Breast Cancer Data Classification using Deep Neural Network

Vipul Sharma[1], Saumendra Kumar Mohapatra[2], and Mihir Narayan Mohanty[3*]

[1,2]Department of Computer Science and Engineering

[3]Department of Electronics and Communication Engineering

ITER, Siksha 'O' Anusandhan(Deemed to be University)

Bhubaneswar, Odisha, India

[1]vipulsharma936@gmail.com,

[2]saumendramohapatra@soa.ac.in,

[3]researchmihir16@gmail.com

Abstract: Artificial neural networks and their variants play an important role in the analysis and classification of different biomedical data. Deep learning is an advanced machine learning approach which has been used in many applications since last few years. Worldwide breast cancer is a major disease for woman and still it is one of the challenging job to detect it at an early stage. The authors in this work have taken an attempt to classify the breast cancer data collected from the UCI machine learning repository. Malignant and Benign two different types of breast cancer tumours are classified using deep neural network (DNN). Before classification two preprocessing steps are done for improving the accuracy. The correlation and one-hot encoding of the dataset was done for getting some relevant features that can be used as the input to the DNN. Around 94% of classification accuracy is achieved by using a six-layer DNN classifier. The result is also compared with some earlier works and it is found that the proposed classifier is providing better results as compare to others.

Keywords: ANN; Deep learning; DNN; Breast cancer; Classification.

1. Introduction

Cancer is a disease when some cells inside the body grow abnormally than healthy cells. It may occur anywhere in the human body. Breast cancer is seen to be an issue nowadays amongst people. It is due to cancer that is developed from breast tissue. Worldwide, it is the second most cancer disease among the woman. Most of the people are affected by this harmful disease globally. It starts at the time of at a time when the breast tissues grow out of control. Generally, in breast cancer, a tumour is formed which is detected through x-ray. Normally tumour is the collection of tissues which are by gathering the abnormal cells. From a survey, it is observed that in the USA, breast cancer is one of the common cancer diagnosed in a woman after skin cancer. It happens when some breast cells grow abnormally. From research, it is estimated that around 5 to 10% of breast cancers are related to gene mutations and having family history (Downs-Holmes & Silverman, 2011; Kelsey, Gammon, & John, 1993; Saritas & Yasar, 2019). The factors that increase the risk factor in breast cancer cases are hormonal imbalance, lifestyle, diet, biological factors, and some environmental aspects (Hulka & Stark, 1995). Research-based on breast cancer is one of the most important areas nowadays. Some effectual techniques were also introduced by the researchers for accurate and early diagnosis of this harmful disease.

As the progress of information technology in the healthcare domain is growing day-by-day, people's expectation is also gradually increasing for better treatment with minimum expenses. Computer-aided automatic disease diagnosis system has the ability to detect the disease at the initial stage before it starts to affect more.

Machine learning-based techniques are capable of detecting and classifying different diseases by strongly analyzing pathological data. This technique can provide better assistance to the medicals sectors for advance treatment. From the literature, it can be observed that numerous algorithms were applied for breast cancer data analysis and classification (Gayathri, Sumathi, & Santhanam, 2013). Still, it is one of the challenging jobs to design a machine learning model for this purpose with minimum error and computational time.

Designing a machine learning classifier with more accuracy for breast cancer classification is one of the most challenging research areas over the past decades. Traditional machine learning algorithms are limited in their capability to process huge amounts of data in their raw form. Deep learning is one of the recent and advanced machine learning-based models that have the capability to extract the features from the raw input data and classify them into the desired class (LeCun, Bengio, & Hinton, 2015). In this work the machine learning approach is considered as deep learning model. The objective is classification of tumors for breast cancer. The Malignant and Benign tumor data is collected from UCI repository and utilized for classification and cancer detection. Data is collected in the form of attributes and observed to decide the preprocessing. Further preprocessing is done by normalization and encoding to avoid misclassification. The tumors are Malignant and Benign. The model is of DNN with four hidden layers. It is found from observed result is better than earlier works. However, the data may be changed and compared i.e. not included here.

