Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2021) 29: 632 – 647

doi:10.3906/elk-2002-31

Automated classification of BI-RADS in textual mammography reports

Mostafa BOROUMANDZADEH, Elham PARVINNIA*
Department of Computer Engineering, Shiraz Branch, Islamic Azad University, Shiraz, Iran

Received: 09.02.2020 • **Accepted/Published Online:** 19.10.2020 • **Final Version:** 30.03.2021

Abstract: The main purpose of this paper is to process key information in medical text records and also classify patients, per different levels of breast imaging-reporting and data system (BI-RADS). The BI-RADS is a scheme for the standardization of breast imaging reports. Therefore, medical text mining is employed to classify mammography reports supported BI-RADS. In this research, a new method is proposed for automated BI-RADS classifications extraction from textual reports and improves the therapeutic procedures. At first, a mammography lexicon is employed for choosing keywords from medical text reports. Word2vec and term frequency inverse document frequency (TFIDF) techniques are used for extracting features, finally, they are combined with the hospital information system (HIS) reports and called With-HIS. The different classifiers like multiclass support vector machine (SVM), naïve Bayesian (NB), extreme gradient boosting (XGBoost), and multilevel fuzzy min-max neural network (MLF) are used so as to compare the accuracy of With-HIS and without HIS (called Without-HIS). The results are confirmed that using HIS beside the proposed approach (Word2vec +TFIDF) encompasses a significant effect on the accuracy of medical text classification. Accuracy within the proposed method with MLF classifier is 0.89% but Without-HIS is 0.85%.

Key words: Breast cancer, patient follow-up, text classification, feature extraction, word2vec

1. Introduction

Breast cancer is one of the most common cancers in females and the main cause of death in cancer illnesses (up to 27% of deaths from all types of cancer) [1–3]. Mammography is the initial imaging method to early detect breast masses. Fast diagnosis with high accuracy is a serious concern for physicians and health centers when they are confronted with certain illnesses [1, 2, 4–6]. However, the accuracy in diagnosis, as well as the quality and speed, can be the boundary between life and death [7]. Breast cancer usually is diagnosed faster than other cancer types, but some of them waste time to diagnose [2, 8–10]. Therefore, a delayed diagnosis may result in an undesirable outcome during the treatment of patients [11–13]. The American College of Radiology (ACR) created the BI-RADS to diminish variation in the radiologists' descriptions of findings used for diagnosis [14]. BI-RADS contains two main sections: (i) a standard lexicon to explain anatomical features available in breast imaging, and (ii) a classification module designed to categorize independently to each breast [7]. Table 1 shows BI-RADS categories and assessment meaning. Seven values of this categorization are from zero to VI and have meanings incomplete, negative, benign findings, probably benign, suspicious abnormality, highly suggestive of malignancy and known biopsy with proven malignancy, respectively. For instance, an incomplete classification, an attempt to ascertain previous imaging, or to call back with the patient for extra views and/or higher quality

*Correspondence: parvinnia@iaushiraz.ac.ir

films are results of this method. Also a BI-RADS classification of IV or V warrants biopsy to further evaluate the lesion [15–17]. In addition to clinical application, BI-RADS is also used in research as a tool for assessment and assuring the quality of health care in mammography, ultrasound imaging, and magnetic resonance imaging (MRI).

Table 1. BI-RADS categories [15].

BI-RADS	Final assessment meaning	Likelihood of cancer
0	Incomplete	Not applicable
I	Negative	Negligible
II	Benign finding	Negligible
III	Probably benign finding	<2%
IV	Suspicious abnormality	(23–34)%
V	Highly suggestive of malignancy	95%
VI	Malignancy confirmed by biopsy	100%

In the two last decades, many studies have shown that the speed and accuracy of the diagnosis of breast cancer can have effective results in treatment. Therefore, automation and improving the accuracy of medical data processing at each stage of the treatment process can be effective results. Some of these studies specialized in deep learning framework techniques [18–21]. In this paper, the prediction process acts on the basis of textual data and using text mining (TM) techniques which useful information can be extracted based on the patients' clinical records at the diagnosis and treatment centers (DTC). It is hypothesized that using TM to analyze textual records from the DTC and Picture archiving and communication system (PACS) may provide clinically reliable data [22–24] to predict BI-RADS levels. TM approaches have been used in medical research findings with variant targets including automatic disease classification of clinical discharges [25], recognition of patients' obesity case [26, 27] and analysis of clinical documents to identify drug-disease associations [28–32].TM covers the gap between structured form of information and free text [23, 24] and uses natural language processing (NLP) techniques, machine learning, and knowledge management to process free text documents. On the other hand, TM has been used to transform essential data from text to logical and numerical format so they can be exported to data storage and analyzed [33]. Here we proposed the hybrid method based on word2vec and TFIDF. At first, we use mammography lexicon to find keyword from medical text reports (called preprocessing), then, word2vec is used to create feature vectors from the selected keywords and secondly, TFIDF calculated and multiplied by the extracted feature vector from word2vec. In this step, each element in the vector of word is added with corresponding elements in the other vectors and calculated the average, we called "feature extraction". Finally, these results are combined with HIS features (called feature engineering). At last, we use SVM, XGBoost, NB, and MLF for automated BI-RADS classifications extraction from breast radiology reports. The real contribution to this paper is as follows: a) Create a dataset using BI-RADS assessments and HIS reports in Namazi Hospital and Saadi Hospital in Fars province, Iran. b) Preprocessing was done and then feature extracted from medical text reports using word2vec and TFIDF (feature extraction). c) Results of step b were combined with HIS (feature engineering). d) Classification using multiclass SVM, XGBOOST, NB, and MLF. e) Comparing the results of the classifier using performance metrics. The rest of this paper is organized as follows: Section 2 provides background information on BI-RADS assessment categories and breast cancer diagnosis, while Section 3 explains the proposed hybrid method. Then we discuss the implementation details

and evaluation results of the proposed algorithm in Section 4. Finally, Sections 5 and 6 conclude the paper and impart insights for future researches.

2. Related works

In recent years, especially between 2015 and 2019, a lot of research has been done to combine TM approaches with medical texts. Some of these studies specialized in machine learning techniques and NLP, others also related directly to decision support systems, feature engineering and feature extraction of textual reports. Some of the most important research was described below:

a) A WF-TF-IDF algorithm to extract keywords from Chinese medical web pages; this study aimed at optimizing TFIDF and improving the precision and recall [34]. b) A symptom extraction system to synthesize the literature on the use of NLP and TM in order to electronic patient-authored text (ePAT) processing [35]. c) Temporal specificity problem was formulated in [23] for text classification such as news documents. d) A method of discovering new keywords based on word embedding, also using word2vec technique to map the words into abstracted n-dimensional vector space; finally, extracting hidden semantic relations between words [36]. e) A developed NLP system to extract all BI-RADS categories from textual radiology reports, using Bayesian, SVM, and PART [7]. f) A classification scheme of hospital treatment texts to classify Japanese medical reports. This study aimed at improving nursing quality; also, the base technique was word2vec [37]. g) An automatic approach to perform a BI-RADS description of density; using multiple kernels hierarchical SVM and a shape-based retrieval strategy [38]. h) A hybrid method based on artificial immune system and fuzzy c-means was proposed for medicine diagnosis such as breast cancer [13]. i) A hybrid comparator system to compare radiologists' comments and output of a density-based image assessment system; using density thresholds and bootstrapping [39]. j) A text classification system based using the naïve Bayesian (NB) learning algorithm to transforms the probability estimation problem into an optimization scheme [40]. According to the papers reviewed, although, most researchers focused on BI-RADS classification and breast cancer diagnosis, using medical text mining, but family history and hospital information are important to predict breast cancer. Papers discussed breast cancer classification and do not consider a hybrid of word2vec and TFIDF feature extraction techniques with HIS. Therefore, using a weighted vector (TFIDF) in the word2vec and were combined them with HIS is the main novelty in our proposed method.

