Breast Cancer Diagnosis Models Using PCA and Different Neural Network Architectures

Md. Mehedi Hasan[1], Md. Rakibul Haque[2], Mir Md. Jahangir Kabir[3]

Department of Computer Science and Engineering

Rajshahi University of Engineering and Technology

Rajshahi, Bangladesh

Email: mmehedihasann@gmail.com[1], rakibulhaq56@gmail.com[2], mmjahangir.kabir@gmail.com[3]

***Abstract*— One of the most wide-spreading diseases among women is Breast Cancer. For this reason, a proper diagnosis is necessary for designating necessary treatment. Using the previous information about patients, diagnosis is being performed by various machine learning algorithms. As the data are getting bigger, it is becoming more necessary to extract the useful information from the huge pile of information. In this paper, we have used the Wisconsin diagnostic breast cancer dataset (WDBC) and SEER 2017 Breast Cancer Dataset. Then we have used Principal component analysis in order to extract useful features. After that, we have classified the reduced datasets using multi-layer perceptron (MLP) and convolution neural network (CNN). Then we have provided a comparative comparison of our model for both the reduced datasets. Our MLP model has achieved an accuracy of 99.1% on reduced WDBC dataset and 89.3% on SEER 2017 Breast Cancer dataset whereas CNN Model has achieved 96.4% on reduced WDBC dataset and 88.3 % on SEER 2017 Breast Cancer Dataset.**

Keywords — Breast Cancer, PCA, MLP, CNN, WDBC, SEER Dataset 2017.

I. Introduction

Breast cancer is one of the most deleterious diseases found in women. In fact, breast cancer is the second most occurring cancer in women's body after lung cancer [1]. Breast cancer is caused by malignant tumor originating in the breasts [2]. In 187 countries, breast cancer victim has increased from 641,000 in 1980 to 1,643,000 in 2010[3]. Breast cancer has contributed to 15% of the total cancer death in the United Kingdom in the year 2015[4].

As the world is progressing, data is generating at a huge rate. We have to reduce the dimension of data to gain knowledge effectively. There are various feature selection and extraction algorithms for reducing the dimension of our data such as PCA [5], LDA [5], wavelet transform [6] ,etc. These algorithms have been effectively used in the field of medical and health science. Wang [7] has used PCA for reducing dimensionality in Wisconsin Breast Cancer Dataset.

Different kinds of traditional data mining classification algorithms such as SVM [8], Decision tree classifier e.g. CART [8], Naïve Bayes [8] have been successfully applied in predicting and diagnosis. Recently with the explosion of data in the biomedical field, deep learning models are having a significant impact on the analysis of different kinds of harmful diseases.

Multi-layer Perceptron is one of the most mentionable neural network architecture that has performed much better than traditional classification algorithms [9]. Another form of neural network architecture is the convolution neural network (CNN) which has been performing very well in classifying image from ImageNet [10] Pim and Hugo [11] have developed a deep convolutional neural network for the purpose of segmenting brain tissue and white matter hyperintensities of presumed vascular origin in MRI. Convolution neural network performs well for one dimension also as it can easily recognize the pattern in the data with numerous filter [12].

In this paper, we have first extracted features from both the dataset using principal component analysis. Then we have analyzed the dataset using our trained MLP and CNN model individually. We have implemented our models on Keras [13] framework which runs on top of Tensorflow [14].

The remainder of the paper is organized as follows. Section II describes the related work on the diagnosis of breast cancer using different kinds of methods. Section III reviews the Principal Component Analysis and the neural network architectures that we have used for the diagnosis of breast cancer. In section IV we have discussed the performance of our architecture in details. Finally, we have drawn some conclusion based on the performance of our architecture.

II. Related works

In order to find useful information among a huge pile of data, different feature extraction methods are having a significant impact on breast cancer risk prediction. Wang [7] has chosen 8 principal components to reduce the dimensionality in Wisconsin breast cancer dataset.