The rest of the paper is organized as follows. Related literature is described in section two. The description of the proposed classification model is presented in section three. The obtained result is discussed in section four. Finally, section five concludes the work.

2. State-of-Art

Many researchers have applied different machine learning techniques for detecting and classifying various types of cancer in the human body. From the literature, it can be found that so many works have been done for breast cancer classification and detection. Artificial neural networks (ANN) and its variants are the most used and popular machine learning-based classifiers for breast cancer data analysis (Saritas & Yasar, 2019). Authors in (Kiyan & Yildirim, 2004) have used four different types of neural networks for cancer data classification. Multi-layer perceptron (MLP), Probabilistic neural network (PNN), Radial basis function network (RBFN), and General Regression Neural Network (GRNN) were used for classifying the data collected from Wisconsin Breast Cancer Dataset (WBCD) database. Among these four types of classifier PNN was providing better results as compared to the other three types of the neural network-based classifier. An ensemble neural network for diagnosis of breast cancer was designed by the authors (Yao & Liu, 1999). They have simultaneously trained more than one simple feed-forward neural network with a negative correlation learning approach. The reason behind using a negative correlation learning approach was to simultaneously train each neural network. The performance of the backpropagation neural network for was enhanced by applying a genetic algorithm (GA) by the authors in (Adam & Omar, 2006). In their work, they have used GA for optimizing the initial weights of the neural network. They have divided the original data set into two different subsets and around 98% accuracy was achieved with the optimized neural network. Authors in (Gharibdousti, Haider, Ouedraogo, & Susan, 2019) have classified breast cancer by using five types of machine learning classifiers. Data used in their work was collected from UCI machine learning repository and processed by applying different data mining techniques. Support vector machine (SVM), Decision tree (DT), Naive Bayes, Neural network, and logistic regression classifiers were used by them for classification. From their obtained result it was observed that SVM was providing better result as compare to other four types of classifiers. Principal component analysis (PCA) is one of the most used data analysis algorithm for reducing the dimension of the dataset. PCA was used for reducing dimension of the cancer data collected from UCI repository in (Sahu, Mohanty, & Rout, 2019). The reduced data was then classified by using ANN, random forest and K-nearest neighbour classifier. ANN with PCA was providing better result as compare to other two types of classifier. Data mining plays an important role in machine learning based classification and detection of different disease. Three data mining based classifiers were taken by researchers for classifying breast cancer in (Aro, Akande, Jibrin, & Jauro, 2019). Support vector machine was providing better result as compare to other classifier for breast cancer data classification. In their work they have also used the bagging and boosting concept for improving the accuracy and computational time of the classifier. A feature ranking based cancer data classification algorithm was design by the authors in (Alam, Rahman, & Rahman, 2019). The ranking algorithms were applied for getting some important features from the original data set. The training and testing data were separated by applying the 10-fold cross validation and random forest classifier was used for the classification purpose. By considering breast tumor pathological images, authors have classified Benign and Malignant tumours using support vector machine classifier in (Chang, Wu, Moon, Chou, & Chen, 2003). They have extracted autocorrelation and auto covariance features from the original image. These features were then used as the input to the SVM model. For comparison purpose they have also used MLP classifier and from the result it can be found that the performance of the SVM is quite better than MLP. Classification was also done for two types of images such as bright and dark images. Classification performance of ANN was compared with Naive bays classifier for breast cancer data classification in (Saritas & Yasar, 2019).

