International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Identification of Breast Tumor Using Hybrid Approach of Independent Component Analysis and Deep Neural Network

Pooja Shah*[1], Trupti Shah[2]

Submitted: 04/08/2021 Accepted : 12/11/2021

Abstract: Among the most prevalent and serious diseases that affect women is breast cancer. A large number of women succumb to breast cancer each year. Breast cancer must be detected in its early stage. To deal with this challenge, Deep Neural Network (DNN) is used to achieve the success. In medical science, DNN has played a vital role in the diagnosis of a wide range of illnesses. In this study, we investigate the use of Regularized Deep Neural Network (R-DNN) for the prediction of breast cancer. A variety of optimization techniques, such as Limited-memory Broyden Fletcher Goldfarb Shanno (L-BFGS), Stochastic Gradient Descant (SGD), Adaptive Moment Estimation (Adam), and activation functions like as Tanh, Sigmoid, and Rectified Linear Unit (ReLu) are used in the simulation of R-DNN. The Independent Component Analysis (ICA) approach is used to identify the most effective features to be used in the study. To measure the efficacy of the model, training and testing of the proposed network is carried out using the Wisconsin Breast Cancer (WBC) (Original) dataset from the University of California at Irvine (UCI) Machine Learning repository. The detailed analysis of the accuracy is carried out and compared to the accuracy of other author's model. We find that the proposed network attains the highest accuracy.

Keywords: *Breast Cancer, Deep Neural Network (DNN), Independent Component Analysis (ICA), k-fold Cross Validation (CV)*

This is an open access article under the CC BY-SA 4.0 license.
(https://creativecommons.org/licenses/by-sa/4.0/)

1. Introduction

Breast cancer is the second most horrible and dangerous disease in the world. It is responsible for the deaths of many women. It is alarming that the incidence rate of breast cancer continuously goes up despite many efforts to reduce the disease. Every year, according to a World Health Organization survey, 2.1 million women are diagnosed with breast cancer, with a significant proportion of women dying as a result of a lack of early identification and treatment. Nearly 627,000 women died from cancer in 2018. According to the World Health Organization, accounting for approximately 15% of all cancer deaths among women worldwide. On the basis of global statistics, it is major public health problem that accounts for a large proportion of cancer-related deaths. Despite major advances in screening and patient management, breast cancer remains the second most common cancer among women in the United States and the second leading cause of cancer death among women in the United Kingdom [1]. In 2019, a total of 268,600 women were diagnosed with breast cancer for the first time in United States. According to the current Cancer Association, 41,760 people died as a result of the disease [1]. Early detection has been proven to be the most effective approach in easing the health and social consequences of this scenario, given the high cost of therapy and the disease's high prevalence among women worldwide. Finding and treating breast cancer in its earliest stages is crucial for successful treatment. With the use of computer-assisted technologies and artificial intelligence, it is now possible to detect this panic condition in its early stages.

Depending on its location, breast cancer can be classed as either invasive or non-invasive. Women over the age of 40 are more likely to have this condition, which is more common in postmenopausal women as well. Swelling of the entire or part of the breast, skin irritation, a typical breast sourness, nipple pain and redness, non-breast milk secretion etc. are some of the signs of breast cancer. For the accurate identification and treatment of many diseases, computer-assisted diagnosis (CAD) is frequently employed. Health care practitioners can use it to scrutinise and determine the phases of disease. Clinical decision support systems (CDSS) are being created to provide considerable improvement and precise choices on patients' situations, supporting medical practitioners in identifying and staging diseases. One of the system's goals is to apply Novel Deep Neural Networks to construct an innovative CAD model that may detect breast cancer in its early stage and deliver more accurate and reliable analysis, thereby assisting patients in preserving their lives.

Support Vector Machine (SVM), Artificial Neural Network (ANN), Extreme Learning Machine (ELM) and Deep Learning (DL) are examples of Machine Learning (ML) algorithms. Artificial Intelligence (AI) is incorporated into machines through the use of these algorithms. The purpose of this paper is to develop a Regularized DNN for the detection of breast cancer. When used in conjunction with other laboratory tests or more investigations, a system like this can aid medical practitioners in making an early diagnosis of breast cancer and commencing therapy accordingly until additional laboratory tests or further investigations are performed. A proposed Regularized Deep Neural Network (R-DNN) is tested on the WBC and the results are obtained. The

[1] *Department of Applied Mathematics, The Maharaja Sayajirao University of Baroda, Vadodara – 390001, Gujarat, India.*
ORCID ID : 0000-0001-6802-6379
[2] *Department of Applied Mathematics, The Maharaja Sayajirao University of Baroda, Vadodara – 390001, Gujarat, India.*
ORCID ID : 0000-0001-7670-0052
** Corresponding Author Email: pooja.s-appmathphd@msubaroda.ac.in*

proposed R-DNN out performed in terms of classification accuracy on the WBC dataset.

The paper is organised as follows. Literature survey is presented in Section 2, while methodology, which involves Deep Neural Networks (DNNs) and Independent Component Analysis (ICA) along with the brief overview of the WBC dataset are presented in Section 3. Section 4 represents the findings of empirical experiments and the results of the proposed approach. The conclusion of this work is contained in Section 5.

2. Literature Survey

For the different dataset like WDBC, WPBC and WBC data set, the authors in [2] employed ANN with GA-based feature selection and a PS (Particle Swarm)-classifier to diagnose breast cancer. PS-classifier is nothing but the Particle Swam Optimization (PSO) technique. The model is evaluated using PSO based classifier. PSO is a bio-inspired algorithm that searches for the best solution in the solution space. It differs from other optimization techniques in that it requires simply the objective function and is not dependent on the gradient or any differential form of the objective. It also has relatively few hyper parameters. In contrast to Genetic Algorithm (GA), PSO lacks evolution operators such as crossover and mutation. In PSO, prospective solutions, known as particles, move through the problem space by following the current optimum particles. PSO is more computationally efficient in terms of both speed and memory needs. One of the undeniable limitations of PSO is that it is less practical and accurate than GA. Because of PSO's benefits and disadvantages, more researchers began to adopt a combined PSO-GA algorithm for optimization. In this they proposed wrapper feature selection method based on genetic algorithm. This method is evaluated using Particle Swarm Optimization algorithm based classifier and obtained 96.70% accuracy for WBC dataset, while for the WDBC data set, they achieved 97.30% accuracy and 79.2% accuracy for WPBC dataset. A deep belief network-based Computer-Aided Detection (CAD) approach for identifying breast cancer was developed by the authors of [3], where they employed Deep Neural Network (DNN) as a classifier model and recursive feature elimination for feature selection. RFE is a feature selection approach that fits a model and removes the weakest feature (or features) until the specified amount of features is reached. RFE seeks to reduce dependencies and collinearity in the model by iteratively deleting a limited number of features per loop. After selection feature they implemented DNN model as a classifier. But their developed system has few limitation regarding training time. Because of the neural network has been intensively trained, the system's constraint is the algorithm's training time. RFE will try to eliminate half of the features if not any number is given. This can be problematic since removing too many or not enough features. As a result, it's critical that to choose number carefully. The experiment was carried out on the WBC dataset and resulted with an accuracy of 99.68%.

On the WBC dataset, the authors of [4] presented an automatic diagnostic system for breast cancer diagnosis that was built on ANNs based on Association Rules (AR). AR is used to reduce the number of dimensions in a breast cancer data set, while NN is utilized to make intelligent classification decisions. The proposed AR+NN model was compared with NN model. AR reduces the size of the input feature space from nine to four dimensions. Potentially AR method is a long processing taking time especially working with large number of data. A huge number of inputs can result in a rule set that is excessively vast and disorderly. In the test set 3-fold cross validation method was applied on WBC dataset and achieved 95.6% of classification accuracy.

On the WBC dataset, the authors of [5] exploited artificial neural networks. Not only is the impact of the number of neurons in the hidden layer on system performance addressed in this study, but also the impact of the number of neurons in the hidden layer and the kind of activation function in the neuron. Feed forward architecture was utilized for breast cancer identification in this study. The number of neurons in the hidden layer was set to 20, 21, 22, 23, and 24, and performance was compared in terms of MSE and Absolute Error. With three hidden layers and 21 neurons in the hidden layer, network achieved 98% accuracy.

