Journal of AI and Data Mining
Vol 8, No 3, 2020, 371-378.
DOI: 10.22044/JADM.2020.8406.1978

Development of an Ensemble Multi-stage Machine for Prediction of Breast Cancer Survivability

M. Salehi, J. Razmara*, Sh. Lotfi

Department of Computer Science, Faculty of Mathematical Sciences, University of Tabriz, Tabriz, Iran.

Received 09 May 2019; Revised 31 March 2020; Accepted 28 April 2020

**Corresponding author: razmara@tabrizu.ac.ir (J. Razmara).*

Abstract

Prediction of cancer survivability using the machine learning techniques has become a common approach in the recent years. In this regard, an important issue is that preparation of some features may require conducting difficult and costly experiments, while these features have less significant impacts on the final decision and can be ignored in the feature set. Therefore, developing a machine for prediction of survivability, which ignores these features for simple cases and yields an acceptable prediction accuracy, has turned into a challenge for the researchers. In this work, we have developed an ensemble multi-stage machine for the survivability prediction, which ignores difficult features for simple cases. This machine employs three basic learners, namely multi-layer perceptron (MLP), support vector machine (SVM), and decision tree (DT), in the first stage in order to predict the survivability using simple features. If the learners agree on the output, the machine makes the final decision in the first stage; otherwise, for difficult cases, where the output of learners is different, the machine makes decision in the second stage using SVM over all features. The developed model is evaluated using the Surveillance, Epidemiology, and End Result (SEER) database. The experimental results obtained reveal that the developed machine is capable of obtaining a considerable accuracy, while it ignores difficult features for most of the input samples.

Keywords: *Breast Cancer Survivability Prediction, Ensemble Learning Machines, Multi-stage Machines, Feature Selection.*

1. Introduction

Breast cancer has become one of the most common malignancies in females, and it has received much attention in the recent decade due to its increasing morbidity and mortality worldwide. This has led the scientists to explore the novel preventive and therapeutic methods against breast cancer [1]. Due to the high frequency of the breast cancer incidence, it has turned into an important challenge for the medical and even non-medical researchers [2-4]. Generally, cancer-related research works are mostly conducted in medical and biological areas [5]. However, considering the increasing requirement in many scientific fields to analyze data and obtain applicable information from the recorded data, the data-driven statistical research and data mining methods have been widely used as a complement to biological research works [6].

Considering cancer, the outcome of the disease can be predicted based on a set of patient's conditions such as tumor type and results of clinical experiments as well as the stage of the tumor, and the treatment strategy is planned regarding the patient's life expectancy. In other words, it is predicted whether a cancer patient will be alive in the years after tumor diagnosis or not. The prediction can be made through

supervised machine learning techniques, and the problem is proposed as the probability of survival during a specific period after diagnosis. This period has been agreed to be five years (60 months) [7]. The outcome of the machine learning techniques highly depends on the appropriateness of the recorded data for training these machines. As a result, the machine can contribute greatly for an accurate prediction of the patient survivability, and therefore, assist physicians to choose the best treatment strategy for patients [8, 9].

Prediction of the patient death or survival has already been done through such statistical methods as the Kaplan-Meier test and the Cox-propositional hazard [10], which investigate the chance of survival in patients by calculation of the conditional probability. However, after a while, it was proven that the machine learning techniques led to the accurate prediction of survivability among cancer patients. Burke *et al.* [11] were the first group who used artificial neural networks (ANNs) to predict cancer survivability. Through the calculation of the area under the curve (AUC $\approx$ 0.77), they proved the applicability of the machine learning methods. Delen *et al.* [7] have investigated the performance of three learning methods, namely logistic regression, ANN, and decision tree (DT), while their gained highest accuracy was 93% by DT. Khan *et al.* [12] have applied fuzzy inference in the structure of DT to improve its performance, improving the accuracy of the basic DT up to 9%. Shin *et al.* [13] have considered the problem as a semi-supervised learning task in order to label the collected data without survival status. Their graph-based technique has the advantage of robustness in a way that it obtains a fixed accuracy of prediction regardless of the primary parameters setup and the training dataset. According to the results of a work carried out by Chao *et al.* [14] over the support vector machine (SVM) and DT methods, SVM with an accuracy of 95% gained a higher performance. In another effort, Henriquez *et al.* [15] have used the SVM, K-nearest neighbor (K-NN), and DT methods, while K-NN with an accuracy value of 81% obtained a better performance compared to the other two methods. Finally, based on a study by Montazeri *et al.* [16] to investigate the performance of seven different machine learning techniques for survivability prediction, random forest was the best with an accuracy of 96% among the methods in terms of classification accuracy.