3. Proposed method

In the first step, text processing techniques are done precisely on medical reports, and the information extracted from this operation, based on the mammography lexicon, is given to the proposed method (using word2vec and TFIDF techniques). The output of this step is used to generate a feature vector for each textual report. Then, in order to classify the resulting vectors using SVM, XGBOOST, NB and MLF algorithms. Through the feature engineering, HIS is added and feature vectors are optimized. The proposed method divided into five main modules which is shown in Figure 1: (1) text report preprocessing module, (2) feature extraction module, (3) feature engineering, (4) prediction module, and (5) evaluation module.

3.1. Text report preprocessing module

This module has five steps, including informative text retrieval, normalization, tokenization, keyword selection and conversion to set of words. Each medical text ended with this statement "conclusion: BI-RADS (level)".

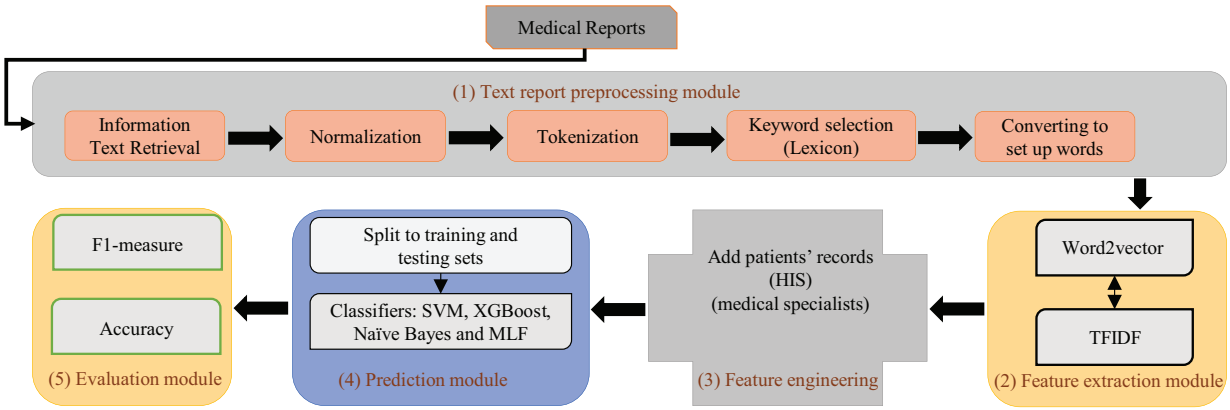


Figure 1. Overview of the main technical road map.

Figure 2 depicts an example of a real report (PACS) in dataset. However, it is just one part of the information employed to examine the patient's condition.

The records were reduced to relevant parts and useful information, and normalization occurred. Punctuation marks and stops words were removed, capital letters substituted, and words were reduced to radicals. Afterward, tokenization was carried out. At this phase, continuous text, breaking down into linguistic units (tokens) such as sentences or words [23, 41]. After identifying the tokens, the words divided into three types; unigrams (i.e. single words), bigrams and trigrams, and these were defined as sequences of one, two or three adjacent words from a list of tokens. The main reference for choosing keywords was the mammography lexicon [42].

```
BILATERAL MAMMOGRAPHY:
Both breasts show scattered areas of fibro glandular density.
There is an 8x15mm circumscribed equal density nodule in upper
central aspect of right breast representing benign nodule.
There is no significant change compared to previous mammography
dated 12.9.92.
There is no skin thickening and nipple retraction.
Reactive axillary lymph nodes in both sides have no clinical
significance.
CONCLUSION: Birads 2 benign noncancerous finding.
```

Figure 2. The sample of medical text reports extracted from the dataset in the PACS system.

3.2. Feature extraction module

In this work, both word2vec and TFIDF techniques are used as a hybrid system for extracting feature vectors. Therefore, their basic concepts are described before describing how to apply them.

3.2.1. Word2vec

Word2vec is an open source tool provided by Google in 2013. The first relevant paper was written by Mikolov [43]. This method is based on neural networks and is used to display words in shape vector view. Word2vec creates a vector distribution in semantic space for each word. This vector can have several dimensions. Word2vec

is a combination of two neural networks called continuous bag-of-words (CBOW) and skip-gram. These networks are trained in such a way that words with a common text context have the same numerical vectors in the semantic space. A sequence of words, called a text context, is shown in Equation 1. Here, k is number of words.

$$Context = \{w_1, w_2, \ldots, w_k\} \tag{1}$$

In CBOW, notable words are predicted from the keywords in their neighborhoods. These neighborhoods are limited and should be determined. Conversely, in skip-gram, words in the neighborhoods of a word are predicted through that word [44].

3.2.2. Term frequency-inverse document frequency

Term frequency-inverse document frequency (TFIDF) is a statistical method that has been considered as an important factor in data retrieval and feature evaluation [45]. TFIDF is the product of two statistics, term frequency (TF), and inverse document frequency (IDF). In TFIDF, the weight is allocated for each word corresponding to the text. Then, the i^{th} words in the j^{th} text are weighted using Equation 2.

$$W_{ij} = tf_{ij} \times \log \frac{N}{df_i} \tag{2}$$

In this regard, W_{ij} represents the weight of i^{th} word in j^{th} text, tf_{ij} indicates the frequency of the i^{th} word in the j^{th} document, N indicates the number of existing documents and df_i illustrates the number of documents containing the i^{th} word [44, 46]. As is clear in Equation 2, since TFIDF only focuses on the repetition of words in the documents, it cannot properly recognize the structural relationship of words in the conceptual medical texts. Therefore, this work tried to combine another learning method with TFIDF.

3.2.3. Hybrid word2vec + TFIDF

In word2vec to display a word, only the element corresponding to the same word takes 1, and other elements are set to 0. An overview of the CBOW and skip-gram networks is shown in Figure 3. This figure shows an example for the vocabulary contains the following sentence: "Breast show scatter area fibroglandular".

