Deep Learning of Gene Expression Data for Breast Cancer Classification

Y.P. Manawadu
Department of Biosystems Technology
Faculty of Technology
Sabaragamuwa University of Sri Lanka
Belihuloya, 70140, Sri Lanka
yahanimanawadu@gmail.com

M.D.R.L. Silva
Department of Computer Science
Faculty of Applied Sciences
University of Sri Jayewardenepura
Nugegoda, 10250, Sri Lanka

Abstract— **Currently, breast cancer is a type of cancer where most of the cancer patients suffer. Identifying the type of breast cancer is an essential step to be done in drug discovery for breast cancers. Thus, the treatments could address the key features of cancer and successfully the breast cancer can be cured. Majority of the related studies on cancer classifications are based on clinical diagnosis, hence affected with restricted classifications. Thus, gene expression data which has obtained via transcription profiling on microarrays have been used as the input for the classification of cancer in this study. The key goal of this study was to research the existing methods of breast cancer classifications and implement an efficient breast cancer classification method based on deep learning using gene expression data which overcomes the defects of existing methodologies of breast cancer classification. Two different deep learning architectures were implemented by this study which is Convolutional Neural Network(CNN) and Deep Belief Network(DBN) using Tensorflow framework to classify breast cancers under the classification based on gene and protein status. Finally, we compared the performance of those two architectures with the deep learning architecture, autoencoder which was implemented before in another study in classifying breast cancer using gene expression data. The two proposed architectures perform better than autoencoder with respect to precision, recall, F1 score and accuracy. In conclusion, CNN is the best supervised deep learning architecture which yielded an accuracy of 63.4395% and DBN is the best unsupervised deep learning architecture which yielded an accuracy of 63.3545% in classifying breast cancers using gene expression data based on gene and protein status.**

Keywords— *autoencoder, breast cancer classification, convolutional neural network, deep learning, deep belief network*

I. Introduction

In the medical field, researches related to cancer is one of the significant research areas [1]. Classification of cancer has always been an interested research area of all researchers within the medical community [2]. Exact prediction of the type of tumour results from best-suited treatment which has minimum toxicity on the patients. To select a rational, more effective and appropriate therapeutic interventions for individualized decision-making process during clinical management of cancer patients, highly accurate predictive tests are essential which will enhance the outcome after therapy. Earlier morphological and clinical based methods were used to classify cancers. These typical cancer classification methods are having several constraints in their diagnostic ability. It is stated that the effectiveness of cancer treatments has been maximized when specifying therapies according to the tumour types identified using pathogenetic patterns[3, 4]. Also, the identified cancer types have been found to be diversified and encompass the diseases that are molecularly noticeable and different clinical courses should be followed to cure different types of cancers.

Due to the advancements in medical sciences and technology, numerous methodologies were found to cure cancers with the availability of massive amounts of data. With these big data, we can easily find insights regarding cancers. The crucial insights for the elementary problems related to biological evolution, ways used to prevent and cure of diseases and drug discovery could be discovered by means of gene expression level. Monitoring millions of genes simultaneously was made possible by the latest advancements in microarray technology which motivates the use of gene expression data to identify the cancer types more accurately.

In biological and clinical applications, incorporation of the knowledge gathered through the correlation between diseased states and gene expression profiles plays a major role [5]. Thus, better insights into disease pathology can be obtained by the comparison of the gene expression profiles of normal tissue with the multiple diseased tissues. To divide gene expression in tumour cells into different types, the above-stated technique has been used. According to the probability of recurrence of diseases and survival after therapy, breast cancer patients were categorized into clusters. Based on the identified molecular basis of cancer, facilitating the patients with better early diagnosis and customized therapeutics using therapeutical tools is a key aim of this study. This study has been proposed a methodology to classify breast cancers by analysis of gene expression profiling data using deep learning approaches in order to gain better insights into the breast cancer classification.

The main goal of this study was to find a deep learning approach for classification of breast cancer using the gene expression data. In this study, we implemented a convolutional neural network, autoencoder, deep belief network to classify breast cancers with notable accuracy. We were able to test the network with higher accuracy than other methods that have been used for breast cancer classification using gene expression data.

II. Related Work

To model the gene expression data, an independent component analysis was employed by De-Shuang Huang et al [6], then apply an optimal scoring algorithm to classify them. This approach can first make full use of the high-order statistical information contained in the gene expression data. Second, this approach also employs regularized regression models to handle the situation of large numbers of correlated predictor variables. Finally, the predictive models are developed for classifying tumours based on the entire gene expression prole. The methodology involves regularizing gene expression data using ICA, followed by the classification applying penalized discriminant method. The main focus of this study lies on regularizing and modelling of gene expression data. But, ICA is not always reproducible when used to analyze gene expression data because ICA algorithm may converge to local optima [7]. This study has used colon cancer data, acute leukaemia data, hepatocellular carcinoma data, high-grade glioma data. Upon this process, they succeeded to obtain an accuracy above 55% in classification of tumours. They have suggested that making full use of the information contained in the gene data can generate more exact prediction of tumour class.

