ISSN 2007-9737

Machine Learning Models for Cancer Type Classification with Unstructured Data

Erick E. Montelongo González, José A. Reyes Ortiz, Beatriz A. González Beltrán

Autonomous Metropolitan University,
Systems Department,
Mexico

{al2181800093, jaro, bgonzalez}@azc.uam.mx

Abstract. Machine learning (ML) techniques have been used to classify cancer types to support physicians in the diagnosis of a disease. Usually, these models are based on structured data obtained from clinical databases. However valuable information given as clinical notes included in patient records are not used frequently. In this paper, an approach to obtain information from clinical notes, based on Natural Language Processing techniques and Paragraph Vectors algorithm is presented. Moreover, Machine Learning models for classification of liver, breast and lung cancer patients are used. Also, a comparison and evaluation process of chosen ML models with varying parameters were conducted to obtain the best one. The ML algorithms chosen are Support Vector Machines (SVM) and Multi-Layer Perceptron (MLP). Results obtained are promising and they show the best model for classification is the MLP model with a precision 0.89 and f1-score 0.87, although the difference in precision between models is minimal (0.02).

Keywords. Machine learning, natural language processing, cancer classification, support vector machines, neural networks, unstructured data.

1 Introduction

Cancer is one of the main causes of deaths in the world and is estimated that will become the most important barrier to overcome to extend the lifespan of people [2]. This disease is defined as an uncontrolled growth of cells that can invade nearby tissue [15].

One of the sub-types of cancer that has the most incidence in the developed world is the breast cancer. Lower breast cancer mortality rates can be archived thanks to early diagnosis and detection. Another sub-type of cancer that has a high incidence rate is Lung Cancer. This type of cancer accounts for 18.4 percent of deaths, and early diagnosis can also drop its mortality rate. Liver cancer represents the 8.2 percent in mortality, thus, early diagnosis of the disease will have a high impact in the outcome for an opportune treatment [2].

Much of the information used by physicians to diagnose cancer comes from clinical data, which is being generated by computer information systems. This information can be either values from laboratory test reported by diagnosis tools or clinical notes written as a free-text document. Moreover, machine learning models can exploit the information for analysis and classification tasks. To be specific, physicians can use this classification as support to improve diagnosis.

Machine Learning is a method of data analysis that automates analytical model building, where the model can learn from experience to improve its performance. The model can be adopted for several tasks, one of which is classification and prediction of new examples(cases). For this reason, machine learning techniques has been used to classify cancer in distinct studies.

This is done to support physicians to improve the accuracy of the diagnosis to lower the mortality rates; however, in the majority of work done, the data used for training the model are structured,

without taking into account the free-text notes written by physicians.

Natural Language Processing is an application area that explores how computers can be utilized to understand and manipulate natural language, for example extracting meaningful information from free-text notes.Different analysis tasks can be done when working with NLP, the more commonly used for ML tasks are lexical and syntactic analysis. Lexical analysis emphasis on the interpretation of individual meaning of words while syntactic analysis emphasis on grammatical structure of sentences.

With both types of analysis a technique for free-text processing can be used to extract meaningful information in documents, this is known as *knowledge extraction*. Therefore, applying *knowledge extraction* in clinical notes, and using this information as input data for the machine learning model creation, can help to exploit a broader set of data contained in patients health records. The obtained model subsequently can be employed by physicians as supplementary information to help improve diagnosis and outcome of patients with breast, lung and liver cancer.

The rest of this paper is organized as follows. Section 2 presents the related work about machine learning approaches and text classification for support in clinical decisions making. In Section 3, our proposed approach is presented. This approach uses machine learning techniques for classifying cancer based on clinical notes with Natural Language Processing (NLP). Section 4 exposes experimental settings and the results obtained by using a well-known clinical database, also a comparison of two machine learning models is presented. Finally, in Section 5, the conclusions and future work of this paper are presented.

2 Related Work

Classification with machine learning techniques has been widely applied to cancer types. The prediction of metastasis and survival of breast cancer is presented in [20]. In their work, the authors used six different machine learning techniques to compare the performance in classification as alive or death patients and the presence or absence of metastasis. The authors used NB, SVM *Least-square SVM (LSSVM)*, *Adabag* and *Logistic Regression (LR)*.

Similarly, in [14], the survivability classification of breast cancer is addressed. The authors used machine learning algorithms to classify breast cancer patients as alive or death. The algorithms used for classification were NB, RF, KNN, *AdaBoost*, SVM, RBFN and *Multilayer Perceptron (MLP)*. The work done in [5] address the early diagnosis of breast cancer using machine learning algorithms. The algorithms implemented in this study are DT, RF and SVM.