As can be seen, the input layer of each network contains one-hot vectors. In the hidden layer of each network, the input vectors are mapped to another space. The number of neurons in the hidden layer indicates the dimensions of the word's vectors (d dimensions). The output layer gives us a one-hot vector for the word (or words) that the study intended to predict. Accordingly, the text content of a given word w with a window size equal to c is determined by Equation 3. The window size is also called the neighborhood distance.

$$Context_c(w_j) = \{w_{j-c}, w_{j-c+1}, \ldots, w_{j-1}, w_{j+1}, w_{j+2}, \ldots, w_{j+c}\} \tag{3}$$

The CBOW aims to maximize $P(w_i|Context_c(w_i))$. To achieve this, a sequence of training words such as $w_1, w_2, \ldots, w_k$ is required. According to Equation 4, applying this sequence will maximize log-likelihood.

$$l_{CBOW} = \frac{1}{k} \sum_{j=1}^{k} \sum_{-c \leq i \leq, i \neq 0} \log P(w_j|w_{j+i}) \tag{4}$$

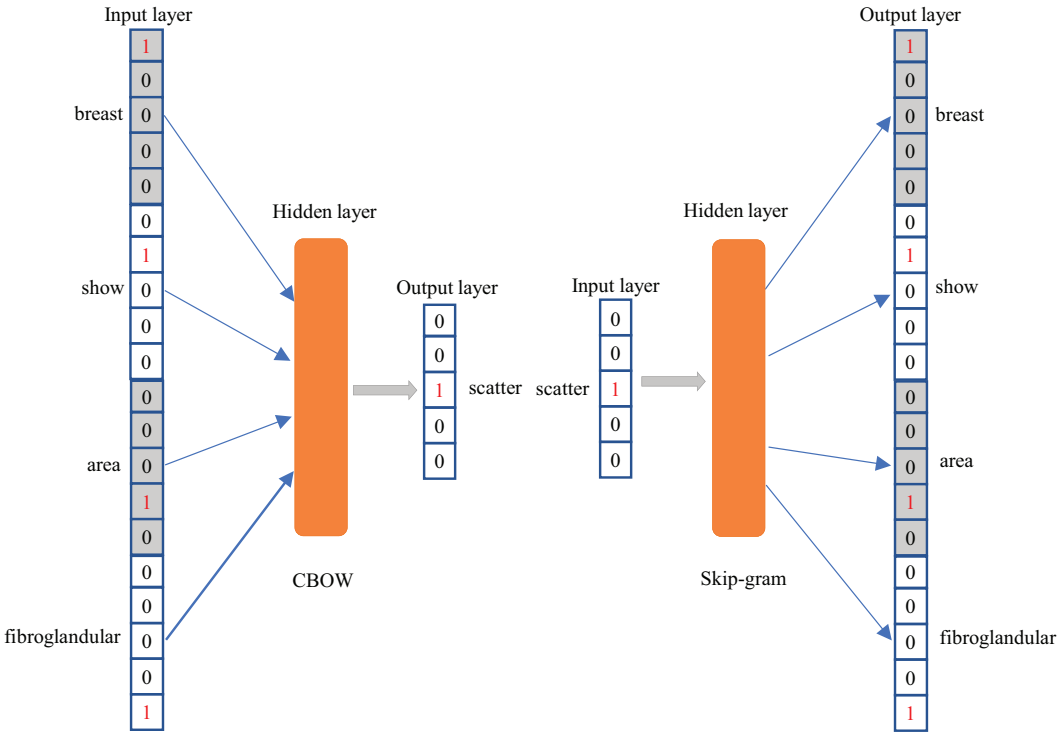


Figure 3. The example of both CBOW and skip-gram structures in a sentence from the medical text report.

At the skip-gram the words in $c(w_j)$ are predicted by w_j and the goal is to maximize $P(Context_c(w_j)|w_j)$. To train this network, a sequence of words such as $w_1, w_2, \ldots, w_k$ is required. To reach the skip-gram goal, according to Equation 5, the mean of log-likelihood must be maximized.

$$l_{Skip-gram} = \frac{1}{k}\sum_{j=1}^{k}\sum_{-c\leq i\leq, i\neq 0} \log P(w_{j+i}|w_j) \tag{5}$$

Generally, $P(w_O|w_I$ is calculated by a function called softmax and based on Equation 6.

$$P(w_O|w_I) = \frac{\exp(V_{w_O}^{'T}.V_{w_I})}{\sum_{w=1}^{W}\exp(V_{w_O}^{'T}.V_{w_I})} \tag{6}$$

$V_{w_O}^{'}$ and V_{w_I} illustrate the output and input vector for the word w , and W denotes the number of words in the vocabulary. There are two methods of training for both the CBOW and skip-gram. These methods are negative sampling and hierarchical softmax [34, 40, 47–51]. Negative sampling works better for repetitive words. In contrast, the second model is suitable for words with low repetition. In this research, the hierarchical softmax method is used, and also according to the trial and error method [49, 52], the number 5 is used for the window size. Now, based on Figure 4, a combination of both TFIDF and word2vec methods has been used to generate numeric vectors for each medical report.

As explained, word2vec generates a numeric vector for each word in the medical text report, then, TFIDF calculated weight for each word. If the text report contains the words $w_1, w_2, \ldots, w_k$, then the vectors generated

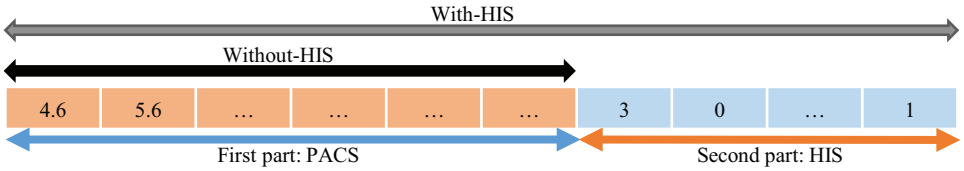


Figure 4. The hybrid system schematic of both TFIDF and word2vec methods to generate report's vectors.

for these words by d dimensions are $v_1, v_2, \ldots, v_k$. Now if the weight of the word i is displayed by TFIDF with t_i, the vector for each textual report derived from the combination of word2vec and TFIDF methods is calculated according to Equation 7.

$$Vectore(report) = \frac{\sum_{i=1}^{k}(t_i \cdot v_i)}{k} = \frac{\sum_{i=1}^{k}(f_i)}{k} \tag{7}$$

Figure 5 depicts an illustration of this interpretation. Therefore, for each medical report, a vector was produced, which was given as input to the SVM, XGBOOST, NB, and MLF algorithms in order to predict the levels of BI-RADS. One of the most important points that distinguish this paper is that selecting keywords and then tries to simultaneously take advantages both of word2vec and TFIDF. In fact, the feature vectors associated with each textual report are based on two important and concurrent points: the content of the report in terms of words order (word2vec), as well as the number of repetitions and the weight of the keywords in a report and all reports (TFIDF).

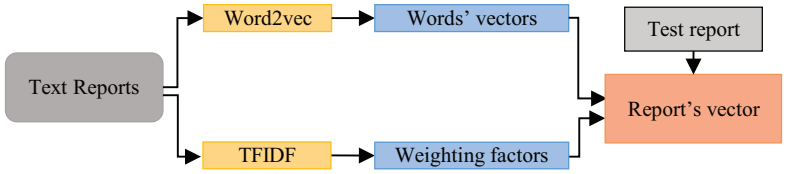


Figure 5. Technical combination of word2vec and TFIDF to generate a numeric vector of each medical report.

3.3. Hospital information system features (HIS)

In order to enhance the quality of recognition and classification, in addition to vector extracted from medical text reports we have extracted HIS from PACS. In this regard, there were 20 features in HIS, we have asked 5 medical specialists to score each feature based on the importance from one to five. Then, based on the average, the 7 most important features were selected.

