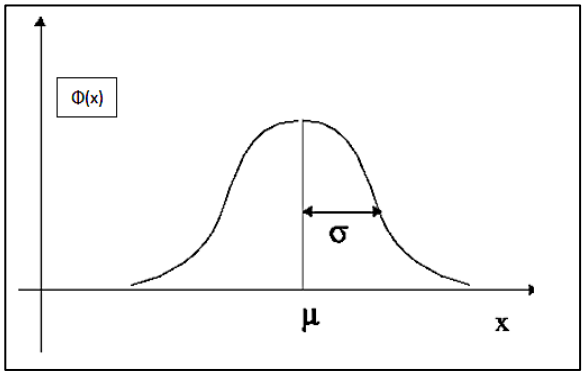

Figure 2 Gaussian neural activation function with center (μ) and width (σ). Taken from [31]

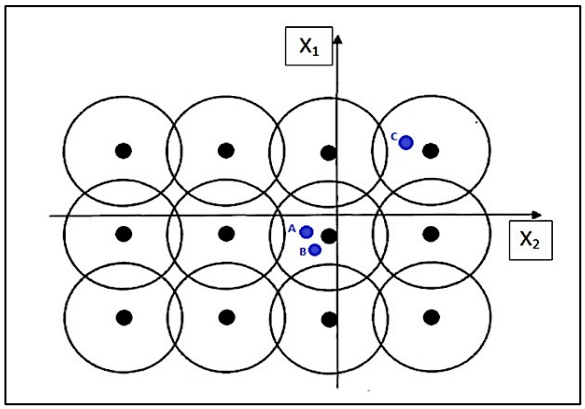

Figure 3 2D vector space is partitioned by 12 Gaussian nodes. Boundary circles represents the range of the receptive of field of neurons. Taken from [31]

The calculation of the kernel width is critical for the performance of the network. If the kernel width is too large, the probability density is affected. Again, if the width is too small,

there can be "over-adaptation" [29]. Thus, a reasonable kernel width is needed to be selected without or less affecting the probability density.

In the second part of the training, after the adaptation of centers and widths, supervised training with LMS or gradient descent method is used to update the connection weights between the hidden and output layer [22].

The most significant advantage of RBFNN is the shallow architecture that reduces the computing speed significantly and has efficient training algorithms compared to MLP with multiple hidden layers [27][22]. RBFNN can also efficiently map complex data and process and solve a wide range of classification problems.

IV. Feature Extraction and Dimensionality Reduction

Multivariate data is represented by a matrix where rows are considered as 'objects' and columns as 'variables'[32]. Both PCA and LDA are widely used for feature extraction and dimension reduction [9]. However, there exist some basic differences between them. PCA transforms the data into a different space and changes the location, and shape of the data for projecting on the most variant variable while LDA focuses more on maximizing the separable region [33].

A. Principal Component Analysis

PCA is a multivariate statistical technique. PCA analyses a dataset and extracts the dominant pattern in a matrix and estimates the correlation among variables [32][7]. Extracted features are expressed as a set of orthogonal variables called principal components. The goals of PCA are to compress the data set by removing unimportant information, extract important information, simplify the data set description and analyze the inter-relation among the variables [7]. This is done by calculating a new variable termed as the first principal component which has the largest variance and represents the most of the data. The second component is calculated orthogonal to the first component. Subsequent components are computed under the constraints of the former components. The values of these new variables are factor scores and are projected onto the principal component. The project of the principal components and distribution of variance are given at Figure 4 and Figure 5.

In the PCA technique, the projection of the vector X by a projection matrix P on a new dimension d' yields the object coordinate T. The column t_a in T is called the score vector and the row p_a in P is called the loading vector. The vector t_a and p_a are orthogonal. The difference between the original coordinates and the projected coordinates is called residuals. The hyperplane goes through the origin. The p_a vector is normalized as such that it equals to 1. The scoring and loading vector with the highest eigenvalue is assumed to have the most relevant information and called the 1st principal component. The rest of the components are expressed in the descending order of eigenvalues. The further values from the 1st principal components are considered as 'noise' and can be discarded [32].

