ISSN 1870-4069

# A Comparative Analysis of Learning Techniques for Cancer Risk Prediction based on Medical Textual Records

Carolina Fócil-Arias, Grigori Sidorov, Alexander Gelbukh,
Miguel A. Sanchez-Perez

Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación,
Mexico City, Mexico

focil.carolina@gmail.com, sidorov@cic.ipn.mx, www.gelbukh.com,
masp1988@hotmail.com

**Abstract.** In this paper, we compare the performance of a variety of machine learning algorithms, including supervised Naïve Bayes, J48, SVM, Random Tree, Random Forest, and non-supervised KNN for determining the type of cancer a patient is suffering using medical textual records. We train these classifiers on different sets of features such as unigrams and bigrams of words, character $n$-grams using tf-idf weighting scheme and binary feature representation. We evaluated performance of the classifiers in terms of accuracy, precision, recall, and F-measure. The obtained results show that Naïve Bayes and SVM achieve the best performance in this task.

**Keywords:** Cancer classification, medical records, supervised learning, SVM, Random Forest, KNN, Naïve Bayes, J48, natural language processing.

## 1 Introduction

Colon cancer is considered to be the major cause of death in the world [13]. According to [16], more than 1.2 million people are being diagnosed with this disease every year. Based on the information provided by American Cancer Society[1] in 2016, 95,270 cases of colon cancer are estimated with 49,190 death cases. Brain cancer has led to 16,050 deaths only in USA during the last year. Cancer is produced when an uncontrollable growth of cells occurs, and there is a spread of abnormal cells [6]. An early detection of these diseases allows significantly increasing survival rates.

One of the most commonly used approaches for classification of any type of data is based on machine learning (ML) techniques. In fact, the machine learning

[1] http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2016/ [last access: 17.07.2016].

strategy coupled with annotated corpora is considered the most efficient method known up to date to solve many natural language processing (NLP) tasks, such as authorship identification related tasks [9, 10, 14], plagiarism detection [21], tasks related to text similarity [23, 24], among many others.

In this work, we apply a machine learning approach using unigrams and bigrams of words and character $n$-grams as features to identify the type of cancer a patient is suffering from. We focus on two types of cancer: colon and brain cancer. We conducted experiments on a large dataset, which was collected and consolidated by Styler *et al.* [26]. We examined the performance of several machine learning algorithms and compared their results.

The rest of the paper is organized as follows. Section 2 presents several studies related to the usage of machine learning techniques for cancer detection. Section 3 describes the materials and methods used for determining whether a person has colon or brain cancer. Section 4 summarizes the results of using various machine learning algorithms. Finally, Section 5 draws the conclusions and points out possible directions of future work.

## 2 Related Work

In this section, we overview several works related to the prediction of cancer at the early stage. We focus on the studies that used machine learning approaches and consider several systems that were able to produce good results in clinical domain.

The system proposed in [12] is able to detect 20 common types of cancer, e.g., bronchus and lung, prostate, colon, breast, pancreas, etc., as well as whether cancer was the cause of death. The system is composed of two following stages: processing natural language pipeline for extracting features and using Support Vector Machines (SVM) algorithm. The authors obtained 94.2% in terms of F-measure.

Sparse Compact Incremental Learning Machine (SCILM) is a method that was proposed in [17] for cancer classification. This algorithm is based on a network structure on small dataset with a high number of dimensions, and it achieved an accuracy of 88.75%. Wang *et al.* [27] applied the same approach in order to identify breast cancer, the most common type of women cancer. The proposed method performed at the accuracy of 96.19%.

The approach proposed by Liu *et al.* [13] is able to detect cancerous colon tissues through the pattern recognition of spectra data. A total number of 60 colon tissues, which form two classes (normal and adenocarcinoma), were selected in order to perform the classification task. This method combines a large variety of features, and uses Principal Component Analysis (PCA) and Fisher's Discriminant Analysis (FDA). The method achieved the classification rate of 90.3%. The study of Rathore *et al.* [20] proposed a CBIC system, which performs a classification of colon cancer using SVM classifier.