Deep learning is one of the advance neural network based classifier and has been used in many machine learning problems. Researchers have also considered deep learning for breast cancer data analysis and classification. Microscopic image plays an important role in breast cancer diagnosis. A convolutional neural network (CNN) based classifier was utilised for classifying histopathology images in (Bayramoglu, Kannala, & Heikkilä, 2016). Two different types of CNN classifier, single task and multi task was used for predicting malignancy and image magnification level of the breast cancer. A back propagation feed forward deep neural network was designed for detecting breast cancer in (Abdel-Zaher & Eldeib, 2016). Their proposed design was validated by using WBC Dataset. The obtained classification accuracy was again compared with some earlier work. For controlling the classification error, an enhanced loss function (ELF) was applied in deep learning based classifier. Before classifying the images they have applied k-means clustering for separating a particular regions of the images. They have achieved a better classification accuracy with 0.0001 learning rate. For boosting the accuracy they have applied the loss function of the SVM classifier. In order to overcome the limitations of some traditional breast cancer data set, a new dataset *BreakHis* was released that satisfies all the clinical requirement of breast cancer diagnosis. Authors in (Benhammou, Achchab, Herrera, & Tabik, 2020) have applied deep learning for classifying breast cancer by using this data. For predicting the breast cancer risk factor a four pair deep neural network (DNN) was designed in (Sun, Tseng, Zheng, & Qian, 2016). Before classification they have divided each image into 100 ROIs with 52*52 pixel size and each ROI was individually trained with DNN. A five by five kernel bank was convolved for every original ROI.

From the literature it can be found that numerous works have been done for breast cancer classification and detection. Different machine learning techniques were adopted for getting a satisfactory result. Some works were also used deep learning based models for classifying the breast cancer images. Here in our work we have used deep neural network (DNN) for classifying breast cancer electronic health report (EHR) data.

3. Proposed Method

Looking at the two types of tumour we can classify between Malignant and Benign tumours to know the tendency of tissue amount present in the person. Therefore, the main goal of the proposed work is these two types of tumour using Deep Neural Networks. The data collected from UCI repository is preprocessed for avoiding the misclassification. Before classification, the whole data set is normalized and one hot encoded (meaning 0 or 1) for Benign and Malignant respectively. The main objective of this work is to achieve maximum accuracy using Deep Neural Networks. The structure of the proposed work is presented in Figure 1.

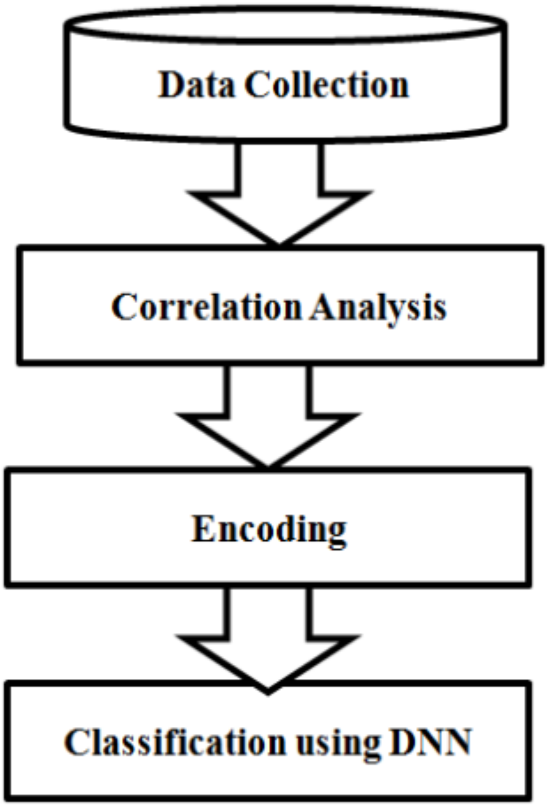

Figure 1: Structure of the proposed work

3.1 Breast Cancer Data

The data set used in this work is Wisconsin(Diagnostic) Data Set available at UCI machine learning repository (Abbass, 2002). Effective features are extracted from the breast mass image by fine needle aspirate (FNA). This features represents the cell nuclei properties of the image. Total 569 instances with 32 attributes per instance and there are no missing value in the data set. The major ten variables are described in Table 1.

Table 1: Description of the dataset

Variable	Description
Radius	Standard deviation of gray-scale values
Area	Area of the cancer tissues
Perimeter	Perimeter of the image
Compactness	Perimeter$_{2}$ / area - 1.0)
Concave Points	Number of concave portions of the contour
Fractal Dimension	Coastline approximation
Smoothness	Local variation in radius lengths
Concavity	Severity of concave portions of the contour
Symmetry	Symmetry of the image
Diagnosis	M = malignant, B = benig

The strength graph for the breast cancer dataset is plotted of the mean which is the strongest amongst texture, radius and perimeter. Thus, the texture had 70 percent data higher than 20 whereas perimeter had 70 percent data higher than 120 in each of their own domain. Figure 2 shows the strength of the dataset.