3.4. Combination of HIS and word2vec+TFIDF

Combining HIS and word2vec+TFIDF is done to obtain feature vectors. This combination is shown in Figure 6. The first part is the numeric vectors extracted from medical text reports using the TFIDF and word2vec methods (shown in Figure 5). In the second part, the values associated with the features specified in the HIS information refer to the same patients. Obviously, feature vectors generated in both methods are independently sent to classification algorithms.

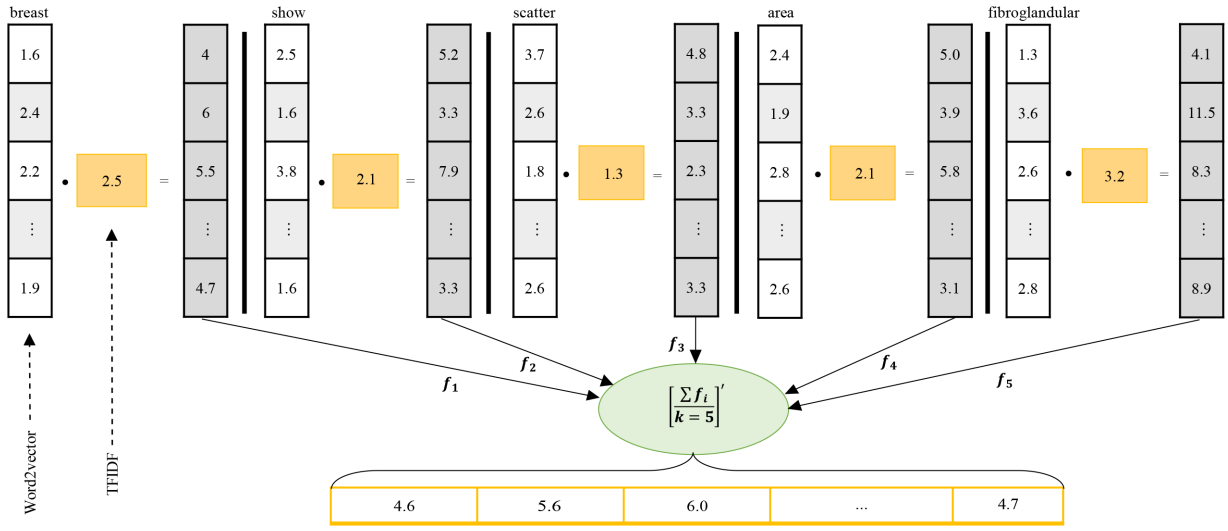


Figure 6. Combined numeric feature vector consisting of text reports of the PACS system, and HIS.

3.5. Prediction module

In this section, feature vectors generated from the previous step are sent to classification algorithms. These algorithms (XGBoost, SVM, NB, and MLF) have been used to predict BI-RADS levels in patient reports.

3.5.1. Multiclass support vector machine (SVM)

SVM widely used in the textual classification algorithms. The underlying idea of the SVM algorithm is achieving to the maximum possible margin. In the case of multiclass classification problems, like what has been done in this paper, one of two common approaches, namely one-against-all and one-against-one, may be preferred to adopt two-class classification to a multiclass case [50]. In this paper, the RBF kernel function was used.

3.5.2. Extreme gradient boosting (XGBoost)

XGBoost is an efficient implementation of the gradient boosting framework and it is proposed by Tianqi Chen [53]. This algorithm includes various functions of regression and classification. XGBoost has high predictive power and uses tree learning algorithms and linear models. Some of the most important goals in using this algorithm include reducing computational time, without changing productivity [54].

3.5.3. Applying multilevel fuzzy min-max neural network (MLF)

MLF is the evolved type of fuzzy min-max neural network (FMNN) [46]. Min-max neural network uses hyper boxes to classify samples. A fuzzy set hyper box is an n-dimensional box defined by a minimum point and a maximum point with a corresponding membership function. Each hyper box belongs to a class. Hyper boxes are created and configured with the arrival of training samples at the time of network training. Equation 8 defines the hyper box.

$$B_j = \{X, V_j, W_j, f(X, V_j, W_j) \forall X \in I^n\} \tag{8}$$

V_j and W_j represent the maximum and minimum points of a hyper box. X represents a sample, and n represents the dimensions of the feature vectors.The size of these *hyper boxes* is controlled by Equation 9.

$$\forall_{i=1\ldots D}(max(w_b^i, x^i) - min(v_b^i, x^i)) \leq \theta \tag{9}$$

In this equation, θ is called the coefficient of expansion. This algorithm consists of three layers. The first layer is related to the inputs, the second layer is related to hyper boxes, and the third layer is related to output or classes [46].

3.5.4. Naïve Bayesian (NB)

In this work, naïve Bayesian is used as a classifier. This technique is useful because it can calculate the probability of an event by making it conditional on the occurrence or nonoccurrence of another event. The naïve Bayesian classification method assumes that the values of the features are conditional independent of each other with the values of the objective function. In other words, this assumption shows that, when the output of the objective function is observed, the probability of observing the features is equal to multiplying the probabilities of each feature separately [40].

4. Experimental setup

A computer with Intel Skylake Core i7-6700 K Processor, 4*8 GB DDR RAM, GTX 1080 VGA and 256GB SSD, 1TB SATA HDD is used for running the proposed method. Our methods are implemented in Python 3.7 in Spyder environment running on Windows 10 to compare results. In order to compare the quantity of Word2vec+TFIDF parameters, are used three basic algorithms: word2vec (With-HIS), word2vec (Without-HIS) and word2vec + TFIDF (Without-HIS). Initially, in addition to the above, two algorithms were considered for (TFIDF with HIS) and (TFIDF without HIS) modes, which have eliminated due to their very low classification accuracy. Finally, we compare the main proposed method (word2vec+TFIDF with HIS) and (word2vec+TFIDF without HIS) in four different classifiers. We used 4 different performance metrics to compare the results such as accuracy, precision, recall, and F1-measure. To calculate these metrics we need to define: true negative, true positive, false positive and false negative. True negative (TN) represents the number of patients whose BI-RADS levels are predicted by the system as "false", and this prediction is accurate. True positive (TP) represents the number of patients whose BI-RADS levels are predicted by the system as "correct", and this prediction is accurate. False positive (FP) represents the number of patients whose BI-RADS levels are predicted by the system as "correct", but this prediction is not accurate. False negative (FN) represents the number of patients whose BI-RADS levels are predicted by the system as "false", but this prediction is not accurate. TN, TP, FP, and FN are typically defined for the binary classification. In this work, in order to calculate recall, precision, F1-measure, and accuracy criteria in multiclass classification, it is necessary to calculate these criteria for each class (relative to other classes) separately and calculate the mean of them. Accordingly, Equations 10– 13 define accuracy, F1-measure, precision and recall metrics, respectively [55, 56].

