Ann Med Res 2020;27(10):2803-6

Annals of Medical Research

DOI: 10.5455/annalsmedres.2020.02.165

Original Article

Prediction of breast cancer subtypes based on proteomic data with deep learning

Seyma Yasar, Cemil Colak, Saim Yologlu

Department of Biostatistics and Medical Informatics, Faculty of Medicine, Inonu University, Malatya, Turkey

Abstract

Aim: Although new advances in diagnosis and treatment have increased, breast cancer is still an important cause of morbidity and mortality today. Proteomics, which collectively deals with relevant information about proteins, is one of the important areas of study that has been emphasized recently. It is a machine learning class that uses many layers of nonlinear processing units for deep learning, feature extraction and conversion. The aim of this study is to classify the molecular subtypes (Basal-like, human epidermal growth factor receptor 2 (HER2)-enriched, Luminal A, Luminal B) of breast cancer with the deep learning algorithm designed by using proteomic data.

Material and Methods: The data set used in this study consists of published Isobaric tags for relative and absolute quantitation (iTRAQ) proteome profiling of 77 breast cancer samples by the Clinical Proteomic Tumor Analysis Consortium (NCI/NIH). The missing values in the data were completed with the mean substitution method. "Lasso Regression Model" was used in the selection of variables and after repeating 100 times with 10 times cross-validation method. Finally, the deep learning algorithm has been used to classify the molecular subtypes of breast cancer.

Results: The overall accuracy rate of the proposed model in classifying breast cancer are found to be 91.53%. The performance of this model for classifying molecular subtypes of breast cancer was calculated as accuracy %96.43, F-score %93.33, MCC %91.29, G-mean %93.54 for Basal-like, accuracy %94.74, F-score %84.21, MCC %81.23, G-mean %92.30 for HER2-enriched, accuracy %98.18, F-score %96.97, MCC %95.76, G-mean %98.71 for Luminal A and accuracy 93.10%, F-score 88.89%, MCC 83.89%, G-mean 91.89% for Luminal B, respectively.

Conclusion: The model designed using the deep learning algorithm has been found to perform quite well in classifying the molecular subtypes of breast cancer. In further studies, different deep learning architectures can be used to classify the molecular subtypes of breast cancer with higher accuracy.

Keywords: Breast cancer; classification; deep learning; proteomics

INTRODUCTION

Breast cancer is one of the most common cancer types that affect women in the world and in our country and causes death in women. Early diagnosis and prevention of breast cancer are very important. Because of the primary tumor metastasis to other organs, many patients succumb to the disease in the advanced stage. Current methods used to detect breast tumors are primarily mammography. However, mammography also has several limitations. Small lesions are often overlooked and may not be seen especially in young women with dense breast tissue (1). Some subtypes of malignant breast tumors have been defined according to their histological and molecular features. The subtypes included in this study are Basal-like, HER2 overexpressed, Luminal A and Luminal B (2).

In recent years, researchers have focused on genomic studies, and have progressed rapidly in the identification of genes especially related to disease and other processes. However, ongoing scientific studies have failed to provide sufficient information about the extent to which the genes identified in an organism use the organism. For this reason, proteomics, which is the continuation of genomic studies, has gained importance in explaining biological functioning. Proteomics approaches explore how proteins are secreted from the cell, what function they perform in the cell, and their mutual communication, how they change after cell damage, and the role they can play as bio-markers of diseases. Proteins are responsible for the functions and cell phenotypes of biological systems in the organism. Cancer cells can be separated from normal cells by proteins that they secrete. Cancerous cell proliferation and excessive protein secretion are often interrelated and can be detected in the blood early period when other methods inadequate (3).

