Penerbit Akademia Baru

Journal of Futuristic Biosciences and Biomedical Engineering

Journal homepage: www.akademiabaru.com/fbbe.html
ISSN: XXXX-XXXX

Molecular Classification of Breast Cancer Subtypes Based on Proteome Data

Open Access

Azian Azamimi Abdullah [1,*], Nur Lili Suraya Ngadiman[1]

[1] Biomedical Electronic Engineering Programme, School of Mechatronic Engineering, Universiti Malaysia Perlis, Pauh Putra Campus, 02600 Arau, Perlis, Malaysia

ARTICLE INFO

Article history:
Received 29 February 2019
Received in revised form 12 April 2019
Accepted 19 April 2019
Available online 21 April 2019

Keywords:
Breast cancer, subtypes, proteome data, machine learning, deep neural network

ABSTRACT

Breast cancer can be classified into different subtypes, which leads to different risk factors. The breast cancer subtypes are expressed based on the genes. To determine prognosis and further treatment, breast cancer subtypes need to be identified at an early stage. In order to classify breast cancer subtypes, the biological information of the cancer cells must be extracted. Besides the normal-like breast cancer subtypes, there are another 4 subtypes, which are Luminal, A, Luminal B, basal-like and HER2 classified according to their biological characteristics of genes from tumor cells. Different subtypes with different biological characteristics result in a variation of treatment quality and clinical outcomes. This study has been conducted to classify proteome data into four molecular subtypes of breast cancer by implementing supervised and unsupervised machine learning methods. Unsupervised machine learning model was constructed based on the K-means algorithm. Meanwhile, supervised machine learning model was constructed based on Random Forest (RF), Gradient Boosting Machine (GBM) and Deep Neural Network (DNN) algorithm. R and Python software has been employed in order to build the machine learning models for the classification. The performance of the machine learning model has been compared in term of accuracy of the classification. Our computational results show that DNN has the highest accuracy (96.9%) compared to RF (89.3%) and GBM (76.1%). Hence, this proves that DNN is the best algorithm to classify breast cancer subtypes based on proteome data.

1. Introduction

The composition of the breast is made up of fat and breast tissue that rich with arteries, veins and small blood vessels to deliver oxygen and nutrients to the cells tissue [1]. Cancer cells begin to develop from mutated cells in which have errors in the growth process. However, the cells did not damage and started to build up to form a mass tissue that commonly called a tumor or lump. The tumor develops from mutated cells and spread to the nearest cells. The tumor formed is where breast

* *Corresponding author.*
E-mail address: azamimi@unimap.com.my (Azian Azamimi Abdullah)

cancer starts to develop and continue spreading to the adjacent cells. A radiologic imaging test such as ultrasound or mammogram can be used to diagnose the tumor.

The severity of breast cancer can be measured based on its stage. Stages of breast cancer range from 0 to 4 and differentiated by TNM systems in which T stands for Tumor size, N is Lymph Node status and M is Metastases. The combination of TNM classifies the breast cancer stage.

Breast cancer can be classified into different subtypes, which leads to different risk factors. The breast cancer subtypes are expressed based on the genes. To determine prognosis and further treatment, breast cancer subtypes need to be identified. In order to classify the molecular of breast cancer into subtypes, the biological information of the cancer cells must be extracted. As seen in Figure 1, biological information can be obtained either through DNA, RNA, protein or metabolites. Formerly, the genetic information is access from DNA and known as genome [2]. The investigation of gene and gene expression had developed until the information of cancer cells can be access through RNA and protein, which is known as transcriptome and proteome respectively. Furthermore, extracting biological information from smaller molecules such as metabolites and macromolecules like sugar, lipid and amino acids had been further investigated. The complete sets of organic metabolites currently are known as metabolome. However, there is a limitation in order to extract biological information through genomes, transcriptomes and metabolomes hence the study in proteome data is needed.

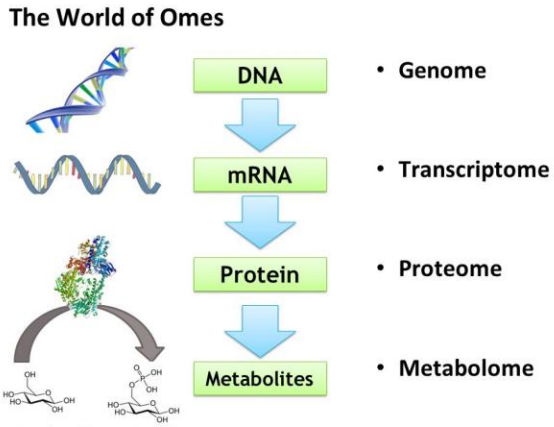


Fig. 1. Visualization of biological information sources

Classification of breast cancer subtypes has a close relation to determining the prognosis among breast cancer patients. Luminal A can be said to have a better prognosis compared to Luminal B. HER2 subtypes has poor prognosis meanwhile basal-like has the worse prognosis among the other subtypes. Even though basal-like subtypes has high response towards chemotherapy, but a high rate of cancer cells mutation has resulted in poor prognosis [3].

As classification using DNA, RNA, and metabolites are identified with few limitations, further research by using proteomic data has to be extended. Early detection of breast cancer is very crucial. In order to come up with an accurate prognosis, an adequate diagnosis of the respective subtypes of breast cancer is required. It needs a reliable source to identify the subtypes. By considering the advantages of proteomic data, a technique of extracting the data needs to be done. In addition, extracting biological features from the data requires a method in order to describe hidden structure from the data and improve data precision. Another approach is needed to carry out for final classification of breast cancer subtypes.