Christos Sotiriou et al [8], focused on Classification of Tumors Samples Based on Clinical Pathologic Characteristics such as ER status, nodal status, tumour size, tumour grade, and menopausal status of the patient affect the behaviour of breast cancers. They have suggested a Hierarchical cluster analysis for the segregation of tumours into two main groups based on their ER status. The parametric t-test results in a p-value of 0.001 suggesting that ER status has a strong association with gene expression. Thus, showed that the expression proles primarily distinguished ER from ER tumours and called them luminal and basal subtypes because of their respective luminal and basal characteristics. However, there is no strong evidence that nodal and menopausal status of the patient or tumour size is associated with the expression proles of the tumours.

The study, quantitative classification of mammographic densities and breast cancer risk, by N.F. Boyd et al [9] was conducted to determine the level of breast cancer risk associated with varying mammographic densities by quantitatively classifying breast density with conventional radiological methods and novel computer-assisted methods. Here they have used mammogram images as the dataset. Statistically significant increases in breast cancer risk associated with increased mammographic density were found by both radiologists and computer-assisted methods. Thus, concluded that increases in the level of breast tissue density as assessed by mammography are associated with increases in risk for breast cancer.

The study, a deep learning approach for cancer detection and relevant gene identification used Stacked Denoising Autoencoder (SDAE) for the extraction of meaningful features from gene expression data that enable the classification of cancer cells while achieving 94.78% accuracy for cancer detection using gene expression data [10]. Results and analysis of this study illustrate that these highly interactive genes could be useful cancer biomarkers for the detection of breast cancer.

Rajendra Rana Bhat et al [2] propose a deep generative machine learning architecture (called Deep Cancer) that learn features from unlabeled microarray data to classify the tissue samples as either being cancerous or non-cancerous. Here they have used 3 types of generative learning models:

i) Restricted Boltzmann Machines
ii) Generative Adversarial Networks
iii) Deep Convolutional GAN

The features are learned through an adversarial feature learning process and then sent as input to a conventional classifier specific to the objective of interest. Based on deep generative learning, the tuned discriminator and generator models learned to differentiate between the gene signatures without any intermediate manual feature handpicking, indicating that much bigger datasets can experiment on the proposed model more seamlessly. They suggest that Deep Cloud model will be a vital aid to the medical imaging community and, ultimately, reduce inflammatory breast cancer and prostate cancer mortality.

Ying Lu et al [1] claims that most previous cancer classification studies are clinical-based and have the limited diagnostic ability. In order to gain deep insight into the cancer classification problem, this study presents a comprehensive overview of various proposed cancer classification methods and evaluate them based on their computation time, classification accuracy and ability to reveal biologically meaningful gene information. Here, they conclude that cancer classification using gene expression data has a promising future in providing a more systematic and unbiased approach in differentiating different tumour types.

Purpose of the study, Gene expression patterns of breast carcinomas distinguish tumour subclasses with clinical implications which were conducted by Therese Sorlie et al [11] was to classify breast carcinomas based on variations in gene expression patterns derived from cDNA microarrays and to correlate tumour characteristics to clinical outcome. For this, they have used hierarchical clustering and it showed significantly different outcomes for the patients belonging to the various groups, including a poor prognosis for the basal-like subtype and a significant difference in outcome for the two estrogen receptor-positive groups.

Gennadi V. Glinsky et al [12] propose a breast cancer classification algorithm taking into account three main prognostic features determined at the time of diagnosis: estrogen receptor (ER) status; lymph node (LN) status; and gene expression signatures associated with distinct therapy outcome. This study evaluated the prognostic power of breast cancer survival predictor signatures alone and in combination with ER and LN status using Kaplan-Meier analysis and it results in P-value of 0.0001 from log-rank test. Thus, conclude that quantitative laboratory tests measuring expression proles of a limited set of identified small gene clusters may be useful in stratification of breast cancer patients at the time of diagnosis into subgroups with a statistically distinct probability of positive outcome after therapy and assisting in the selection of optimal treatment strategies.