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (10)$$

$$Precision = \frac{TP}{TP + FP} \quad (11)$$

$$Recall = \frac{TP}{TP + FN} \quad (12)$$

$$F1 - Measure = \frac{2 \times Precision \times Recall}{recision + Recall} \quad (13)$$

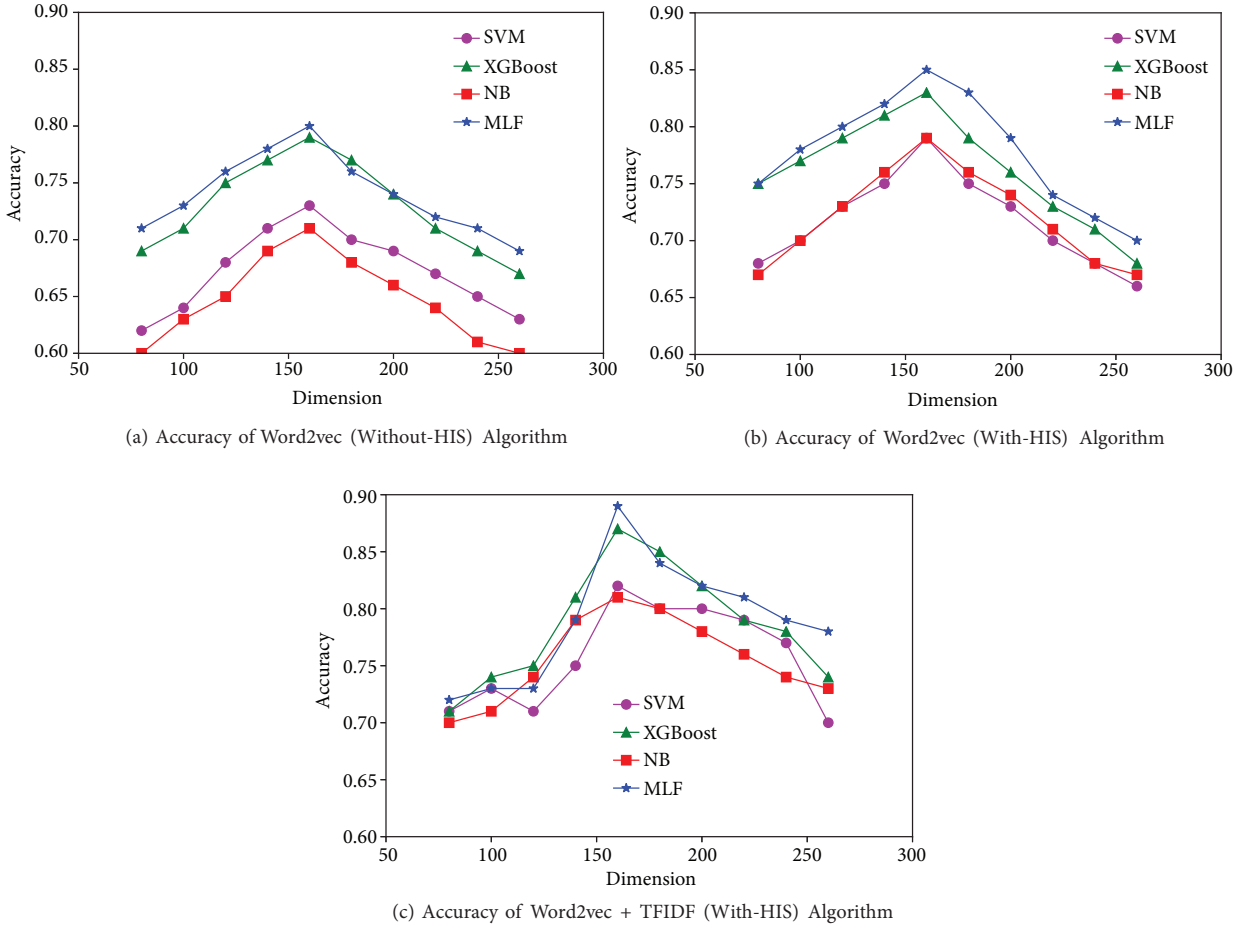
4.1. Dataset

Our dataset contains two major resources: medical text reports and HIS, extracted from the PACS and patient records respectively. The PACS includes electronic records for storing and retrieving medical images and related documentation and reports. The HIS is an integrated information system, created to cover all aspects of the hospital's operation, such as financial, administrative, medical and legal issues services. Dataset uses information from the PACS systems at Namazi Hospital and Saadi Hospital in Fars province, Iran. This dataset includes medical reports of 5076 patients. Some of the extracted key features and elements presented in medical text reports and lexicon key words were: density (fat, low, equal or high), asymmetry (asymmetry, global, focal or developing), associated features (skin retraction, nipple retraction, trabecular thickening, parenchyma with no visible mass or etc.), distribution (diffuse, regional, grouped or etc.), typically benign (bilateral, right or bilateral), suspicious (for instance, coarse heterogeneous, fine pleomorphic, amorphous, fine linear or etc.), size (for instance, 15, 19 or etc.), breast-quad (N or Y), margin (circumscribed, obscured, microlobulated, indistinct or speculated), shape (oval, round or irregular), composition (A: entirely fatty, B: scattered areas of fibro-glandular density or etc.) and so forth. Also, some of the features reviewed by medical experts and patients' records (related to HIS) were: menopausal (0 or 1), lactation history (for instance, 0, 2 or etc.), sports activities (0 or 1), pregnancy history (for instance, 0, 2 or etc.), marital status (single or married), age (for instance, 39, 57 or etc.), cancer family history (0 or 1) and so forth. It is observed that the proportions of 25.02% (1270 patients), 8.75% (444 patients), 30.00% (1523 patients), 7.49% (380 patients), 12.51% (635 patients), 7.49% (380 patients), and 8.75% (444 patients) of the patients under study, were placed in the levels of BI-RADS 0 to BI-RADS VI, respectively. The highest proportion was related to BI-RADS II and the lowest ratio was related to BI-RADS III.

5. Results and discussion

Here, we have processed medical reports to extract BI-RADS using NLP, TFIDF, and HIS. While several papers on the useful extraction of clinical information from mammography reports have been published through various NLP systems [7, 35], but a system based on medical texts and HIS that can detect BIRADS has not been studied. Therefore, in this work, using word2vec and TFIDF, the feature vectors have been extracted from the medical texts and then the important features of HIS have been selected according to the medical specialist's opinion, and with the vectors of the previous stage, it has been used for classification. In this section, the results of the proposed method are discussed in the details. Figure 7 focuses on the effect of different dimensions of the feature vectors on the accuracy of classifiers. In fact, when mapping words into vector space, using word2vec, there is no single rule for determining the number of vectors' dimensions [52]. In fact, the selection of appropriate dimensions is based on frequent experiments. Therefore, to obtain the best results in the dataset used, the trial and error method was used, and vector dimensions of 80 to 260 were evaluated based on the papers. The accuracy of the classifiers has increased with increasing the dimensions to reach a maximum in 160. Then the accuracy decreases with increasing dimensions. SVM, NB, XGBoost, and MLF classifiers have reached to maximum accuracy at 160. So, we selected the vector with 160 dimensions. Decreased accuracy is due to increased dimensions due to computational error [52].