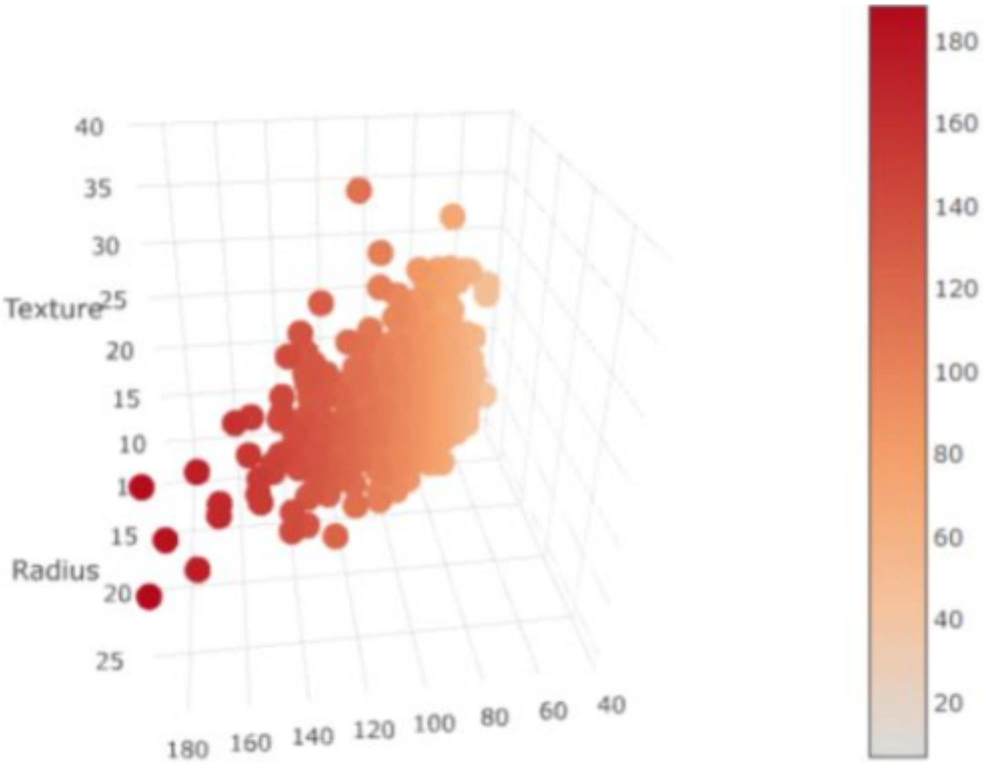

Figure 2: Strength graph.

3.2 Correlation Analysis

To check how each of the variables are correlated, summary of the data set is being used. Through the summary the mean factor is calculated for all the variables. This gives a strong correlation which can be grouped together later to increase the accuracy of the model. Also, it reduces the feature space substantially, making the model more generalizable. The correlation plot that follows this section uses clustering technique that make easy to observe that which variable is strongly correlated with another. The color of the line in Figure 3 displays the correlation trend, whereas the shaded line and thickness characterize the relationship strength.