The study, molecular classification of cancer: class discovery and class prediction by gene expression monitoring [13] employed neighbourhood analysis to classify acute leukaemia into two types namely Acute Myeloid Leukemia (AML) and Acute lymphoblastic leukaemia (ALL) using gene expression data as the dataset. It results in a 1% significance level for the number of genes within corresponding neighbourhoods of the randomly permuted class distinctions concluding the proposed method can classify leukaemia into types accurately using gene expression data.

III. METHODOLOGY

Since we have decided to classify breast cancer based on gene and protein status, we needed a dataset which consists of data for both ER and HER2 status for breast cancer. But, a dataset which consists of both ER and HER2 status wasn't available. Therefore, two datasets have been utilized to classify breast cancer where 1 dataset consists of gene expression data to classify ER status and the other dataset consist of gene expression data to classify HER2 status. Fig 1 shows how we can classify breast cancers under gene and protein status if we know the ER status and HER2 status of the dataset.

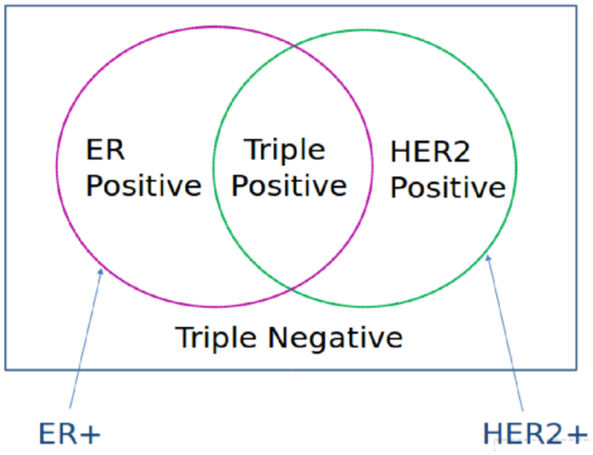

Fig. 1. Classification of Breast Cancers under gene and protein status

A. Data Collection and Preprocessing

These 2 gene expression datasets had been collected from the Gene Expression Omnibus (GEO) data repository which stores curated gene expression datasets as well as original series and platform records. The gene expression dataset to determine ER status was taken from the GEO accession ID 4057. This dataset consists of gene expression data for 25760 number of unique genes of breast cancer patients. There are 113 features for each unique genes in this dataset. Other dataset which we use to determine the HER2 status was extracted from the GEO accession ID 4069. This dataset consists of gene expression data for 21780 number of unique genes of breast cancer patients. There are 18 features for each unique genes in this dataset.

We had to download data from the GEO repository in the form of text files where each le consist of gene expression data for 20 unique genes. From those text files, we had to extract useful data into a table format. For that, we have used a macro in Excel application of Microsoft office package, where we had to use the programming language Visual Basic. Since that dataset gave some errors in training, we had to validate the dataset. For the validation process of the dataset, we have used a macro in Excel where the programming language was Visual Basic while the concept of regular expressions was used invalidating. Subsequently, the dataset was partitioned into 3 groups which are training, validation and testing with the proportions of 60%, 20% and 20% respectively.

B. Algorithms

Among the supervised deep learning techniques, we have selected a convolutional neural network because only CNN can handle high dimensionality feature of gene expression dataset. Among the unsupervised deep learning techniques, we have selected other 2 deep learning algorithms. With considering the literature regarding the deep learning used for cancer detection using gene expression data, they have used Restricted Boltzmann Machines (RBMs) to detect cancers. Since Deep Belief Network (DBN) is a stack of RBMs, we have selected DBN to classify breast cancers using gene expression data. Since autoencoder has the ability to capture non-linear relationships of the datasets with the feature high dimensionality, we have selected autoencoder to classify breast cancers using gene expression data.

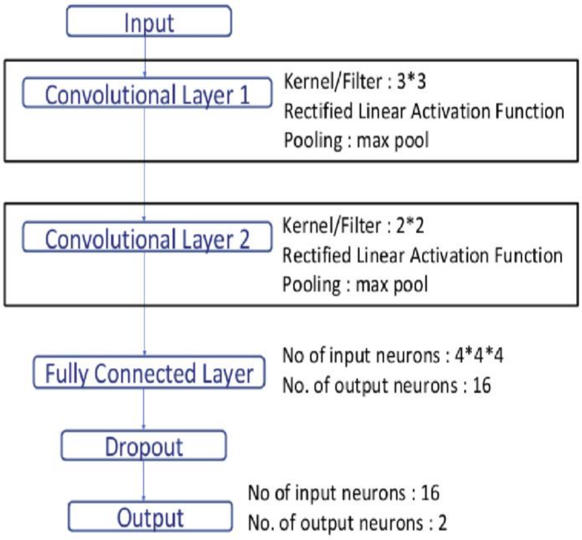

Fig. 2. CNN Implementation Architecture

Every algorithm was designed to classify 2 classes of data. We have used these networks to classify ER dataset into 2 classes known as ER-positive and ER-negative and HER2 dataset into 2 classes known as HER2-positive and HER2-negative. Each network was trained for 30 epoch because network shows less convergence after 30 epoch. All networks use the same base learning rate of 0.01 with exponentially decaying with the epoch and weight decay of 0.0005 with solver momentum of 0.9.