In [11], the authors predict the outcome of patients with lung cancer, using machine learning techniques.The machine learning algorithm used are LR, DT, RF, *Generalized Boosting Machines (GBM)*, SVM and a custom method which combines all algorithms used in a vote scheme.

Prediction of mortality caused by radical cystectomy is presented in [21]. The machine learning for and *Back-propagation Networks (BPN) Radial Basis Function Network (RBFN) Extreme Learning Machine (ELM) Regularized ELM (RELM)*, *K-Nearest Neighbour (KNN)* and NB.

Information extraction from clinical free-text notes is used for diverse classification tasks, data analysis and data mining tasks. In [16], clinical data analysis for correlating mammography and pathology findings is presented. The authors used an enterprise database to obtain mammogram readings and its corresponding pathology reports. Next they developed an NLP algorithm to automatically extract mammographic and pathological findings from free text. The correlation from these two data sources was used to extract information about the breast cancer sub-type.

Another study about breast cancer is presented in [1]. The authors designed a pipeline to predict the probability of malignancy based on analysis of mammographic reports. The pipeline consisted in NLP analysis to extract characteristics from the reports, then used as input variables to a *Bayesian Network* which provides the probability in a report of being malignant.

In [9], the authors exploit the information contained in death certificates to automatically

extract information of cancer sub-type (common and rare) to create statistics. In this work, the author employed two techniques for classification of cancer sub-types which were *Support Vector Machines (SVM)* and a rule based approach. The pipeline followed consisted in extracting detailed features with NLP (n-grams, SNOMED CT codes and ICD-O properties). Then this information was used as input variables for SVM and the rule-based approach to classify the reports in ICD-10 codes.

Electronic Health Records (EHR) contains a large amount of data which can be exploited and mined for distinct types of analysis. In [8], the authors address the problem of high cost of lung cancer treatment, derived from unnecessary visits to ER and unscheduled appointments. The analysis is developed from transcripts of a service of Thelehealthcare Phone Service (TPS) attended by oncology nurses, located at the medical oncology clinic and EHR from patients.Through NLP the authors extract meaningful information and stores it in a relational database, this information is then used to perform statistical analysis to profile the patients who employ the TPS service.

In [7], the authors' goal is to develop a system that can automatically classify radiology reports. The pipeline followed by the authors is NLP information extraction from CT reports. The information is then used by machine learning classifiers, which *Naive Bayes (NB)* and *Desition Trees (DT)*.

The current paper presents a supervised classification of clinical reports as a free-text representation of liver, breast and lung cancer. The process includes the pre-processing of clinical reports, information extraction and vector space representation of the documents. These will be used by a supervised machine learning algorithm to classify a particular document between the types of cancer mentioned.

3 Proposed Approach

We propose a pipeline for *information extraction* from clinical notes using *Natural Language Processing (NLP)* techniques, as presented in Figure 1. Initially, we extract clinical notes from patients with breast, lung or liver cancer. Next, the notes are used in the NLP pipeline to transform them into a vector space representation. After, we use these vectors as machine learning model inputs to train them for classification. Finally, based on the evaluation and scoring of these models, we select the best one for this classification task.

For vector space representation we used the *Paragraph Vectors* model. Also, the machine learning algorithms chosen to create the models are *Support Vector Machines (SVM)* and *Multi-Layer Perceptron*. The *dl4j* API [4] for Java was used for implementation of the *Paragraph Vectors* algorithm, and *scikit-learn* [17] for classification algorithms.

Natural Language Processing pipeline used in our proposed approach for transforming raw text into cleaned text is depicted below.

3.1 Text Pre-Processing, Tokenization, Lemmatization and Lexicon Creation

Clinical notes written by physicians have the evolution, interpretation of lab tests, evaluation and diagnosis of patients. These notes are written as free-text notes that can be considered as unstructured data to work with. The text contained in these notes contains valuable information for classification task, but can, in addition, contain ”garbage” text that does not contain significant information. Because of this, a procedure to ”clean” the text of this unnecessary information has to be done.

At the beginning, a decision is made about what characters does not give meaningful information to the analysis, then these characters are removed from the text. They can be special characters; i.e. (& % $), punctuation signs (although, in this work, the comma and point remain in the text for sentence boundary detection), etc.