978-1-7281-3060-6/19/$31.00 ©2019 IEEE

Hiba and Asria [8] has provided a detailed comparison among SVM, Naïve Bayes, and Decision Tree on Wisconsin Breast Cancer Dataset. P.Dhivyapriya [15] has achieved an accuracy of 96% in predicting the malignant or benign tumor in the same dataset. Wang and Zheng [16] have used a support vector based ensemble learning algorithm where they hybridized 12 different support vector machine model in order to analyze the Wisconsin and SEER 2017 Breast Cancer dataset.

Recently, Multi-layer Perceptron is also being used to provide diagnosis of breast cancer. Abdelghan [17] have applied a multi-layer backpropagation neural network in order to analyze the SEER 2002 Breast Cancer dataset in which he has achieved 86% accuracy. Wang and Sang [7] has also used a multi-layer perceptron in order to analyze the Wisconsin dataset.

Our work is inspired by [8] and [16] where we have used multi-layer perceptron and convolutional neural network in order to provide a diagnosis for breast cancer dataset. In our MLP network on WDBC dataset, we have achieved an accuracy of 99.1% which is higher compared to traditional machine learning algorithms on the same dataset [8].

III. Methodology

A. *Principal Component Analysis(PCA)*

Too much dimensionality of data sometimes makes the predicting model more complex and less accurate. Because, as the volume of the feature space increases the available data becomes sparse and sparsity is a problem for any kind of model that requires statistical significance PCA provides a quite significant solution to this problem. For any dataset or input matrix that is represented by an n-dimensional vector or attribute PCA searches for k n-dimensional orthogonal vectors that can be used to represent the entire input matrix where $k \leq n$. The basic steps for finding k n-dimensional orthogonal are described as follows [19].

Step 1: Normalize the input data matrix so that all the features fall into the same range.

Step 2: Compute K-orthonormal vectors that provide a basis for the input data. The orthonormal vectors are unit vector and perpendicular to each other and known as a principal component. The input matrix is a linear combination of principal components.

Step 3: Sort the principal component in decreasing order of their “significance”. Here significance refers to the variance of the dataset covered by the components.

Step 4: Eliminate the weaker components with low significance value, that is, those with lower variance.

After following these steps we have been able to get the principal components that reduced the dimensionality of both the datasets used for the purpose of providing a diagnosis of breast cancer.

B. *Multi-layer Perceptron*

One of the most famous and widely used neural network for prediction is multi-layer perceptron (MLP) with backpropagation [20-22]. Neural Network tries to mimic brain unit known as neuron at lower level hence the name. Brains neuron works by combining input signals from multiple dendrites and if the signal crosses some threshold value, the neuron fires. This same Principal is used in multi-layer perceptron. Generally, multi-layer perceptron has 3 layers in total namely: input layer (number of nodes depending on the feature of the data on which MLP is being trained on), hidden layer combining of multiple inputs multiplied with weights and output layer consisting of required number of nodes to accommodate the number of output classes [22]. Brains neurons junction point where inputs from other neurons come is known as synapses which may be great at accumulating signifying it has a higher weight or may not be effective that much signifying it has a lower weight.

C. *Convolutional Neural Network (CNN)*

Convolutional neural network is one of the emerging neural network architecture in the present era. Convolution neural network has performed with high precision in pattern recognition [10]. For this particular reason, we have used CNN in our analysis of finding breast cancer so that it can find the pattern in our datasets in order to pre-calculate the risk and provide a diagnosis. Convolution neural network has three significant layers [23]. First one is the convolution layer in which the input feature matrix is convoluted with various filters. Filter or kernel is used for detecting specific pattern or correlation in the input features. The number of filters used in the layer depends on the data matrix we are working with. The second one is pooling layer in which the goal is to reduce the size of spatial representation in order to reduce the parameter and computation of the network. pooling layer operates on feature maps which is produced as the output of the convolution layer. Then there is a fully connected layer at the end of CNN which is connected with all the activation function of its previous layer.