As one can see, machine learning techniques are successfully used in medical domain. However, the question which classifier performs better in this task is still an open research question and the main motivation of the current work.

## 3 Materials and Methods

In this study, the idea is to classify whether a patient has symptoms associated with colon or brain cancer. The approach consists of seven following steps:

1. Accessing the clinical data obtained from Semantic Evaluation Exercises 2016 [26] (SEMEVAL is based on series of evaluations to explore meaning in language).
2. Data preprocessing for improving the quality of the dataset.
3. Extraction of features: unigrams and bigrams of words.
4. Construction of the vector space model.
5. Selection of machine learning algorithms based on the state of the art.
6. Evaluation of the results.
7. Comparison of the machine learning algorithms used to perform this task.

Figure 1 shows the methodology for conducting machine learning experiments.

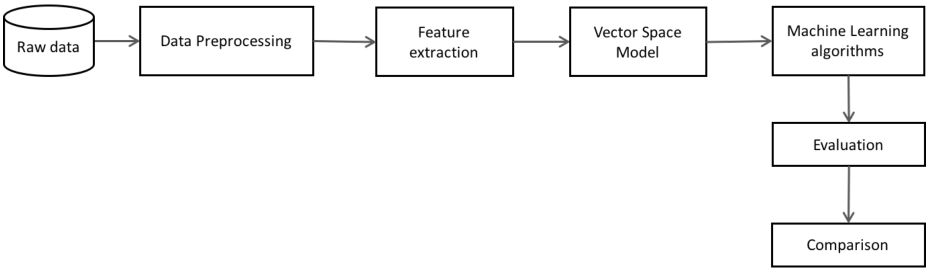

Fig. 1: Machine learning basic steps.

In our research, the corpus is divided into two subsets: training and test. A hold out approach was used to estimate the performance of machine learning classifiers. In this validation method, a certain amount for training and test data is reserved, and it is often used with independent test set [28].

### 3.1 Dataset

The dataset used in this research is the THYME (Temporal Histories of your Medical Events) [26] corpus, which was designed by clinic Mayo, University of Colorado, and the Harvard Medical School/Boston Children's Hospital. The corpus consists of a total number of 1200 documents describing 400 patients and is divided into two major diseases within oncology: colon and brain cancer.

Each patient is associated with three types of documents: clinical notes, radiology files, and pathology reports. However, in this research, we use only pathology reports, since they provide an important information concerning the patients [26]. For example, the confirmation of a benign or malignant tumor.

Thus, the number of pathology reports used in this research is 400, where 200 documents correspond to brain cancer and 200 correspond to colon cancer.

To the best of our knowledge THYME corpus is used for a classification task for the firs time.

### 3.2 Preprocessing Steps

An important step for improving data quality is data preprocessing, which has been used in several natural language processing tasks [10, 15, 18, 25] in order to increase the efficiency of the classifiers. Unlike conventional data preprocessing methods applied in the field of pattern recognition, the area of natural language processing (NLP) proposes alternative techniques, such as stop words extraction, stemming, detection of sections, etc. to enhance the quality of input data representation to be fed into a machine learning algorithm.

In our case, first, each pathological document in the corpus was segmented into sections. Then, we expanded the contractions (e.g., can't → cannot, I've → I have, etc.), which was required to normalize certain words. Afterwards, word span tokenizer [3] was applied using regular expressions, since some words were not correctly separate by a white space. Then, all the stop words were omitted, since they are used in all of the documents and do not provide useful information for classifiers. For instance, we removed the words "the", "my", "can", "ever", "by", "of", "now", and others that showed a high frequency in all the documents.

### 3.3 Classification

In this study, to determine the importance of a word in a document, the metric known as *tf-idf* is used. The *tf* (term frequency) value is the frequency of the word in document, when *idf* (inverse document frequency) is the inverse proportion of the frequency of the word in a set of documents [29]. We also try binary feature representation, that is, whether a feature exists or no exists in the corpus.