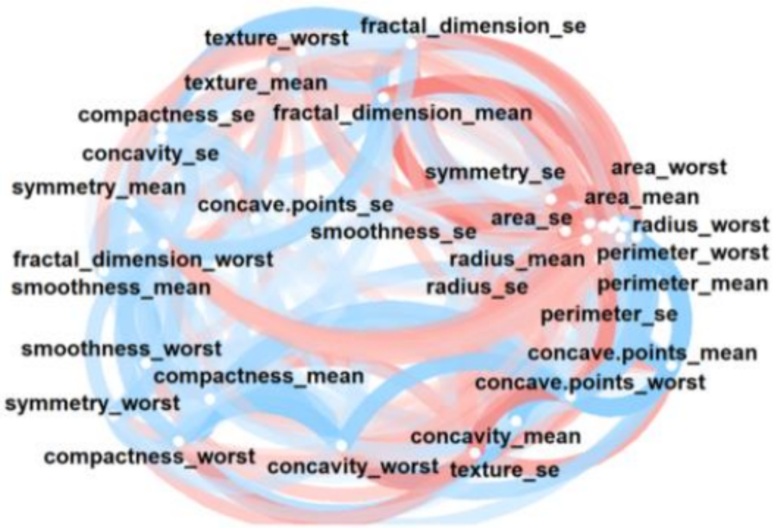

Figure 3: Correlation of the dataset.

3.3. One Hot Encoding

The target column of the data matrix is represented as string value. 62.74 percent of the label column comprises of 'B' (Benign) and the rest 37.26 percent is 'M' (Malignant). Input is given in terms of feature matrix and is encoded to target column separately. To obtain an accurate feature vector for the proposed DNN classifier, it is a major task to encode the target column separately. Therefore, to avoid the classification error, the target column of the dataset is encoded. The target class 'M' is represented as 1 and 'B' is represented as 0.

3.4. Data Normalization

The features of the dataset are normalized by scaling them which basically is zero mean and unit variance on each of these features. The main aim of normalization is to transform the values of numeric columns in the data set to use a common scale, without altering differences in the ranges of values or losing information.

3.5. Classification Model

A feed forward deep neural network model is considered for the classification purpose. The proposed neural network has, three four layers with 16, 8, 6, and 4 nodes each. The output layer will have 2 nodes which are the labels itself. The overall structure of the neural network is displayed in Figure 4.

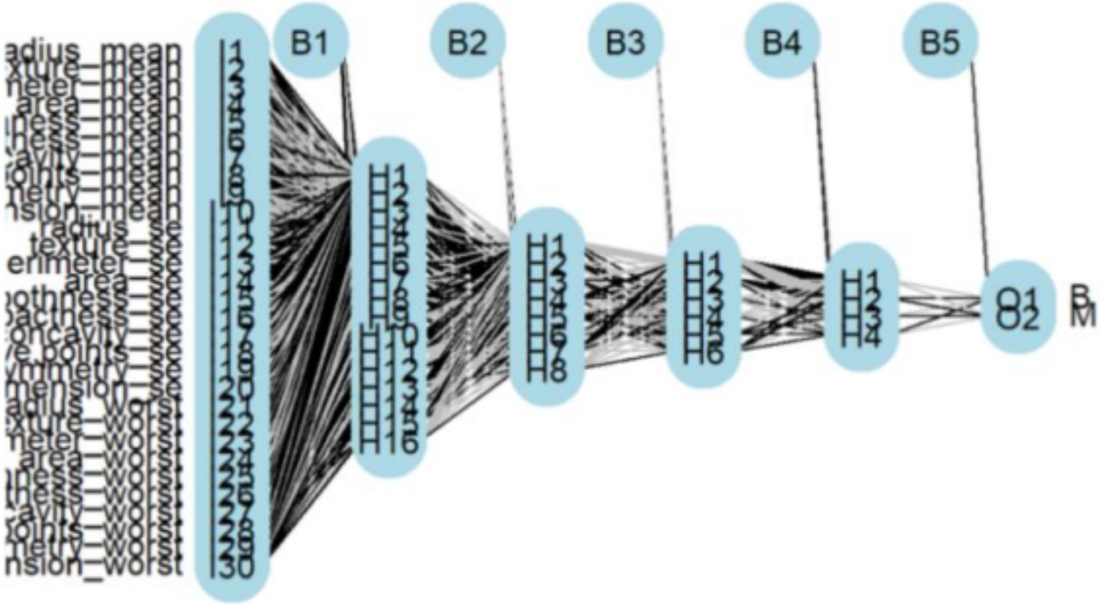

Figure 4: Proposed DNN structure for breast cancer data classification.