The sluggish convergence and constant being caught at the local minima are two major drawbacks of the artificial neural network (ANN) classifier. To solve this challenge, the differential evolution algorithm (DE) was employed by authors [6] to find the best or near-best ANN parameter values. DE can significantly increase ANN learning. However, some issues with the DE approach remain, including a longer training time and lower classification accuracy. A concept based on islands has been developed in this system to address these issues. Each island in the island model approach runs a standard sequential evolutionary algorithm. A migration process ensures communication between subpopulations. Depending on a communication topology, some randomly selected individuals (migration size) migrate from one island to another after a particular number of generations (migration interval) (migration topology). The migration size, which shows the number of individuals travelling and governs the quantitative aspect of migration, and the migration interval, which denotes the frequency of migration, are the two essential and most sensitive elements of the island model approach. Using the WBC dataset, the authors proposed and tested an island-based training model with an ANN, in which they found 99.97% classification accuracy [6].

The author of [7], proposed fusion at classification level between MLP and Gradient based classifiers to get the most suitable multi-classifier approach for each data set like WBC, WDBC and WPBC. On a WBC dataset, the authors of [6] employed a mixture of MLP and the feature reduction technique PCA. Results showed that Learning Machine Neural Networks (LM ANN) has better generalization classifier model than BP ANN. However, the standard Gradient-Based Back Propagation Artificial Neural Networks (BP ANN). and achieved accuracy of 97.6% with their method.

On a WBC dataset, the authors of [8] utilized ANN with Extreme Learning Machine and a 5-fold cross validation technique. The performance of ELM with conventional BP ANN with gradient descent based learning algorithms are compared. In terms of breast cancer diagnosis, ELM ANN performed significantly better than BP ANN. Despite the fact that the specificity rate was slightly lower than that of BP ANN, ELM ANN significantly improved the sensitivity and accuracy rates. ELM ANN has a superior generalization model than BP ANN when it comes to identifying breast cancer using the Wisconsin Breast Cancer Dataset and achieved 96.40% accuracy.

According to the authors of [9], they employed a DNN with REF-feature selection technique, multiple training-testing sets. DNN with multiple layers of processing attained higher classification rate than SVM. In this proposed technique first they proposed with noisy data and then four features were selected by applying logistic regression model with recursive feature elimination process. RFE ranked and extracted the best feature and eliminates other features. Dataset were classified using DNN. And on the WBC dataset they

obtained classification accuracy of 98.62%.

The authors of [10] developed a process of diagnosis and prognosis using ANN with supervised learning techniques like perceptron, cascade-forward back propagation and feed-forward back propagation for WBC dataset. Backpropagation with Feed-Forward Layers make up a Feed-Forward network. A link to the network input is made by the first layer. Each layer after that has a link to the layer before it. The network's output is generated by the last layer. Any type of input to output mapping can be done with feed-forward networks. Cascade-forward neural networks are similar to feed-forward neural networks in that they feature a connection from the input and every previous layer to the next layers. In a three-layer network, the output layer, in addition to the hidden layer, is connected directly to the input layer. Cascade-Forward Back propagation algorithm achieved good result for the diagnosis of breast cancer.

The suggested technique by authors of [11] is put to the test on the WBC Dataset from the UCI Machine Learning Databases Repository. 10-fold cross validation was used during the categorization phase. On the same database, the proposed algorithm was compared to various classification algorithms like Random Forest (RF), K-Nearest Neighbors, Naïve Bayes (NB) and Neural Networks. The suggested algorithm's evaluation findings show that sensitivity and F measure have improved accuracy. A prediction model for the WBC dataset with 11 features, proposed by the authors of [11], was shown to be highly accurate (97.00% classification accuracy) when combining ANN with feature reduction technique - PCA. The primary aim of PCA is to detect patterns in datasets and determine similarities and differences between individual attributes.

Abien Fred M. Agarap [12] demonstrated six ML methods on the (WDBC) dataset: Gated Recurrent Unit (GRU) with SVM, Linear Regression, Multilayer Perception (MLP), Nearest Neighbor (NN) search, Softmax Regression, and SVM by assessing classification test accuracy, sensitivity, and specificity values. However, GRU models continue to have issues such as sluggish convergence rate and low learning efficiency, resulting in excessive training time and even under-fitting. All of the classifiers' hyper parameters were manually assigned. With a test accuracy of 99.04%, the MLP algorithm stands out among the applied algorithms.

The authors of [13] uses a 10-fold cross validation method to compare the classification accuracy of the different classifiers Decision Tree (J48), MLP, NB, Sequential Minimal Optimization (SMO), and Instance Based for K-Nearest Neighbor (IBK) on three different databases of breast cancer WBC, WDBC, and WPBC. In addition, they combined these classifiers at the classification level to obtain the best multi-classifier strategy for each data set. Their experimental results suggest that fusion of MLP and J48 with PCA is superior to other classifiers utilizing the WBC data set in classification. The following characteristics were chosen using PCA: Uniformity of Cell Size, Mitoses, Clump Thickness, Bare Nuclei, Single Epithelial Cell Size, Marginal Adhesion, Bland Chromatin, and Class. The results of the WDBC data set reveal that classification using SMO alone or fusion of SMO and MLP or SMO and IBK is superior to the other classifiers. The results of the WPBC data set reveal that the classification utilizing the fusion of MLP, J48, SMO, and IBK is superior to the other classifiers. Their all experiments were carried out using the WEKA data mining software.

The authors [14] employed SVM, ANN, and NB to diagnose breast cancer, and they found that they were effective. For reducing the dimension of features, they used Linear Discriminant Analysis (LDA). LDA is a feature extraction technique that computes transformation by maximizing between-class scatter and decreasing within-class scatter. This is done concurrently, and the utmost level of discrimination is reached. Eigen decomposition on covariance matrices is used in LDA to find the best transformation. LDA determines the characteristics that account for the most variation between classes. Unlike PCA, LDA may yield better results. They preferred SVM-LDA over NN-LDA because NN-LDA requires more processing time. It achieved 98.82 percent classification accuracy, 98.41 percent sensitivity, 99.07 percent specificity, and an area under the receiver operating characteristic curve of 0.9994. They were able to get classification accuracy of 97.64% for NN using PCA and 98.82% for NN using LDA.

The authors of [15] employed SVM and ANN with correlation feature selection strategies to reach classification accuracy of 97.14% and 96.71%, respectively, for their classification algorithms. Disadvantage of SVM is that, they have very complex algorithms and they need more memory and power for computation. Whereas, in ANN it is able to solve complex problem with less memory and power. But ANN is like “Black Box” i.e. solution behind the function is unknown and it has also overfitting issue.

According to [16], the authors suggested an SVM-based classifier and compared it to other classifiers such as the ANN and the Bayesian classifiers. Bayesian Nets are a set of probabilistic computing methods for most issues with uncertainty. A Bayesian Net classifier is defined as a theoretically sound method of displaying probability distributions in a graphical format that is concise and understandable. Authors of [16] proposed two different approach as: Classification of the WPBC patient data based on the disease free or recurrence time (prognosis) and Automated diagnosis of breast cancer based on the WDBC patient data. In first approach the WPBC instances were divided over four classes, namely C1, C2, C3 and C4, according to the value of the recurrence or the DFS (disease-free) time. Based on that SVM has applied for each of C1, C2, C3 and C4. In second approach, SVM model were experimented with two kernels namely polynomial and Gaussian. Also experiments were performed on Bayes net, Naïve Bayes and ANN.