For example, n vectors $(x_1, x_{2,}, \ldots . x_n)$ from a d dimension is projected on d' dimension such that $(x'_1, x'_{2,}, \ldots . x'_n)$. Then from [34],

$$x'_i = \sum_{k=1}^{d'} a_{k,i} e_k \, , \, d' \leq d \qquad (4)$$

where, e_k are eigenvectors in scatter matrix S and principal components $a_{k,i}$ are the projection of the original vectors x_i on e_k

The scatter matrix S is defined as [34]

$$S = E\,[\, x_i x_i^T \,], \; for\; i - 1\; to\; n \qquad (5)$$

where, $E\,[\, x_i x_i^T \,]$ is the statistical expectation operator. The variance of the principal component is given by [34]

$$\sigma^2\,(e_k) = E\left[\, a_{k,i}^2 \right] = \; e_k^T S\, e_k \qquad (6)$$

where, e_k in d by 1 vector $[(e_{1,k}\,, e_{2,k}\,, \ldots e_{d,k}\,]$. When $d' < d$ is satisfied, the orginal data set can be given by [34]

$$E_{d'} = \frac{1}{2} \sum_{i=d'+1}^{d} \lambda_k \qquad (7)$$

where λ_k are the eigenvalues of the matrix S. From the eqn 7, it is seen that the largest value of eigenvalues represents the important features of the entire data table [34].

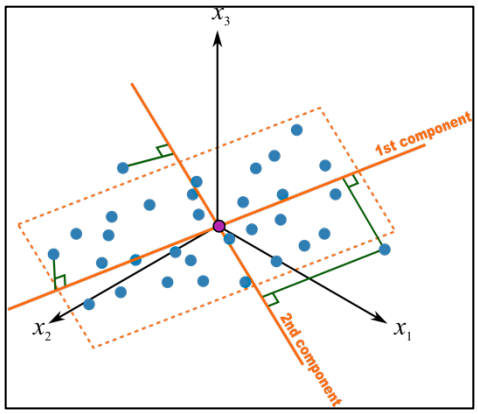

Figure 4 1st and 2nd component of PCA. Taken from [35]

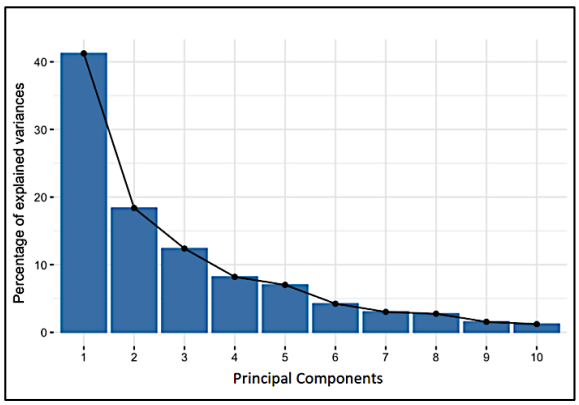

Figure 5 PCA scree plot. Percentage of variation reducing by each succeeding Taken from [36]

B. *Linear Discriminant Analysis*

LDA projects the data in a such way that the ratio between the distance in between-class to the distance in within-class is maximized and spread the discrimination among classes [37] (refer to Figure 6). LDA reduces the data dimension by compression the multi-dimensional data into one-dimensional line [9].

Considering a dataset $X = \{ x_1, x_2, \ldots . x_n \}$ of N samples with M classes, LDA performs a linear transformation of G over X such that the ratio between two scatter matrices, between-class, S_b and within-class S_w is maximized [8][37], where

$$S_w = \sum_{c=1}^{M} \sum_{x \in \{ x_c \}} (x - \bar{x_c}) (x - \bar{x_c})^T \quad (8)$$

and

$$S_b = \sum_{c=1}^{M} n_c (\bar{x_c} - \bar{x})(\bar{x_c} - \bar{x})^T \quad (9)$$

Where, n_c is the number of samples in class c, $\bar{x}$ is the mean vector of the dataset X, such that $\left(1/N\right) \sum_{i=1}^{N} x_i$, and $\bar{x_c}$ is the mean vector of class c.

Now, the transformation matrix G is given by eigenvalue decomposition matrix such that $D = S_w^{-1} S_b$ [8].

G is a matrix whose column represents the discriminate eigenvector. Vectors with small eigenvalues can be discarded keeping the higher values [8][37][33].

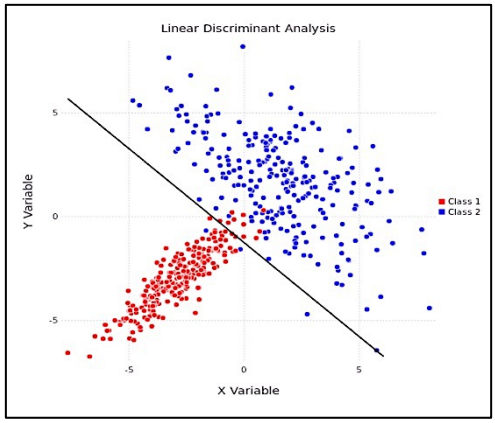

Figure 6 Separation of classes by LDA. Taken from [38]

V. METHODOLOGY

The flow chart for designing this experiment is given in Figure 7. Dataset (description of the dataset is given in the next section) was processed using the Pandas data frame. The dimension of the data was independently reduced using PCA and LDA. PCA was used to find the most variant data feature while LDA was used to maximize the separability of the data classes. The reduced dataset was split into training and testing data for 3 cases: 50-50, 70-30 and 80-20. The accuracy, sensitivity specificity and f-score of the classifier were analyzed to understand the efficacy and validity of the models.