In order to identify the subtypes, it requires a machine-learning algorithm to come out with the classification. It needs to find the correct features of protein breast cancer sets using k-means clustering technique to improve prediction precision. Consequently, supervised machine learning methods can perform the classification of breast cancer subtypes according to Luminal A, Luminal B, basal-like and HER2 subtypes.

Implementation of the machine learning model on proteome data is conducted through R and Python software [4]–[6]. Pre-processing is employed to remove any unwanted data that can cause misleading results. Feature extraction is carried out using unsupervised machine learning. Meanwhile, the classification of breast cancer subtypes is conducted using supervised machine learning.

2. Methodology

This research will specify on classifying the proteome data into breast cancer subtypes using the approaching of machine learning, which are unsupervised and supervised machine learning methods. Both unsupervised and supervised machine learning is conducted with different purposes. There are 12553 protein features from different genes of 80 breast cancer patients. These genes of breast cancer need to be classified into four different subtypes which are Basal-like, Her2-enriched, Luminal A and Luminal B. Method used for unsupervised machine learning to cluster the proteomes data is K-means. Meanwhile, Random Forest (RF), Gradient Boosting Machine (GBM) and Deep Neural Network (DNN) are used to classify using supervised machine learning [7]–[10]. The flow of the research will be done as shown in Figure 2.

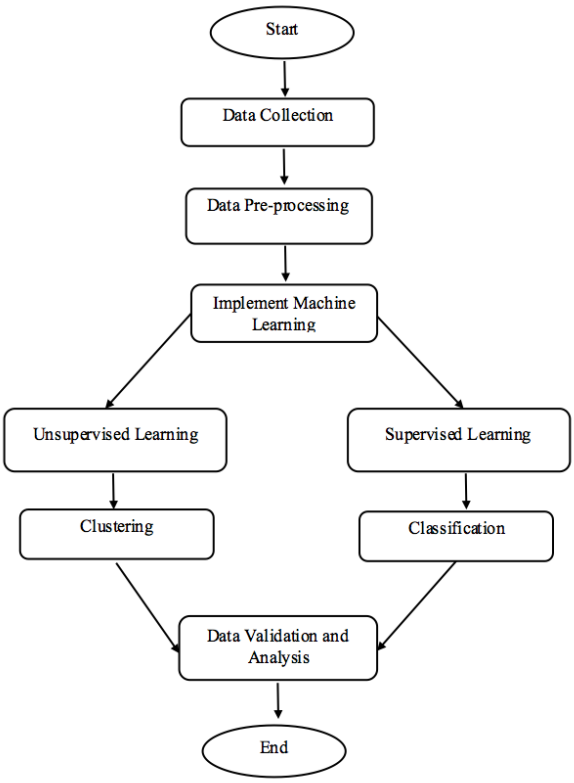


Fig. 2. Visualization of the research flow

2.1 Data Collection

This data set contains published iTRAQ proteome profiling of 77 breast cancer samples generated by the Clinical Proteomic Tumor Analysis Consortium (NCI/NIH) [11,12]. It contains expression values for ~12.000 proteins for each sample, with missing values present when a given protein could not be quantified in a given sample. The first data set contains 12553 genes of 80 patients with breast cancer. All genes are expressed in the form of numbers after going through mass spectrometry procedure. The second data set contains all clinical information of those particular patients [13].

2.2 Data Pre-Processing

2.2.1 For unsupervised machine learning

Pre-processing takes place on both datasets before begin to analyze. This is to ensure any unwanted data that can produce misleading results is removed. Proteomes data is raw data that contains missing values. The objective of pre-processing is to handle the missing value in the data. Thus, a method of handling missing value in data mining must be applied to the proteomes data. Imputation method is implemented as pre-processing. This method is handling missing values by getting the mean value of those particular features. The mean value is assigned and replaced the missing value.

2.2.2 For supervised machine learning

A different approach of data pre-processing is applied to data for supervised machine learning. The goal of supervised machine learning is to classify 80 samples with 12553 attributes of unique proteins into 4 intrinsic subtypes specifically as Basal-like, HER2-enriched, Luminal A and Luminal B. Despite the data is big and will consume a lot of memory and processing time for a machine learning model to learn, it has a lot of missing values. A proper pre-processing technique is needed to handle the data in order to get an accurate result with less error. Dimension reduction technique is used as the pre-processing approach. Shrinkage method known as Lasso function is employed as a dimensional reduction technique to the proteome data with 12553 protein attributes by using R software. Lasso works by regulating the coefficient estimates [14]–[16]. It shrinks the coefficient estimates towards zero. Hence, attributes with a coefficient estimate of zero are excluded from the final model. This method has cleaned the data and reduces the dimension. From 12553 protein attributes, only 56 attributes are used after implementing Lasso. The 56 proteins are chosen based on the least missing value. For classification purpose in supervised machine learning, these 56 proteins are used as the attribute to classify 80 samples into target molecular subtypes of breast cancer.