The authors in [17] studied the feature reduction approach ICA using K-NN, ANN, SVM, and RBFNN as well as other neural networks. Features were reduced up to one feature out of 20 using ICA feature reduction technique. ICA considers that each data sample is a collection of distinct components, and it seeks to identify these independent components. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

The model was tested on a number of occasions and found to be accurate. ANN, C5.0, SVM, and a combination of these approaches were employed as classifiers by the authors of [18]. In a tree structure, Decision Tree (DT) is one of the supervised learning approaches used for classification and regression. The goal is to use the developed model to build this tree structure that predicts the label of a target variable. C5.0 is one of Clementine's rule induction techniques for generating a decision tree. It gives the option of seeing the rules in two separate formats: decision tree presentation and rule set presentation. A typical neural network is made up of multiple interconnected neurons that are stacked in layers to form networks. The network's ability to learn patterns and interrelationships in data is due to the connections between the neurons. A supervised learning model with associated learning methods is the SVM. It tries to classify results by mapping data to a higher-dimensional feature space and categorizing data points.

The authors of [18] achieved classification accuracy of 98.77% for the ensemble methodology and 97.54% for the ANN, 98.07 for SVM and 98.07% for C5.0.

3. Methodology

3.1. Deep Neural Network (DNN):

The Perceptron model, suggested by Frank Resenberg in 1957-1958, was implemented by Rosenblatt between 1965 and 1968, and the Multi-Layer Perceptron (MLP) model was introduced in 1968. An analogous model can be found in modern Deep Neural Networks (DNNs) or modern Feed Forward Neural Networks (FFNNs). DNNs can be thought of as stacked neural networks, that is, networks that are built of many layers. DNNs are incredibly robust networks based on the design of Artificial Neural Networks (ANN). Bengio (2009) [19] describes DNN as having numerous deep hidden layers between the input and output layers. Refer Fig. 1 to see the Deep Neural Network model that was used in this investigation. The back propagation algorithm is used to train the network [20]. Fig. 1 shows the difference between Simple Neural Network and Deep Neural Network.

Arrows in Fig. 1 indicate the direction of data flows. The first layer is the input layer, and nodes are made up of features or variables. The second layer is the hidden layer, which accepts input from the input layer and sends it to the output layer. Following the application of the activation function, the neurons in the output layer hold the output vector. The final layer is the output layer, which is made up of neurons that represent the desired results. Each layer is linked to the next layer by weights, bias, and an activation function, and the bias is added to the weighted total of inputs.

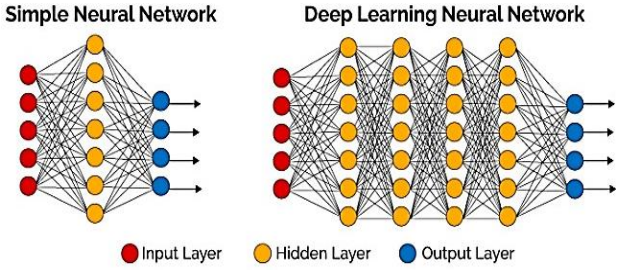

Fig. 1. Architecture of Simple Neural Network vs. Deep Neural Network (DNN) (Source: https://www.researchgate.net/publication/332165523_Survey_on_deep_learning_with_class_imbalance/figures?lo=1)

In this case, bias is defined as $b = +1$, which is the conventional value. Weights and bias are initially assigned at random and evaluated for the entire network. After the evaluation of the difference between the network output and the desired output, the error of the objective function is minimized by updating weights.

The regularization technique has been employed to address the overfitting problem. To increase model efficiency, we introduced Regularized Deep Neural Network (R-DNN) by incorporating a regularize term in the loss function as indicated in (1). In the proposed model, the performance of the classification model is measured by log loss error function as shown in (10).

$$loss + \sum_{j=1}^{n} \|w_j\|^2 \frac{\lambda}{2m} \quad (1)$$

Where, $n =$ number of layers
$w_j =$ weight matrix for the j^{th} layer
$m =$ number of inputs
$\lambda =$ regularization parameter

3.2. Optimization technique:

In order to reduce the error of an objective function by changing parameters, there are several optimization approaches to choose from. It is possible to train DNN model using Gradient Descent (GD), Stochastic Gradient Descent (SGD), Momentum based GD, Nesterove Accelerated GD, Adagrad, Adam, RmsProp, and other techniques. Several optimization approaches, including Stochastic Gradient Descent (SGD), Adaptive Moment Estimation (Adam), and L-BFGS, were employed in our model.

3.2.1. Stochastic Gradient Descant (SGD);

Iterative optimization techniques such as Gradient Descent (GD) are used to update network parameters such as weights and biases in order to reduce error in the objective function. SGD divides a dataset into a small number of batches, after which it identifies the gradient using a single batch and updates the parameters for each and every data point. For each subsequent set of instances, the process is repeated in the same manner until convergence is reached. Weights are updated using following (2):

$$w_{t+1} = w_t - \eta \Delta w_t \quad (2)$$

Where, η is learning rate.

3.2.2. Adaptive Moment Estimation (Adam):

Adam is yet another optimization strategy to be considered. This is regarded to be a more comprehensive variant of SGD. During parameter updating, it takes into account the cumulative history of the gradient. This algorithm takes into account both the first momentum and the second momentum of the gradient and it produces a successful outcome during the training of the model [21]. It also delivers a faster convergence rate than the SGD algorithm.

β_1 and β_2 are two values on which Adam relies. For the first moment estimation, β_1 is the exponential decay rate which is 0.9 and for second moment estimation, β_2 is the exponential decay rate with 0.999. On a given iteration t, m_t and v_t are calculated using (3) and (4) which are moving averages i.e. m_t is exponential average of gradients along w and v_t is exponential squared average of gradients along w. g_t is gradient w.r.t. objective function or cost function or loss function at time step t.

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \quad (3)$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \quad (4)$$

The bias correction of moving averages is calculated by (5).

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t} \text{ and } \hat{v}_t = \frac{v_t}{1-\beta_2^t} \quad (5)$$

Parameters are updated based on the calculated moving averages with learning rate η using following (6):

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t \quad (6)$$

Where, $\hat{m}_t$ and $\hat{v}_t$ are moving averages, and η is learning rate.

3.2.3. Limited-memory Broydern Fletcher Goldfarb Shanno (L-BFGS):

L-BFGS is a modified version of the Broyden Fletcher Goldfarb Shanno algorithm in the family of Quasi-Newton methods that approximates the BFGS algorithm using a limited amount of computer memory [22]. L-BFGS is a second-order optimization approach that requires significantly less memory than BFGS does. It takes into account both gradient and curvature information when

updating the parameter [22]. Iterative approximation of gradients is achieved by using the Hessian matrix. Compute the gradient of the function, Δf_i at the point X_i using $S_i = -[B_i]\Delta f_i$. Update the Hessian matrix as given in (7):

$$[B_{i+1}] = [B_i] + \left(1 + \frac{g_i^T [B_i] g_i}{d_i^T g_i}\right) \frac{d_i d_i^T}{d_i^T g_i} \quad (7)$$

Where, $[B_i]$ is an approximation to the Hessian matrix.

g_i and d_i are symmetric rank-one matrix.

3.3. Independent Component Analysis (ICA):

An analysis technique known as Independent Components Analysis (ICA) which is used to take a huge data set including many variables and break it down into smaller number dimensions that can be interpreted as self-organized functional networks (Beckmann & Smith, 2004) [19, 23]. i.e. ICA is used for feature reduction. When compared to Principal Component Analysis (PCA), which assumes that the components are uncorrelated in both the spatial and temporal domains, ICA components are maximum statistically independent in only one of the two domains. ICA is a linear dimension reduction technique that divides a dataset into columns of independent components. In this case, each sample of data is assumed to be a mixture of independent components, and the objective is to identify these independent components.

3.4. Data description:

Wisconsin Breast Cancer (WBC) (Original) dataset from the University of California at Irvine (UCI) Machine Learning repository [24] is used to test the proposed model. Dr. William H. Wolberg of the University of Wisconsin Hospitals in Madison provided this breast cancer database. He evaluated breast tumor biopsies for 699 patients up to July 15, 1992; each of nine features (attributes) was scored on a scale of 1 to 10. There are 699 rows and 11 columns in all. Among the 699 samples in the WBC dataset, 458 were benign and 241 were malignant. The WBC dataset has nine features and contains 699 samples. The patient's ID and class properties are included in all of these aspects. Nine characteristics (attributes) of WBCD are shown in Table 1.