A sequential model is used with hidden layers having RELU activation function while the output activation function is sigmoid. For compiling the model, the optimizer algorithm used is RMSPROP. The loss is done by binary cross entropy method. Finally the validation metric used is accuracy. The layers here are dense layers each of which runs on the rectified linear unit (ReLU) activation and the final layer has the Sigmoid activation function for predicting the labels. The output of the hidden layer can be calculated as;

$$C_j^{i,a} = \sigma\left(d_a + \sum_{m=1}^{M} w_m^a x_{i+m-1}^{o_a} \right) \quad (1)$$

where x_i^0= (x_1, x_2, x_3,....., x_m) is the input vector and m is the total number of segments. j is the layer index and d is the bias of the feature map. σ is the activation function. M is the filter size. w_m^a is the weight for mth filter index. In the proposed model two types of filter is used. In the hidden layer RELU activation function is used. The main advantage of using this activation function is that it does not activate all the neurons at the same time and converts all the negative input into zero so that the neuron does not get activated. It can be represented by;

$$f(x) = \max(0, x) \quad (2)$$

where x is the input data and $f(x)$ is the output function that returns the maximum value between 0 and input data. Again Softmax activation function is used in the output layer. In softmax activation function the exponential (e-power) of the given input value and the sum of exponential values of all the values in the inputs is computed. Then the ratio of the exponential of the input value and the sum of exponential values is the output of the softmax function. The main advantage of using it, is the output probabilities range. The range varies between 0 and 1 and the sum of all the probabilities will be equal to one. If the softmax function used for multi-classification model it returns the probabilities of each class and the target class will have a high probability. Mathematically softmax activation function can be represented as:

$$s_j = \frac{e^{x_j}}{\sum_{i=1}^{n} e^{x_n}} \quad (3)$$

where the input is x and the output value of s is between 0 and 1 and their sum is equal to 1.

For training, at each step of gradient descent in the hidden layers, we do forward propagation with the current model parameters to get an output. At each layer of the network, we compute a matrix multiplication of the output from the previous layer and the weights connecting the previous layer to the current layer. After this the bias is added to improve the value at each step. Completion the summation, an activation function like Sigmoid is applied, which converts the weighted result into a value between 0 and 1, indicating the classification at that layer. Then back propagation algorithm is applied to adjust the parameters to minimize loss from that output. In Table 2 a complete representation of the parameters considered for the proposed DNN is presented.

Table 2: Parameters used in the proposed DNN classifier

Parameter	Specification
Hidden layer	Four
Hidden layer nodes	16,8,6,4
Model type	Sequential
Optimization	RMSPROP
Loss function	Binary cross entropy
Learning rate	0.001
Training algorithm	Back propagation

4. Results and Discussion

After collecting the data, it is separated into training and testing set. Separating data into training and testing sets is an important part of evaluating data mining models. 70 % of the original data is separated into training set and rest 30% data is used for validation purpose. After successfully completing the training, validation of the model is done for making the prediction against the test set data. The hyper parameters used for fitting the model were learning Rate as 0.001, epochs which were 200 iterations and a batch size of 1. Model is trained using backpropagation algorithm. It is observed that the loss decreases gradually with increasing of epochs. Similarly the training accuracy increases with the increasing number of epochs. Further the network is validated with the testing data. The corresponding loss and accuracy is shown in Figure 5 and 6. The confusion matrix obtained at the time of validation is presented in Table 3.