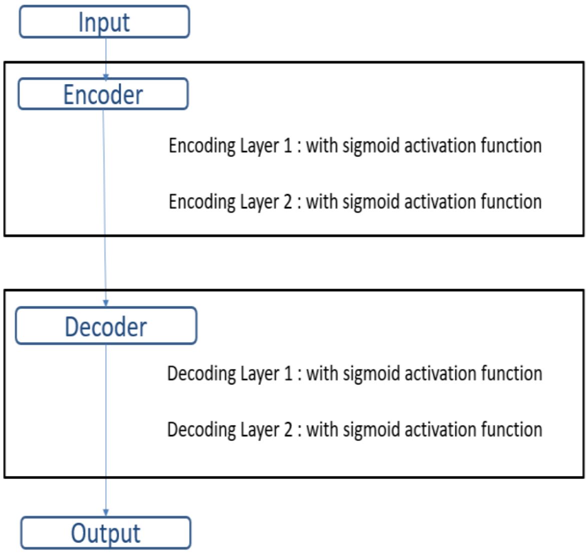
Input
Encoder
Encoding Layer 1 : with sigmoid activation function
Encoding Layer 2 : with sigmoid activation function
Decoder
Decoding Layer 1 : with sigmoid activation function
Decoding Layer 2 : with sigmoid activation function
Output

Fig. 3. Autoencoder Implementation Architecture

C. Tools and Technologies

A personal computer of Dell model was used for this study in which the CPU with Intel Core i5 dual-core, DDR3 Memory with 8GB capacity and DDR5 GPU with Nvidia GeForce 650 processor; where GPU was used to train the deep learning networks. Since the GPU consists of 1GB memory, it became the major reason to limit the network size and input size. The main reason for using GPU is the fast computation time of deep learning implementations. The GPU which used for this process has only 384 CUDA cores and a base clock of 1058 Megahertz. All these factors made GPU computation faster than the CPU that we had. Since the majority of machine learning libraries support Linux based platforms more than the other platforms, majority of tools used in deep learning works in good quality on Linux based systems, Linux's huge support for GPU computation libraries and Ubuntu is also a Linux platform; 64bit Ubuntu 16.04 was used as the implementation platform of this study.

Here, we have implemented the Convolutional Neural Network and Deep Belief Network for breast cancer classification using gene expression data using the framework Tensorflow. Then we have implemented autoencoder on Tensorflow framework and compared the performance concerning execution time, precision, recall, F1 score and accuracy. Also, as the deep learning primary step, we have implemented Multi-Layer Perceptron on Tensorflow framework to classify breast cancer using gene expression data.

D. Classification of Data

To classify new data from the trained model, Tensorboard interface of tensorflow or the terminal can be used. Tensorflow provides a Python implementation to classify new data using the trained model with the terminal after downloading the model form of tensorflow. After completion of training in Tensorflow, a suitable model can be downloaded from any epoch and by uploading new data to tensorflow using the interface, classifications can be done. Tensorflow interface provides useful information like activation of each layer when conducting classification via a user interface.

IV. Results

Table 1 shows the performance measures concerning the precision, recall, F1 score and accuracy and Table 2 shows performance concerning execution time for the tensorflow implementation of CNN to classify breast cancer using gene expression data. The CNN that we have trained, performed somewhat well by obtaining approximately 63% accuracy with high precision.

TABLE I. Performance measures of the best CNN model

Accuracy	63.4395%
Precision	0.806
Recall	0.498
F1 score	0.38105

TABLE II. Temporal values for training the best CNN model

Real	0m 36.0395s
User	0m 42.876s
Sys	0m 21.636s

Autoencoder has shown poor accuracy for the classification of breast cancers and with considerable precision. Table III shows the performance measures concerning the precision, recall, F1 score and accuracy and Table IV shows performance with respect to execution time for the tensorflow implementation of Autoencoder to classify breast cancer using gene expression data.