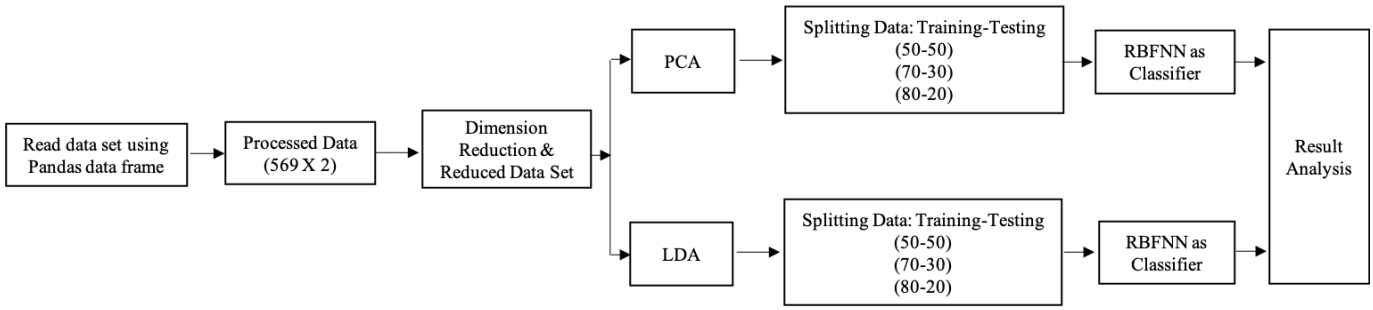

Figure 7 Study methodology flowchart

VI. IMPLEMENTATION

A. Dataset

The Wisconsin dataset on breast cancer diagnosis was created and mined in 1991 [39]. The breast cancer database was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg [40]. The dataset is multivariate with 569 instances. Features were extracted from digitized images of needle aspirate of a breast mass. The images of the breast mass are given in Figure 8 a and 8b. The features in the dataset represent the cell nuclei characteristics of the breast mass images. The separating plane was acquired with Multi-surface Method-Tree [40]. The dataset has 32 attributes of which two attributes were the ID Number and Diagnosis Classification (M for Malignant, B for Benign). Rest 30 attributes were features obtained from the breast mass images. The number of cases with two classes in the dataset is given in figure 9.

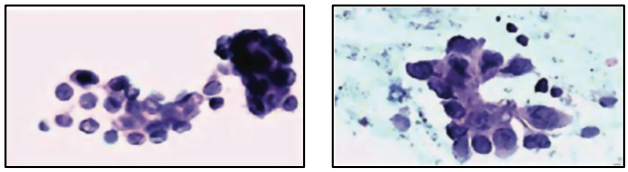

Figure 8a (left) Image of the benign breast mass (left). 8b (right) Image of the malignant breast mass. Taken from

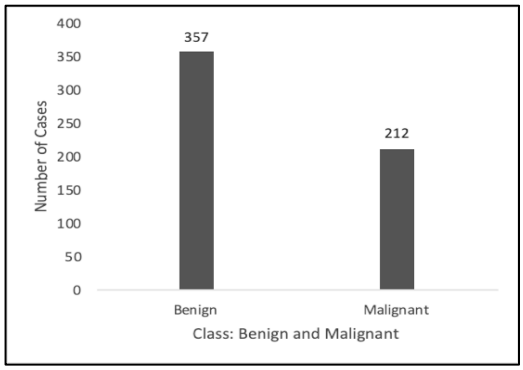

Figure 9 WDBC Dataset class frequency

B. Classification

In this study, the classification was binary in nature. The data features corresponding to "Malignant" and "Benign" were transformed to "1" and "0". The features of the dataset were extracted and dimensions were reduced using PCA and LDA.

PCA was implemented on the data to extract features, and the first 5 principals components were considered for the extraction. 5 covariances were plotted on the scattered plot to understand the effect of principal components. LDA was used to maximize the class separability of the dataset. Data were projected and transformed into 2 dimensions. In both cases, the dimensionally reduced dataset was feed to RBFNN for training.