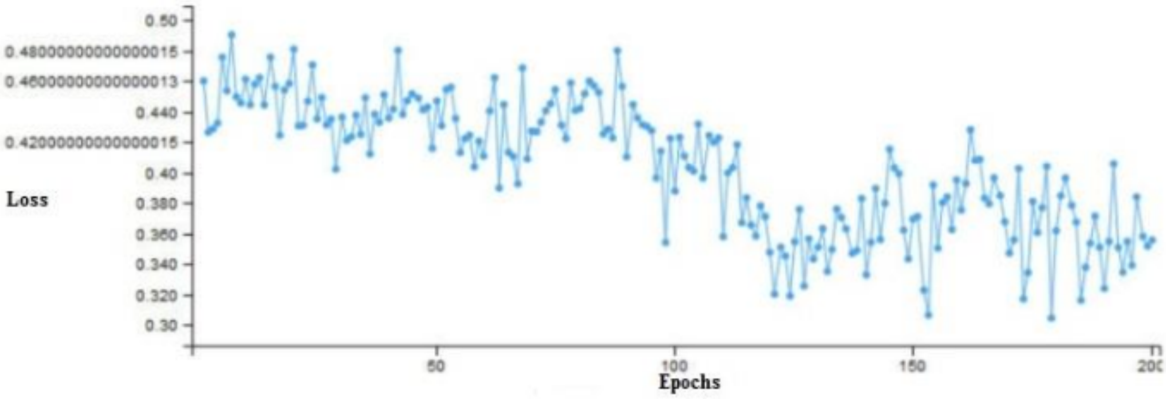

Figure 5: Error of the DNN model

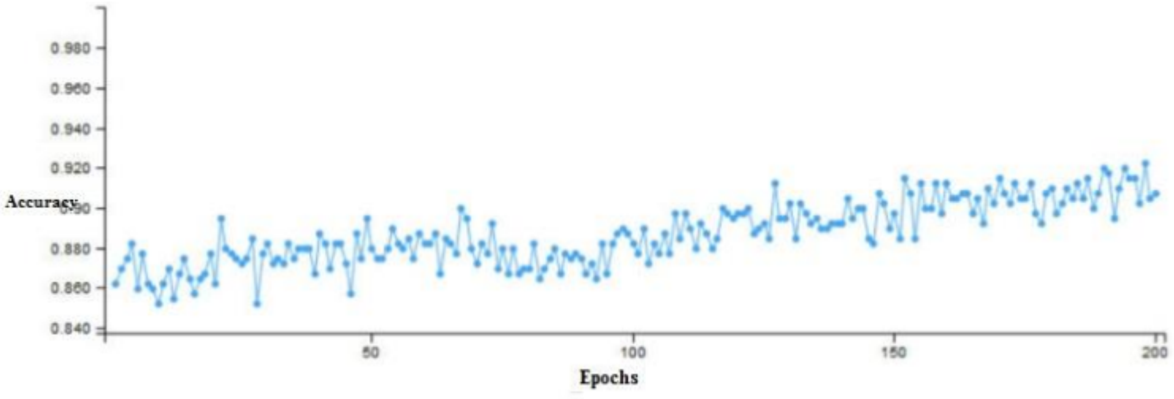

Figure 6: Accuracy of the DNN model for breast cancer classification

Table 3: Confusion matrix of the DNN classifier

	Benign	Malignant
Benign	99(TP)	8(FP)
Malignant	3(FN)	61(TN)

Standard measures of sensitivity, specificity, and accuracy are considered for calculating the performance of the classifier. Sensitivity is the number of correct positive predictions divided by the total number of positives (N). Specificity is another performance measuring parameter and can be calculated as the number of correct negative predictions (TN, FN) divided by the total number of negatives. Performance of the proposed method is presented in Table 4. The comparison of obtained result with different activation function in hidden layer is shown in Table 5.

$$Accuracy = \frac{TP + TN}{N} \tag{4}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

Table 5: Classification performance of the proposed DNN classifier

Measuring Parameter	Percentage(%)
Accuracy	93.56
Sensitivity	94.05
Specificity	88.40

Table 6: Classification accuracy with different activation functions

Activation Function	Accuracy
Tanh	92%
Softmax	93%
Sigmoid	92.1%
ReLu	**93.56%**

Table 2: Comparison of the proposed classifier with earlier works

Reference	**Classifier**	**Accuracy**
(Gharibdousti et al., 2019)	Decision Tree	91%
(Gharibdousti et al., 2019)	ANN	50%
(Sahu et al., 2019)	Naive Bayes	91%
(Sahu et al., 2019)	KNN	93%

(Saritas & Yasar, 2019)	ANN	83.54%
(Saritas & Yasar, 2019)	Naive Bayes	86.95%
Proposed work	**DNN**	**93.56%**

5. Conclusion

Nowadays breast cancer is one of the major cause of death for woman. Computer aided diagnosis system can help the physicians for the early diagnosis of this disease. Here in the proposed work DNN is used for the classification of breast cancer data. Two types of breast cancer are classified using the deep learning base classifier and around 94% accuracy is achieved. In future, the data may be changed and compared for obtaining more accurate classifier.