Table 1. Wisconsin Breast Cancer (WBC) Dataset

No.	Attributes	Domain
1	Clump Thickness (CT)	1-10
2	Uniformity of cell size (UCS)	1-10
3	Uniformity of cell shape (UCSH)	1-10
4	Marginal Adhesion (MA)	1-10
5	Single Epithelial cell size (SEC)	1-10
6	Bare Nuclei (BN)	1-10
7	Bland Chromatin (BC)	1-10
8	Normal Nucleoli (NN)	1-10
9	Mitoses (Mit)	1-10
	Class	2 for benign 4 for malignant

Description of data is given as follows [25, 26]: In terms of CT, benign cells tend to be clustered in monolayers, but malignant cells are frequently grouped in multilayers. In the UCS and UCSH, the size and form of cancer cells might vary. Because of this, these factors are useful in determining whether or not the cells are malignant. In MA, Normal cells have a proclivity to adhere to one another. This capacity is often lost in cancer cells. As a result, adhesion loss is an indication of cancer. SEC has something to do with the previously mentioned homogeneity. Significantly expanded epithelial cells may be cancerous. Nuclei that are not surrounded by cytoplasm are referred to as BNs (the rest of the cell). These are most commonly found in benign tumours. BC describes the homogeneous "texture" of the nucleus seen in benign cells. The chromatin in cancer cells is coarser. Nucleoli are tiny structures that can be seen within the nucleus. The nucleolus is normally relatively tiny, if present at all, in normal cells. The nucleoli grow more visible, and there are sometimes more of them, in cancer cells.

4. Experiments and Results:

4.1. Experiments:

R-DNN with ICA is the foundation of the suggested model. The proposed model incorporates four major steps for training R-DNN, which are depicted in Fig. 2. There are five parts in this process: data collection, data pre-processing, feature reduction, separating the data into train and test sets and finally evaluating the model's performance by defined optimization techniques and stated activation functions.

After loading the WBC dataset into the network, we normalize dataset using a pre-processing technique to ensure that no attributes were missing or that the data was consistent. The following (8) is used to normalize the data.

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (8)$$

In the third stage, we employed a feature reduction strategy to decrease multi-collinearity from the model and increase its overall performance and accuracy. We utilize ICA to minimize the number of features in the WBC dataset. The number of features was decreased to three with no loss of generality and all of the original information from the data was kept. Following that, a 10-fold CV was applied to the data to separate it into a train-test set. With this technique, data sets are divided into 10-folds (i.e. groups), with each group serving as both a training and testing set to determine the model's overall effectiveness in general.

Working steps of proposed R-DNN:

i. Input dataset i.e. feature vector of WBC i.e. inputs $X = \{x_1, x_2, \ldots, x_9\}$ are pass through in the 1st layer. Then the data are sent to hidden layers with weights assigned. There can be as many as hidden layers.

ii. All calculation is done in the hidden layer after the inputs have been passed on. i.e. all input vectors are multiplied by weight vector W and added to the bias b : $y = W * X + b$. Weights are randomly initialized in forward propagation.

iii. The activation function is then applied to linear equation y in the step ii. as shown in (9). The activation function is a nonlinear transformation applied to the input before it is sent to the next layer of neurons. The activation function's significance is that it introduces nonlinearity into the model.

$$y = \varphi\left(\sum_{i=1}^{n} w_i x_i + b\right) = \varphi(W^T X + b) \quad (9)$$

Where, W is weight vectors, X is the input vectors, b is the bias and φ is the non-linear activation function.

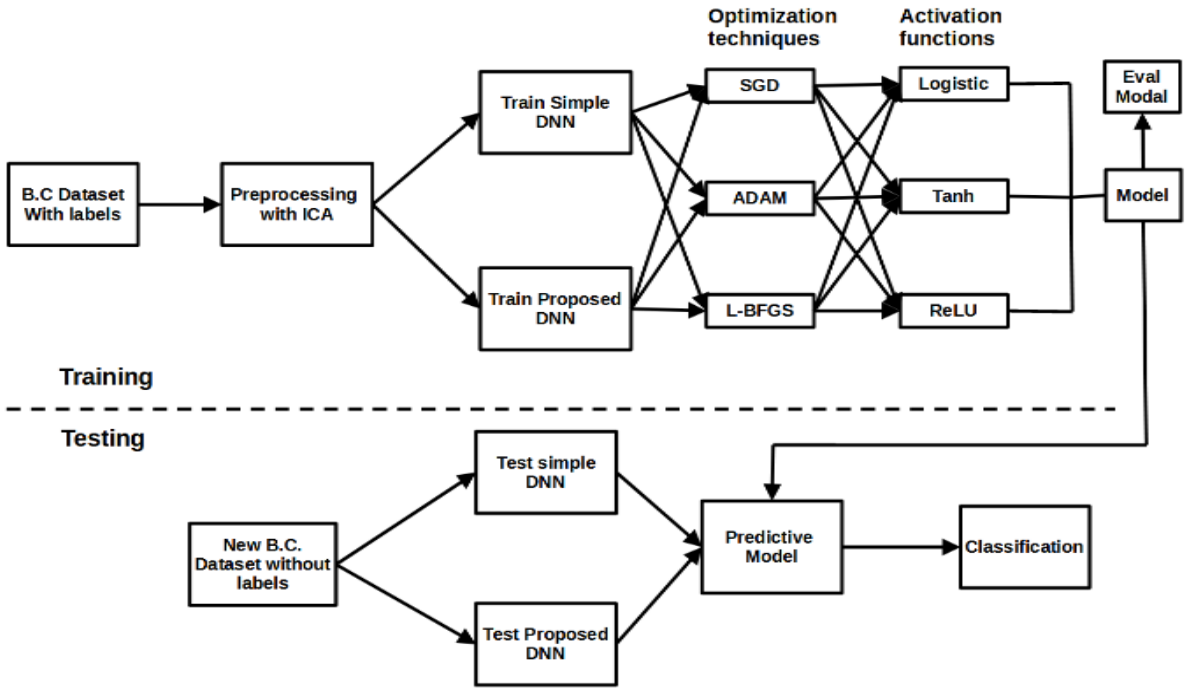

Fig. 2. Flowchart of proposed R-DNN with ICA

iv. Each hidden layer goes through the entire process stated in step iii. After we have passed through all of the hidden layers, we go to the final layer, which is our output layer, which provides us with the final output.

v. The error is calculated after receiving the output layer's predictions, which is the difference between the actual and predicted output. If the error is large, actions are taken to minimize the error using Backpropagation. The error or loss is calculated using log loss error function as shown in (10):

$$loss = -\frac{1}{N}\sum_{i=1}^{n} y_i ln\big(p(y_i)\big) + (1-y_i)\ln\,(1-p(y_i)) \quad (10)$$

Where, y_i is actual class and $ln\big(p(y_i)\big)$ is probability of actual class.

vi. The weights are updated with different optimizers methods using (2), (6) and (7).

The suggested DNN was trained and evaluated using the optimizers like SGD, Adam, and L-BFGS algorithms, which used activation functions such as Logistic, Tanh, and ReLU to train and test the network. We used the log-loss function, also known as the binary cross-entropy function, to calculate the difference in error between the network output and the desired output and to compare the two results.

We include an additional L2 Regularization term in the log-loss error function as shown in (1) to make it more robust. To reduce error by fitting a function adequately on the provided training set and avoiding overfitting, the regularization term is employed. This regularization technique helps to reduce variance in our model by penalizing for complexity. By adding L2 regularization to our model, we're effectively giving over some of our model's capacity to fit the training data well in exchange for the ability to generalize the model to data which hasn't been seen before i.e. on test data. Large weights are penalized by adding L2 regularization into log loss error function.

We evaluated our proposed model for various λ values such as 0.1,0.01,0.001,0.0001 and 0.00001 and achieved 100% classification accuracy for $\lambda = 0.001$.

Following the training of the Deep Neural Network, test data is collected in order to evaluate the model's performance in the classification of breast cancer as benign or malignant, respectively. Model performance and efficiency are determined by the following parameters: accuracy, sensitivity, specificity, precision, and F-score from the confusion matrix.