In order to predict survivability by the machine learning techniques, it is necessary to perform some medical experiments on a cancer patient or a tumor in order to measure the value of some features. These experiments may be difficult or expensive, and may cause some side-effects for patients [17]. The main purpose of this work is to propose a method for predicting the survivability of cancer patients without using the difficult experiments. To this end, the features within the dataset are divided into two groups, difficult and simple. The dataset was used to develop a multi-stage learning machine by combining three machine learning technique including multi-layer perceptron (MLP), SVM, and DT with a minimum need for the features from the difficult dataset preserving approximately the same prediction accuracy. Performance of the developed multi-stage machine is compared with two feature selection methods, namely sequential backward selection and basic individual selection, in order to investigate its accuracy and effectiveness regarding the removed features.

The rest of this article is organized in the following sections. In Section 2, first, the SEER database is introduced including the statistics of survivability for cancer patients, and then the structure of the multi-stage learning machine is explained, which is a combination of three machine learning techniques, namely MLP, SVM, and DT. In Section 3, the results obtained from different learning machines are explained and analyzed. Finally, Section 4 discusses and concludes the results of this paper.

2. Methods

The developed multi-stage learning machine to predict the survivability of breast cancer patients is introduced in this section.

2.1. Dataset preparation

The SEER (2000-2013) database is one of the most comprehensive databases worldwide including the statistical information related to the cancer patient survivability in the United States. Besides the survival or non-survival of cancer patients, this database provides a set of useful information such as the patient's race and living conditions, and the tumor stage. The comprehensiveness and availability of the database have attracted many researchers to use its data in their studies and have turned it into a

reliable source for predicting survivability among cancer patients [18—20].

Among information of cancer patients in the SEER database, a file is dedicated to breast cancer cases including the 505,731 instances with 109 features. After a pre-processing to remove useless records and features, a set of 50000 instances of breast cancer with 35 features was prepared.

Among these instances, the 22,930 cases led to the death of a patient due to a cancer-related reason during five years after the tumor diagnosis, and the 27,070 cases led to the survival of the cancer patient after five years [21].

The selected features from the dataset and their brief description are given in table 1. The detailed description of this dataset is available in [21]. As it can be seen in this table, the values of the last ten features (26 to 35) have been calculated using the SSF and AJCC algorithms [22, 23]. These features were added to the SEER database in 2004. In order to calculate the input of these two algorithms, it is necessary to perform the medical experiments on the tumor and on the lymph nodes of the patients [22].

Calculation of these ten features may not be available for patients in some countries, and thus they were considered as difficult or costly features, while the rest of the features were grouped as simple features. The main focus of this paper was to develop a machine that could predict the survivability of a cancer patient yielding an acceptable accuracy without using these ten costly features.

Table 1. Summary of the selected features.