2.3 Unsupervised Machine Learning

After data pre-processing stage, feature extraction needs to be done through unsupervised machine learning. The unsupervised algorithm learns from unlabeled data. The goal of this learning is to find a pattern or structure of the data set and group it according to their similarities. The clustering method is commonly used in unsupervised machine learning. It helps to discover the similar parameter of the data set and it is widely used in the analysis of gene expression data [17]. One of the clustering tools used in this analysis is k-means [18,19]. The algorithm begins with selecting k cluster center known as centroids. By assigning a point to the closest centroids, the data

is computed. The objects from the corresponding cluster remain close to each other while remains far from another cluster [20]. The process is continued until no change in the centroids coordinates is detected. K means can be said to be the chosen algorithm to reduce the dimension of a data set. This can result in lowering the usage of memory and a short time in running the data analysis [21, 22].

2.4 Supervised Machine Learning

Supervised machine learning is used to predict the input classification based on the labeled output [23]. For supervised machine learning, Random Forest (RF), Gradient Boosting Machine (GBM) and Deep Neural Network (DNN) technique are used. However, to increase the precision of supervised classification, cross-validation is employed for the data. From Figure 3, the supervised machine learning algorithm begins with train and test data. 3 different methods are employed to the classification model before class prediction takes place. In the end, data can be classified after being tested and trained and the accuracy of the respective methods is calculated.

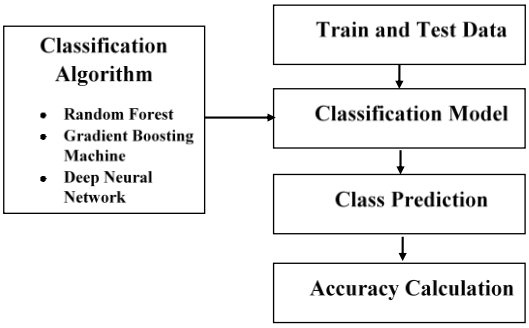


Fig. 3. Supervised machine learning algorithm workflow

2.4.1 Random Forest (RF)

Random Forest is usually used to solve problems regarding classification and regression. Forest is created with a number of trees in this algorithm. Instances or attributes of the sample is placed as the root of the trees. Data in each attribute will be split into a subset for training. Training model can be used to predict class or target variables by learning from training data. The data will be continuously split for each subset until leaf nodes are created at all branches of the tree [24]. Each internal node of the tree represents the attributes and each leaf node represents a class label. RF is able to handle unbalanced data in which data contains missing values [25]. Another advantage of RF is this method can handle multi-dimensional data [26]. Higher accuracy can be achieved by increasing the number of trees in the forest.

2.4.2 Gradient Boosting Machine (GBM)

There are three elements involved in GBM algorithm which are optimizing loss function, make predictions by using weak learner, and development of an additive model to add weak learner in order to reduce loss function [27]. The goal of GBM is to construct a prediction model based on weak learner such as decision trees. GBM can fit the training data instantly and reduce overfitting to improve performance. To enhance GBM performance, a method of tree constraints, shrinkage, random sampling, penalized learning can be applied [28-30].

2.4.3 Deep Neural Network (DNN)

Meanwhile, DNN use feed forward network with hidden layers. Usually, a neural network is divided into three layers which are input layer, hidden layer, and output layer. Basically, the architecture of the deep neural network is similar with the conventional neural network but the main difference between them is that DNN consists of more than one hidden layers, which enable it to perform deeper and wider learning [10]. Different layers carry out different roles. Hidden layer neural network needs an activation function to calculate the hidden layer output by using the input layers during training [31]. The activation function is used in DNN to learn the data and mapping it into input and output variables.

2.5 Cross Validation

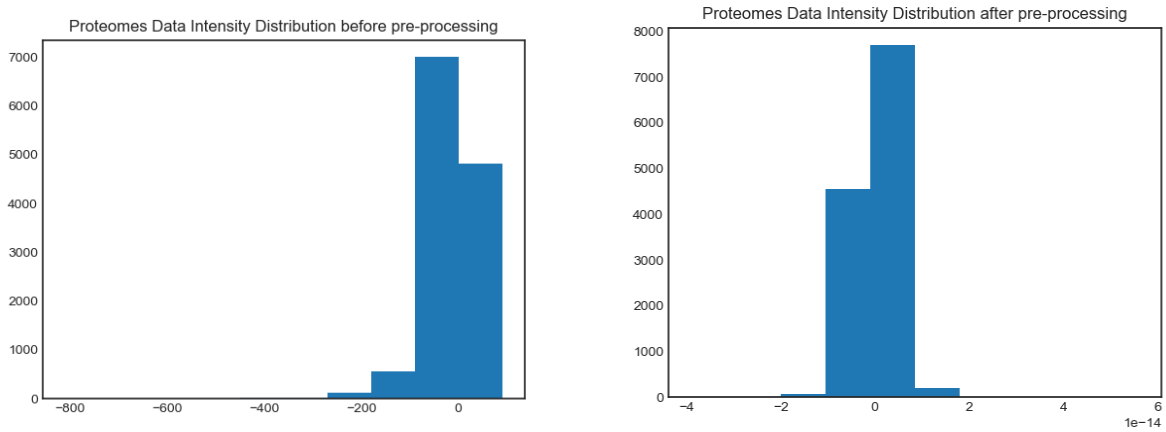
To conduct classification using supervised machine learning, it is advisable to use cross validation in order to measure the accuracy and prediction error [32]. Higher accuracy and lower prediction error indicate a better classification method. K-fold cross validation is used in this study, where K-fold is employed by dividing the dataset into *k* different folds [33]. Next, the *k-1* subset is trained and the rest are used for testing. The process is repeated until all folds are being trained. By using RF, GBM and DNN, molecular of breast cancer can be classified into the corresponding subtypes which are Luminal A, Luminal B, Basal-like, HER2-enriched. The input will be sorted out according to respective characteristics that fit the breast cancer subtypes that remark as output. As the final step, the accuracy of RF, GBM and DNN are determined and compared.