We examine five machine learning algorithms: Naïve Bayes, SVM, Random Tree, J48, Random Forest, and KNN with $K = 3$. According to [8, 27], the most commonly used algorithms in the field are SVM and KNN. However, we also examine Random Forest, Random Tree, Naïve Bayes, and J48, since these algorithms are considered among the best ones to tackle classification tasks.

Machine learning is a branch of artificial intelligence which aims at detecting meaningful patterns in data and is based on statistics and computer science. This field is divided into several subfields dealing with different types of learning tasks: supervised, unsupervised, active, passive and others [22].

The algorithms used in this study employ supervised learning. This means that the algorithm can estimate the success of prediction using the labeled training data. A brief description of each of the classifiers is presented below.

Naïve Bayes is commonly used in classification tasks. This algorithm belongs to the probabilistic classifiers, where features are conditionally independent. This algorithm ignores possible dependencies among the inputs and reduces a multivariate problem to a group of univariate problems [1]. It is based on Bayesian theorem, which shows a relation between marginal probabilities and conditional probabilities [11].

Support Vector Machine (SVM) is a powerful classification algorithm, which belongs to linear model. SVM allows separating the data when it is tend to be linearly or non-linearly separable through a linear decision surface (hyperplane) based on the maximum distance found between the surface and the nearest points of the two classes [20]. The distance between the hyperplane and the closest examples should be maximized to separate its input space into two classes. Also, this algorithm can be used in a non-linear classification using a non-linear kernel, which is a mathematical function that transforms input data to a high dimensional feature space [7].

J48 belongs to decision tree family. It is considered being a powerful classifier and hierarchical structure for supervised learning. This classifier can be used for both classification and regression, even though they it is more frequently for classification [1]. The learning process is based on splitting the labels of training data into subsets according to statistical tests under divide-and-conquer heuristic [1, 22].

Random Forest [4] consists of many decision trees for making a decision based on a response of each decision tree. This algorithm is able to handle the missing values, as well as to compute generalization errors, and to identify relevant variables. Also, it is considered as a potential algorithm for building classifiers due to the selection of a random subset of input features [6, 19]. It has shown to work well on a large corpus with a large number of features. This classifier has a collection of decision trees [22] and combines multiple Random Tree, which is built with a number of random features and stochastic process [28].

K-Nearest Neighbor is a non-supervised classifier based on pattern recognition. The basic idea is to classify a new pattern with the most probable class according to its $k$ nearest neighbors [2].

A simplified representation of several machine learning algorithms is given in Figure 2. Table shows the advantages and limitations of the algorithms used in this study [5, 7, 22].

A pictorial depiction of the classification task performed in this study is shown in Figure 3.

## 4 Experimental Results

The aim of this study is to determine with a high accuracy the type of cancer a patient is suffering from. This means whether he/she has colon or brain

Table 1: A comparison among the algorithms used in this study.

| **Supervised algorithm** | **Advantages** | **Limitations** |
|---|---|---|
| SVM | – Lower risk of over-fitting.<br>– Can achieve a nonlinear separating hyperplane.<br>– Computational complexity reduced to quadratic optimization problem. | – Training can be slow.<br>– Difficult when training data is not linearly separable.<br>– The structure of algorithm is difficult.<br>– Lack the transparency of results.<br>– Speed and size for both training and test.<br>– Selection of kernel function parameters.<br>– High complexity and extensive memory requirements. |
| KNN $K = 3$ | – Fast classification of instances.<br>– The cost of learning process is zero.<br>– The local approximation is used to learn complex concepts.<br>– Tolerant with missing values and noise.<br>– Assumes similar classification when the instances have similar features.<br>– Can be used with categorical features. | – Computationally expensive complex when number of attributes increases.<br>– The performance depends on the number of dimensions.<br>– Assumes that attributes will be equally relevant.<br>– Slower to update. |
| Naïve Bayes | – Efficient training algorithm.<br>– Consider the relationships between attributes.<br>– Handles discrete, real data and streaming.<br>– Fast to classify instances.<br>– Irrelevant attributes do not affect the performance. | – Assumes independence of features.<br>– Classes must be mutually exclusive.<br>– Frequency of attributes and classes can affect the performance. |
| Decision Tree | – Very intuitive predictors.<br>– Very simple to understand and to interpret<br>– Discover nonlinear relations and interactions.<br>– Can generate rules for helping the knowledge.<br>– Are not affected by outliers. | – Computationally hard to learn.<br>– Not guarantee to return the globally optimal decision tree.<br>– Can be complex and time consuming with large decision tree.<br>– Large trees are not intelligible.<br>– The cost of analysis can be an expensive option. |