No.	Name	Description	Unique values
1	Marital status	Patient's marital status at the time of diagnosis	7
2	Race	Recoded race of the patient	36
3	Age	Age at the diagnosis	101
4	Seq. number	Number of all reportable malignant, *in situ*, benign, and borderline primary tumors	2
5	Primary site	Identifies the site in which the primary tumor originated	9
6	Laterality	Side of the body on which the reportable tumor originated	5
7	Histology ICD_O_3	Microscopic composition of cells and/or tissue for a specific primary	100
8	Behavioral ICD_O_3	Malignancies with *in situ* as described in ICD_O_3	2
9	Grade	Tumor similarity to high or low aggressive tumors	5
10	Regional nodes positive	The exact number of regional lymph nodes examined by the pathologist that were found to contain metastases	60
11	Regional nodes examined	Total number of regional lymph nodes that were removed and examined by the pathologist	67
12	CS tumor size	Information on tumor size	210
13	CS extension	Information on the extension of the tumor	40
14	CS lymph nodes	Information on the involvement of lymph nodes	39
15	CS mets	Information on distant metastasis	9
16	RX surg prime site	Describes a surgical procedure that removes and/or destroys the tissue of the primary site	47
17	RX scope reg	Describes the procedure of removal, biopsy or aspiration of regional lymph nodes	9
18	RX surg/reg	Describes the surgical removal of distant beyond the primary site	8
19	Record ICD_9	The ICD_9 recorded primary site and morphology	17
20	ER status	Combining information from Tumor marker 1 and CS site-specific factor 1	4
21	PR status	Combining information from Tumor marker 2 and CS site-specific factor 2	4
22	SEER Stage	It is a simplified version of the stage	5
23	First malignant prime	The first primary tumor is malignant or not	2
24	SS 2000	Stage coding for year 2000 and after that	7
25	Primary by international rules	These rules are created using IARC multiple primary rules	
26	SSF2		2
27	SSF3	Each CS site-specific factor (SSF) is scheme-dependent. They provide the information	7
28	SSF4	required to stage the case, clinically relevant information or prognostic information	60
29	SSF5		5
30	SSF6		4
31	AJCC stage group	Stage of the tumor using the CS algorithm	8
32	AJCC M	Describes metastasis of the tumor using the CS algorithm	12
33	AJCC T	Describes the size of the tumor using the CS algorithm	4
34	AJCC N	Describes lymph node involvement using the CS algorithm	16
35	Breast AJCC N	Breast cancer lymph node involvement	20

2.2. Standard predictors

MLP is a typical ANN, which is widely used in different supervised learning problems. The most widely used type of MLP, which has obtained the best trade-off between complexity and adaptability, is organized in a three-layered fully connected network [24]. The layers are called the input, hidden, and output layers, whereas each layer consists of a number of neurons based on the problem definition and its complexity. The first layer receives the features of an input sample and propagates them to the neurons in the hidden layer. After processing the input signals, the neurons in the hidden layer transfer the results to the output layer. Finally, the output layer neurons calculate the output of the network. In this work, the back-propagation algorithm was used to train MLP. Generally, the performance of an MLP machine depends on its structure, primary weights, and the utilized activation function. The MLP classifier used in this work consisted of 35, 20, and 2 neurons in the input, hidden, and output layers based on the number of input features and output classes, respectively. The number of neurons in the hidden layer was chosen based on the highest accuracy obtained by different examined number of neurons. The sigmoid activation function was used to calculate the output of each neuron.

The SVM classifier tries to find an optimum border between classes through drawing a hyperplane, which has the highest margin from samples of each class [25]. The machine first uses a linear or non-linear kernel function to transfer the feature vector of the training samples to another feature space where the samples are more separable. Then the algorithm looks for the coordinates of a hyperplane that can separate two (or more) classes present in the new space with the least amount of error tolerance and the maximum margin from samples of each class. The SVM algorithm developed for the proposed multi-stage machine was implemented using the radial basis function as the kernel function.

DT is a hierarchical arrangement of internal decision nodes leading finally to leaves [26, 27]. During the learning of a decision tree, a divide-and-conquer algorithm is applied in a top-down manner in a way that produces a tree starting from its root. The impurity of each feature is calculated and the feature providing the minimum impurity is chosen at each node to make internal decisions until a fully pure node is reached.

2.3. Multi-stage classifier

The main objective in this paper was to develop a classifier to predict the survivability of the breast cancer patients with an acceptable accuracy possibly without using the last ten features of the feature vector. To this end, the samples within the dataset were divided into two groups. The first group includes the samples whose class can be determined certainly using only 25 simple features, while the second group consists of the samples whose class is predicted using 25 features with a high risk, and we would better to use all 35 features to predict their classes. The ensemble machine developed in this work categorizes the simple cases using only the simple features, while it uses all features for categorizing the difficult cases. In order to achieve this goal, a multi-stage classifier machine was developed that works in two stages. Figure 1 represents the block diagram of this multi-stage classifier. The machine in the first stage employs three base-learners simultaneously including MLP, SVM, and DT. These learners were trained separately on a subset of dataset including only the simple features. In the second stage, the machine employs the SVM base-learner, which was trained on all 35 features of the dataset. Selection of the SVM learner was due to obtaining the highest accuracy by this classifier compared to the other two.