Received: 23.02.2020 **Accepted:** 28.09.2020 **Available online:** 22.10.2020
Corresponding Author: Seyma Yasar, Department of Biostatistics and Medical Informatics, Faculty of Medicine, Inonu University, Malatya, Turkey **E-mail:** seyma.yasar@inonu.edu.tr

Deep learning (also deep structured learning, hierarchical learning or deep machine learning) is the work area that includes artificial neural networks and similar machine learning algorithms containing one or more hidden layers. It uses many non-linear processing unit layers for deep learning, feature extraction, and conversion. Each consecutive layer takes the output from the previous layer as input. There are many algorithms used in deep learning. Feedforward networks used in this study are a type of multi-layer artificial neural networks. Feedforward neural networks allow for unidirectional signal flow. The input layer moves through the hidden layers to the output layer and the information is always transmitted to the front layer. The input layer transmits the information it receives from outside to the hidden layer neurons without making any changes. The information is processed in the hidden and output layer, and the network output is determined (4).

The aim of this study is to classify the molecular subtypes of breast cancer based on proteomic data with the help of the model created with the deep learning approach.

MATERIAL and METHODS

Dataset

The data set used in the study consists of the published iTRAQ proteome profiling of the sample from 77 breast cancer patients created by Clinical Proteomic Tumor Analysis Consortium (NCI/NIH). The data set contains expression values for approximately 12,000 proteins per sample. The sample contains a missing value if it does not contain protein (5).

Feature Selection

In determining the missing values in the data set used in the study, the percentage of the missing value of each variable are first determined. Then, variables with this missing percentage greater than 0.25 are excluded from the study. After this process, the "mean substitution" method was used to complete the missing observations in the remaining variables in the data set. In the mean substitution method, the average of the observations available for each variable is assigned to the place of the missing data for this variable, thereby achieving a complete data set. For each variable, as a result of this assignment, the average of the complete observations is equal to the average obtained from the completed data set (6).

The approach used in variable selection is the Penalized Logistic Regression model. Penalized Logistic Regression adds a penalty parameter (also known as regularization) to logistic regression, which has a lot of variables so that the coefficients of variables that contribute less to the model are zero. The most frequently used punished regression model is Lasso regression. The coefficients of some variables that contribute less to the Lasso regression are forced to be exactly zero. Only the most important variables are kept in the final model. The lasso logistic regression model established in the study has been applied 100 times, and variables that included more than 15 in the model have been included in the study (7).

Deep Learning

The deep learning model constructed in this study is based on a multi-layer feed-forward artificial neural network trained with stochastic gradient descent using the back-propagation technique. In multi-layer feed-forward neural networks, the information between neurons is in one-way regular layers from input to output. A layer is connected only to the layer that comes after it, and the output of a neuron can only be input to the layer of the neuron that comes after it. There is no connection to another neuron in the peer layer (4). The created model consists of 4 layers, one input, two hidden and one output. The dropout values in the input and hidden layers are 0.50, 0.50 and 0.50, respectively. In the model based on deep learning, "MaxoutWithDropout" is used as an activation function and "ADADELTA" is used as the optimization method (8). Hyperparameter values related to the ADADELTA optimization method are given in Table 1.

Table 1. Hyperparameters for ADADELTA algorithm

Hyperparameter	Value
epsilon	0.00000001
rho	0.99

Cross-Entropy function is used as a loss function in the output layer of the created model. The Loss function, which measures the error rate of the designed model, is also a function that measures the performance of the model. The Loss function basically calculates how much the model's estimate differs from the ground truth (9). For the validation of the model is used 10-fold cross-validation method (10).

Evaluation Metrics

The performance of the deep learning model used in the study was evaluated using accuracy, F-Score, Matthew's Correlation Coefficient (MMC) and G-mean performance metrics. DTROC: Diagnostic Tests and ROC Analysis Software was used to calculate these performance metrics and 95% confidence intervals (11). The formulas for these performance criteria are given below.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

$$\text{G-mean} = (\text{Sensitivity}^{*}\text{Specificity})^{1/2}$$

$$\text{F-score} = 2^{*}TP/(2^{*}TP+FP+FN)$$

$$\text{MCC} = (TP^{*}TN-FP*FN)\backslash ((TP+FP)^{*}(TP+FN)^{*}(TN+FP)^{*}(TN+FN))^{1/2}$$

In the formulas above, TP represents the true positive number, TN defines the true negative number, FP explains the false positive number and FN describes the false negative number.