3. Results

This section discusses the results obtained from machine learning classification results. The results for data pre-processing, exploratory data analysis, unsupervised and supervised machine learning methods are discussed in the next subsection.

3.1 Data Pre-Processing

The dimension of the raw data will be reduced, result in smaller data dimension with complete data information ready for further analysis, as shown in Figure 4.



(a) (b)

Fig. 4. Intensity of the proteome data (a) before pre-processing (b) after pre-processing

3.2 Exploratory Data Analysis

Clinical data of 80 patients who have been diagnosed with breast cancer are compiled. From the data, the average age of breast cancer patients at their initial pathologic diagnosis with breast cancer is 58 that vary from 36 until 88 years old. There are 2 male patients diagnosed with breast cancer out of 80 patients. Referring to Figure 5, there are total 19 patients, 13 patients, 22 patients and 26 patients for Basal-like, HER2-enriched, Luminal A and Luminal B subtypes respectively with both 2 male patients are diagnosed with Luminal B breast cancer subtypes.

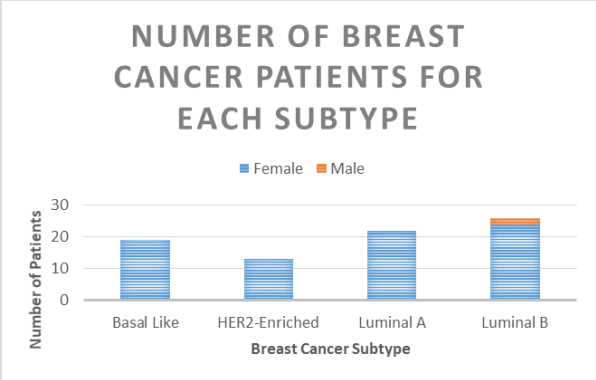


Fig. 5. Number of breast cancer patients for each subtype

Figure 6 has shown that majority of the patients are initially diagnosed with breast cancer at stage 2. Based on the TNM classification, tumor size start to develop up until 5cm and lymph node metastases start to present. Some of the patients are diagnosed at stage 3 and 1 patient is initially diagnosed with stage 4 of breast cancer. Stage 4 of breast cancer indicates that the cancer cells have spread to another area of the body. Once the patients have been diagnosed with breast cancer, patients need to immediately undergo specific treatment regardless of the breast cancer stage at the moment. However, breast cancer treatment differs from one patient with another in term of their subtypes.

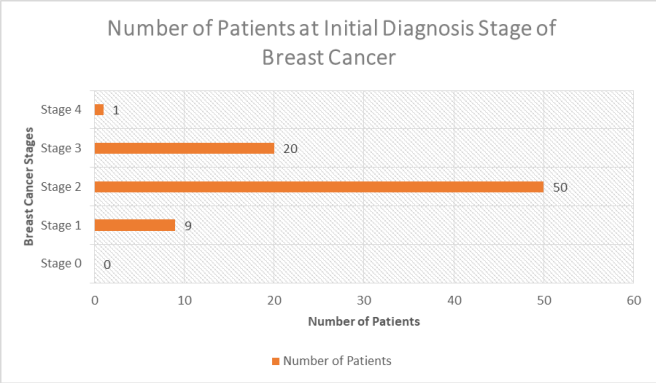


Fig. 6. Number of patients at initial diagnosis stage of breast cancer

As stated earlier, molecular subtypes of breast cancer are divided into 4 types which are Basal-like, HER2-enriched, Luminal A and Luminal B. all these subtypes differ from one another based on their protein characteristics such as the absence or presence of hormone-receptor like estrogen and progesterone. The protein level in each subtype can determine the prognosis of breast cancer patients. Generally, Luminal A has the best prognosis compared to other subtypes. Vital status of

breast cancer patients is shown in Table 1. Deceased patients occur among patients with Basal-like, Luminal A and Luminal B molecular subtypes. There are no deceased patients recorded for HER2-enriched. Hence, it makes the survival rate of HER-enriched patients is the highest which is 100%. Luminal A has a survival rate of 90.9% while Basal-like has a survival rate of 89.7%. The least survival rate is coming from patients with Luminal B which is 88.46%.

Table 1
Vital status and survival rate

Vital Status	*Basal-like*	*HER2*	*Luminal A*	*Luminal B*
Living	17	13	20	23
Deceased	2	0	2	3
Survival Rate	89.47%	100%	90.9%	88.46%

Identification of breast cancer molecular subtypes at early diagnosis is very important for assigning a specific treatment. Each subtype has its own unique protein characteristics that can determine further treatment for the patients. This is why early detection is crucial in order to increase the survival rate among breast cancer patients.