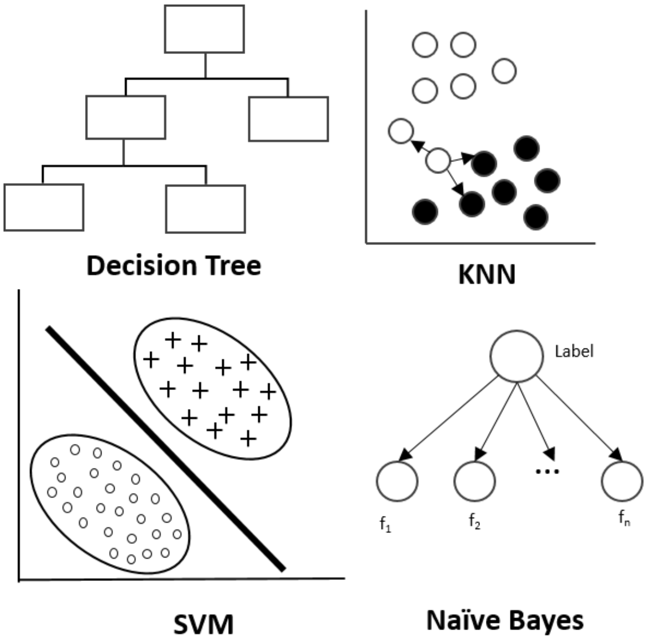

Fig. 2: An example of the algorithms used in this study.

cancer according to pathological reports [26]. The evaluation of the classifiers was carried out in terms of the following metrics: accuracy (A) provides the number of instances that are correctly predicted; precision (P) is the number of retrieved documents that are relevant; recall (R) gives the number of relevant documents that are retrieved; F-measure (F1) denotes a combination of precision and recall [28].

The experiments were carried out on the test set using the well-known data mining tool, WEKA (Waikato Environment for Knowledge Analysis) [28], which has a large collection of implemented machine learning algorithms to perform classification tasks.

We conducted four series of experiments. The first set of experiments consisted in using bag-of-words (BoW) approach with a total 5,860 features. Based on this approach, Naïve Bayes and J48 achieved a classification accuracy of 100.0% as can be seen in Table 2. SVM, Random Forest, Random Tree and KNN also showed high accuracy of 99.01%, 99.01%, 97.03% and 94.06%, respectively. For classification task, a proximity baseline was used. Each document was predicted to be "Brain cancer", which is the majority class.

In the next stage of experiments, we used bigrams of words as features to perform the task. The total number of such features for our dataset is 22,501. The usage of bigrams showed the same accuracy scores for predicting two types of cancer as when using unigrams of words as features. Based on the classifier