We have used a deep convolution neural network for our CNN model which consists of a total of five layers. We have used two convolution layers and two pooling layers. The fully connected layer at the end had been connected to an output layer with one neuron with sigmoid activation function. The weights of the network were first initialized with Gaussian distribution. After each epoch, the loss was measured using binary cross entropy [24] which was then minimized using Adam optimizer [25] by backpropagation of the weights through the network and thus adjusting the weights. We have also used the dropout method [26] in order to prevent overfitting.

IV. Results

A. *Dataset Description*

We have used two publicly available datasets in order to evaluate our breast cancer diagnosis models. The first dataset is Wisconsin Diagnostic Breast Cancer Dataset (WDBC) [27] and the second one is SEER 2017 Breast Cancer Dataset [28].

- *Wisconsin Breast Cancer Dataset*

This dataset is mainly collected from the University of Wisconsin Hospital in 1995. This dataset Contains 569 samples and 32 patient attributes which includes patient ID, 30 attributes about tumor diagnosis and One diagnosis result saying if the tumor is benign or malignant [27].

- *SEER Breast Cancer Dataset (2017)*

The dataset has 4024 rows and 14 columns. The features that were included in the prediction of patient survivability includes Patient's Age, Race, Marital Status, T STAGE, N STAGE, 6TH STAGE, GRADE, A STAGE, Tumor Size, Estrogen Status, Regional Node Examined, Regional Node-Positive, Survival month[28].

B. *Design of experiment*

We have performed an extensive preprocessing on both the datasets. In the case of WDBC dataset, we have performed standardization on the feature space so that the feature or attribute with larger value does not have a huge impact on our diagnosis models. For the case of SEER breast cancer dataset, as it contained both numerical and categorical variables, we have converted the categorical variable into numerical variables by adding dummy variables. After performing encoding we have got a 21-dimensional feature space for SEER dataset.

After that, we performed Principal component analysis on the feature space of both datasets. For WDBC dataset we have chosen 20 Principal components as it covers 99.32% of the whole dataset. And for the SEER Breast Cancer dataset, 15 Principal components were chosen from 21 Principal components as 15 Principal components have covered 99.6% variance of the dataset.

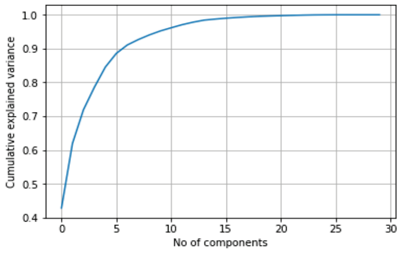

Fig. I. Cumulative explained variance for WDBC.

After extracting features from both the dataset, we have individually divided both the reduced dataset into an 80:20 training ratio. 20 percentage of the training dataset was used as a validation data for both datasets. We have chosen this splitting ratio because the size of our dataset sample was not very large. In the case of WDBC, the number of the epoch is 50 and for CNN, it is 10.Epochs has been chosen based on the training and validation loss in order to avoid overfitting.

C. *Result Analysis*

In case of both the dataset, the task of the models is binary classification. So we have used confusion matrix [30] for measuring the performance of our classification models. We have analyzed the performance of our model based on the evaluation metric that can be generated using the confusion matrix [29]. In the case of WDBC dataset the number of benign or malignant sample was distributed almost evenly. So all the evaluation measure from the confusion matrix for generated by both MLP and CNN model got a higher value.

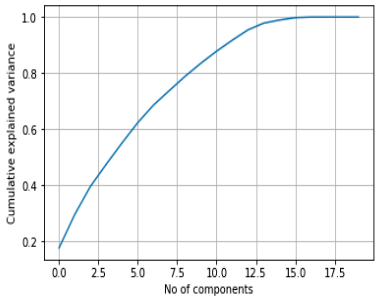

Fig. II. Cumulative explained variance for SEER.

TABLE I. PERFORMANCE ANALYSIS OF MLP AND CNN ON WDBC BASED ON EVALUATION MEASURES OF CONFUSION MATRIX.