3.3 Unsupervised Machine Learning for Clustering

K-means algorithm will classify the data into the desired clustering according to their similar characteristics. The desired cluster needs to be fixed. To classify into breast cancer subtypes, k=5 is set. Initialization takes place by choosing the centroids. The classification will examine the point in the data set and assign them to the nearest centroid. Grouping based minimum distance to the centroids will continue until no change in centroid coordinates. Value of K is chosen depends on the number of output predicted. K-Means is carried out using 3 different features specifically as ER status, PR status and HER2 status as in Figure 7, Figure 8 and Figure 9, respectively.

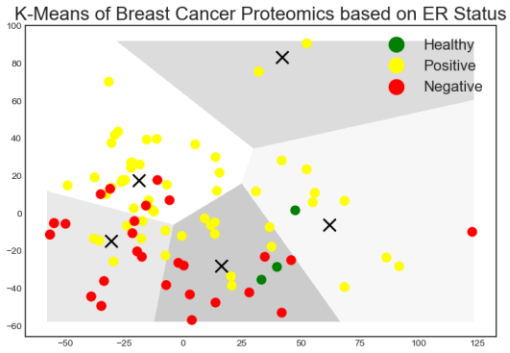


Fig. 7. 5 clustering groups are generated based on the ER Status using K-means

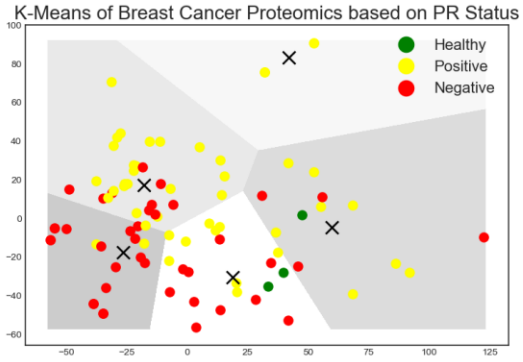


Fig. 8. 5 clustering groups are generated based on the PR Status using K-means

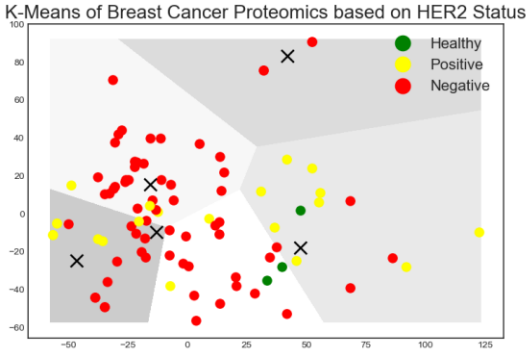


Fig. 9. 5 clustering groups are generated based on the HER2 Status using K-means

Five clustering groups are remarked with 'X' sign indicates 5 centroids. Every protein will be assigned to the respective 'X' sign according to their similar characteristics values. There are 3 protein values from the healthy individual act as a marker for further analysis and comparison. K-means will not classify the proteins into the output target, hence Luminal A, Luminal B, Basal-like and HER2-enriched subtypes cannot be determined yet. However, the centroid value is chosen based on the output predicted that is 4 subtypes and healthy protein type. For K-means using ER status in Figure 7, the distribution of ER hormone present in the protein is clearly separated. PR status in Fig.8 also has quite a similar pattern of distribution. In Fig.9, the distribution pattern of HER2 status is quite mingled but still in an organized clustering pattern. The information that can be analyzed through Figure 7, Figure 8 and Figure 9 is the presence of ER status, PR status and HER2 status can determine the protein clustering. It means the protein characteristics from four intrinsic subtypes also depends on ER, PR and HER2 status. Hence, this three status can be assigned as the correct features in order to cluster protein based on their similar behavior or characteristics.

3.4 Supervised Machine Learning for Classification

Classification of breast cancer molecular subtypes needs to be carried out using supervised machine learning. A machine learning model needs to build in order to train and test the data. Each machine learning model has its own algorithm in classifying the data. Three supervised machine learning methods are implemented to analyze and train the proteome data. 12553 of unique proteins

attributes has been minimized to 56 proteins during pre-processing. 80% of the samples are trained and another 20% is spared for testing for 80 samples of breast cancer patients. Based on the subtyped specified as output, the training data will learn the protein characteristics for each patient. The characteristics will be matched with the specified subtypes. Next, the training of data takes place. Learning during training data is applied during testing to get the output. The output attained is compared to the output specify where the accuracy of the learning algorithm is determined. Random Forest (RF), Gradient Boosting Machine (GBM) and Deep Neural Network (DNN) have been chosen as an algorithm used to build the machine learning model in order to classify the subtypes.

Results from the algorithm are achieved in the form of confusion matrix of training data, predicted error and cross validation accuracy. Cross validation (CV) takes place by dividing the data into test and training. The model will run the data once again to get the accuracy for 5 times as K is set as 5 times in K-folds. Results from the analysis are shown below.

3.4.1 Random Forest (RF)

RF model is built according to its components. The components contain in RF are a number of trees and number of internal trees. Number of trees in this RF model is 50 and the number of internal trees is 200. Table 2 shows the accuracy of every validation folds. Cross validation in the first fold is 0.875, the second fold is 1.0, the third fold is 0.842, the fourth fold is 0.867 and fifth fold is 0.882. Mean accuracy after cross validation in RF is 0.893 which equivalent to 89.3%. As RF requires a number of trees to build the machine learning model, training classification error with respect to a number of trees has been calculated in Figure 10. This has proved that the higher number of trees in RF can increase the accuracy of classification precision [27].