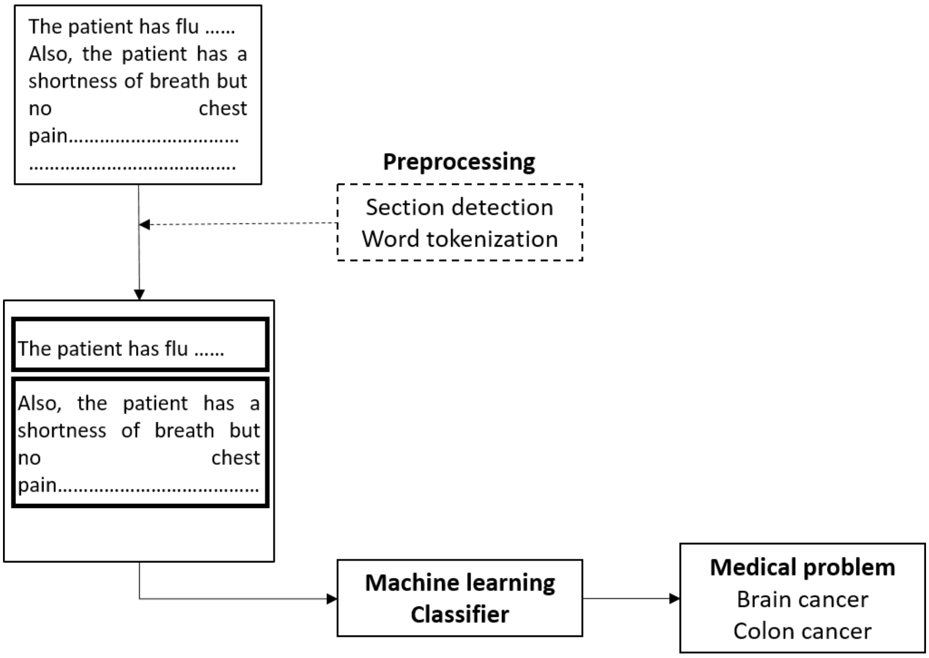
The patient has flu ......
Also, the patient has a shortness of breath but no chest pain.................................
.......................................
Preprocessing
Section detection
Word tokenization
The patient has flu ......
Also, the patient has a shortness of breath but no chest pain.......................................
Machine learning Classifier
Medical problem
Brain cancer
Colon cancer

Fig. 3: An example of document review process.

Table 2: Classification results using bag-of-words (BoW) model and tf-idf weighting scheme.

| Classifier | Acc., % | Precision | | Recall | | F-measure | |
|---|---|---|---|---|---|---|---|
| | | Brain | Cancer | Brain | Cancer | Brain | Cancer |
| **Naïve Bayes** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** |
| **J48** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** |
| SVM | 99.01 | 98.00 | 100.00 | 100.00 | 98.00 | 99.00 | 99.00 |
| Random Forest | 99.01 | 98.00 | 100.00 | 100.00 | 98.00 | 99.00 | 99.00 |
| Random Tree | 97.03 | 98.00 | 96.20 | 96.00 | 98.00 | 97.00 | 97.00 |
| KNN $K = 3$ | 94.06 | 89.30 | 100.00 | 100.00 | 88.20 | 94.30 | 93.06 |
| Baseline | 0.5016 | 1 | 0 | 0.5016 | 0 | 0.6680 | 0 |

performance on the test data, Naïve Bayes and J48 achieved 100.0% of accuracy for detecting two types of cancer (colon and brain), followed by SVM, Random Forest, Random Tree and KNN (see Table 3) with 99.01%, 98.02%, 95.05% and 81.19%, respectively.

Next, another set of experiments are conducted using character $n$-grams ($n = 3$) as features. The results for this experiment are shown in Table 4. Here, Naïve Bayes and J48 outperformed the other classifiers, achieving 100% of accuracy. SVM, KNN, and Random Forest showed 99.01% of accuracy followed by Random Tree with 97.03% of accuracy.

Table 3: Classification results using bigrams of words as features and tf-idf weighting scheme.

| **Classifier** | **Acc., %** | **Precision** | | **Recall** | | **F-measure** | |
|---|---|---|---|---|---|---|---|
| | | Brain | Cancer | Brain | Cancer | Brain | Cancer |
| **Naïve Bayes** | **99.01** | **98.00** | **100.00** | **100.00** | **98.00** | **99.00** | **99.00** |
| **SVM** | **99.01** | **98.00** | **100.00** | **100.00** | **98.00** | **99.00** | **99.00** |
| **Random Forest** | **99.01** | **98.00** | **100.00** | **100.00** | **98.00** | **99.00** | **99.00** |
| J48 | 98.02 | 96.20 | 100.00 | 100.00 | 96.10 | 98.00 | 98.00 |
| Random Tree | 95.05 | 94.10 | 96.00 | 96.00 | 94.10 | 95.00 | 95.00 |
| KNN, $K = 3$ | 81.19 | 72.50 | 100.00 | 100.00 | 62.70 | 84.00 | 77.10 |