Then, the texts are divided by white spaces to obtain individual words in each sentence called *token*. These tokens are then ”lemmatized” to bring each individual word to its base dictionary representation. With each lemmatized token, the process to eliminate stop-words is done.

These words are considered useless in the context of a given task.

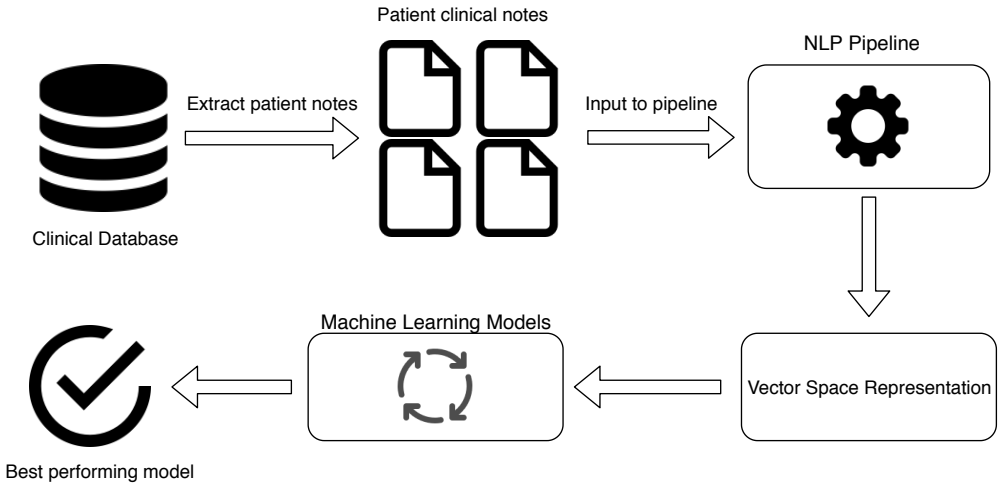

Fig. 1. Proposed approach for cancer type classification using free-text clinical notes

The mentioned process is performed to all of the documents considered for the classification task. In this case, all the clinical notes considered for the patients that presents lung, breast and liver cancer.

When all documents have been pre-processed, the distinct tokens are considered for the *lexicon* creation. The lexicon contains the unique words that appear in all the documents and is used for the following task in the pipeline, which is the vector space representation. This, along with the vector space representation, corresponds to the NLP pipeline in Figure 1.

3.2 Vector Space Representation

The representation of a set of documents as vectors in a common space is known as the *vector space model*. Many machine learning models used for classification require that input are encoded in a fixed-length vector. One of the most used fixed-length vector representation of text is the Bag-of-Words(BoW) model.

The BoW model produces a representation of the occurrence of a word within a document using a lexicon for the unique words in the *corpus* and weights each word for all the documents in the corpus. The problem for this representation is that doesn't take into account the ordering or meaning of words. In recent years a new model was proposed to represent variable-length text seen as *Paragraphs* [10], this approach was partially based on the Word2Vec model for representing words [12]. The Paragraph Vector approach is based on learning the vector representation of words and an additional matrix in which each column is the vector representation of each paragraph. In the training phase of word vectors, the objective is to predict a word given other words in the context. Formally, given a set of training words w_1, w_2, w_T, maximize the average log probability as shown in (1):

$$\frac{1}{T}\sum_{t=k}^{T-k} log\ p(w_t \mid w_{t-k}, \ldots, w_{t+k}). \tag{1}$$

The second phase is the prediction task, which is done by a multiclass classifier. The authors suggest the use of softmax, as shown in (2):

$$p(w_t \mid w_{t-k}, \ldots, w_{t+k}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}}. \tag{2}$$

In equation (2) the y_i corresponds to a unnormalized log-probability for each word i obtained with (3):

$$y = b + Uh(w_{t-k}, \ldots, w_{t+k}W). \tag{3}$$

In equation (3) U, b are softmax parameters, h is constructed by the average or concatenation of paragraph matrix D and word vector matrix W.

The training of word and paragraph vectors is done using stochastic gradient decent. In each step of the stochastic gradient decent, a sample of fixed-length context is extracted from a random paragraph. The error of the gradient is computed, and the weights are updated based on this error.

When the training is completed, the paragraph vectors can then be used as a feature vector for that paragraph. Eventually, these vectors can be provided to conventional machine learning models to do the classification task.