TABLE III. Performance Measures for the best autoencoder model

Accuracy	51.923%
Precision	0.8
Recall	0.475
F1 score	0.5905

TABLE IV. Temporal values for training the best autoencoder model

Real	0m 3.2115s
User	0m 4.434s
Sys	0m 4.452s

Deep Belief Network has shown good accuracy for the classification of breast cancers and with considerable precision. Table V shows the performance measures with respect to precision, recall, F1 score and accuracy and Table VI shows performance with respect to execution time for the tensorflow implementation of Deep Belief Network to classify breast cancer using gene expression data.

TABLE V. Performance Measures for the best DBN model

Accuracy	63.3545%
Precision	0.605
Recall	0.725
F1 score	0.63615

TABLE VI. Temporal values for training the best DBN model

Real	0m 63.852s
User	1m 45.7805s
Sys	0m 5.193s

V. Conclusion

We can clearly observe that both convolutional neural network and deep belief network shows better performance than autoencoder for the breast cancer classification using gene expression data. Since Convolutional Neural Network is the only supervised deep learning method used here, CNN is the best performing model out of these 3 models in classifying breast cancers using gene expression data. Although Deep Belief Network is an unsupervised deep learning technique it also gives a very good accuracy which is very close to CNN's accuracy. But, Deep Belief Network consumes more execution time than autoencoder and Convolutional Neural Network. Autoencoder, which is another unsupervised deep learning technique shows a poor overall performance than CNN and DBN for the breast cancer classification using gene expression data. But, autoencoder takes very small time to execute than CNN and DBN. Therefore, Convolutional Neural Network is the best deep learning technique to classify breast cancers using gene expression data.

VI. Future Works

We were able to implement accurate and efficient methods to classify Breast Cancers using gene expression data, our study was limited for only 3 deep learning techniques and our method only focused on classification of breast cancers based on protein and gene status. There are several extensions that can be added to tackle those requirements.

The most important extension will be a combination of the classification with the detection capability to identify breast cancers using gene expression data. There are several deep learning methods that can be used for the detection of cancers using gene expression data.

These classification algorithms were used to classify breast cancers using 2 different datasets where we have used one dataset to classify ER status and other dataset to classify HER2 status. Therefore, If any dataset with both ER and HER2 status for breast cancers is available, it can be used to take precise measures for breast cancer classification. Also, the dataset can be increased and tested in order to increase test accuracy.

Apart from the above-mentioned extension we can also change the network architecture and tested in an environment with more processing power to reduce the train time and increase the accuracy of the proposed model.

REFERENCES

[1] Jiawei Han Ying Lu. Cancer classification using gene expression data. Information Systems - Special issue: Data management in bioinformatics, 28(4):243268, 2003. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Xiaolin Li. Rajendra Rana Bhat, Vivek Viswanath. deep cancer: detecting cancer through gene expressions via deep generative learning, 2016.

[3] M.B.Eisen et al A.A. Alizadeh. Distinct types of diffuse large bcell lymphoma identified by gene expression profiling. ELSEVIER, 403(6769):503511, 2000.

[4] L.Penland et al J.DeRisi. Use of a cdna microarray to analyse gene expression patterns in human cancer. Nature Genetics, 14(4):45760, 1996.

[5] D.Gilbert A.C.Tan. Ensemble machine learning on gene expression data for cancer classification. Applied Bioinformatics, 2(3):7583, 2003.

[6] Chun-Hou Zheng De-Shuang Huang. Independent component analysis based penalized discriminant method for tumour classification using gene expression data. Bioinformatics, 22(15):18551862, 2006.

[7] Ming Zhanet al Huai Li. Identifying conserved and divergent transcriptional modules by cross-species matrix decomposition on microarray data. Bioinformatics, 2 (14):117125, 2009.

[8] M.Lisa Christos Sotiriou, Soek-Ying Neo. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. PNAS, 100(18): 18551862, 2003.

[9] J.W.Byng et al N.F.Boyd. Quantitative classification of mammographic densities and breast cancer risk: Results from the Canadian national breast screening study. Journal of the National Cancer Institute, 87(9):670675, 1995.

[10] David A Hendrix Padideh Danaee, Reza Ghaeini. A deep learning approach for cancer detection and relevant gene identification. Pacic Symposium on Biocomputing, 22(219):1724, 2016.

[11] Charles M. Peroua Therese Sorlie. Gene expression patterns of breast carcinomas distinguish tumour subclasses with clinical implications. PNAS, 98(19):1086910874, 2001.

[12] Anna B. Glinskii Gennadi V. Glinsky, Takuya Higashiyama. Classification of human breast cancer using gene expression profiling as a component of the survival predictor algorithm. Clinical Cancer Research, 10(7):22722283, 2004.

[13] D. K. Slonim et al T. R. Golub. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Clinical Cancer Research, 286 (5439):531537, 1999.