As mentioned above, dimensions 160 is selected for the vectors because it has the best accuracy in any Algorithm. Figure 7a depicts the accuracy of the word2vec (Without-HIS) Algorithm. Figures 7b and 7c, respectively, clearly show the effect of HIS interference and the use of TFIDF in the proposed method in increasing the accuracy of classifiers compared to Figure 7a. The distribution of patient proportions in the



(a) Accuracy of Word2vec (Without-HIS) Algorithm

(b) Accuracy of Word2vec (With-HIS) Algorithm

(c) Accuracy of Word2vec + TFIDF (With-HIS) Algorithm

Figure 7. Comparison of the classifiers' accuracy in different algorithms with various vectors' dimensions.

employed approach is consistent with the actual levels of BIRADS. Table 2 compares the accuracy and F1-measure values of NB, SVM, XGBOOST, and MLF classifiers for classifying BIRADS levels with and without HIS contributions. The results demonstrate that the MLF algorithm is more efficient and differs slightly from the XGBOOST method. The accuracy of the NB, SVM, XGBOOST, and MLF methods for word2vec + TFIDF (Without-HIS) are 76%, 77%, 82%, and 85%, respectively. Accuracy for word2vec + TFIDF (With-HIS) in different classifier is NB (81%), SVM (82%), XGBOOST (87%) and MLF (89%). Also, a comparison of the NB, SVM, XGBOOST with precision, recall and F1-measure metrics in word2vec + TFIDF (With-HIS) is denoted in Table 2. The results illustrate that the MLF algorithm, with respect to the extracted vector of the combined method, is more efficient.

Table2 shows the average of precisions for all classes. As it shown above the word2vec + TFIDF (With-HIS) method performs better than others. Precision is the ratio of classified samples by the classifier in a given class, to the total number of samples the classifier has classified in that class, either correctly or incorrectly. As it turns out from Equation 11, the precision shows what proportion of the detected positives are really positive. The NB, SVM, XGBoost, and MLF classrooms have 75%, 77%, 85%, and 86% accuracy, respectively. This shows that MLF is more accurate than other classes because of fast one-shot training. Another parameter

Table 2. Comparison of evaluation metrics in different algorithms and the proposed method.

Metrics	Algorithms	NB	SVM	XGBoost	MLF
Accuracy	Word2vec (Without-HIS)	71%	73%	79%	80%
	Word2vec (With-HIS)	79%	79%	83%	85%
	Word2vec + TFIDF (Without-HIS)	76%	77%	82%	85%
	Word2vec + TFIDF (With-HIS)	81%	82%	87%	89%
Precision	Word2vec + TFIDF (With-HIS)	75%	77%	85%	86%
Recall	Word2vec + TFIDF (With-HIS)	77%	79%	86%	87%
F1-measure	Word2vec + TFIDF (With-HIS)	76%	78%	85%	86%

shown in Table2 is the average of recall for classes. This parameter for word2vec + TFIDF (With-HIS) is 87%, and in other methods i.e. NB = 77%, SVM = 79%, and XGBoost = 86%, therefore, this indicates that the MLF performs better than other methods. The recall shows the ratio of true classification of samples in a given classes by the classifier to the number of samples in that class. So, the recall shows what proportion of true positives are correctly identified as positive. Therefore, like precision, MLF predict better each classes. Positive here actually represents each class of BI-RADS, so by combining the precision values and recall and calculate the F1-measure, we can conclude that MLF is more accurate than other methods in detecting different types of BI-RADS. As it is mentioned, since we are faced with a multiclass mode in this work, Tables3, represent the actual classes and the predicted classes. It can be deduced that the boundary between the classes defined for BI-RADS is very sensitive.

Table 3. Confusion matrixes related to the a) NB, b) SVM, c) MLF, and d) XGBoost classifiers.

a. NB–predicted classes

		Without-His							With-His						
	BI-RADS	VI	V	IV	III	II	I	0	VI	V	IV	III	II	I	0
Actual classes	0	2	4	3	2	5	10	227	5	4	3	2	2	1	236
	I	9	3	2	2	5	66	2	2	0	2	3	5	75	2
	II	2	1	3	5	284	9	1	3	2	1	2	295	1	1
	III	5	3	2	49	1	14	2	5	0	2	52	1	14	2
	IV	2	3	115	1	3	3	0	2	1	117	1	3	4	3
	V	2	57	2	2	1	9	3	3	58	2	2	1	9	1
	VI	79	3	0	3	0	2	2	76	3	0	3	3	2	2

b. SVM–predicted classes

		Without-His							With-His						
	BI-RADS	VI	V	IV	III	II	I	0	VI	V	IV	III	II	I	0
Actual classes	0	10	5	11	14	8	10	195	5	5	3	6	4	5	225
	I	5	2	6	5	1	67	3	3	3	4	1	6	69	3
	II	10	8	13	8	245	16	5	5	6	5	6	262	12	9
	III	4	5	4	55	4	3	1	3	0	4	62	1	4	2
	IV	0	1	105	6	8	3	1	9	4	101	3	2	4	4
	V	2	59	2	2	2	3	6	4	59	5	2	2	4	0
	VI	57	4	5	6	3	8	6	61	4	4	1	9	6	4

c. MLF–predicted classes

		Without-His							With-His						
	BI-RADS	VI	V	IV	III	II	I	0	VI	V	IV	III	II	I	0
Actual classes	0	3	4	6	13	7	6	214	3	4	3	3	4	2	234
	I	1	2	2	2	2	77	3	4	0	2	3	1	76	3
	II	5	3	6	9	271	5	6	8	4	5	5	277	2	4
	III	2	4	4	61	1	1	3	2	0	5	64	2	1	2
	IV	6	3	105	5	0	5	3	3	2	114	1	5	2	0
	V	1	68	1	1	2	2	1	0	68	3	2	1	2	0
	VI	68	3	3	5	4	4	2	72	1	6	1	4	4	1

d. XGBoost–predicted classes

		Without-His							With-His						
	BI-RADS	VI	V	IV	III	II	I	0	VI	V	IV	III	II	I	0
Actual classes	0	5	5	1	2	1	6	233	0	1	1	2	0	1	248
	I	2	1	2	2	5	75	2	1	0	2	2	5	77	2
	II	2	1	2	1	296	2	1	2	1	2	1	296	2	1
	III	3	0	3	63	1	3	3	3	0	2	65	1	3	2
	IV	6	3	112	1	3	1	1	1	3	117	1	3	1	1
	V	2	63	2	3	1	2	3	2	63	2	3	1	2	3
	VI	73	3	0	4	1	5	3	79	3	0	3	0	2	2

5.1. Limitations

Here, we used only the standard packages in natural language processing and we do not use a medical specialized dictionary for text processing. Otherwise, the generalization of the study may also be limited to the BI-RADS annotator because the annotator detection phase relies on the precision of preprocessing steps. This method may have the capacity to serve as a basis for future studies that will reinforce the BI-RADS labeling to assists the radiologists and physicians as part of the learning health system. Also using molecular subtypes and deep image mining for improving accuracy is proposed for future works.

6. Conclusion

Hybrid word2vec technique with TFIDF can increase the accuracy of text classification, but the medical history of patients is important in diagnosing disease and can improve accuracy. Therefore, we proposed HIS beside word2vec and TFIDF using feature engineering. In order to evaluate the proposed method, we have used four classifiers such as SVM, NB, XGBoost, and MLF. In this context, MLF is more accurate than other classifiers (accuracy = 89%). Precision, recall, and F1-measure are also assessed for all classifiers. F1-measure for the SVM, NB, XGBoost, and MLF are 78%, 76%, 85%, and 86%, respectively. The results suggest that, just focus on medical reports and do not use other clinical information and medical history of patients because of human error in writing can cause of errors in results, therefore the use of HIS beside medical text reports can improve BI-RADS classification and cause a positive effect on treatment procedures.