In order to determine whether a cancer model is capable of appropriately classifying cancer as malignant or benign, the Receiver Operating characteristic (ROC) [23] curve is utilized. A True Positive Rate (TPR) is calculated by comparing the rate of correctly classified cases (True Positive Rate) against the rate of wrongly classified instances (False Positive Rate). The True Positive Rate (TPR) is between 0 and 1. A different trade-off between a correctly diagnosed tumor being classified as benign or malignant is represented by each dot on the curve.

A dataset of WBC has been used as an input in this experiment, with the desired result being either benign or malignant. Nine input neurons, two hidden layers, and one output neuron are used to construct the model. It is necessary to increase or decrease the number of neurons in hidden layers in order to attain the best accuracy. It is trained with two hidden layers, each of which has seven pairs of hidden neurons, such as 100-0, 100-100, 250-100, 250-250, 500-100, 500-500, and 1000-1000, respectively. We used learning rates of 0.01, 0.001, and 0.0001 to train our network in order to accelerate the convergence of our model. DNN without ICA and 10-fold CV were used in the tests, as were DNN with ICA and 10-fold CV in the case of DNN with ICA.

4.2. Results

4.2.1. Case 1: Results for DNN without ICA and 10-fold CV (Simple DNN model):

In this scenario, the dataset is divided into two parts: a train set and a test set. In this case, 80% of data are taken as training set and 20% of data are taken as testing test. There are no approaches utilized in this case because DNN is a simple model.

The log loss error is used to assess performance. It depicts error vs. epoch for the training dataset, as seen in Fig. 3, Fig. 4 and Fig. 5. Although the error decreases with additional training epochs, the

error on the validation dataset may begin to climb when the network becomes over-fitted to the training data as a result of overfitting. For e.g., in the default configuration, the training is terminated after six consecutive increases in validation error, and the highest performance is obtained from the epoch with the lowest validation error. In this particular instance, the error is assessed in the cross-entropy function.

The Fig. 3, Fig. 4 and Fig. 5 depicts the varied learning rates for the training error. After training, the version of the network that performed the best on the validation set was used.

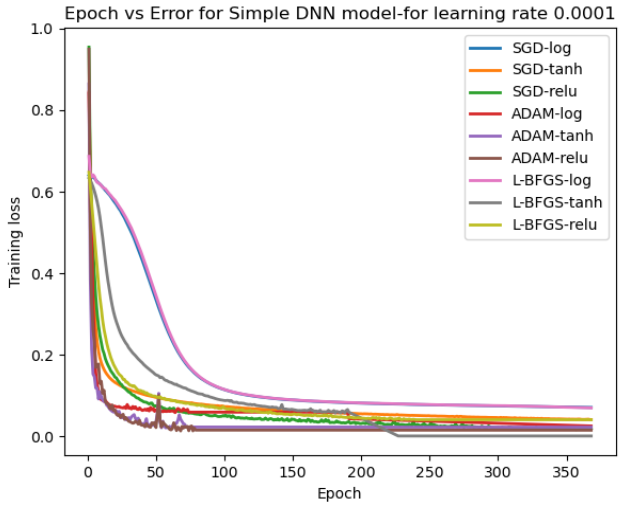

Fig. 3. Epoch vs Error for Simple DNN model-for learning rate 0.0001

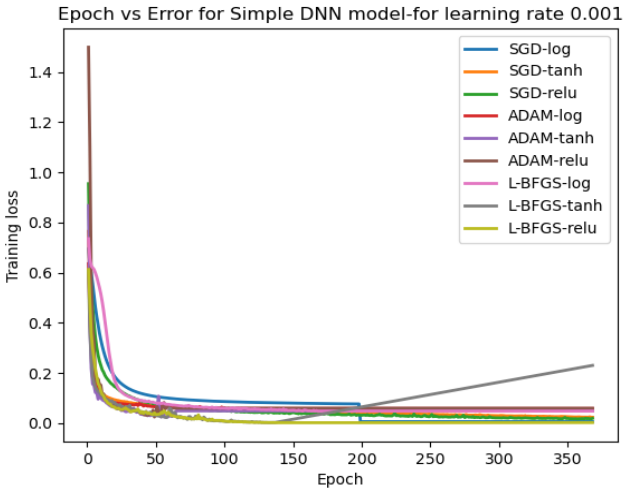

Fig. 4. Epoch vs Error for Simple DNN model-for learning rate 0.001

On the basis of train-test data, Fig. 6 depicts the confusion matrix for DNN without ICA and without a 10-fold CV for training data. When using a trained network, the Fig. 6 reveals that 97.5% of samples are accurately detected. For the first and second rows of the diagonal block, it is shown that 33.47% of the samples, or 160 samples, are accurately classified as benign and 64.02%, or 306 samples, are correctly classified as malignant. Both the benign and malignant instances represented by the other two blocks were misdiagnosed. A total of 0.21% i.e. only one malignant sample is mistaken as a benign sample and 2.30% i.e. 11 benign samples is mistaken as malignant samples.

Fig. 7 depicts the confusion matrix for testing data. 97.1% success rate for correctly diagnosing tumors is shown Fig. 7 for the network under consideration. According to the first two diagonal blocks, 36.59%, or 75 samples, are accurately classified as benign, whereas 60.49%, or 124 samples, are appropriately classified as malignant.

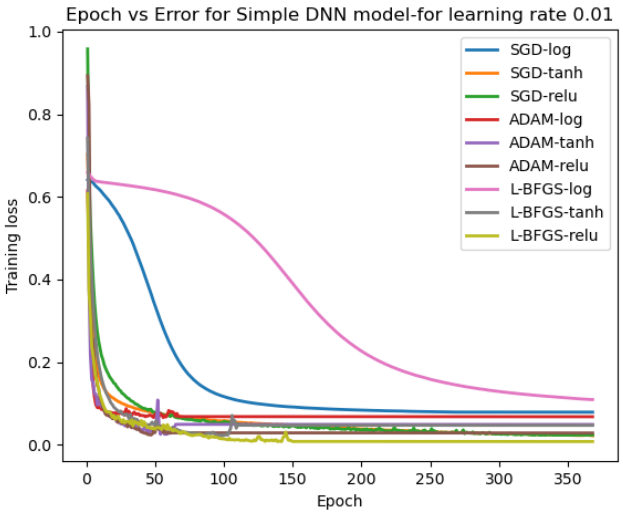

Fig. 5. Epoch vs Error for Simple DNN model-for learning rate 0.01

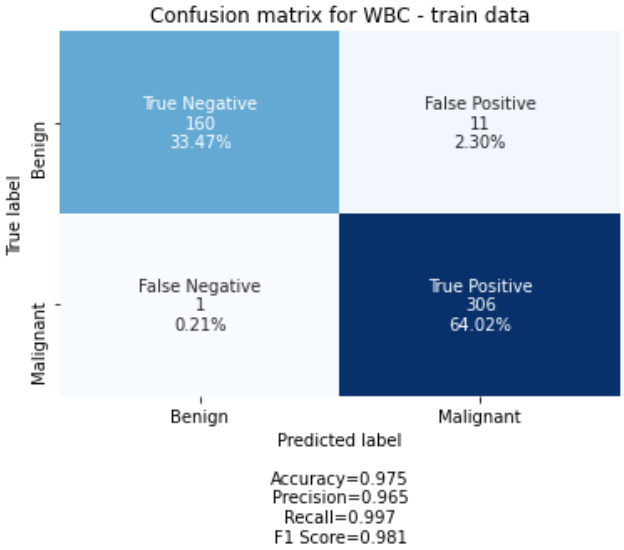

Fig. 6. Confusion matrix for train dataset - Case 1

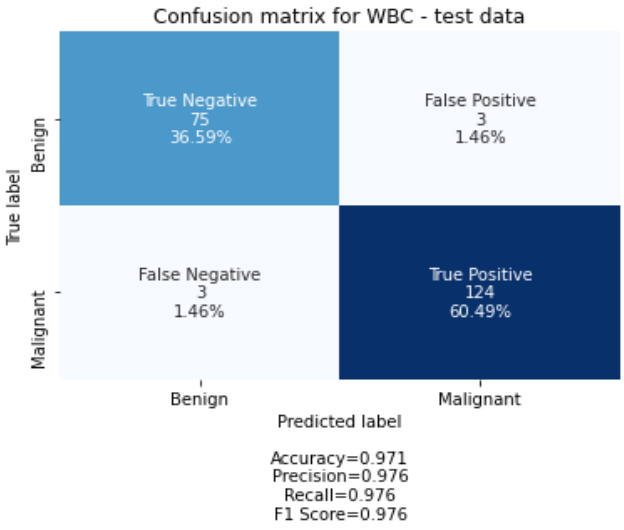

Fig. 7. Confusion matrix for test dataset - Case 1

Comparative analysis of different optimizers with different activation functions is shown in Table 2. DNN obtained the

greatest accuracy of 97.1% with 97.6% precision and 97.6% recall for Adam algorithm with logistic activation function in 1.47 seconds with 100-100 neurons at two hidden layers.