Table 2
RF accuracy for 5-fold cross-validation

	Mean	***CV1***	***CV2***	***CV3***	***CV4***	***CV5***
Accuracy	0.893	0.875	1.000	0.842	0.867	0.882

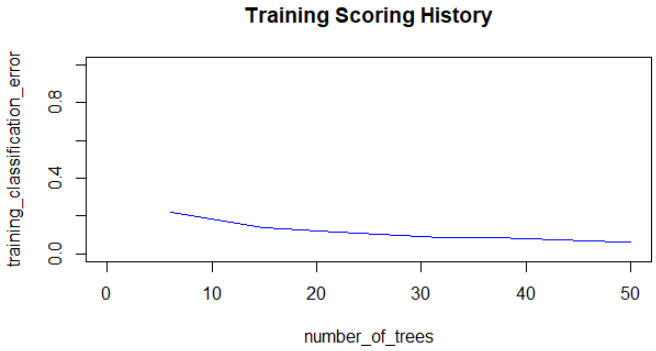


Fig. 10. Training classification error in RF with respect to the number of trees

3.4.2 Gradient Boosting Machine (GBM)

As GBM operates with weak learner such as decision tree, a number of trees is used as the component consists of this model. Similar to RF, a number of trees in this model is 50 and the number of internal trees is 200. Table 3 shows the accuracy for every validation folds occurred in GBM model. Cross validation in the first fold is 0.727, second fold is 0.500, the third fold is 0.882, the fourth fold

is 0.850 and fifth fold is 0.846. Mean accuracy after cross validation in RF is 0.761 which equivalent to 76.1%. Like RF, GBM also requires a number of trees to build its machine learning model. Hence, training classification error with respect to a number of trees has been calculated in Figure 11. With a slightly difference compared to RF, GBM training score graph in Figure 11 can be said as inversely proportional at first then remain constant. It abides the hypothesis that a higher number of trees in weak learner in GBM can increase the accuracy of classification precision. However, it is only applicable for a small number of trees. As the numbers of trees getting higher, the error is constant.

Table 3
GBM accuracy for 5-fold cross-validation

	Mean	***CV1***	***CV2***	***CV3***	***CV4***	***CV5***
Accuracy	0.761	0.727	0.500	0.882	0.850	0.846

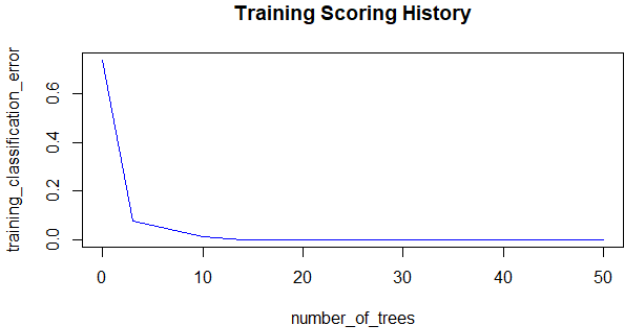


Fig. 11. Training classification error in GBM with respect to the number of trees

3.4.3 Deep Neural Network (DNN)

DNN needs activation function in order to learn the data and mapping it into input and output variables. In this model, Maxout activation function has been used along with epoch 1000. Accuracy for 5 folds of cross validation takes place in DNN model is recorded. According to Table 4, the accuracy of first, second, third, fourth, and fifth folds are 1.000, 0.846, 0.822, 1.000 and 1.000 respectively. The mean accuracy after cross validation in DNN is 0.969 which equivalent to 96.9%. DNN operates with epoch set for every model. In this DNN machine learning model, the epoch is set for 1000. Figure 12 displays training classification error vary with a number of the epoch.

Table 4
DNN accuracy for 5-fold cross-validation

	Mean	***CV1***	***CV2***	***CV3***	***CV4***	***CV5***
Accuracy	0.969	1.000	0.846	0.882	1.000	1.000

Fig. 12. Training classification error in DNN with respect to the number of epochs

3.5 Discussion

Dealing with pre-processing is very crucial before data can be used for clustering or classification. Raw data are subjected to be pre-processed using a specific method. Manipulation of data depends on how the objectives need to be achieved. Data imputation has been used for unsupervised machine learning while data reduction is used for supervised machine learning. After data is being pre-processed, then it is allowed to further with the subsequent method. Clustering and classification of proteome data have been conducted through a machine learning method. Unsupervised machine learning is applied to proteome data by finding correct features. Meanwhile, supervised machine learning satisfies the need for classifying proteome data into four intrinsic subtypes which are Basal-like, HER2-enriched, Luminal A and Luminal B.

Unsupervised machine learning learns with the attributes without output. From the attributes characteristics, it clusters the samples based on similar behavior of attributes. In this project, 12553 unique protein are used as attributes for 80 samples of breast cancer patients. K-mean is implemented to find the correct features of protein clustering. Value of centroid K is set to 5 indicate 5 predictive clusters which are 4 subtypes and a healthy protein. K-means model is employed to proteome data with ER status, PR status and HER2 status features as variables. This result satisfies the condition that the expression of ER, PR and HER2 can be as indicators for particular subtypes of breast cancer [6].