Acknowledgment

We are grateful to the Department of Computer Engineering, Islamic Azad University, Branch of Shiraz, Iran, and also Dr. Sepideh Sefidbakht, from the Department of Radiology, Shiraz University of Medical Sciences, Shiraz, Iran.

Compliance with ethical standards

This paper does not contain any studies with human participants performed by any of the authors.

References

[1] Koo MM, Von Wagner C, Abel GA, McPhail S, Rubin GP et al. Typical and atypical presenting symptoms of breast cancer and their associations with diagnostic intervals: Evidence from a national audit of cancer diagnosis. Cancer Epidemiology 2017; 48: 140-146. doi: 10.1016/j.canep.2017.04.010

[2] Redaniel MT, Martin RM, Ridd MJ, Wade J, Jeffreys M. Diagnostic intervals and its association with breast, prostate, lung and colorectal cancer survival in England: historical cohort study using the Clinical Practice Research Datalink. PLoS ONE 2015; 10 (5). doi: 10.1371/journal.pone.0126608

[3] Wang M, Yang Z, Liu C, Yan J, Zhang W et al. Differential diagnosis of breast category 3 and 4 nodules through BI-RADS classification in conjunction with shear wave elastography. Ultrasound in Medicine and Biology 2017; 43 (3): 601-606. doi: 10.1016/j.ultrasmedbio.2016.10.004

[4] Lucini FR, Fogliatto FS, Da Silveira GJ, Neyeloff JL, Anzanello MJ et al. Text mining approach to predict hospital admissions using early medical records from the emergency department. International journal of medical informatics 2017; 100: 1-8. doi: 10.1016/j.ijmedinf.2017.01.001

[5] Savoie B, Nagy P. PACS and the potential for medical errors. Journal of the American College of Radiology 2012; 9 (10): 756-758. doi: 10.1016/j.jacr.2012.06.021

[6] Shan J, Alam SK, Garra B, Zhang Y, Ahmed T. Computer-aided diagnosis for breast ultrasound using computerized BI-RADS features and machine learning methods. Ultrasound in medicine and biology 2016; 42 (4): 980-988. doi: 10.1016/j.ultrasmedbio.2015.11.016

[7] Castro SM, Tseytlin E, Medvedeva O, Mitchell K, Visweswaran S et al. Automated annotation and classification of BI RADS assessment from radiology reports. Journal of Biomedical Informatics 2017; 69: 177-187. doi: 10.1016/j.jbi.2017.04.011

[8] Lyratzopoulos G, Abel G, McPhail S, Neal R, Rubin G. Measures of promptness of cancer diagnosis in primary care: secondary analysis of national audit data on patients with 18 common and rarer cancers. British Journal of Cancer 2013; 108 (3): 686-690. doi: 10.1038/bjc.2013.1

[9] Hansen RP, Vedsted P, Sokolowski I, Søndergaard J, Olesen F. Time intervals from first symptom to treatment of cancer: a cohort study of 2,212 newly diagnosed cancer patients. BMC Health Services Research 2011; 11 (1): 284. doi: 10.1186/1472-6963-11-284

[10] Neal R, Din N, Hamilton W, Ukoumunne O, Carter B et al. Comparison of cancer diagnostic intervals before and after implementation of NICE guidelines: analysis of data from the UK General Practice Research Database. British Journal of Cancer 2014; 110 (3): 584-592. doi: 10.1038/bjc.2013.791

[11] Webber C, Jiang L, Grunfeld E, Groome PA. Identifying predictors of delayed diagnoses in symptomatic breast cancer: a scoping review. European Journal of Cancer Care 2017; 26 (2): e12483. doi: 10.1111/ecc.12483

[12] Mendonca SC, Abel GA, Saunders CL, Wardle J, Lyratzopoulos G. Pre-referral general practitioner consultations and subsequent experience of cancer care: evidence from the English cancer patient experience survey. European Journal of Cancer Care 2016; 25 (3): 478-490. doi: 10.1111/ecc.12353

[13] Özşen S, Ceylan R. Comparison of AIS and fuzzy c-means clustering methods on the classification of breast cancer and diabetes datasets. Turkish Journal of Electrical Engineering & Computer Silences 2014; 22 (5): 1241-1254. doi: 10.3906/elk-1210-62

[14] Gürüf A, Öztürk M, Bayrak İK, Polat A. Shear wave versus strain elastography in the differentiation of benign and malignant breast lesions. Turkish Journal of Medical Sciences 2019; 49 (5): 1509-1517. doi: 10.3906/sag-1905-15

[15] Levy L, Suissa M, Chiche J, Teman G, Martin B. BIRADS ultrasonography. European Journal of Radiology 2007; 61 (2): 202-211. doi: 10.1016/j.ejrad.2006.08.035

[16] Mundinger A, Wilson A, Weismann C, Madjar H, Heindel W et al. E5. Breast ultrasound-update. European Journal of Cancer Supplements 2010; 8 (3): 11-14. doi: 10.1016/S1359-6349(10)70009-4

[17] Boyer B, Canale S, Arfi-Rouche J, Monzani Q, Khaled W et al. Variability and errors when applying the BIRADS mammography classification. European Journal of Radiology 2013; 82 (3): 388-397. doi: 10.1016/j.ejrad.2012.02.005

[18] Fei G, Hyunsoo Y, Wu T, Xianghua C. A feature transfer enabled multi-task deep learning model on medical imaging. Expert Systems with Applications 2020; 143: 1-11. doi: 10.1016/j.eswa.2019.112957

[19] Karim AM, Güzel MS, Tolun MR, Kaya H, Çelebi FV. A new generalized deep learning framework combining sparse autoencoder and Taguchi method for novel data classification and processing. Mathematical Problems in Engineering 2018; 6: 1-13. doi: 10.1155/2018/3145947

[20] Karim AM, Güzel MS, Tolun MR, Kaya H, Çelebi FV. A new framework using deep auto-encoder and energy spectral density for medical waveform data classification and processing. Biocybernetics and Biomedical Engineering 2018; 39 (1): 148-159. doi: 10.1016/j.bbe.2018.11.004

[21] Zobeidi S, Naderan M, Alavi SE. Opinion mining in Persian language using a hybrid feature extraction approach based on convolutional neural network. Multimedia Tools and Applications 2019; 78 (22): 32357-32378. doi: 10.1007/s11042-019-07993-4

[22] Yao H, Zhang B, Zhang P, Li M. A novel kernel for text classification based on semantic and statistical information. Computing and Informatics 2018; 37 (4): 992-1010. doi: 10.4149/cai 2018 4 992

[23] Khan S, Islam A, Aleem M, Iqbal M. Temporal specificity-based text classification for information retrieval. Turkish Journal of Electrical Engineering & Computer Silences 2018; 26 (6): 2916-2927. doi: 10.3906/elk-1711-136

[24] Spasić I, Livsey J, Keane JA, Nenadić G. Text mining of cancer-related information: review of current status and future directions. International Journal of Medical Informatics 2014; 83 (9): 605-623. doi: 10.1016/j.ijmedinf.2014.06.009