The trained machine is used to predict the survivability of an unknown patient in the recall phase. The machine first calculates the output of three learners in the first stage using 25 simple features. If three learners in this stage agree on the output, it is considered as the final decision for prediction of the patient survivability; otherwise, for an inconsistent output between three learners at the first stage when one of the learners produces a different output, the input sample is considered as a difficult case, and the final decision is depended on the second stage. In the second stage, it is necessary to evaluate ten costly features for this patient and consider all the 35 features for making the final decision. This form of decision-making in two stages guarantees a more cost-effective way to predict survivability for simple cases without the side-effects originated by difficult experiments.

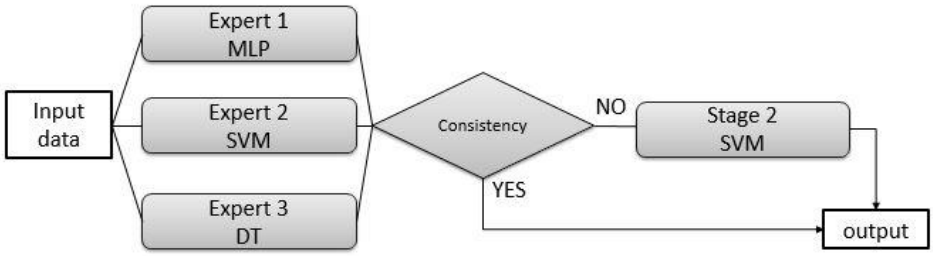

Figure 1. Structure of the multi-stage classifier.

3. Results

The developed multi-stage classifier was investigated to examine its performance. In this section, the experimental setup and the obtained results are presented.

3.1. Setting up experiments

The 10-fold cross-validation technique was used to investigate the performance of the predictors [28]. Since the dataset of the previous studies is not available, comparing the results of the developed machine with those of the previously proposed methods was not possible. For the same reason, the previous studies on the breast cancer survivability prediction have not compared their performance directly with those of other studies. Therefore, in most studies, either standard machines such as SVM and MLP have been used [7, 14, 15] or a specific machine has been developed and the results obtained by that machines have been compared with those of standard machines [12, 13]. In this paper, the performance of the proposed multi-stage machine was compared with that of the SVM, MLP, and DT standard machines.

Three evaluation criteria were used to assess the performance of the predictors including the sensitivity, specificity, and accuracy. Sensitivity is the criterion, which shows the reliability of a machine for the samples belonging to the survival class, while specificity is the same criterion for the samples from the non-survival class. Moreover, accuracy is the main criterion calculating the general performance of the machine for the correct prediction of a sample's class.

3.2. Evaluation of multi-stage machine

As described in the "methods" section, the major aim in the development of the multi-stage machine was for the prediction of survivability of breast cancer patients without using difficult features for simple cases. In order to investigate the performance of the developed machine, three introduced standard classifying methods, namely MLP, SVM, and DT, were used for prediction on both 25 simple features and all features through a 10-fold cross-validation. The results obtained are represented in table 2. Additionally, the results of the 10-fold cross-validation of the multi-stage machine are shown in table 3.

In table 2, it could be seen that the highest accuracy was obtained by SVM for both the 25 and 35 feature sets with an accuracy of 83.74% and 84.32%, respectively. In addition, the slight difference between the accuracy of machines for the two 25 and 35 feature sets reveals that costly features can be ignored for simple cases in order to reduce the cost of survivability prediction.

Comparing the results in tables 2 and 3 reveals that the multi-stage machine obtained an average accuracy of 84.34%, which is higher than other single predictors. Additionally, about 80.85% of the samples were classified in the first stage without using difficult features. Consequently, the developed multi-stage machine significantly leads to ignore difficult features without their side-effects and save cost for four samples out of the five input samples.

