IOP Conf. Series: Materials Science and Engineering **864** (2020) 012079 doi:10.1088/1757-899X/864/1/012079

Application of Deep Learning Method in Facilitating the Detection of Breast Cancer

Azurah A Samah[1, 2], Dewi Nasien[3], Haslina Hashim[1, 2], Julia Sahar[1], Hairudin Abdul Majid[1, 2], Yusliza Yusoff[2], Zuraini Ali Shah[1, 2]

[1] Artificial Intelligence and Bioinformatics Research Group, School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia.
[2] School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia.
[3] Institut Bisnis dan Teknologi Pelita Indonesia, No. 78-88
JI Jendral Ahmad Yani. 28127, Pekanbaru, Riau, Indonesia.

E-mail: azurah@utm.my

Abstract. Breast cancer is a type of tumour that could be treated if the disease is identified at an earlier stage. Early diagnosis is crucial when it comes to reducing the mortality rate. In this study, deep neural network method is applied to facilitate the detection of breast cancer. The aim of this study is to implement deep neural network in breast cancer classification models that can produce high classification accuracy. Deep Neural Network (DNN) with multiple hidden layers was applied to learn deep features of the breast cancer data. Dataset used in this study was obtained from the UCI Machine Learning Repository which consists of Wisconsin Breast Cancer Dataset (WBCD) and used for the original and diagnostic dataset. The performance of the proposed DNN method was compared against previous machine learning classifier in terms of accuracy. From the results, the accuracy obtained for the original dataset was 97.14% and 97.66% for the diagnostic dataset, which is better than previous SVM method.

1. Introduction

Breast cancer are usually detected and diagnosed after symptoms appear but many women with cancer have no symptoms at all. This explains why regular breast cancer screening is so important. Early detection of breast cancer enables the disease to be treated and to reduce the rate of mortality. A cancerous tumour is known as a malignant cell. Difficulties in detecting the malignancy cancerous cells lead to a rise in the death rate around the world. It has long become one of the major lethal diseases which leads to around 8.2 million, or 14.6% of all human deaths each year [1]. One significant difference between normal cells and cancerous cells is that cancer cells are less specialised than normal cells. Normal cells grow into very distinct cell types with specific functions while cancer cells do not [2]. This is one reason that, unlike normal cells, cancer cells continue to divide without stopping. The most effective way to reduce cancer deaths is by having an exact and a solid diagnosis procedure that could be utilised by biologist and computer scientist to identify malignant cells in an early stage [3]. Classification technique plays the main role in the medical field for diagnosing and identifying early treatment [4]. Deep learning architecture is a new and powerful method used by the researches to measure the performance of cancer classifiers particularly in terms of accuracy. In this study, a Deep Neural Network (DNN) approach is implemented to prove its ability and efficiency in diagnosing breast cancer.

From previous studies, there are many techniques used in cancer diagnosis, which involves machine learning techniques. Machine learning consists of four main types of algorithm which are supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning [5]. However, the most popular methods of machine learning are supervised and unsupervised learning.

Several machine learning methods have been applied in classifying breast cancer using WBCD (Original) dataset. The methods include Support Vector Machine (SVM) which obtained 96.87% accuracy [6], Decision Trees (DT) with 95% accuracy [7], Bayesian Network (BN) with 91.31% accuracy [4] and Artificial Neural Network (ANN) with 92.1% accuracy [1]. The results of these techniques shown that SVM implemented by Chen *et* al., 2011 leads the accuracy performance compared to other classifiers. Thus, in this study comparative analysis is carried out between SVM and the proposed DNN method in terms of accuracy.

DNN follow the structure of a typical artificial neural network with a complex network model. It helps us to create a model and define its complex hierarchies in a simple form [8]. It has 'n' hidden layers and processes the data from the previous layer called the input layer. After every epoch, the error rate of the input data will be gradually reduced by adjusting the weights of every node, back-propagating the network and continues until it reaches better results [9]. In this study, 9 and 16 number of inputs is assigned as input nodes in the input layer. The output layer will determine either it is malignant (cancerous) or benign (non-cancerous) class. This paper is organized into 4 sections. Section 2 describes the materials and method. This is followed by Section 3 which presents the experimental results. This paper ends with a conclusion included in Section 4.

2. Materials and Method

2.1. Dataset

The dataset used in this study is Wisconsin Breast Cancer Dataset (WBCD) obtained from the University of Wisconsin Hospital, Madison by Dr William H. Wolberg publicly available at UCI machine learning repository https://archive.ics.uci.edu/ml/datasets. They provide two types of breast cancer datasets which are WBCD (Original) and WBCD (Diagnostic) that has changed in the number of attributes and number of samples. The proposed DNN algorithm was measured against different classifier algorithms on the same database. The first dataset applied in DNN model was WBCD (Original) which consist of 699 samples and 11 attributes including sample code number and class with 16 missing values, as shown in Table 1. The second dataset applied was WBCD (Diagnosis) which has 569 samples with 32 attributes including id and class diagnosis. All the attributes composed of ten real-valued features which are computed for each cell nucleus including texture, radius, perimeter, smoothness, area, concavity, compactness, symmetry, concave points and fractal dimension. These features describe the characteristics in the image of the cell nuclei [10]. The features' mean, standard error (SE), along with the mean of the three largest values were calculated for every image resulting in a total of 30 features. The class distribution is 62.7% (357 instances) for benign and 37.3% (212 instances) for malignant. However, after pre-processing the data, only 16 attributes are used to be trained and tested in the model. Table 2 provides the attribute information for WBCD (Diagnostic) dataset after pre-processed.