[25] Solt I, Tikk D, Gál V, Kardkovács ZT. Semantic classification of diseases in discharge summaries using a context-aware rule-based classifier. Journal of the American Medical Informatics Association 2009; 16 (4): 580-584. doi: 10.1197/jamia.M3087

[26] Yang H, Spasic I, Keane JA, Nenadic G. A text mining approach to the prediction of disease status from clinical discharge summaries. Journal of the American Medical Informatics Association 2009; 16 (4): 596-600. doi: 10.1197/jamia.M3096

[27] Ambert KH, Cohen AM. A system for classifying disease comorbidity status from medical discharge summaries using automated hotspot and negated concept detection. Journal of the American Medical Informatics Association 2009; 16 (4): 590-595. doi: 10.1197/jamia.M3095

[28] Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated acquisition of disease-drug knowledge from biomedical and clinical documents: an initial study. Journal of the American Medical Informatics Association 2008; 15 (1): 87-98. doi: 10.1197/jamia.M2401

[29] Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. Journal of the American Medical Informatics Association 2009; 16 (3): 328-337. doi: 10.1197/jamia.M3028

[30] Michelson JD, Pariseau JS, Paganelli WC. Assessing surgical site infection risk factors using electronic medical records and text mining. American Journal of Infection Control 2014; 42 (3): 333-336. doi: 10.1016/j.ajic.2013.09.007

[31] Yang M, Kiang M, Shang W. Filtering big data from social media-Building an early warning system for adverse drug reactions. Journal of Biomedical Informatics 2015; 54: 230-240. doi: 10.1016/j.jbi.2015.01.011

[32] Vallmuur K. Machine learning approaches to analysing textual injury surveillance data: a systematic review. Accident Analysis & Prevention 2015; 79: 41-49. doi: 10.1016/j.aap.2015.03.018

[33] Günal S. Hybrid feature selection for text classification. Turkish Journal of Electrical Engineering & Computer Silences 2012; 20 (2): 1296-1311. doi: 10.3906/elk-1101-1064

[34] Sun P, Wang L, Xia Q. The keyword extraction of Chinese medical web page based on WF-TF-IDF algorithm. In: International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC); Nanjing, China; 2017. p. 193-198. doi: 10.1109/CyberC.2017.40

[35] Dreisbach C, Koleck TA, Bourne PE, Bakken S. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. International Journal of Medical Informatics 2019; 125: 37-46. doi: 10.1016/j.ijmedinf.2019.02.008

[36] Ding X, Zhang X. Research on text structuralization in medical field. In: 2nd International Conference on Cloud Computing and Internet of Things (CCIOT); Dalian, China; 2016. p. 155-161. doi: 10.1109/CCIOT.2016.7868324

[37] Nii M, Tuchida Y, Iwamoto T, Uchinuno A, Sakashita R. Nursing-care text evaluation using word vector representations realized by word2vec. In: IEEE International Conference on Fuzzy Systems (FUZZ-IEEE); Vancouver, BC, Canada; 2016. pp. 2165-2169. doi: 10.1109/FUZZ-IEEE.2016.7737960

[38] Narváez F, Díaz G, Poveda C, Romero E. An automatic BI-RADS description of mammographic masses by fusing multiresolution features. Expert Systems with Applications 2017; 74: 82-95. doi: 10.1016/j.eswa.2016.11.031

[39] Østerås BH, Martinsen ACT, Brandal SHB, Chaudhry KN, Eben E et al. BI-RADS density classification from areometric and volumetric automatic breast density measurements. Academic Radiology 2016; 23 (4): 468-478. doi: 10.1016/j.acra.2015.12.016

[40] Diab DM, El Hindi KM. Using differential evolution for fine tuning naïve Bayesian classifiers and its application for text classification. Applied Soft Computing 2017; 54: 183-199. doi: 10.1016/j.asoc.2016.12.043

[41] Zhang L, Hu X. Word combination kernel for text classification with support vector machines. Computing and Informatics 2014; 32 (4): 877-896. doi: 10.1016/j.irbm.2016.03.002

[42] Phillips J, Fein-Zachary V, Slanetz P. Pearls and pitfalls of contrast-enhanced mammography. Journal of Breast Imaging 2019; 1 (1): 64-72. doi: 10.1093/jbi/wby013

[43] Hu K, Luo Q, Qi K, Yang S, Mao J et al. Understanding the topic evolution of scientific literatures like an evolving city: using Google word2Vec model and spatial autocorrelation analysis. Information Processing & Management 2019; 56 (4): 1185-1203. doi: 10.1016/j.ipm.2019.02.014

[44] Yang L, Liu B, Lin H, Lin Y. Combining local and global information for product feature extraction in opinion documents. Information Processing Letters 2016; 116 (10): 623-627. doi: 10.1016/j.ipl.2016.04.009

[45] Ittoo A, Bouma G. Term extraction from sparse, ungrammatical domain-specific documents. Expert Systems with Applications 2013; 40 (7): 2530-2540. doi: 10.1016/j.eswa.2012.10.067

[46] Davtalab R, Dezfoulian MH, Mansoorizadeh M. Multi-level fuzzy min-max neural network classifier. IEEE Transactions on Neural Networks and Learning Systems 2013; 25 (3): 470-482. doi: 10.1109/TNNLS.2013.2275937

[47] Barzegar S, Davis B, Handschuh S, Freitas A. Classification of composite semantic relations by a distributional-relational model. Data & Knowledge Engineering 2018; 117: 319-335. doi: 10.1016/j.datak.2018.06.005

[48] Ali I, Asif M, Shahbaz M, Khalid A, Rehman M et al. Text categorization approach for secure design pattern selection using software requirement specification. IEEE Access 2018; 6: 73928-73939. doi: 10.1109/AC-CESS.2018.2883077

[49] Kim D, Seo D, Cho S, Kang P. Multi-co-training for document classification using various document representations: TF-IDF, LDA, and Doc2Vec. Information Sciences 2019; 477: 15-29. doi: 10.1016/j.ins.2018.10.006

[50] Tang F, Adam L, Si B. Group feature selection with multiclass support vector machine. Neurocomputing 2018; 317: 42-49. doi: 10.1016/j.neucom.2018.07.012

[51] Raiskin Y, Eickhoff C, Beeler PE. Categorization of free-text drug orders using character-level recurrent neural networks. International Journal of Medical Informatics 2019; 129: 20-28. doi: 10.1016/j.ijmedinf.2019.05.020

[52] Pennington J, Socher R, Manning CD. Glove: global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) 2014; 1532-1543. doi: 10.3115/v1/D14-1162

[53] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016; 1: 785-794. doi: 10.1145/2939672.2939785

[54] Tianqi C, Tong H. Comparison of neuron-based, kernel-based, tree-based and curve-based machine learning models for predicting daily reference evapotranspiration. PLoS One 2019; 14 (5): 1-27. doi: 10.1371/journal.pone.0217520

[55] Brucker F, Benites F, Sapozhnikova E. Multi-label classification and extracting predicted class hierarchies. Pattern Recognition 2011; 44 (3): 724-738. doi: 10.1016/j.patcog.2010.09.010

[56] Chaudhary A, Kolhe S, Kamal R. A hybrid ensemble for classification in multiclass datasets:An application to oilseed disease dataset. Computers and Electronics in Agriculture 2016; 124: 65-72. doi: 10.1016/j.compag.2016.03.026