Prediction of molecular subtypes is important to increase mortality and survival rate. Besides, prognosis performance and specific treatment for respective subtypes also can be determined. Identification of breast cancer subtypes might be useful to discover new therapy. The study of protein to improve breast cancer therapy can lead to prolong survival rate. This shows the importance of protein study in predicting molecular subtypes among breast cancer patients.

Implementation of supervised machine learning to classify proteome data is done though three proposed method specifically as RF, GBM and DNN. These three method differ to one another in term of components of the model and approach used to analyze proteome data. However, all these three methods are applicable to achieve the goal. These three methods used in the machine learning model classify proteome data by optimizing training and testing data through K-fold cross validation where K is set to 5 times. Hence, the data will be split into 5 part. 80% from the samples (4 folds) which equal to 64 samples are undergoing training at one time and the rest 20 % of the samples (1 fold) which equivalent to 16 samples are undergoing testing at the same time. To increase the class prediction precision, cross validation will be continuously running for 5 times until all folds have undergone training and testing.

Results from three classification models are produced in the form of accuracy. The confusion matrix is a quick reference to describe the performance of a classifier. It shows the error occurs in the training algorithm on the applicable methods. Based on the training confusion matrix generated for respective models, it can be analyzed that there is no error occur if the proteome data are trained by using GBM and DNN. Data training using RF model has 0.625 of the predicted error value. After the data undergo cross validation process, accuracy for every fold is generated. A number of folds used in all three machine learning model is set to 5 times. Hence, every model has 5 accuracy value and mean accuracy is determined. The highest accuracy of testing and training data happens in DNN model. The accuracy of the DNN model is the highest which is 96.9% followed by GBM (76.1%) and RF (89.3%). This can be supported that DNN has the highest accuracy of testing and training model for proteome data [33].

4. Conclusions

Breast cancer developed from mutated cells that spread in the body. The mutated cells are not damaged hence spread to the adjacent cells. It continuously damages the normal cells if no prevention action or treatment is taken. The limitation of treatment based on breast cancer staging has led to the importance of identifying breast cancer subtypes. Predicting subtypes is crucial for clinical therapy as well as to increase the survival rate.

In respect to the limitation identified, molecular subtypes of breast cancer need to determine. The main purpose of this project is to classify proteomes data based on the protein features to obtain subtypes characteristics. The machine learning method is implemented along the analysis.

Pre-processing is employed to proteome data to remove unwanted values that can cause misleading in the result. Missing values are handled by using imputation method of mean substitution. In order to classify 80 patients with 12553 unique protein features, the machine learning model is introduced. Unsupervised learning and supervised learning are used to cluster and classify the proteomes data respectively.

Clustering does not require subtypes as output as the clustering algorithm is based on similar features is grouped together. By using the K-means algorithm, unsupervised learning is employed to the proteome data to find the correct features. Finding correct features is important to improve prediction precision. On the other hand, classification requires subtypes as the output for training and testing purpose. To increase the precision of the machine learning model of classification, K-fold is implemented during testing and training. Random Forest, Gradient Boosting Machine and Deep Neural Network are three methods used in the classification of molecular subtypes of breast cancer using proteome data.

Proteome data undergo dimension reduction before being processed with supervised machine learning. Lasso function helps to choose 56 protein attributes with the least missing values. Supervised machine learning takes place on 56 protein to classify into 4 subtypes. The evaluation of the result is observed by comparing the accuracy of the machine learning model to identify the best method for protein classification. DNN has the highest accuracy compared to GBM and RF. This shown proteome data learned best in DNN algorithm.

For conclusion, the purpose of this study is to build a machine learning model that able to learn protein behavior in order to predict the subtypes. As breast cancer therapy plays an important role in increasing the mortality of breast cancer function, the study of breast cancer molecular subtypes need to be extended through protein study. Machine learning model needs to build in order to learn protein behavior for classification purpose. In the end, the most suitable machine learning model is successfully determined to learn protein characteristics in a better way to predict subtypes. Hence,

proteome data able to classify into Luminal A, Luminal B, basal-like and HER2 subtypes through machine learning application. The objectives of this project have been achieved.

The limitation of this study is it contains large raw data with a lot of missing values. Besides consuming a lot of memory during the machine learning process, the missing values need to handle properly to avoid any misleading result. Raw data requires a distinct method to be employed as pre-processing before the data can be used for machine learning analysis. Hence, the decision of method used it must based on the machine learning model that will be created. Different machine learning approach requires the different technique of data pre-processing.

This study can be further extended by designing a system for subtype detection. By learning the protein characteristics of each subtype, the system of subtype identification can be developed. This model can work by introducing the patient's protein and generate the subtypes based on the proteomes data. The range of every protein intensity must be determined in order for the model to predict the output.