Table 2. DNN without ICA and 10-fold CV - Case 1

Optimization algorithm	Activation function	Accuracy (%)	Time (seconds)	Iteration
Stochastic Gradient Descent (SGD)	Logistic	95.61	0.87	197
	Tanh	96.07	42.65	192
	ReLU	96.59	35.89	191
Adam	Logistic	97.10	1.47	154
	Tanh	96.09	1.309	258
	ReLU	97.07	2.84	47
L-BFGS	Logistic	97.07	54.02	177
	Tanh	95.61	2.26	60
	ReLU	95.61	2.89	122

Precision-Recall curve and ROC curve for a simple model are shown in Fig. 8 and Fig. 9, which are used to assess the efficiency and ability of the model and to determine if it is great, good, or awful in working with test data. Due to the fact that it is not closer to the top left corner of the ROC curve, this curve is not the best for classification. The attainment of a desirable outcome of network prediction is represented by the precision-recall curves.

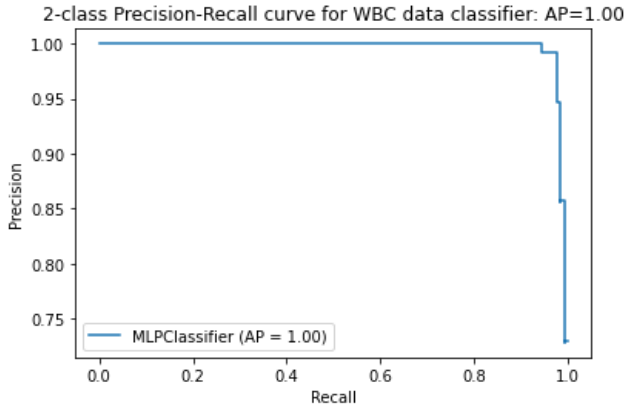

Fig. 8. 2-class Precision-Recall Curve - Case 1

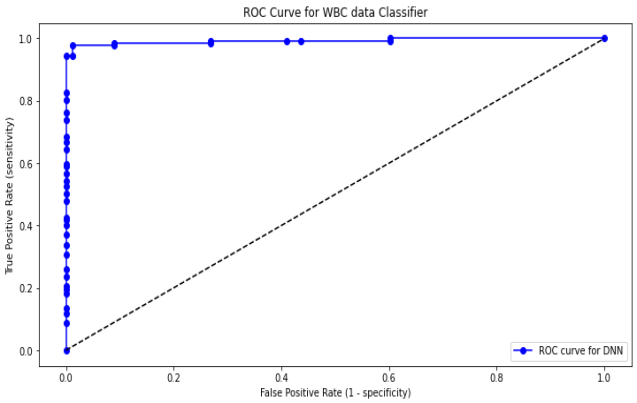

Fig. 9. ROC Curve-Case 1

4.2.2. Case 2: Results for R-DNN with ICA and 10-fold CV (Proposed R-DNN model):

In this scenario, data is separated into 10 folds for the purposes of training and testing the model. Each fold is treated as a training and testing set. The Fig. 10, Fig. 11 and Fig. 12 demonstrates the varied learning rates for the training error. After training, the version of the network that performed the best on the validation set was used.

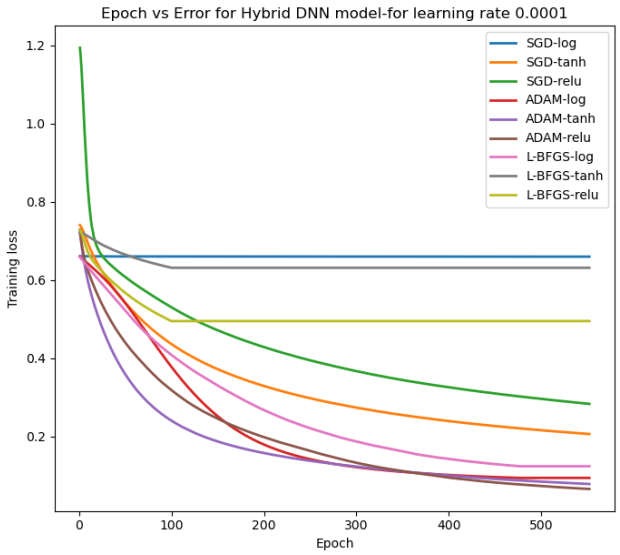

Fig. 10. Epoch vs Error for Proposed R-DNN model-for learning rate 0.0001

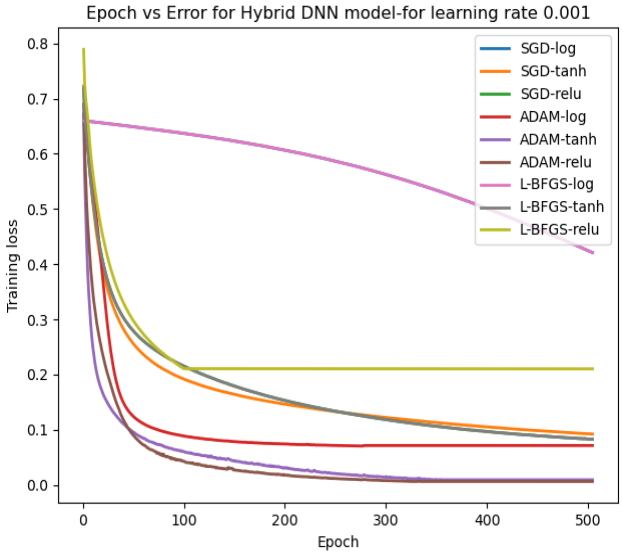

Fig. 11. Epoch vs Error for Proposed R-DNN model-for learning rate 0.001

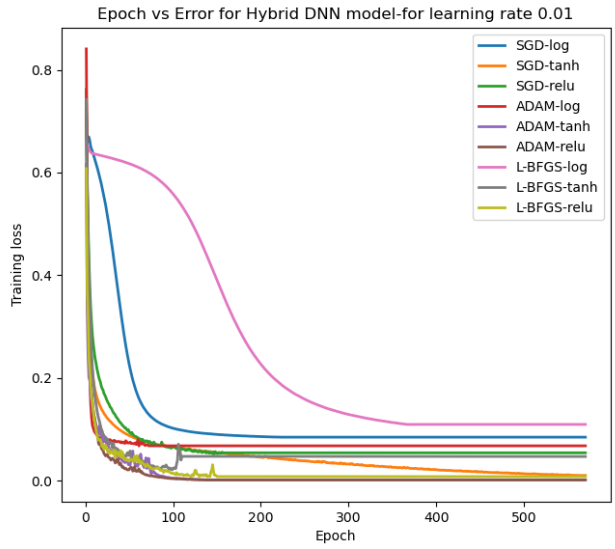

Fig. 12. Epoch vs Error for Proposed R-DNN model-for learning rate 0.01

This is the case in which the 10-fold CV and ICA are applied on WBC data. Table 3 compares the performance of various optimization algorithms with varying activation functions. With Logistic, Tanh and ReLU. L-BFGS optimizers fluctuate a lot and also didn't reach to optimal solution and leads to divergence. Such that L-BFGS mislead to the solution. As it can be observed in Fig. 10, Fig. 11 and Fig. 12. With Logistic, Tanh, and ReLU activation functions, SGD and Adam optimizer provide 100% accurate predication. Adam optimizer with ReLU activation function provides the maximum accuracy for 100-100 neurons at each hidden layers, 0.001 learning rate with reduced training time, and 163 numbers of iterations, which is superior to the accuracy produced by other optimizers. Out of all other trials with varied hidden neurons, regularisation parameter, and learning rate, we received the best results when the regularisation value was set to 0.0001 and the learning rate was set to 0.001.