RESULTS

The highest accuracy value was obtained as 0.91 in the training phase of the deep learning model based on the proteomic data of breast cancer. The classification performances of the constructed model

in the test data set are given in Table 2. Table 2 shows performance metrics such as accuracy, F-score, Matthews Correlation Coefficient (MCC) and G-mean with 95% confidence interval for the deep learning model.

Table 2. Performance metrics for the model with 95% confidence interval

Metrics	Performance Metric Value (%) (95% CI)			
	Basal-like	HER2-enriched	Luminal A	Luminal B
Accuracy	94.74 (89.04-100.01)	96.43 (91.69-100.01)	96.43 (91.69-100.01)	94.74 (89.04-100.01)
F-Score	89.66 (81.88-97.43)	90.00 (82.35-97.65)	94.12 (88.11-100.00)	91.43 (84.29-98.57)
MCC	86.22 (77.43-95.02)	87.83 (79.48-96.17)	91.55 (84.46-98.65)	87.71 (79.33-96.09)
G-Mean	91.98 (85.05-98.91)	93.83 (87.69-99.97)	95.76 (90.62-100.01)	94.56 (88.77-100.00)

In the deep learning model based on proteomics data, the protein code and relative importances of the top 10 proteins, which are also important in classifying subtypes of breast cancer, are given in Table 3.

Table 3. The protein code and relative importances of the top 10 proteins found important in classifying subtypes of breast cancer in the model

Protein Code	Relative Importance	Percentage
NP_005484	1.000	0.03057
NP_001257810	0.889	0.02719
NP_660208	0.864	0.02644
NP_004243	0.805	0.02464
NP_001229372	0.799	0.02445
NP_001371	0.794	0.02427
NP_004522	0.793	0.02425
NP_004453	0.792	0.02421
NP_003144	0.781	0.02390
NP_653081	0.771	0.02360

DISCUSSION

Cancer is one of the health problems that are increasing all over the world and result in death. While the most common cancer rate in men is known as lung cancer, it is known as breast cancer in women. There are many factors that cause breast cancer. These factors consist of exchangeable factors such as diet, having children, breastfeeding, smoking, alcohol use, as well as unchangeable factors such as gender, family history, age, race, dense breast tissue, genetic predisposition. However, pain, unusual shape, tenderness in the nipples, change in breast skin tone can be listed as symptoms that give information about the presence of breast cancer. However, most of these symptoms are not seen in the first stage of cancer. Therefore, it is possible that the cancer will be in an advanced stage when breast cancer is diagnosed (12).

The aim of this study is to classify the molecular subclasses of breast cancer using a deep learning algorithm based on proteomic data. According to the experimental results, when the performance metrics obtained in classifying the molecular subtypes of breast cancer are analyzed, the classification accuracy values of the subtypes are found to be quite high (greater than 94%).

Deep learning algorithms created using images obtained with different imaging techniques related to breast cancer are frequently used in breast cancer classifications. In a previous study, an AlexNet deep learning algorithm based on thermal images was created and breast cancer classification (normal/suspicious) are made with this model. The proposed model has classified the thermal images of 140 people with an accuracy rate over 90% (13).

In another study, it was aimed to predict the metastasis of axillary lymph nodes (ALN) with the help of breast ultrasound with the model created using the convolutional neural networks, one of the deep learning algorithms. According to the results obtained, the model of convolutional neural networks performed better than radiological models in predicting ALN metastasis of breast cancer (14).

In another study, the architectures and the performances of these two deep learning architectures are compared. As a result of the study, classification of histopathological images according to the performance metric used, the Xception algorithm gave better results (15).