3.3 Machine Learning Models

First proposed by Vapnik et al. [3], Support Vector Machines (SVM) is a supervised learning algorithm that finds a hyperplane in a N-dimensional space (N representing the number of features on its input). The found hyperplane is used to separate the data into the desired classes. The objective is to find a plane that has the maximum margin (maximum distance between points). If linear separation is not possible, this algorithm uses kernel methods to obtain a mapping to a feature space. The principal parameters used for this model are:

— **C**: Penalty parameter that trades off correct classification against maximization of the decision function margin.

— **Gamma**: Define how far the influence of a training example reaches. Used for *Radial Basis Function (RBF)* kernel.

The model originally conceived by *F. Rosenblatt* [18] describes the basic unit for learning called *Perceptron*, later Minsky et al. [13] adopted this model and generalized it for computational use, introducing the concept of weights and a mechanism to learn those weights, nowadays is the base unit for the *Neural Network* models. The multi-layer perceptron is a supervised learning algorithm of interconnected perceptrons that learns a function *f* over a train dataset to produce a non-linear mapping between input and output vectors [6]. The parameters used in this paper for training the model are:

— **Activation function**: It is a function that maps an input signal of a neuron to an output signal. The types of activation function used for this paper are *identity*, *logistic*, *tanh*, and *relu*.

— **Hidden Layer Size**: Number of neurons presents in the layer.

— **Maximum Iterations**: Maximum number of iteration (or epochs) for training the model.

— **Solver**: Correspond to the method employed for weight optimization. The ones used are *Limited-Memory Broyden-Fletcher-Goldfarb-Shanno (lbfgs)*, *Stochastic Gradient Descent (sgd)* and a first-order gradient-based stochastic optimization *Adam*.

4 Experimentation and Results

4.1 Score and Evaluation of models

In supervised learning, given input and output labels for each record, the model is trained to classify each input to the desired output. In this case, the input for the model is a paragraph vector for each patient record and the output corresponds to the cancer type considered for this study (lung, breast and liver). The reliability of this classification is evaluated by three standard performance measures for multiclass classification: precision, recall and f1-score. Each measure formula can be seen in 4, 5 and 6:

$$Precision = \frac{TP}{TP + FP}, \tag{4}$$

$$Recall = \frac{TP}{TP + FN}, \tag{5}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{6}$$

where TP, FP and FN stands for true positive, false positive and false negative respectively.

To prevent over-fitting or under-fitting the model, we used the cross-validation strategy for experiments. This process divides the original dataset by the desired number of subsets, which contains a balanced class distribution. Next,

the process of cross-validation alternately uses one subset for evaluation of the model, and the remaining ones are employed for training purposes. This process is replicated the desired number of times, and each time a different subset is used for evaluation. With this, the standard deviation for the performance of each model is calculated.

Finally to explore the parameter configuration for each machine learning model and select the best parameter combination, we used the GridSeachCV implemented in scikit-learn [17]. This method conducts an exhaustive search over specified parameter values for a machine learning model, validates trough cross-validation and reports the best configuration that maximizes the specified objective value (precision, recall).

4.2 Results

In this work the MIMIC II clinical database [19] was used. This database contains clinical information from patients admitted into the intensive-care unit in the **Beth Israel Deaconess Medical Center**. Each patient is categorized via ICD-9 code, and his records contain, among other data, clinical notes written by physicians during patient stay. The dataset employed contained 10,518 clinical notes among 225 patients, the total and average number of clinical notes per cancer type is presented in table 1. With this data, the percentage for training and evaluation for each model was 20% and 80% respectively.

Two machine learning models were used to classify and predict new instances of documents represented as *Paragraph Vector*. These models were trained under different parameters for each one, and five fold cross-validation was applied. The parameters used to train the SVM models, along with the precision value and standard deviation are presented in table 2.

In table 2, the best result presented a classification precision of 0.870. Although the best with minimum standard deviation provided a precision of 0.854, showing high gamma value. The results present little difference between precision values with the maximum difference being 0.02.

Table 1. Average and total clinical notes for patients

Cancer type	# of patients	# of notes	Average of notes per patient
Lung	75	3653	48
Breast	75	2157	28
Liver	75	4708	62

Table 2. Top 10 results for SVM cancer type classification

Case #	C value	γ value	Kernel	Precision	σ
1	100	0.8	RBF	**0.870**	0.072
2	100	0.6	RBF	0.868	0.080
3	100	0.7	RBF	0.866	0.084
4	100	0.3	RBF	0.864	0.093
5	1	0.5	RBF	0.861	0.059
6	100	0.4	RBF	0.860	0.084
7	100	0.9	RBF	0.857	0.082
8	1	0.8	RBF	0.855	0.052
9	1	0.9	RBF	0.854	0.051
10	1	0.4	RBF	0.850	0.074

In standard deviation, the maximum difference among the values is 0.042. Little variation can be observed in precision values. Top results maintain consistent C values and variations in gamma values results in minimum loss or gain in precision.