Table 1. Attribute information of WBCD (original).

Breast Cancer Dataset			Breast Cancer Dataset		
No.	Attributes	Domain	No.	Attributes	Domain
1	Sample Code Number	id number	7	Bare Nuclei	1 – 10
2	Clump Thickness	1 – 10	8	Bland Chromatin	1 – 10
3	Uniformity of Cell Size	1 – 10	9	Normal Nucleoli	1 – 10
4	Uniformity of Cell Shape	1 – 10	10	Mitoses	1 – 10
5	Marginal Adhesion	1 – 10	11	Class	2 – Benign 4 – Malignant

Table 2. Attribute information of WBCD (diagnostic) after pre-processed.

No.	Attributes	No.	Attributes
1	Texture_mean	9	Smoothness_se
2	Area_mean	10	Concavity_se
3	Smoothness _mean	11	Symmetry_se
4	Concavity_mean	12	Fractal_dimension_ se
5	Symmetry_mean	13	Smoothness_worst
6	Fractal_dimension_mean	14	Concavity_worst
7	Texture_se	15	Symmetry_worst
8	Area_se	16	Fractal_dimension_ Worst

2.2. Pre-processing and Normalization

Data preprocessing is an essential step as datasets highly influenced by the negative elements such as the presence of noise, missing values, inconsistent and superfluous data. The raw dataset was preprocessed using the tools available in Weka 3.8 to produce well-form data that suitable for training and testing process. Data cleaning is a step involved in data preparation to make sure the breast cancer datasets are free from noise and irrelevant data. These data can affect the accuracy of the classifier and the validity of the outcomes. In order to maintain the significant value of the data, modification, deletion, or replacing the missing values were done There were16 missing values in the WBCD (Original) dataset for attributes bare nuclei. The 16 missing values were replaced with the mean of the numerical distribution using the Replace Missing Values filter provided in Weka. When the data was free from missing value, a hot encoding was applied to convert categorical data into numerical data. The label encoding was performed on both datasets. After the data has been pre-processed, the data acquired need to be normalized so the values are in the range of 0 to 1. This will shorten the time required for the learning rate of the model. The minimum and maximum values for each dataset need to be known for the calculation of normalization. Normalization needs to be done to ensure the tested data will not over fitted or under fitted.

2.3. Deep Neural Network Method

In this study, DNN classifier has been implemented to classify breast cancer diagnosis. The proposed DNN model has three hidden layers. DNN architecture is shown in Figure 1. The important phase of DNN algorithm is to define the number of nodes in each layer. The number of input node is equal to the number of variables in the dataset. In this study, the number of variables used for the WBCD (Original) dataset is 9 while in the WBCD (Diagnostic) is 16 attributes which mean the number of input node used for both datasets is 9 and 16 nodes respectively. The input nodes and output nodes rely on the value of input variables and the output class of the dataset. The output node represents the malignant class of diagnosis. The node in the hidden layer is determined by using the formula $h = n$, $h = 2n$, and $h = n/2$ where n refers to input features count and h refers to hidden layer node.

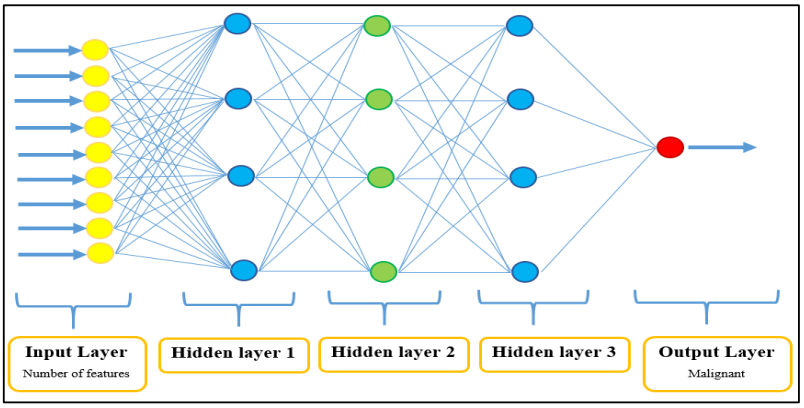

Figure 1. DNN architecture.

3. Experimental Results

To validate the effectiveness of the proposed classifier performance, the results obtained from the implementation of DNN model for both original and diagnostic datasets are compared with SVM classifier, as shown in Table 3.