Table 4: Classification results using character-level trigrams and tf-idf weighting scheme.

| **Classifier** | **Acc., %** | **Precision** | | **Recall** | | **F-measure** | |
|---|---|---|---|---|---|---|---|
| | | Brain | Cancer | Brain | Cancer | Brain | Cancer |
| **Naïve Bayes** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** |
| **J48** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** |
| SVM | 99.01 | 98.00 | 100.00 | 100.00 | 98.00 | 99.00 | 99.00 |
| KNN $K = 3$ | 99.01 | 98.00 | 100.00 | 100.00 | 98.00 | 99.00 | 99.00 |
| Random Forest | 99.01 | 98.00 | 100.00 | 100.00 | 98.00 | 99.00 | 99.00 |
| Random Tree | 97.03 | 94.30 | 100 | 100.00 | 94.10 | 97.10 | 97.00 |

Finally, another set of experiments was conducted using bag-of-words and binary feature representation (see Table 5). J48 yielded the best accuracy of 100%, followed by Naïve Bayes, SVM, Random Forest, and Random Tree with an accuracy of 99.01%, 99.01%, 99.01%, 95.04% and 78.21%, respectively.

Table 5: Classification results using bag-of-words model and binary feature representation.

| **Classifier** | **Acc., %** | **Precision** | | **Recall** | | **F-measure** | |
|---|---|---|---|---|---|---|---|
| | | Brain | Colon | Brain | Colon | Brain | Colon |
| **J48** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** |
| Naïve Bayes | 99.01 | 98.00 | 100.00 | 100.00 | 98.00 | 99.00 | 99.00 |
| SVM | 99.01 | 98.00 | 100.00 | 100.00 | 98.00 | 99.00 | 99.00 |
| Random Forest | 99.01 | 98.00 | 100.00 | 100.00 | 98.00 | 99.00 | 99.00 |
| Random Tree | 95.04 | 90.90 | 100.00 | 100.00 | 90.20 | 95.20 | 94.80 |
| KNN, $K = 3$ | 78.21 | 69.40 | 100.00 | 100.00 | 56.90 | 82.00 | 72.50 |

The results presented in Tables 2 – 5 indicate that it is possible to identify the type of cancer from the pathology reports using either unigrams, bigrams of words, or character $n$-grams as features for machine learning algorithms.

As can be seen in Figure 4, the results are high for the majority of the examined classifiers. It can be explained by the fact that we considered just two classes representing colon and brain cancer. The very interesting result is that character trigrams with unsupervised KNN algorithm obtained practically the same results as the supervised algorithms (99%). It shows the importance of character $n$-grams for these type of tasks.

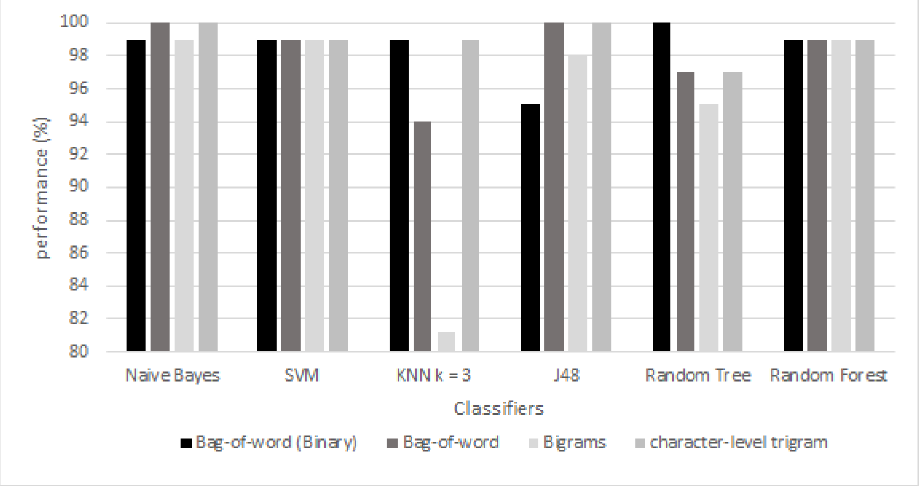

Fig. 4: Results in terms of accuracy.