The classification report for the best SVM model found can be seen in table 3.

Table 4 presents the parameters for the Multi-Layer Perceptron model. As seen in this table, the best result presented for average precision is 0.890 and the minimum value for standard deviation being 0.034 with a precision of 0.858. It can be observed that the best results are

Table 3. Best SVM model found for cancer type classification

	Precision	Recall	F1-Score
Breast	0.92	0.80	0.86
Liver	0.87	1.0	0.93
Lung	0.83	0.83	0.83
Average	0.87	0.87	0.87

Table 4. Top 10 results for MLP cancer type classification

Case #	Activation function	Hidden layer size	Max. iter.	Precision	σ
1	Relu	300	500	**0.890**	0.065
2	Relu	300	200	0.865	0.109
3	Relu	400	800	0.864	0.057
4	Logistic	300	600	0.862	0.082
5	Tanh	400	100	0.861	0.045
6	Logistic	500	700	0.860	0.064
7	Relu	500	100	0.859	0.042
8	Tanh	300	800	0.858	0.034
9	Identity	400	800	0.857	0.053
10	Logistic	400	200	0.856	0.079

Table 5. Best MLP model found for cancer type classification

	Precision	Recall	F1-score
Liver	0.95	0.90	0.93
Breast	0.73	1.00	0.85
Lung	0.91	0.71	0.80
Average	0.89	0.87	0.87

presented for the activation function *relu* and layer sizes 300 and 400.

This corresponds with the fact that the Paragraph Vectors number of dimensions is 300. For the solver, *lbfgs* is the one that presented the best results. The maximum difference in the average precision is 0.034 which is more than the one found for SVM. For the standard deviation it can be observed that the maximum difference is 0.075, being higher than SVM model. The best MLP model classification report is presented in table 5.

Although the average precision for the MLP model is higher than the SVM model. Additionally, MLP model has problems with the classification of breast cancer patients and excels on the classification of the other types of cancer. Taking into account that the values for recall and f1-score are the same, the decision for the best model is based on precision. Because of this, MLP is chosen as the best model.

5 Conclusions and Future Work

This paper has presented machine learning models for classifying clinical texts in cancer domain based on natural language processing. The complete approach includes a natural language processing pipeline and a machine learning phase. NLP pipeline consists of a text pre-processing, tokenization, lemmatization and lexicon creation. On the other hand, the machine learning phase includes a vector space representation by using paragraph vector model and the use of a suitable machine learning model.

The primary contributions of this work are: a) the comparison based on results of two machine learning models for classifying clinical notes; b) the NLP pipeline for text processing to adapt the clinical notes for the feature extraction to be represented in the paragraph vector; c) the experimentation with SVM and MLP classifiers by using several configurations.

The machine learning models were evaluated by using well-known metrics as precision, recall and f-measure. The best result was achieved with MLP classifier using *relu* as activation function, 300 hidden layer size and 500 maximum iterations. This experimental configuration has achieved 0.890 precision. The main benefits for physicians are reflected in a decrease of the error rate in diagnoses and prevent the manual analysis of

clinical notes that is a tedious and time-consuming task.

As future work, complement with structured data like lab tests, microbiology test, demographics, etc. can be used to test if results are improved. Also, work to integrate these models to an information system to assist physicians can be done. This implementation can be a web or mobile application for hospital and clinic use. With some of the advantages of this applications (low specs needed, reach, etc.) different areas of the hospitals can benefit from notes analysis and at the same time, generate more data for the models. Furthermore, with some modifications, this approach can be adapted for other diseases i.e. diabetes, heart diseases, infectious diseases, etc.

Acknowledgements

We would like to thank the Autonomous Metropolitan University, Azcapotzalco. This work is funded by CBI-UAM at SI001-18 project.