Table 3. Performance comparison of DNN and SVM classifier.

Datasets	Accuracy of Classifier (%)		
	DNN	SVM	Improvement
WBCD (Original)	97.14	96.19	0.95
WBCD (Diagnostic)	97.66	94.15	3.51

From this finding, the WBCD (Original) data obtained 96.19% accuracy in SVM and has improved about 0.95% accuracy in DNN while for the second dataset which is WBCD (Diagnostic), it only obtained 94.15% in SVM and DNN achieved 3.51% improvement. Based on the results above, it can be seen that both datasets achieved higher accuracies in DNN compared to SVM. To make it clearer, Figure 2 presents a comparison bar chart for DNN and SVM classifier using both datasets.

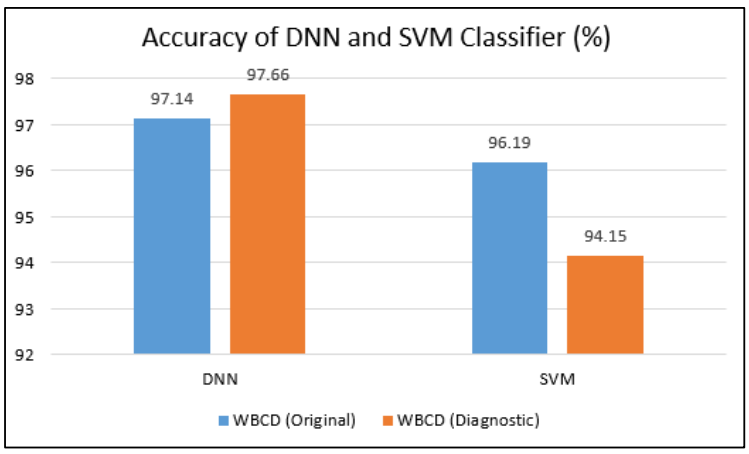

Figure 2. Summary of performance comparison between DNN and SVM classifier.

In conclusion, the proposed DNN classifier performs better than SVM in solving the classification problem of the case study. It proved by an average of 97% accuracy achieved in both breast cancer

dataset. This is mainly due to multiple hidden layers in DNN which allows this architecture to learn by creating more abstract representation of data as the network grows deeper. As a result, the DNN model automatically extracts feature and yields higher accuracy results.

4. Conclusion and Future Works

In this study, the main objective related to the development of an accurate classification model using deep learning to facilitate the detection of breast cancer has been achieved with satisfying results. Based on the 97% accuracy of the result obtained for both WBCD datasets has proved that the proposed DNN classifier can produce high classification accuracy for diagnosing breast cancer. Despite this achievement, few works can be carried out in future to expand the research further and improve existing results. One of it is, the DNN algorithm has several hidden layers with many neurons in each layer that can propagate and parse the input through several layers. Since the structure of this network is very complex, the duration of training is inevitable. Therefore, in order to further improve the DNN classification performance, other researchers may perform recursive feature elimination or particle swarm optimisation techniques used in feature selection. The accuracy of the general model can be further increased by selecting features based on their fitness values. Other than that, the power of Deep Learning performance, especially on big data, is undeniable. Thus, by applying bigger datasets of other types of cancers may produce another discovery of deep learning classification performance. It is also suggested to use a cloud-assisted virtual machine or graphics processing unit to optimise the computational model effectiveness. This will reduce the time needed to train the model and make the model to be computationally inexpensive.

Acknowledgments

The authors would like to express gratitude to the reviewers and editors for helpful suggestions and Universiti Teknologi Malaysia for sponsoring this research by the GUP Research Grant (Grant Number: Q.J130000.2528.19H17) and TDR Research Grant (Grant Number: Q.J130000.3551.05G42). The research is also sponsored by the School of Computing, Faculty of Engineering, UTM.

References

[1] Saabith S, Sundararajan E and Azuraliza A A 2014 *J. Computer Science And Mobile* Computing **3** 10 185-191.
[2] Mandal S and Banerjee I 2015, *J. Emerging Engineering* **3** 7 172-178.
[3] Huang M L, Hung Y H and Chen W Y *et al* 2010 *J.* Medical Systems 34(5) pp 865-873.
[4] Minar M R 2018 *Recent Advances in Deep Learning : An Overview*, 0-31.
[5] Ayodele T O 2010 *New Advances in Machine Learning* **3** 19-48
[6] Chen H L, Yang B, Liu J and Liu D Y 2011 *Expert Systems with Applications* **38** 7 9014-9022.
[7] Elsayad A M and Elsalamony H A 2013 *J. Computer Applications* **83** 5 19-29.
[8] Borges L R 1989 *Proceedings of XI Workshop de Visão Computacional* 15-19.
[9] Hamsagayathri P and Sampath P 2017 *J. Current Pharmaceutical Research* **9** 2 19-25.
[10] Utomo C P, Kardiana A and Yuliwulandari R 2014 *J. Advanced Research in Artificial Intelligence* **3** 7 10-14.