3.3. Comparing with feature selection methods

Another investigation was carried out in order to study the impact of each feature on the prediction accuracy. The study was performed using two feature selection methods including sequence backward selection (SBS) [29] and basic individual selection (BIS) [30]. Based on the SBS method, one of the features was removed each time and the SVM classifier was trained using a 10-fold cross-validation. The features whose removal results in the minimum decrease in the prediction accuracy are

removed from the features. This procedure was run to remove 10 out of 35 features. Furthermore, through the basic individual selection (BIS), the correlation coefficient between all features was calculated, and then 10 features having the minimum correlation coefficient were removed from the dataset [31]. Table 4 shows the features that are removed using the two SBS and BIS methods. After removing 10 features from the dataset, a 10-fold cross-validation was used to train the SVM machine. Table 5 shows the results produced by the SVM machine for the two feature sets selected using the SBS and BIS methods.

Table 2. Summary of the results obtained by standard predictors using 10-fold cross-validation on two 35 and 25 feature sets

Predictor		**MLP**			**SVM**			**Decision tree**	
35 features	**SEN (%)**	**SPE (%)**	**ACC (%)**	**SEN (%)**	**SPE (%)**	**ACC (%)**	**SEN (%)**	**SPE (%)**	**ACC (%)**
fold-1	82.46	87.6	84.48	83.3	86.86	84.74	80.78	76.59	78.9
fold-2	81.11	87,9	83.74	83.25	86.66	84.66	81.13	77.17	79.32
fold-3	81.31	89.07	84.34	82.63	87.26	84.54	80.96	80.18	80.62
fold-4	81.2	88.08	84.04	83.29	86.1	84.44	81.65	78.45	80.24
fold-5	81.24	87.93	83.82	83.23	86.16	84.44	81.31	79.41	80.48
fold-6	81.89	88.39	84.44	83.36	86.89	84.82	81.87	80.03	81.06
fold-7	79.78	86.96	82.62	80.91	86.04	83.02	79.51	77.85	78.76
fold-8	80.7	88.74	83.96	82.11	87.18	84.28	80.72	80.83	80.78
fold-9	80.84	87.16	83.36	82.47	85.72	83.84	80.37	77.88	79.24
fold-10	75.15	94.24	80.92	83.16	86.17	84.42	81.5	79.1	80.42
Average	80.56	**88.6**	83.57	**82.77**	86.5	**84.32**	80.98	78.74	79.98
25 features	**SEN (%)**	**SPE (%)**	**ACC (%)**	**SEN (%)**	**SPE (%)**	**ACC (%)**	**SEN (%)**	**SPE (%)**	**ACC (%)**
fold-1	81.36	88.09	83.92	83.14	85.63	84.16	81.17	77.59	79.58
fold-2	81.21	89.26	84.28	82.36	86.33	83.98	81.79	78.08	80.1
fold-3	81.2	88.26	83.98	81.56	86.4	83.54	81.06	79.41	80.32
fold-4	79.73	89.51	83.22	82.76	85.76	83.98	81.9	78.01	80.16
fold-5	80.53	88.85	83.64	82.21	86.03	83.76	82.16	79.38	80.92
fold-6	80.94	88.76	83.92	82.79	86.77	84.42	81.97	80.09	81.14
fold-7	78.14	88.82	82.08	80.42	85.4	82.46	80.28	78.64	79.54
fold-8	77.53	91.77	82.66	81.44	86.76	83.7	79.42	79.72	79.56
fold-9	81.1	86.27	83.2	82.22	86.25	83.9	80.19	77.66	79.04
fold-10	77.2	92.6	82.28	81.98	85.77	83.54	80.94	77.28	79.26
Average	79.89	**89.21**	83.31	**82.08**	86.11	**83.74**	81.08	78.58	79.96

Table 3. Summary of the results obtained by the multi-stage predictor using 10-fold cross-validation: the last column shows the percentage of the test samples that is predicted in stage 1 using 25 features.