K-means clustering was used to obtain the centers and the widths of the Gaussian functions for the RBFNN classifier. The spread of the network was varied to understand the effect of the spread on the classification accuracy. ADAM optimizer was used for optimizing the connection weights between the hidden and output layer. RBFNN network was trained with 300 epochs. The accuracy, sensitivity, specificity and F1 score were recorded for the varying spread between 10 and 100. The results were also recorded with the varying splits of training and testing data. Google Colaboratory was used to run the codes and TensorFlow was used at the backend.

VII. RESULT AND DISCUSSIONS

The heatmap showing the correlation among the features of the dataset is given in Figure 10. As seen from map, the features are mostly correlating and the brighter diagonal line demonstrates a perfect correlation. The symmetry about the diagonal indicates that similar features are paired.

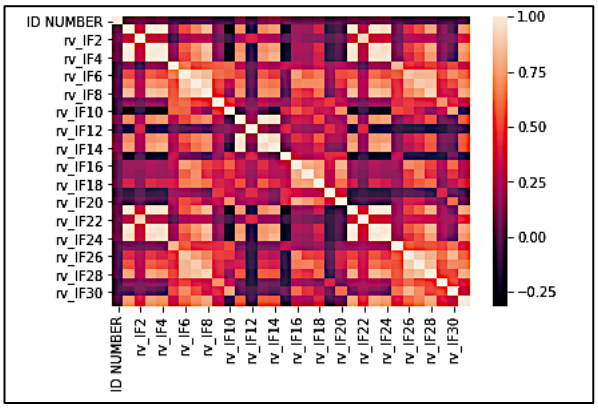

Figure 10 Heatmap of the dataset

A. Classification Accuracy for Different Training-Testing Splitting

The training-testing data partition was varied as 50-50, 70-30, and 80-20. The accuracy, specificity, and sensitivity of the model for varied partitioned are given in Table 1. It is found that the correct splitting of training and testing data is an important parameter for the classification model. The prediction accuracy increased gradually for the increment in the training dataset. While the sensitivity of all the models (RBFNN, PCA-RBFNN, and LDA-RBFNN) remained almost equal, the specificity improved significantly for the increase of the training dataset. It signifies that the model becomes reliable, by achieving high sensitivity and specificity, when the ratio of the training-testing split is higher.

TABLE I
THE VALUE OF STATISTICAL PARAMETERS FOR RBFNN, PCA-RBFNN AND LDA-RBFNN MODELS

Method	Accuracy	Specificity	Sensitivity
Training-testing : 50-50			
RBFNN	.74	.32	.98
PCA-RBFNN	.80	.48	.99
LDA-RBFNN	.90	.92	.97
Training-testing : 70-30			
RBFNN	.84	.65	.96
PCA-RBFNN	.93	.87	.97
LDA-RBFNN	.95	.90	.99
Training-testing : 80-20			
RBFNN	.90	.76	.97
PCA-RBFNN	.94	.85	.98
LDA-RBFNN	.97	.92	.99

B. *RBFNN Model without Data Dimension Reduction*

In the case of classification with RBFNN (training-testing partition at 80-20) without dimension reduction by PCA and LDA, the accuracy of prediction was 90% (refer to Figure 11). The sensitivity of the model was calculated as 97% while specificity was 76%. This means that the model can classify positive cases with higher accuracy than the negative cases as negative. The F 1 score of the model was calculated as 92%. It shows that RBFNN, as a classifier, cannot predict better when the data is not dimensionally reduced or data features are extracted.

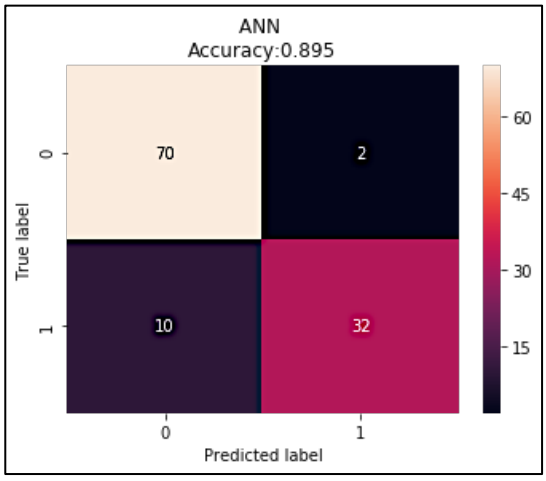

Figure 11 Confusion matric for RBFN with training-testing partition of 80-20

C. *PCA-RBFNN Model*

The Implementation of PCA on the dataset has yielded better accuracy with RBFNN. The accuracy of the classifier increased to 94% (refer to Figure 12). The sensitivity of the model was 98%, indicating that it could predict positive cases with better accuracy than the model without any data extraction. The specificity also increased to 85% from 76%, but it still lacks satisfactory accuracy in predicting negative cases. The F1 score of the model was 95%, making it a comparatively better model than the model with only RBFNN.