## 5 Conclusions and Future Work

The main goal of this study consisted in predicting two deadliest types of cancer in the world: colon and brain cancer. The effectiveness of six machine learning classifiers (J48, Naïve Bayes, SVM, Random Forest, Random Tree, and KNN) was examine using different features and feature representations.

We empirically showed that these algorithms are effective for classification task of the leading types of cancer in the world. Naïve Bayes and J48 produced the highest results in the experiments, achieving an accuracy of 100% for detecting colon or brain cancer, when bag-of-words (BoW) and character $n$-grams with tf-idf weighting scheme were used. However, J48 generated the best result with 100 % when bag-of-words model and binary representation is used followed by Naïve Bayes, SVM, Random Forest with 99.01% of accuracy.

Overall, we believe that KNN uses all features for determining the class of a patterns assuming that attributes are equally relevant, unlike decision tree algorithm and Naïve Bayes, which use features that distinguish a disease. This means that Naïve Bayes reduces the number of parameters that must be estimated to learn. It is very interesting that the performance of KNN increases drastically using character trigrams as features (from 84% ot 99%).

The focus of this research aimed at predicting cancer risk using a corpus with pathological information without regard to the limitations in the corpus, which has two classes. This study can provide a great help to physicians to detect colon and brain cancer at the early stage, which can contribute to assign a curative treatment on time and save lives.

In future work, we will conduct experiments with more classes to make the task more challenging, that is, we will include other types of cancer to be classified. Furthermore, we will apply Latent Semantic Analysis (LSA) in order to reduce the number of dimensions in the vector space model.

## References

1. Alpaydin, E.: Introduction to Machine Learning. The MIT Press, 2nd edn. (2010)
2. Bhuvaneswari, P., Therese, A.: Detection of cancer in lung with k-nn classification using genetic algorithm. Procedia Materials Science 10, 433–440 (2015)
3. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly Media, Inc., 1st edn. (2009)
4. Breiman, L.: Random forests. Mach. Learn. 45(1), 5–32 (Oct 2001), http://dx.doi.org/10.1023/A:1010933404324
5. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. 2(2), 121–167 (Jun 1998), http://dx.doi.org/10.1023/A:1009715923555
6. Chen, H., Lin, Z., Wu, H., Wang, L., Wu, T., Tan, C.: Diagnosis of colorectal cancer by near-infrared optical fiber spectroscopy and random forest. Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy 135, 185–191 (2015)
7. Cruz, J.A., Wishart, D.S.: Applications of machine learning in cancer prediction and prognosis. Cancer Informatics 2, 59–77 (2006)
8. Durgalakshmi, B., Vijayakumar, V.: Progonosis and modelling of breast cancer and its growth novel naïve bayes. Procedia Computer Science 50, 551–553 (2015)
9. Gómez-Adorno, H., Markov, I., Sidorov, G., Posadas-Durán, J., Fócil-Arias, C.: Compilación de un lexicón de redes sociales para la identificación de perfiles de autor [Compiling a lexicon of social media for the author profiling task] (in Spanish, abstract in English). Research in Computing Science 115 (2016)
10. Gómez-Adorno, H., Markov, I., Sidorov, G., Posadas-Durán, J., Sanchez-Perez, M.A., Chanona-Hernandez, L.: Improving feature representation based on a neural network for author profiling in social media texts. Computational Intelligence and Neuroscience 2016 (2016)
11. Karabatak, M.: A new classifier for breast cancer detection based on naïve bayesian. In: Measurement: Journal of the International Measurement Confederation. vol. 72, pp. 32–36 (2015)
12. Koopman, B., Zuccon, G., Nguyen, A., Bergheim, A., Grayson, N.: Automatic icd-10 classification of cancers from free-text death certificates. International Journal of Medical Informatics 84(11), 956–965 (2015)
13. Liu, L., Nie, Y., Lin, L., Li, W., Huang, Z., Xie, S., Li, B.: Pattern recognition of multiple excitation autofluorescence spectra for colon tissue classification. Photodiagnosis and Photodynamic Therapy 10(2), 111–119 (2013)