	SEN (%)	**SPE (%)**	**ACC (%)**	**Stage1 (%)**
fold-1	83.27	87.48	84.96	80.32
fold-2	82.83	87.18	84.6	80.7
fold-3	82.49	87.23	84.44	81.8
fold-4	83.42	85.84	84.42	81.84
fold-5	82.96	86.59	84.44	89.9
fold-6	83.23	87.57	85	82.68
fold-7	80.57	86.37	82.92	79.74
fold-8	81.72	87.82	84.28	81.44
fold-9	82.5	85.88	83.92	80.24
fold-10	82.95	86.68	84.5	78.88
Average	82.59	86.86	84.34	80.85

It can be seen in table 5 that the SVM classifier obtained a higher accuracy when the SBS method was used for feature selection. However, considering the results in table 3, the average accuracy of the developed multi-stage machine is higher than the SVM classifier using the two feature selection methods. In addition, considering the list of the removed features by two feature selection methods in table 4, SBS and BIS removed only three out of ten features from the set of difficult features, respectively, while most of the difficult features still remained and were used for the final decision. Therefore, it can be concluded that the multi-stage

machine gains a better performance than the two feature selection methods.

4. Conclusions

In this work, a multi-stage predictor machine was developed in order to predict cancer survivability. The major aim was to reduce the dimension of the feature set through removing the difficult features with an ignorable decrease in the accuracy of prediction. To this aim, 10 difficult features were determined and kept for the second stage of the machine. The developed machine simultaneously uses the SVM, MLP, and DT base-learners to predict survivability for a patient based on 25 simple features. If all machines agree on the output, the machine produces the answer to the input sample. However, if the machines disagree on the output, the second stage of the machine uses all features of the dataset including 10 difficult features to make the final decision. Based on the performance investigation through a 10-fold cross-validation, the multi-stage machine obtained an accuracy of 84.34%, which was the highest accuracy compared to the standard classifiers. In addition, for 80.85% of the patients, there was no need to use the difficult features. In conclusion, the results of the experiments indicate that the developed multi-stage method efficiently produces highly accurate results than the other classifiers, while it ignores the difficult features.

Table 4. Removed features using the two feature selection methods.

SBS		BIS	
Feature number	**Feature name**	**Feature number**	**Feature name**
24	SS 2000	17	RX Scope reg
17	RX Scope reg	9	Grade
19	Record ICD_9	2	Record ICD_9
7	Histology ICD_O_3	6	Laterality
18	RX surg/reg	4	Sequence number
11	Regional nodes examined	26	SSF2
28	SSF4	24	SS 2000
35	Breast AJCC N	1	Marital status
10	Regional nodes positive	16	RX surg prime site
29	SSF5	3	Age

Table 5. Performance of the SVM machine using the selected features by SBS and BIS.

	SBS			BIS		
	SEN (%)	**SPE (%)**	**ACC (%)**	**SEN (%)**	**SPE (%)**	**ACC (%)**
fold-1	83.61	87.14	85.04	82.62	86.4	84.14
fold-2	82.87	86.08	84.2	82.13	85.62	83.56
fold-3	82.07	87.35	84.22	81.69	86.72	83.74
fold-4	83.64	86.47	84.8	83.02	85.3	83.96
fold-5	83.08	86.49	84.48	82.32	85.49	83.62
fold-6	82.91	87.41	84.74	82.64	85.83	83.96
fold-7	80.99	85.7	82.94	80.01	84.65	81.92
fold-8	81.96	87.47	84.3	81.87	86.7	83.94
fold-9	82.47	85.72	83.84	81.76	84.72	83
fold-10	82.92	86.03	84.22	82.28	85.56	83.64
Average	**82.65**	**86.58**	**84.27**	82.03	85.69	83.54

References

[1] Siegel, R. L., Miller, K. D. & Jemal, A. (2020). Cancer statistics, 2020. CA: A Cancer Journal for Clinicians, vol. 70, no. 1, pp. 7-30.

[2] Waks, A. G. & Winer, E. P. (2019). Breast cancer treatment: a review. Jama, vol. 321, no. 3, pp. 288-300.

[3] Parvizpour, S., Razmara, J., Pourseif, & M., Omidi, Y. (2019). In silico design of a triple-negative breast cancer vaccine by targeting cancer testis antigens. Bioimpacts, vol. 9, pp. 45–56.