Refer to the ROC curve of PCA (see Figure 13), the AUC is 0.91 which means that the model has a good measure of separability for the classification problem on the dataset. The shape of the curve goes almost up, and then a gradual right-up towards 1. The shape signifies that PCA-RBFFN is a better classification model.

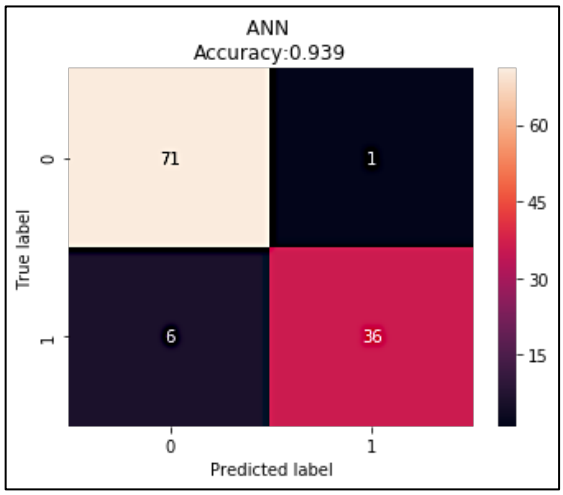

Figure 12 Confusion matric for PCA-RBFN with training-testing partition of 80-20

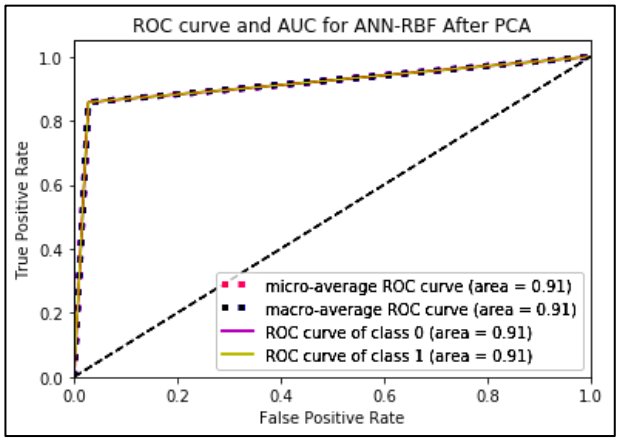

Figure 13 ROC Curve of PCA-RBFNN

D. LDA-RBFNN Model

LDA was used to maximize the data class separability. The implementation of LDA on the dataset has resulted in better accuracy with RBFNN. The accuracy of the classifier increased to 97% (refer to Figure 14). The sensitivity of the model is increased to 99%, indicating that it could predict positive cases better than the PCA-RBFNN model, or RBFNN model. The specificity also increased to 92%, indicating better accuracy in predicting negative cases. The F1 score of the LDA-RBFNN model was 97%. Compared to RBFNN and PCA-RBFNN, this model performs superior classification for the dataset.

Refer to the ROC curve of LDA (Figure 15), the AUC is 0.96, meaning that the model has high accuracy and separability for the classification. The shape of the curve signifies that the LDA-RBFFN is a better classification model than the PCA-RBFNN model.