14. Markov, I., Gómez-Adorno, H., Sidorov, G., Gelbukh, A.: Adapting cross-genre author profiling to language and corpus. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings, vol. 1609, pp. 947–955. CLEF and CEUR-WS.org (2016)
15. Meystre, S., Haug, P.: Natural language processing to extract medical problems from electronic clinical documents: Performance evaluation. Journal of Biomedical Informatics 39(6), 589–599 (2006)
16. Mohammed, A., El-tanni, H., El-khatib, H., Mirza, A., El-kashif, A.: Molecular classification of colorectal cancer: Current perspectives and controversies. Journal of the Egyptian National Cancer Institute (2016)
17. Nayyeri, M., Sharifi Noghabi, H.: Cancer classification by correntropy-based sparse compact incremental learning machine. Cold Spring Harbor Labs Journals (2015)
18. Olson, D.L., Delen, D.: Advanced Data Mining Techniques. Springer Publishing Company, Incorporated, 1st edn. (2008)
19. Rastghalam, R., Pourghassem, H.: Breast cancer detection using mrf-based probable texture feature and decision-level fusion-based classification using hmm on thermography images. Pattern Recognition 51, 176–186 (2014)
20. Rathore, S., Hussain, M., Aksam Iftikhar, M., Jalil, A.: Ensemble classification of colon biopsy images based on information rich hybrid features. Computers in Biology and Medicine 47(1), 76–92 (2014)
21. Sánchez-Pérez, M., Sidorov, G., Gelbukh, A.: The winning approach to text alignment for text reuse detection at pan 2014, in working notes of clef 2014. In: Conference and Labs of the Evaluation forum. CEUR Workshop Proceedings, vol. 1180, pp. 1004–1011. CEUR (2015)
22. Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, New York, NY, USA (2014)
23. Sidorov, G., Ibarra Romero, M., Markov, I., Guzman-Cabrera, R., Chanona-Hernández, L., Velásquez, F.: Detección automática de similitud entre programas del lenguaje de programación Karel basada en técnicas de procesamiento de lenguaje natural [Automatic detection of similarity of programs in Karel programming language based on natural language processing techniques (in Spanish, abstract in English)]. Computación y Sistemas 20(2), 279–288 (2016)
24. Sidorov, G., Ibarra Romero, M., Markov, I., Guzman-Cabrera, R., Chanona-Hernández, L., Velásquez, F.: Measuring similarity between Karel programs using character and word n-grams. Programming and Computer Software 43, (in press) (2017)
25. Singh, G., Samavedham, L.: Unsupervised learning based feature extraction for differential diagnosis of neurodegenerative diseases: A case study on early-stage diagnosis of parkinson disease. Journal of Neuroscience Methods 256, 30–40 (2015)
26. Styler, W., Savova, G., Palmer, M., Pustejovsky, J., O'Gorman, T., Groen, P.: Thyme annotation guidelines (2014)
27. Wang, P., Hu, X., Li, Y., Liu, Q., Zhu, X.: Automatic cell nuclei segmentation and classification of breast cancer histopathology images. Signal Processing 122, 1–13 (2016)
28. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2005)
29. Yun-tao, Z., Ling, G., Yong-cheng, W.: An improved tf-idf approach for text classification. Journal of Zhejiang University-SCIENCE A 6(1), 49–55 (2005)