[4] Mavaddat, N., Michailidou, K., Dennis, J., Lush, M., Fachal, L., Lee, A., Tyrer, J. P., Chen, T. H., Wang, Q., Bolla, M.K. & Yang, X. (2019). Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. The American Journal of Human Genetics, vol. 104, no. 1, pp. 21-34.

[5] Parvizpour, S., Razmara, J. & Omidi, Y. (2018) Breast cancer vaccination comes to age: impacts of bioinformatics. Bioimpacts, vol. 8, no. 3, 223-235.

[6] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, vol. 13, pp. 8-17.

[7] Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, vol. 34, no. 2, pp. 113-127.

[8] Wiens, J. & Shenoy, E.S. (2018). Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clinical Infectious Diseases, vol. 66, no. 1, pp. 149-153.

[9] Wong, D. & Yip, S. (2018). Machine learning classifies cancer. Nature, pp. 446-447.

[10] Cox DR (2018) Analysis of survival data. Routledge, New York

[11] Burke, H. B., Goodman, P. H., Rosen, D. B., Henson, D. E., Weinstein, J.N., Harrell Jr, F.E., Marks, J.R., Winchester, D.P. & Bostwick, D.G., (1997). Artificial neural networks improve the accuracy of cancer survival prediction. Cancer, vol. 79, no. 4, pp. 857-862.

[12] Khan, U., Shin, H., Choi, J. P., & Kim, M. (2008). wFDT: weighted fuzzy decision trees for prognosis of breast cancer survivability. Australian Computer Society, Australia, 2008.

[13] Park, K., Kim, D., An, Y., Kim, M., & Shin, H. (2013). Robust predictive model for evaluating breast cancer survivability. Engineering Applications of Artificial Intelligence, vol. 26, no. 9, pp. 2194-2205.

[14] Chao, C. M., Yu, Y. W., Cheng, B. W., & Kuo, Y. L. (2014). Construction the model on the breast cancer survival analysis use support vector machine, logistic regression and decision tree. Journal of medical systems, vol. 38, no. 10, pp. 106.

[15] García-Laencina, P. J., Abreu, P. H., Abreu, M. H., & Afonoso, N. (2015). Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. Computers in biology and medicine, vol. 59, pp. 125-133.

[16] Montazeri, M., Montazeri, M., Montazeri, M., & Beigzadeh, A. (2016). Machine learning models in breast cancer survival prediction. Technology and Health Care, vol. 24, no. 1, pp. 31-42.

[17] Greenstein JP (2016) Biochemistry of cancer. Elsevier, Amsterdam.

[18] Feuer, E. J., Rabin, B. A., Zou, Z., Wang, Z., Xiong, X., Ellis, J. L., ... & Hankey, B. F. (2014). The surveillance, epidemiology, and end results cancer survival calculator SEER* CSC: validation in a managed care setting. Journal of the National Cancer Institute Monographs, vol. 2014, no. 49, pp. 265-274.

[19] Hwang, K.T., Kim, J., Jung, J., Chang, J.H., Chai, Y.J., Oh, S.W., Oh, S., Kim, Y.A., Park, S.B. & Hwang, K.R. (2019). Impact of breast cancer subtypes on prognosis of women with operable invasive breast cancer: a population-based study using SEER database. Clinical Cancer Research, vol. 25, no. 6, pp. 1970-1979.

[20] Khalid, S.I., Kelly, R., Adogwa, O., Carlton, A., Tam, E., Naqvi, S., Kushkuley, J., Ahmad, S., Woodward, J., Khanna, R. & Davison, M. (2019). Pediatric brainstem gliomas: a retrospective study of 180 patients from the SEER database. Pediatric neurosurgery, vol. 54, no. 3, pp. 151-164.

[21] Salehi, M., Razmara, J. & Lotfi, S. (2019). A Novel Data Mining on Breast Cancer Survivability Using MLP Ensemble Learners. The Computer Journal.

[22] Giuliano, A. E., Edge, S. B., & Hortobagyi, G. N. (2018). of the AJCC cancer staging manual: breast cancer. Annals of surgical oncology, vol. 25, no. 7, pp. 1783-1785.