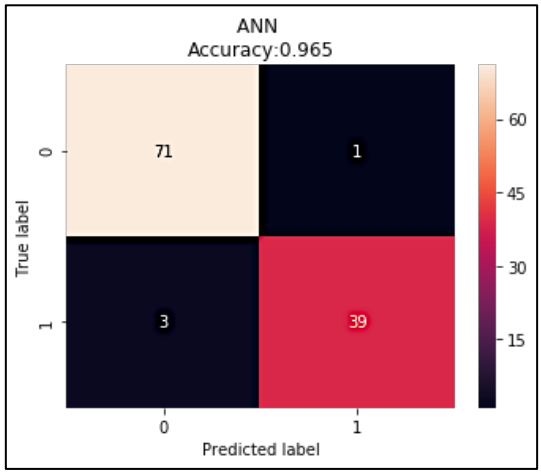

Figure 14 Confusion matric for LDA-RBFN with training-testing partition of 80-20

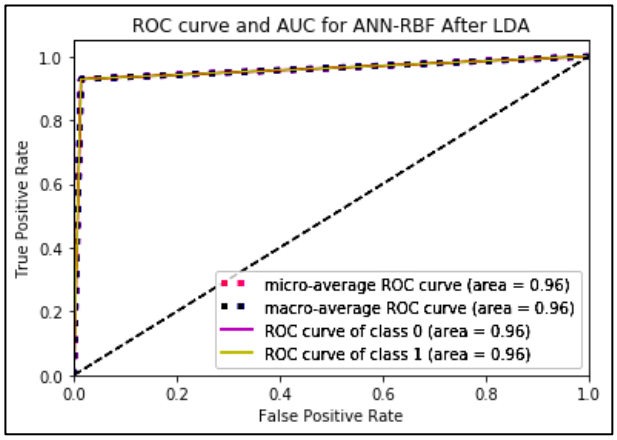

Figure 15 ROC Curve of LDA-RBFNN

E. Effect of Spread on the Classification Accuracy

In the RBFNN model, the spread is an important parameter for the classification performance of the network. Too large spread causes all the hidden layer nodes to respond to input, and loses the classification accuracy. On the other hand, for small spread, each hidden node responses only to a specific input. The result of different spread from 10 to 100 in the increment of 10 for RBFN, PCA-RBFNN and LDA-RBFNN are given in Figure 16a, 16b, and 16c.

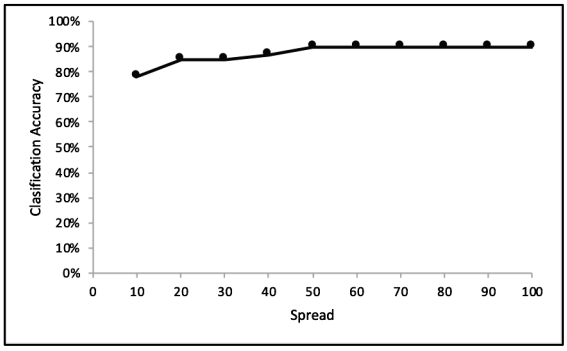

Figure 16a Spread versus Classification accuracy for RBFNN

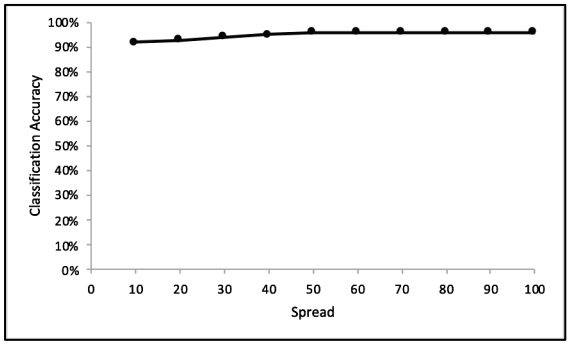

Figure 16b Spread versus Classification accuracy for PCA-RBFNN

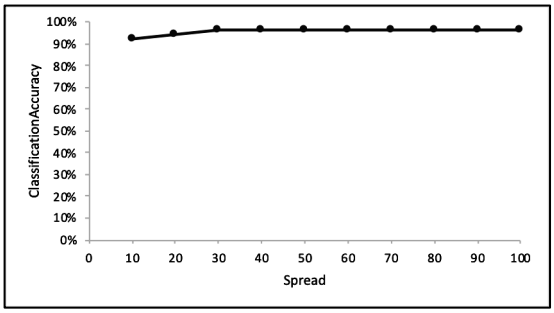

Figure 16c Spread versus Classification accuracy for LDA-RBFN

The classification accuracy, 78%, at the spread of 10 rose to 90% for the spread of 50 for the RBFNN model without data dimension reduction. For the spread between 50 and 100 the classification accuracy of 90% remained steady (refer to Figure 16a). After the feature extraction and dimension reduction by PCA, the classification accuracy increased to 92% for the spread of 10. After rising in every 10 spread variation,

classification accuracy reached 96% and remained steady up to the spread of 100 (refer to Figure 16b). For the dimension reduction with LDA, the classification accuracy improves to 96% for the spread of 30 (refer to Figure 16c). There was an insignificant variation in the classification accuracy with regards to changing spread values. Analysis of the results shows that the classification accuracy is significantly affected by the spread value where the accuracy increases up to a certain level with the increase of spread, but it does not increase beyond that point for the increase of spread.

From the comparison of the three models (RBFNN, PCA-RBFNN, and LDA-RBFNN), the LDA-RBFNN model performs better in terms of accuracy and reliability. This is because of the very fundamental difference in the approach of dimension reduction between PCA and LDA. While PCA focuses more on the most variant variables, LDA focuses more on increasing the gap between the classes. Thus, for the classification models, LDA performs better.