Evaluation Measure	Neural Network Architecture	
	MLP	CNN
Sensitivity	1.000	0.97
Specificity	0.975	0.950
Precision	0.986	0.972
Accuracy	0.991	0.964
F-score	0.993	0.973
AUC	1.000	0.993

From Table I, we can say that our MLP has achieved an accuracy of 99.1% compared to 96.4% achieved by our CNN model. Multi-layer perceptron has outperformed CNN because of the number of samples available for training. However, both our models on WDBC dataset have achieved better accuracy than traditional Machine learning approaches [18].

In the case of SEER Breast Cancer dataset, the record of persons who are alive is much more than the person those are dead. So the dataset is skewed and for that reason, some of the evaluation measures of confusion matrix got lower values.

TABLE II. PERFORMANCE ANALYSIS OF MLP AND CNN ON SERR 2017 BREAST CANCER DATASET

Evaluation Measure	Neural Network Architecture	
	MLP	CNN
Sensitivity	0.982	0.970
Specificity	0.417	0.417
Precision	0.900	0.898
Accuracy	0.893	0.883
F-score	0.939	0.933
AUC	0.863	0.863

We have achieved an accuracy of 89.3% using MLP model and 88.3% using the CNN model. Again, MLP has outperformed CNN for its simple design and more connection. They have outperformed the Hybrid SVM model [20]. We have used Principal Component Analysis which has given us an advantage over the previous work where they hadn't use any Feature extraction technique or SEER 2017 dataset.

MLP has performed much better than CNN for WDBC dataset as the number of samples is less. But for SEER dataset when the number of samples has increased CNN has performed similar to MLP with less number of epochs. For this reason, we can say that CNN performs well only when there are enough samples available to train the model.

V. CONCLUSION

Breast cancer is one of the most found cancers in women of the present era. Detecting breast cancer in early stage can save thousands of lives. In this paper, we have used two public datasets in order to analyze breast cancer. At first, we have used principal component analysis to reduce the dimension of feature space. Then we have applied MLP and CNN models for classification. MLP has achieved an accuracy of 99.1% in predicting benign or malignant tumor and 89.3% in predicting the survivability status of a breast cancer patient. CNN has also achieved similar performance. For future work, we would like to use more architectures and compare their performance with the above models on this two datasets and also on other datasets in order to provide more accurate diagnosis model for analyzing breast cancer.

VI. REFERENCES

[1]. US Cancer Statistics Working Group, "United States cancer statistics: 1999–2012 incidence and mortality web-based report." *Atlanta (GA): department of health and human services, centers for disease control and prevention, and national cancer institute* (2015).

[2]. Shajahaan, S. S., Shanthi, S., and ManoChitra, V., "Application of data mining techniques to model breast cancer data." *International Journal of Emerging Technology and Advanced Engineering* 3.11 (2013): 362-369.

[3]. Forouzanfar, M. H., Foreman, K. J., Delossantos, A. M., Lozano, R., Lopez, A. D., Murray, C. J., & Naghavi, M., "Breast and cervical cancer in 187 countries between 1980 and 2010: a systematic analysis." *The lancet* 378.9801 (2011): 1461-1484.

[4]. Siegel, R. L., Miller, K. D., & Jemal, A., "Cancer statistics, 2015." *CA: a cancer journal for clinicians*65.1 (2015): 5-29.

[5]. Tominaga, Y., "Comparative study of class data analysis with PCA-LDA, SIMCA, PLS, ANNs, and k-NN." *Chemometrics and Intelligent Laboratory Systems* 49.1 (1999): 105-115.

[6]. Bruce, L. M., Koger, C. H., & Li, J., "Dimensionality reduction of hyperspectral data using discrete wavelet transform feature extraction." *IEEE Transactions on geoscience and remote sensing* 40.10 (2002): 2331-2338.

[7]. Wang, H., & Yoon, S. W.,"Breast cancer prediction using data mining method." *IIE Annu. Conf. Expo*. 2015.