[23] Amin, M. B., Greene, F. L., Edge, S. B., Compton, C. C., Gershenwald, J. E., Brookland, R. K., ... & Winchester, D. P. (2017). The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. CA: a cancer journal for clinicians, vol. 67, no. 2, pp. 93-99.

[24] Alpaydin, E., (2020). Introduction to machine learning. MIT press.

[25] Mohri, M., Rostamizadeh, A. and Talwalkar, A., (2018). Foundations of machine learning. MIT press.

[26] Roiger RJ (2017) Data mining: a tutorial-based primer. Chapman and Hall/CRC, London.

[27] Breiman L (2017) Classification and regression trees. Routledge, New York.

[28] Wong, T. T. (2015), Performance evaluation of classification algorithms by k-fold and leave-one-out cross-validation. Pattern Recognition, vol. 48, no. 9, pp. 2839-2846.

[29] Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, vol. 40, no. 1, pp. 16-28.

[30] Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J. & Liu, H., (2017). Feature selection: A data perspective. ACM Computing Surveys (CSUR), vol. 50, no. 6, pp. 1-45.

[31] Chatfield C (2018) Introduction to multivariate analysis. Routledge, New York.

نشریه هوش مصنوعی و داده کاوی

طراحی یک ماشین یادگیر جمعی چندسطحی برای پیش‌بینی بقای بیماران سرطان سینه

محسن صالحی، جعفر رزم‌آرا* و شهریار لطفی

گروه علوم کامپیوتر، دانشکده ریاضی، دانشگاه تبریز، تبریز، ایران.

ارسال ۲۰۱۹/۰۵/۰۹؛ بازنگری ۲۰۲۰/۰۳/۳۱؛ پذیرش ۲۰۲۰/۰۴/۲۸

چکیده:

در سال‌های اخیر، استفاده از تکنیک‌های یادگیری ماشین برای پیش‌بینی بقاء بیماران سرطانی بسیار مرسوم بوده و باعث افزایش دقت این پیش‌بینی‌ها شده است. در این میان، باید توجه کرد که آماده سازی و ثبت برخی از ویژگی‌ها برای انجام پیش‌بینی نیازمند آزمایش‌های سنگین و پرهزینه‌ای برای بیماران است، در حالی که این ویژگی‌ها تاثیر چندانی در تصمیم‌گیری نهائی ندارند و می‌توان بدون استفاده از این ویژگی‌های هزینه‌بر، پیش‌بینی بقاء بسیاری از بیماران را انجام داد. بنابراین، طراحی یک ماشین برای پیش‌بینی بقاء بیماران سرطانی که بتواند برای نمونه‌های ساده پیش‌بینی را با دقت بالا و بدون نیاز به استفاده از ویژگی‌های هزینه‌بر انجام دهد به یک هدف برای محققین تبدیل شده است. در این مقاله، یک ماشین یادگیر جمعی چند سطحی برای این منظور طراحی شده است. ماشین پیشنهادی از ترکیب سه ماشین مبتنی بر شبکه عصبی مصنوعی، ماشین بردار پشتیبان و درخت تصمیم در سطح اول ساخته شده است که پیش‌بینی بقاء بیماران سرطان سینه را برای نمونه‌های ساده و بدون استفاده از ویژگی‌های هزینه‌بر انجام می‌دهند. در صورتی که ماشین‌های سطح اول اتفاق نظر در تصمیم‌گیری نداشته باشند پیش‌بینی در سطح دوم و با اضافه کردن ویژگی‌های هزینه‌بر توسط یک روش مبتنی بر ماشین بردار پشتیبان انجام می‌گیرد. بر اساس نتاج به دست آمده، ماشین طراحی شده دقت قابل توجهی در این پیش‌بینی ارائه می‌دهد که در مقایسه با روش‌های دیگر بالاترین میزان دقت می‌باشد. همچنین، تصمیم‌گیری در سطح اول باعث می‌شود تا نیاز به استفاده از ویژگی‌های هزینه‌بر برای بسیاری از بیماران وجود نداشته باشد.

کلمات کلیدی: پیش‌بینی بقای بیماران سرطان سینه، ماشین‌های یادگیر جمعی، ماشین‌های چندسطحی، انتخاب ویژگی.