Table 3. R-DNN with ICA and with 10-fold CV - Case 2

Optimization algorithm	Activation function	Accuracy (%)	Time (seconds)	Iteration
Stochastic Gradient Descent (SGD)	**Logistic**	100.00	4.17	1449
	Tanh	100.00	3.38	505
	ReLU	100.00	9.31	605
Adam	**Logistic**	100.00	2.03	149
	Tanh	100.00	2.4	136
	ReLU	100.00	1.81	163
L-BFGS	**Logistic**	98.51	1.39	120
	Tanh	98.51	0.98	78
	ReLU	97.06	1.79	121

Confusion matrix for WBC train-test data is shown in Fig. 13 and Fig. 14. The first diagonal block in the Fig. 13 represents the trained network's correct benign and malignant diagnosis rate. The first two diagonal blocks indicate that out of 699 samples, 35.40%, or 218 samples, are accurately classified as benign, while 61.30%, or 377 samples, are correctly classified as malignant. 1.95%, or 12 instances, receive an incorrect diagnosis of malignancy, while 1.30%, or 8 samples, receive an incorrect diagnosis of benign. Similarly, the confusion matrix for the WBC test dataset is illustrated in the Fig. 14. 55 samples, or 80.88% of 699 total cases, are accurately classified as benign, while 13 samples are appropriately classified as malignant (4).

The remainder of the block represents the number of benign and malignant tumours misdiagnosed. No specimens are misdiagnosed in benign or malignant. Hence, we claim that the proposed model worked exceptionally well, achieving a perfect accuracy rate of 100%.

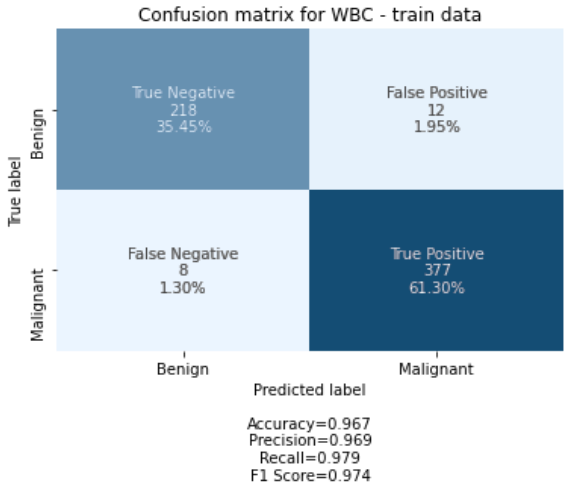

Fig. 13. Confusion matrix for train dataset - Case 2

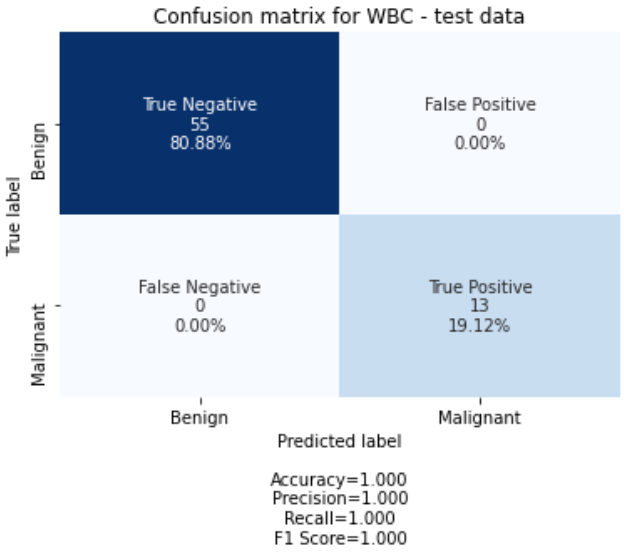

Fig. 14. Confusion matrix for test dataset - Case 2

Overall, 100% of predictions were properly classified, and 100% of precision, recall, and F-score were attained for the diagnosis of a cancerous tumour. The Precision-Recall curve for a 2-class system is depicted in Fig. 15.

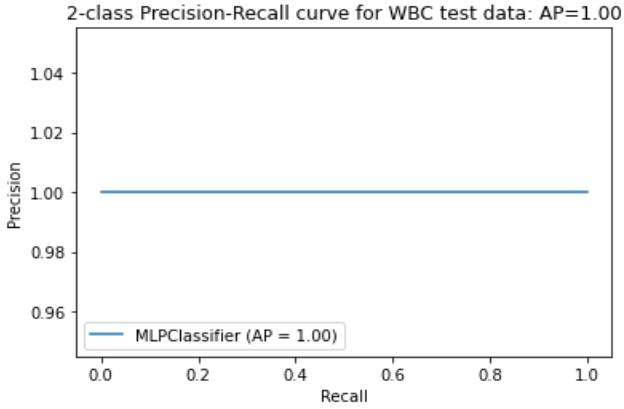

Fig. 15. 2-class Precision-Recall Curve - Case 2

Fig. 16 depicts a Receiver Operating Characteristic curve (ROC curve), which indicates that the model can accurately and perfectly diagnose WBC data. In this case, AUC=1 indicates that all test data results in the proper classification of benign and malignant tumours for the provided model.

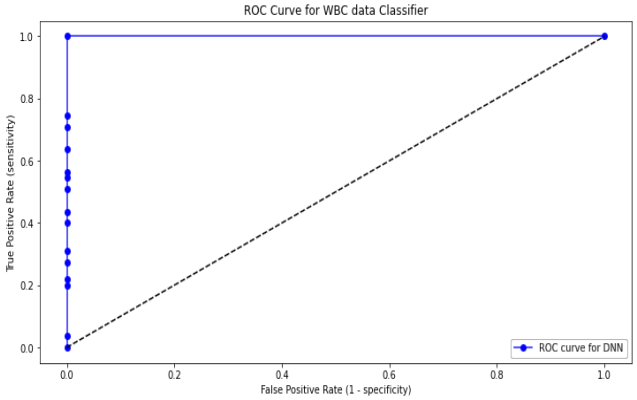

Fig. 16. ROC Curve - Case 2

A comparison of categorization approaches with the work of other authors is shown in Table 3. In the illustration, you can see a graphical depiction in Fig. 17 of Table 4.

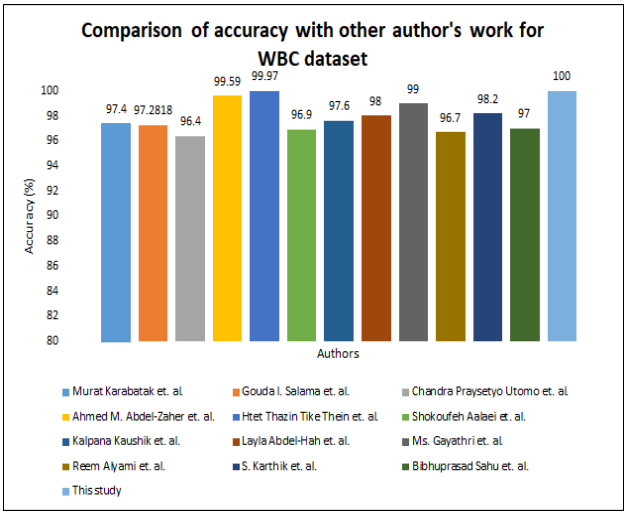

Fig. 17. Performance analysis of the proposed R-DNN+ICA model with others authors