References

1. **Bozkurt, S., Gimenez, F., Burnside, E. S., Gulkesen, K. H., & Rubin, D. L. (2016).** Using automatically extracted information from mammography reports for decision-support. *Journal of Biomedical Informatics*, Vol. 62, pp. 224–231.
2. **Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018).** Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*.
3. **Cortes, C. & Vapnik, V. (1995).** Support-vector networks. *Machine learning*, Vol. 20, No. 3, pp. 273–297.
4. **Eclipse Deeplearning4j Development Team (2019).** Deeplearning4j: Open-source distributed deep learning for the JVM.
5. **Farooqui, N. A. & Ritika (2018).** A study on early prevention and detection of breast cancer using three-machine learning techniques. *International Journal of Advanced Research in Computer Science U6 - Journal Article*, Vol. 9, No. Special Issue 2, pp. 37.
6. **Gardner, M. W. & Dorling, S. (1998).** Artificial neural networks (the multilayer perceptron)–a review of applications in the atmospheric sciences. *Atmospheric environment*, Vol. 32, No. 14-15, pp. 2627–2636.
7. **Gerevini, A. E., Lavelli, A., Maffi, A., Maroldi, R., Minard, A.-L., Serina, I., & Squassina, G. (2018).** Automatic classification of radiological reports for clinical care. *Artificial Intelligence in Medicine*, Vol. 91, pp. 72–81.
8. **Goulart, B. H. L., Silgard, E., Baik, C. S., Bansal, A., Greenwood-Hickman, M. A., Hanson, A., Ramsey, S. D., & Schwartz, S. (2017).** Validation of natural language processing (NLP) for automated ascertainment of EGFR and ALK tests in SEER cases of non-small cell lung cancer (NSCLC). *Journal of Clinical Oncology*, Vol. 35, pp. 6528–6528.
9. **Koopman, B., Zuccon, G., Nguyen, A., Bergheim, A., & Grayson, N. (2018).** Extracting cancer mortality statistics from death certificates: A hybrid machine learning and rule-based approach for common and rare cancers. *Artificial Intelligence In Medicine*, Vol. 89, pp. 1–9.
10. **Le, Q. & Mikolov, T. (2013).** Distributed representations of sentences and documents. *International conference on machine learning*, pp. 1188–1196.
11. **Lynch, C. M., Abdollahi, B., Fuqua, J. D., de Carlo, A. R., Bartholomai, J. A., Balgemann, R. N., van Berkel, V. H., & Frieboes, H. B. (2017).** Prediction of lung cancer patient survival via supervised machine learning classification techniques. *International Journal of Medical Informatics*, Vol. 108, pp. 1–8.
12. **Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013).** Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
13. **Minsky, M. & Papert, S. (1969).** An introduction to computational geometry. *Cambridge tiass., HIT*.
14. **Montazeri, M., Montazeri, M., Montazeri, M., & Beigzadeh, A. (2016).** Machine learning models in breast cancer survival prediction. *Technology and health care: official journal of the European Society for Engineering and Medicine*, Vol. 24, No. 1, pp. 31–42.
15. **National Cancer Institute (2018).** National Cancer Institute NCI dictionary of cancer terms. Accessed: 2018-09-13.

16. **Patel, T. A., Puppala, M., Ogunti, R. O., Ensor, J. E., He, T., Shewale, J. B., Ankerst, D. P., Kaklamani, V. G., Rodriguez, A. A., Wong, S. T. C., & Chang, J. C. (2017).** Correlating mammographic and pathologic findings in clinical decision support using natural language processing and data mining methods: Natural language processing. *Cancer*, Vol. 123, No. 1, pp. 114–121.
17. **Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011).** Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, Vol. 12, pp. 2825–2830.
18. **Rosenblatt, F. (1958).** The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, Vol. 65, No. 6, pp. 386.
19. **Saeed, M., Villarroel, M., Reisner, A. T., Clifford, G., Lehman, L.-W., Moody, G., Heldt, T., Kyaw, T. H., Moody, B., & Mark, R. G. (2011).** Multiparameter intelligent monitoring in intensive care II: A public-access intensive care unit database. *Critical Care Medicine*, Vol. 39, No. 5, pp. 952–960.
20. **Tapak, L., Shirmohammadi-Khorram, N., Amini, P., Alafchi, B., Hamidi, O., & Poorolajal, J. (2018).** Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. *Clinical Epidemiology and Global Health*.
21. **Wang, G., Lam, K.-M., Deng, Z., & Choi, K.-S. (2015).** Prediction of mortality after radical cystectomy for bladder cancer by machine learning techniques. *Computers in Biology and Medicine*, Vol. 63, pp. 124–132.

Article received on 29/10/2019; accepted on 03/05/2020.
Corresponding author is José A. Reyes-Ortiz.