VIII. Conclusions

Breast Cancer is considered one of the deadliest diseases among women from all family backgrounds. Early detection and prediction of cancer can tremendously ameliorate cancer treatment, and reduce the death count. Ironically, in the absence of any early symptoms, the process of cancer detection is highly complex, involving the evaluation of a large amount of non-linear data. ANN and ML techniques have been proven to be a great aid in the diagnosis decision making. RBFNN, as a special class of feedforward networks, has the advantage of a narrow architecture with radial basis function. It performs a nonlinear transformation between the input space and kernel space and a linear transformation between the hidden and output layer. Nonlinear transformation with Gaussian function spreads the data in high-dimensional space which makes the separability of the classes more distinct. Thus, RBFNN has a better classification efficacy for the complex data. In this work, the RBFNN as a classification model without any data dimension reduction yields comparatively a poor result. PCA and LDA based approach for selecting the most representative features of the dataset has shown to improve the classification performance of the RBFNN. However, LDA implemented RBFNN presents a better classification model than PCA implemented RBFNN model. It is noticed that the data set feature extraction is of immense importance for the stable and reliable classification model. Therefore, future work may focus on feature extraction using evolutionary computation based algorithms and swarm intelligence based algorithms like ACO, PSO, and ABC.

Acknowledgment

I am grateful to Prof. Haitham Masaud Amar for encouraging me to study this interesting and insightful topic. I also acknowledge the continuous support of Mahmoud Mohamed Nasr Mohamed.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.," *CA. Cancer J. Clin.*, vol. 68, no. 6, pp. 394–424, 2018.

[2] A. Bhardwaj and A. Tiwari, "Expert Systems with Applications Breast cancer diagnosis using Genetically Optimized Neural Network model," *Expert Syst. Appl.*, vol. 42, no. 10, pp. 4611–4620, 2015.

[3] M. A. Devi, S. Ravi, J. Vaishnavi, and S. Punitha, "Classification of Cervical Cancer Using Artificial Neural Networks," *Procedia Comput. Sci.*, vol. 89, pp. 465–472, 2016.

[4] S. Kaymak, A. Helwan, and D. Uzun, "Breast cancer image classification using artificial neural networks," *Procedia Computer Science*, vol. 120. pp. 126–131, 2017.

[5] S. Agrawal and J. Agrawal, "Neural network techniques for cancer prediction: A survey," *Procedia Comput. Sci.*, vol. 60, no. 1, pp. 769–774, 2015.

[6] S. B. Cho and H. H. Won, "Cancer classification using ensemble of neural networks with multiple significant gene subsets," *Appl. Intell.*, vol. 26, no. 3, pp. 243–250, 2007.

[7] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 2, no. 4, pp. 433–459, Jul. 2010.

[8] S. Pang, S. Ozawa, and N. Kasabov, "Incremental Linear Discriminant Analysis for Classification of Data Streams," *IEEE Trans. Syst. Man Cybern. Part B*, vol. 35, no. 5, pp. 905–914, Oct. 2005.

[9] A. Subasi and M. I. Gursoy, "EEG signal classification using PCA, ICA, LDA and support vector machines," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8659–8666, 2010.

[10] A. M. Abdel-Zaher and A. M. Eldeib, "Breast cancer classification using deep belief networks," *Expert Syst. Appl.*, vol. 46, pp. 139–144, 2016.

[11] A. Mert, N. Kılıç, E. Bilgili, and A. Akan, "Breast Cancer Detection with Reduced Feature Set," *Comput. Math. Methods Med.*, vol. 2015, pp. 1–11, 2015.

[12] T. Ayer, O. Alagoz, J. Chhatwal, J. W. Shavlik, C. E. Kahn, and E. S. Burnside, "Breast cancer risk estimation with artificial neural networks revisited," *Cancer*, vol. 116, no. 14, pp. 3310–3321, Apr. 2010.

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

[14] A. T. Azar and S. A. El-Said, "Probabilistic neural network for breast cancer classification," *Neural Comput. Appl.*, vol. 23, no. 6, pp. 1737–1751, 2013.

[15] X. Xu, Y. Zhang, L. Zou, M. Wang, and A. Li, "A gene signature for breast cancer prognosis using support vector machine," *2012 5th Int. Conf. Biomed. Eng. Informatics, BMEI 2012*, no. Bmei, pp. 928–931, 2012.

[16] Y. C. Chen, W. C. Ke, and H. W. Chiu, "Risk classification of cancer survival using ANN with gene expression data from multiple laboratories," *Comput. Biol. Med.*, vol. 48, no. 1, pp. 1–7, 2014.

[17] B. A. Garro, R. A. Vazquez, and K. Rodríguez, "Classification of DNA Microarrays Using Artificial Bee Colony (ABC) Algorithm," vol. 38, pp. 207–214, 2014.

[18] J. Dheeba, N. Albert Singh, and S. Tamil Selvi, "Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach," *J. Biomed. Inform.*, vol. 49, pp. 45–52, Jun. 2014.

[19] F. Ahmad, N. A. Mat Isa, Z. Hussain, M. K. Osman, and S. N. Sulaiman, "A GA-based feature selection and parameter optimization of an ANN in diagnosing breast cancer," *Pattern Anal. Appl.*, vol. 18, no. 4, pp. 861–870, 2015.