References

Abbass, H. A. (2002). An evolutionary artificial neural networks approach for breast cancer diagnosis. *Artificial intelligence in Medicine, 25*(3), 265-281.

Abdel-Zaher, A. M., & Eldeib, A. M. (2016). Breast cancer classification using deep belief networks. *Expert Systems with Applications, 46*, 139-144.

Adam, A., & Omar, K. (2006). *Computerized breast cancer diagnosis with Genetic Algorithm and Neural Network.* Paper presented at the Proc. of the 3rd International Conference on Artificial Intelligence and Engineering Technology (ICAIET).

Alam, M. Z., Rahman, M. S., & Rahman, M. S. (2019). A Random Forest based predictor for medical data classification using feature ranking. *Informatics in Medicine Unlocked, 15*, 100180.

Aro, T. O., Akande, H. B., Jibrin, M. B., & Jauro, U. A. (2019). Homogenous Ensembles on Data Mining Techniques for Breast Cancer Diagnosis. *Daffodil international university journal of science and technology, 14*(1).

Bayramoglu, N., Kannala, J., & Heikkilä, J. (2016). *Deep learning for magnification independent breast cancer histopathology image classification.* Paper presented at the 2016 23rd International conference on pattern recognition (ICPR).

Benhammou, Y., Achchab, B., Herrera, F., & Tabik, S. (2020). BreakHis based breast cancer automatic diagnosis using deep learning: Taxonomy, survey and insights. *Neurocomputing, 375*, 9-24.

Chang, R.-F., Wu, W.-J., Moon, W. K., Chou, Y.-H., & Chen, D.-R. (2003). Support vector machines for diagnosis of breast tumors on US images. *Academic radiology, 10*(2), 189-197.

Downs-Holmes, C., & Silverman, P. (2011). Breast cancer: overview & updates. *The Nurse Practitioner, 36*(12), 20-26.

Gayathri, B., Sumathi, C., & Santhanam, T. (2013). Breast cancer diagnosis using machine learning algorithms-a survey. *International Journal of Distributed and Parallel Systems, 4*(3), 105.

Gharibdousti, M. S., Haider, S. M., Ouedraogo, D., & Susan, L. (2019). Breast cancer diagnosis using feature extraction techniques with supervised and unsupervised classification algorithms. *Applied Medical Informatics., 41*(1), 40-52.

Hulka, B. S., & Stark, A. T. (1995). Breast cancer: cause and prevention. *The Lancet, 346*(8979), 883-887.

Kelsey, J. L., Gammon, M. D., & John, E. M. (1993). Reproductive factors and breast cancer. *Epidemiologic reviews, 15*(1), 36.

Kiyan, T., & Yildirim, T. (2004). Breast cancer diagnosis using statistical neural networks. *Istanbul University-Journal of Electrical & Electronics Engineering, 4*(2), 1149-1153.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature, 521*(7553), 436-444.

Sahu, B., Mohanty, S., & Rout, S. (2019). A hybrid approach for breast cancer classification and diagnosis. *EAI Endorsed Transactions on Scalable Information Systems, 6*(20).

Saritas, M. M., & Yasar, A. (2019). Performance analysis of ANN and Naive Bayes classification algorithm for data classification. *International Journal of Intelligent Systems and Applications in Engineering, 7*(2), 88-91.

Sun, W., Tseng, T.-L. B., Zheng, B., & Qian, W. (2016). *A preliminary study on breast cancer risk analysis using deep neural network.* Paper presented at the International Workshop on Breast Imaging.

Yao, X., & Liu, Y. (1999). *Neural networks for breast cancer diagnosis.* Paper presented at the Proceedings of the 1999 Congress on Evolutionary Computation-CEC99 (Cat. No. 99TH8406).