Table 4. Performance analysis of the proposed R-DNN+ICA model

Authors	Year	Methods	Accuracy (%)
Murat Karabatak et. al.	2009	ANN + AR1(Association Rules)	97.40
Gouda I. Salama et. al.	2012	Fusion of SMO +IBK + NB + J48	97.28
Chandra Praysetyo Utomo et. al.	2014	ELM+ANN	96.40
Ahmed M. Abdel-Zaher et. al.	2015	DBN-NN(Conjugate gradient BP)	99.59
		RIW-BPNN(Conjugate gradient BP)	98.86
		DBN-NN (Levenberg Marquardt)	99.68
		RIW-BPNN (Levenberg Marquardt)	99.03
Htet Thazin Tike Thein et. al.	2015	ANN	99.97
Shokoufeh Aalaei et. al.	2016	ANN	96.70
		PS-Classifier	96.90
		GA-Classifier	96.60
Kalpana Kaushik et. al.	2016	ANN	97.60
Layla Abdel-Hah et. al.	2017	ANN	98.0
Ms. Gayathri et. al.	2017	Cascade-forward BP	99.00
Reem Alyami et. al.	2017	ANN	96.70
		SVM	97.14
S. Karthik et. al.	2018	DNN+RFF	98.20
Bibhuprasad Sahu et. al.	2019	ANN	97.00
This Study	**2021**	**R-DNN+ICA**	**100.00**

5. Conclusion:

In this study, we investigated and observed the influence of ICA and 10-fold CV for the diagnosis of breast cancer. We compared the results to a simple model with proposed hybrid R-DNN model. Using the WBC dataset, we discovered that R-DNN with ICA and 10-fold CV provide the best accuracy rate, as well as 100% precision and recall. By introducing additional term L2 regularization, the error is controlled so that the coefficients do not take on extreme values when the function is excessively fluctuating. By modifying the weights of the penalty term, the hyper parameter $\lambda = 0.001$ adjusts the trade-off between how well the data fits and how complex the model is. If λ is increased, the model complexity will contribute more to the cost. Because the minimal cost hypothesis has been chosen. This means that a higher λ will favour the model with the lowest complexity. Dimensions were decreased to only three features without sacrificing generality and classification accuracy attained to 100%. In proposed model, the sensitivity and specificity are 100%, resulting in the ROC curve of 1, indicating that there is no erroneous prediction in either the benign or malignant cases studied.

Acknowledgement

The authors express their gratitude to the university administration for providing financial assistance through the "DST-PURSE Program Phase-II".

References

[1] R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, 2019, CA. Cancer J. Clin., vol. 69, pp. 7-34, 2019, 10.3322/caac.21551.

[2] S. Aalaei, H. Shahraki, A. Rowhanimanesh and S. Eslami, "Feature selection using genetic algorithm for breast cancer diagnosis: Experiment on three different datasets", Iran. J. Basic Med. Sci., vol. 19, pp. 476–482, March 2016, 10.22038/ijbms.2016.6931.

[3] A.M. Abdel-Zaher and A.M. Eldeib, B reast cancer classification using deep belief networks, Expert Syst. Appl., vol. 46, pp. 139–144, 2015. https://doi.org/10.1016/j.eswa.2015.10.015.

[4] M. Karabatak, M.C. Ince, An expert system for detection of breast cancer based on association rules and neural network, Expert Syst. Appl., vol. 36, pp. 3465–3469, 2009, 10.1016/j.eswa.2008.02.064.

[5] L. Abdel-Ilah, H. Šahinbegovic, Using machine learning tool in classification of breast cancer, IFMBE Proceedings, vol. 62, pp. 3-8, 2017. 10.1007/978-981-10-4166-2.

[6] H. Tike Thein and K.M. Mo Tun, An Approach for Breast Cancer Diagnosis Classification Using Neural Network, Adv. Comput. An Int. J., vol. 6, pp. 1–11, 2015, 10.5121/acij.2015.6101.

[7] K. Kaushik and A. Arora, Breast Cancer diagnosis using Artificial Neural Network, International Journal of Latest Trends in Engineering and Technology (IJLTET), vol. 7, pp. 41-48, 2016, 10.21172/1.72.507.

[8] C. Prasetyo Utomo, A. Kardiana and R. Yuliwulandari, Breast Cancer Diagnosis using Artificial Neural Networks with Extreme Learning Techniques, Int. J. Adv. Res. Artif. Intell., vol. 3, pp. 10–14, 2014, 10.14569/ijarai.2014.030703.

[9] S. Karthik, R. Srinivasa Perumal and P.V.S.S.R. Chandra Mouli, Breast cancer classification using deep neural networks, Knowl. Comput. Its Appl., vol. 1, pp. 227–241, 2018, 10.1007/978-981-10-6680-1_12.

[10] Ms. M. Gayathri and Mrs. A. Shahin, Performance Evaluation Using Supervised Learning Algorithms for Breast Cancer Diagnosis, Int. Res. J. Eng. Technol., vol. 4, pp. 1339-1345, 2017, https://irjet.net/archives/V4/i6/IRJET-V4I6247.pdf.

[11] B. Sahu, S. Mohanty and S. Rout, A Hybrid Approach for Breast Cancer Classification and Diagnosis, ICST Trans. Scalable Inf. Syst., vol. 6, 2019.

[12] A.F.M. Agarap, On breast cancer detection: An application of machine learning algorithms on the Wisconsin diagnostic dataset, Proc. 2nd Int. Conf. Mach. Learn. Soft Comput. 2018., pp. 5-9, 2018, 10.1145/3184066.3184080.

[13] G.I. Salama, M.B. Abdelhalim and M.A. Zeid, Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers, Int. J. Comput. Inf. Technol., vol. 1, pp. 36-42, 2012, www.ijcit.com.

[14] D.A. Omondiagbe, S. Veeramani, A.S. Sidhu, Machine Learning Classification Techniques for Breast Cancer Diagnosis, IOP Conf. Ser. Mater. Sci. Eng., vol. 495, pp. 1-6, 2019, 10.1088/1757-

899X/495/1/012033.

[15] R. Alyami, J. Alhajjaj, B. Alnajrani, I. Elaalami, A. Alqahtani, N. Aldhafferi, T.O. Owolabi and S.O. Olatunji, Investigating the effect of correlation based feature selection on breast cancer diagnosis using artificial neural network and support vector machines, 2017 Int. Conf. Informatics, Heal. Technol. ICIHT 2017, pp. 1-7, 2017. 10.1109/ICIHT.2017.7899011.

[16] I. Maglogiannis, E. Zafiropoulos and I. Anagnostopoulos, An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers, Appl. Intell., vol. 30, pp. 24-36, 2009, 10.1007/s10489-007-0073-z.

[17] A. Mert, N. Kiliç, E. Bilgili and A. Akan, Breast cancer detection with reduced feature set, Comput. Math. Methods Med. 2015, pp. 1-11, 2015, 10.1155/2015/265138.

[18] M.A. Gokhan Zorluoglu, Diagnosis of Breast Cancer Using Ensemble of Data Mining Classification Methods, Int. J. Bioinforma. Biomed. Eng., vol. 1, pp. 318-322, 2015, 10.1007/978-981-10-6680-1_12.

[19] A. Hyvärinen and E. Oja, Independent Component Analysis Algorithms and Applications, Neural Networks, vol. 13, issue 4-5, pp. 411-430, 2000.

[20] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT press, 2016. http://www.deeplearningbook.org.

[21] D.P. Kingma and J.L. Ba, Adam: A method for stochastic optimization, 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., pp 1-15, 2015.

[22] R. Fletcher, Practical Methods of Optimization, A Wiley-Interscience publication, New-York, 1999.

[23] C.F. Beckmann, S.M. Smith, Probabilistic Independent Component Analysis for Functional Magnetic Resonance Imaging, IEEE Trans. Med. Imaging., vol. 23, pp. 137-152, 2004, 10.1109/TMI.2003.822821.

[24] D. Dua, Graff Casey, UCI Repository of Machine learning databases, (2017). http://archive.ics.uci.edu/ml.

[25] W. H. Wolberg, O. L. Mangasarian, Multisurface method of pattern separation for medical diagnosis applied to breast cytology, Proceedings of the National Academy of Sciences, vol. 87, pp. 9193-9196, 1990.

[26] J. Zang, Selecting typical instances in instance-based learning, Proceedings of the Ninth International Machine Learning Conference, pp. 470-479, 1992.