In the studies summarized above, breast cancer classification was made with deep learning models created with the help of data obtained using different imaging techniques. However, studies classifying subtypes of breast cancer using proteomic technologies using deep learning algorithms are limited. Proteomic technologies that enable to obtain molecular information specific to each patient play an important role in the clinical early diagnosis of breast cancer. Today, proteomics technologies make important contributions to the detection of the disease at a treatable stage, to regulate timely treatment protocols, and most importantly to the cancer drug development process. The development of new treatment procedures for breast cancer is expensive and time consuming, as well as requiring clinical trials that often require a large number of patients. Therefore, the analysis of proteomics data with artificial intelligence and deep learning architectures has gained great importance in recent years in the development of new approaches to the diagnosis, treatment and follow-up of diseases, and their clinical practice continues to increase (16).

CONCLUSION

According to the experimental results of the current study, it was concluded that the deep learning model created on breast cancer proteomic data in this study is very successful in classifying subtypes of breast cancer. In further studies, classification can be made by different artificial intelligence and other deep learning approaches using proteomic data of other cancer types.

Competing interests: The authors declare that they have no competing interest.

Financial Disclosure: There are no financial supports.

Ethical approval: Ethics committee approval is not required in this study.

REFERENCES

1. Carney PA, Miglioretti DL, Yankaskas BC, et al. Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. Ann Intern Med 2003;138:168-75.
2. Ünçel M, Aköz G, Yıldırım Z, ve ark. Meme kanserinin klinikopatolojik özelliklerinin moleküler alt tipe göre değerlendirilmesi. Tepecik Eğit Hast Derg 2015.
3. Özenoğlu S, Yıldızhan H, Özel-demiralp D, ve ark. Farklı biyolojik organizmalarda proteomik uygulamalar. Turk Hij Den Biyol Derg 2016;73.
4. Öztürk K, Şahin ME. Yapay sinir ağları ve yapay zekâ'ya genel bir bakış. Takvim-i Vekayi 2018;6:25-36.
5. Breast Cancer Proteomes: Dividing breast cancer patients into separate sub-classes. Available from:https://www.kaggle.com/piotrgrabo/breastcancerproteomes.
6. Somasundaram R, Nedunchezhian R. Missing value imputation using refined mean substitution. IJCSI 2012;9:306.
7. Fonti V, Belitser E. Feature selection using lasso. VU Amsterdam Research Paper in Business Analytics 2017:1-25.
8. Zeiler MD. Adadelta: an adaptive learning rate method. arXiv preprint arXiv:12125701. 2012.
9. De Boer P-T, Kroese DP, Mannor S, et al. A tutorial on the cross-entropy method. Ann Oper Res. 2005;134:19-67.
10. Fushiki T. Estimation of prediction error by using K-fold cross-validation. Stat Comput 2011;21:137-46.
11. Yasar Seyma, Arslan AK, Yologlu S, et al. DTROC: Tanı Testleri ve ROC Analizi Yazılımı [Web-tabanlı yazılım] 2019 [07.17.2019]. Available from: http://biostatapps.inonu.edu.tr/DTROC/
12. Dean A. Primary breast cancer: risk factors, diagnosis and management. Nursing Standard 2008;22.
13. Ekici S, Ünal F. Termografİ ve derİn transfer öğrenme ile meme kanserİ teşhisi.
14. Sun Q, Lin X, Zhao Y, et al. Deep learning vs. radiomics for predicting axillary lymph node metastasis of breast cancer using ultrasound images: don't forget the peritumoral region. Front Oncol 2020;10:53.
15. Yilmaz F, Kose O, Demir A, editors. Comparison of two different deep learning architectures on breast cancer. 2019 Medical Technologies Congress (TIPTEKNO); 2019: IEEE.
16. Alcantara D, Leal MP, García-Bocanegra I, et al. Molecular imaging of breast cancer: present and future directions. Front Chem 2014;2:112.