References

[1] M. Banning, "The experience of living with breast cancer," *Breast Journal*, vol. 14, no. 3. pp. 317–318, 2008.
[2] T. McKee and J. R. McKee, "The Molecular Basis of Life, Fifth Edition," in *The Molecular Basis of Life*, 2012, pp. 1–60.
[3] S. E. Barnato and W. J. Gradishar, "Biological subtypes of breast cancer," *Med. Radiol.*, vol. 63, 2015.
[4] A. Chang, "R for Machine Learning," *Predict. Mach. Learn. Stat.*, 2012.
[5] J. Monogan, "An Introduction to R," *User Man.*, no. October, pp. 1–12, 2009.
[6] en.wikibooks.org, "Python Programming," *Supercomputer*, no. February, p. 137, 2007.
[7] Breiman, Leo. "Random forests." *Machine learning* 45, no. 1 (2001): 5-32.
[8] Natekin, Alexey, and Alois Knoll. "Gradient boosting machines, a tutorial." *Frontiers in neurorobotics* 7 (2013): 21.
[9] Ma, Junshui, Robert P. Sheridan, Andy Liaw, George E. Dahl, and Vladimir Svetnik. "Deep neural nets as a method for quantitative structure–activity relationships." *Journal of chemical information and modeling* 55, no. 2 (2015): 263-274.
[10] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." *nature* 521, no. 7553 (2015): 436.
[11] Edwards, Nathan J., Mauricio Oberti, Ratna R. Thangudu, Shuang Cai, Peter B. McGarvey, Shine Jacob, Subha Madhavan, and Karen A. Ketchum. "The CPTAC data portal: a resource for cancer proteomics research." *Journal of proteome research* 14, no. 6 (2015): 2707-2713.
[12] Rudnick, Paul A., Sanford P. Markey, Jeri Roth, Yuri Mirokhin, Xinjian Yan, Dmitrii V. Tchekhovskoi, Nathan J. Edwards et al. "A description of the clinical proteomic tumor analysis consortium (CPTAC) common data analysis pipeline." *Journal of proteome research* 15, no. 3 (2016): 1023-1032.
[13] Mertins, Philipp, D. R. Mani, Kelly V. Ruggles, Michael A. Gillette, Karl R. Clauser, Pei Wang, Xianlong Wang et al. "Proteogenomics connects somatic mutations to signalling in breast cancer." *Nature* 534, no. 7605 (2016): 55.
[14] Tibshirani, Robert. "The lasso method for variable selection in the Cox model." *Statistics in medicine* 16, no. 4 (1997): 385-395.
[15] Wang, Hansheng, Guodong Li, and Chih-Ling Tsai. "Regression coefficient and autoregressive order shrinkage and selection via the lasso." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69, no. 1 (2007): 63-78.
[16] Chen, Si-Bao, Chris HQ Ding, and Bin Luo. "An algorithm framework of sparse minimization for positive definite quadratic forms." *Neurocomputing* 151 (2015): 223-230.
[17] Dai, Xiaofeng, Liangjian Xiang, Ting Li, and Zhonghu Bai. "Cancer hallmarks, biomarkers and breast cancer molecular subtypes." *Journal of cancer* 7, no. 10 (2016): 1281.
[18] Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28, no. 1 (1979): 100-108.
[19] S. Mannor, X. Jin, J. Han, X. Jin, J. Han, X. Jin, J. Han, and X. Zhang, "K-Means Clustering," in *Encyclopedia of Machine Learning*, 2011, pp. 563–564.
[20] M. M. Kantardzic and J. Zurada, *Next generation of data-mining applications*. 2005.
[21] Ghosh, Soumi, and Sanjay Kumar Dubey. "Comparative analysis of k-means and fuzzy c-means algorithms." *International Journal of Advanced Computer Science and Applications* 4, no. 4 (2013).
[22] Dubey, Ashutosh Kumar, Umesh Gupta, and Sonal Jain. "Comparative Study of K-means and Fuzzy C-means Algorithms on The Breast Cancer Data." *International Journal on Advanced Science, Engineering and Information*

Technology 8, no. 1 (2018): 18-29.
[23] D. Bzdok, M. Krzywinski, and N. Altman, "Machine learning: supervised methods," *Nat. Methods*, vol. 15, p. 5, 2018.
[24] S. T. Selvi., P. Karthikeyan, A. Vincent, V. Abinaya, G. Neeraja, and R. Deepika, "Random Forest algorithm," *Gait Posture*, vol. 38, pp. S42–S43, 2013.
[25] Tang, Fei, and Hemant Ishwaran. "Random forest missing data algorithms." *Statistical Analysis and Data Mining: The ASA Data Science Journal* 10, no. 6 (2017): 363-377.
[26] Pal, Mahesh. "Random forest classifier for remote sensing classification." *International Journal of Remote Sensing* 26, no. 1 (2005): 217-222.
[27] Friedman, Jerome H. "Greedy function approximation: a gradient boosting machine." *Annals of statistics* (2001): 1189-1232.
[28] Brownlee, Jason. "A gentle introduction to the gradient boosting algorithm for machine learning." *Machine Learning Mastery. Nov* 9 (2016).
[29] Chen, Xing, Li Huang, Di Xie, and Qi Zhao. "EGBMMDA: extreme gradient boosting machine for MiRNA-disease association prediction." *Cell death & disease* 9, no. 1 (2018): 3.
[30] Babajide Mustapha, Ismail, and Faisal Saeed. "Bioactive molecule prediction using extreme gradient boosting." *Molecules* 21, no. 8 (2016): 983.
[31] Koutsoukas, Alexios, Keith J. Monaghan, Xiaoli Li, and Jun Huan. "Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data." *Journal of cheminformatics* 9, no. 1 (2017): 42.
[32] B. Cooil, R. S. Winer, and D. L. Rados, "Cross-Validation for Prediction," *J. Mark. Res.*, vol. 24, no. 3, pp. 271–279, 1987.
[33] Browne, Michael W. "Cross-validation methods." *Journal of mathematical psychology* 44, no. 1 (2000): 108-132.