[8]. Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T., "Using machine learning algorithms for breast cancer risk prediction and diagnosis." *Procedia Computer Science* 83 (2016): 1064-1069.

[9]. Delen, D., Walker, G., & Kadam, A., "Predicting breast cancer survivability: a comparison of three data mining methods." *Artificial intelligence in medicine* 34.2 (2005): 113-127.

[10]. Krizhevsky, A., Sutskever, I., & Hinton, G. E.,"Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.

[11]. Moeskops, P., de Bresser, J., Kuijf, H. J., Mendrik, A. M., Biessels, G. J., Pluim, J. P., & Išgum, I., Evaluation of a deep learning approach for the segmentation of brain tissues and white matter hyperintensities of presumed vascular origin in MRI." *NeuroImage: Clinical* 17 (2018): 251-262.

[12]. Kalchbrenner, N., Grefenstette, E., & Blunsom, P.,"A convolutional neural network for modelling sentences." *arXiv preprint arXiv:1404.2188* (2014).

[13]. Chollet, F., "Keras." (2015).

[14]. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M. and Kudlur, M., "Tensorflow: A system for large-scale machine learning." *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*. 2016.

[15]. Dhivyapriya, P., Sivakumar, S. Dr., "Classification of Cancer Dataset in Data Mining Algorithms Using R Tool", *International Journal of Computer Science Trends and Technology (IJCST).* Vol. 5. Jan – Feb 2017.

[16]. Wang, H., Zheng, B., Yoon, S. W., & Ko, H. S., "A support vector machine-based ensemble algorithm for breast cancer diagnosis." *European Journal of Operational Research* 267.2 (2018): 687-699.

[17]. Sarvestani, A. S., Safavi, A. A., Parandeh, N. M., & Salehi, M. , "Predicting breast cancer survivability using data mining techniques." *2010 2nd International Conference on Software Technology and Engineering*. Vol. 2. IEEE, 2010.

[18]. Ma, W., Gong, C., Hu, Y., Meng, P., & Xu, F.,"The Hughes phenomenon in hyperspectral classification based on the ground spectrum of grasslands in the region around Qinghai Lake." *International Symposium on Photoelectronic Detection and Imaging 2013: Imaging Spectrometer Technologies and Applications*. Vol. 8910. International Society for Optics and Photonics, 2013.

[19]. Han, J., Pei, J., & Kamber, M., *Data mining: concepts and techniques*. Elsevier, 2011, ch. 3, pp. 102-103.

[20]. Rumelhart, D. E., Hinton, G. E., & Williams, R. J. "Learning representations by back-propagating errors." *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[21]. Maier, H. R., & Dandy, G. C., "The effect of internal parameters and geometry on the performance of back-propagation neural networks: an empirical study." *Environmental Modelling & Software* 13.2 (1998): 193-209.

[22]. Maier, H. R., & Dandy, G. C., "Neural network based modelling of environmental variables: a systematic approach." *Mathematical and Computer Modelling* 33.6-7 (2001): 669-682.

[23]. Ouyang, X., Zhou, P., Li, C. H., & Liu, L., "Sentiment analysis using convolutional neural network." *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*. IEEE, 2015.

[24]. Brink, A. D., & Pendock, N. E., "Minimum cross-entropy threshold selection." *Pattern recognition* 29.1 (1996): 179-188

[25]. D.P. Kingma, J. Ba,"Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980*, 2014.

[26]. N. Srivastava, Hinton, A.Krizhevsky,I. Sutskever,R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *The Journal of Machine Learning Research* , 2014.

[27]. 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.

[28]. "SEER Breast Cancer Data", IEEE Dataport,2019. [Online]. Available: http://dx.doi.org/10.21227/a9qy-ph35. Accessed: Mar. 07, 2019.

[29]. Sokolova, M., & Lapalme, G., "A systematic analysis of performance measures for classification tasks." *Information Processing & Management* 45.4 (2009): 427-437.