[20] N. Kayarvizhy, S. Kanmani, and R. V. Uthariaraj, "ANN models optimized using swarm intelligence algorithms," *WSEAS Transactions on Computers*, vol. 13. pp. 501–519, 2014.

[21] S. Fisher, Ronald Aylmer, "138: The Use of Multiple

Measurements in Taxonomic Problems.,” *Ann. Eugen.*, vol. 7, pp. 179–188, 1936.

[22] F. O. Karray and C. W. De Silva, *Soft computing and intelligent systems design : theory, tools, and applications*. Pearson/Addison Wesley, 2004.

[23] F. Ahmadizar, K. Soltanian, F. Akhlaghiantab, and I. Tsoulos, “Artificial neural network development by means of a novel combination of grammatical evolution and genetic algorithm,” *Eng. Appl. Artif. Intell.*, vol. 39, pp. 1–13, 2015.

[24] D. J. Montana and L. Davis, “Training Feedforward Neural Networks Using Genetic Algorithms,” *Proc. 11th Int. Jt. Conf. Artif. Intell. - Vol. 1*, vol. 89, pp. 762–767, 1989.

[25] M. Karnan, K. Thangavel, R. Sivakuar, and K. Geetha, “Ant colony Optimization for Feature Selection and Classification of Microcalcifications in Digital Mammograms,” in *2006 International Conference on Advanced Computing and Communications*, 2006, pp. 298–303.

[26] M. H. Aghdam, J. Tanha, A. R. Naghsh-Nilchi, and M. E. Basiri, “Combination of Ant Colony Optimization and Bayesian Classification for Feature Selection in a Bioinformatics Dataset,” *J. Comput. Sci. Syst. Biol.*, vol. 02, no. 03, pp. 186–199, Jun. 2009.

[27] A. Roy, S. Govil, and R. Miranda, “An algorithm to generate radial basis function (RBF)-like nets for classification problems,” *Neural Networks*, vol. 8, no. 2, pp. 179–201, Jan. 1995.

[28] D. Lowe, “Multi-variable functional interpolation and adaptive networks,” *Complex Syst.*, vol. 2, pp. 321–355.

[29] F. Schwenker, H. A. Kestler, and G. Palm, “Three learning phases for radial-basis-function networks,” *Neural Networks*, vol. 14, no. 4–5, pp. 439–458, May 2001.

[30] S. Elanayar V.T. and Y. C. Shin, “Radial basis function neural network for approximation and estimation of nonlinear stochastic dynamic systems,” *IEEE Trans. Neural Networks*, vol. 5, no. 4, pp. 594–603, Jul. 1994.

[31] “Radial Basis Function Network | HackerEarth Blog.” [Online]. Available: https://www.hackerearth.com/blog/developers/radial-basis-function-network/. [Accessed: 05-Aug-2019].

[32] S. Wold, K. Esbensen, and P. Geladi, “Principal component analysis,” *Chemom. Intell. Lab. Syst.*, vol. 2, no. 1–3, pp. 37–52, Aug. 1987.

[33] S. Balakrishnama and A. Ganapathiraju, “Linear Discriminant Analysis—A Brief Tutorial,” *Computer (Long. Beach. Calif).*, vol. 11, 1998.

[34] A. Malhi and R. X. Gao, “PCA-Based Feature Selection Scheme for Machine Defect Classification,” *IEEE Trans. Instrum. Meas.*, vol. 53, no. 6, pp. 1517–1525, Dec. 2004.

[35] “geometric-PCA-8-both-components-with-plane.png (2387×2048).” [Online]. Available: https://learnche.org/pid/_images/geometric-PCA-8-both-components-with-plane.png. [Accessed: 06-Aug-2019].

[36] “A step by step explanation of Principal Component Analysis.” [Online]. Available: https://towardsdatascience.com/a-step-by-step-explanation-of-principal-component-analysis-b836fb9c97e2. [Accessed: 06-Aug-2019].

[37] J. Ye, R. Janardan, and Q. Li, “Two-Dimensional Linear Discriminant Analysis,” in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. MIT Press, 2005, pp. 1569–1576.

[38] “Introduction — Discriminant Analysis 1 documentation.” [Online]. Available: https://discriminantanalysis.readthedocs.io/en/latest/. [Accessed: 06-Aug-2019].

[39] “Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http:// archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.,” 2019.

[40] W. H. Wolberg and O. L. Mangasarian, “Multisurface method of pattern separation for medical diagnosis applied to breast cytology.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 87, no. 23, pp. 9193–6, Dec. 1990.