# Stratifying risk of coronary artery disease using discriminative knowledge-guided medical concept pairings from clinical notes

**Abstract.** Document classification (DC) is one of the broadly investigated natural language processing tasks. Medical document classification can support doctors in making decision and improve medical services. Since the data in document classification often appear in raw form such as medical discharge notes, extracting meaningful information to use as features is a challenging task. There are many specialized words and expressions in medical documents which make them more challenging to analyze. The classification accuracy of available methods in medical field is not good enough. This work aims to improve the quality of the input feature sets to increase the accuracy. A new three-stage approach is proposed. In the first stage, the Unified Medical Language System (UMLS) which is a medical-specific dictionary is used to extract the meaningful phrases by considering disease or symptom concepts. In the second stage, all the possible pairs of the extracted concepts are created as new features. In the third stage, Particle Swarm Optimisation (PSO) is employed to select features from the extracted and constructed features in the previous stages. The experimental results show that the proposed three-stage method achieved substantial improvement than the existing medical DC approaches.

**Keywords:** Medical Text Classification · Particle Swarm Optimization · Feature Selection · Feature Construction · Conceptualization · Ontology.

## 1 Introduction

Document classification has many important application such as filtering spam emails, labeling client queries and tagging patient reports. In general, text mining includes preprocessing, representing text, weighting features, selecting features, training, testing and evaluating.

There is a principal difference between clinical text mining and standard text mining in terms of text terminology and their frequency. In clinical text mining, the text describes a set of clinical events within a narrative, with the goal of producing an explanation as precise and comprehensive as possible when describing the health status of a patient. Generally, such text heavily uses domain specific terminology and acronyms, making clinical text analysis very different from standard text mining. Moreover, various combinations of domain-specific medical events in a clinical report can describe patients conditions totally differently. Hence, extracting meaningful information to analyze medical discharge notes is very important.

Information extraction (IE) task targets to extract structured information from the unstructured and semi-structured texts Sun et al. ((2018)). The process involves transforming an unstructured text or a collection of texts into structured data that can be used in a database. As our society became more data oriented, many different communities of researchers bring in techniques from machine learning, databases, information retrieval, and computational linguistics for various aspects of the information extraction problem in different fields such as the medical domain.

In medical document classification, there are thousands of features and often there are redundant and irrelevant features which can make noise in the training step to create a model. Consequently, the obtained model may have poor classification accuracy. This issue can be addressed by utilizing feature engineering approaches such as feature selection Bai et al. ((2018)) and feature construction to improve the quality of features by removing irrelevant and noisy features.

Most previous approaches for document classification are not effective enough for feature extraction due to a larger number of redundant features Bai et al. ((2018)). To solve this issue and improve the performance of document classification, this paper proposes a three-stage method by using discriminative knowledge-guided medical concept pairings from clinical notes for stratifying risk of coronary artery disease (CAD).

In this method, a tool is employed to extract concepts and detect most related features to the candidate classification problem. As medical domain is the main focus, a domain specific ontology is used for feature extraction. After extracting features from the documents, all the possible pairs of the extracted features are constructed to create new features. Then, particle swarm optimization (PSO) is utilized for feature selection. This paper aims to investigate the following research questions:

1. Whether the concept pairs can construct meaningful features from the extracted information of document set;
2. Whether PSO can reduce the number of features and keep the meaningful features; and
3. Whether the suggested approach can increase the classification accuracy in the aimed clinical notes classification.

The rest of the paper is organized as follows: Section 2 gives the problem description and related works. The proposed method is described in Section 3. The experiment design and results are presented in Section 4 and Section 5. At the end, the conclusions and future works are showed in Section 6.

## 2 Background

### 2.1 Document classification in medical domain

The first application of classifier models in predicting medical research results was presented by Bellazzi in Bellazzi and Zupan ((2008)). In this study, the authors tried to make use of data mining in the field of medicine. Yoo et al. investigated the advantages and disadvantages of using data mining algorithms

in the biomedical field Yoo et al. ((2012)), in which the proposed medical features include prediction health costs, prognosis and diagnosis, hidden knowledge from biomedicine data, relationship among diseases and among drugs are tested using data mining methods, and the extracted information is used in prediction. In another study Wagholikar et al. ((2012)) more than ten methods have been used to identify more than ten types of diseases. Based on the results of this study, the efficacy of these methods is better for some diseases such as gastroenterology, oncology and cardiovascular.

### 2.2 Information extraction in medical document classification

There has been research on using statistical methods from the distribution of the features in document classification problems for ranking features Shah and Patel ((2016)). Existing methods employed metrics associated with word frequency, information gain, mutual information, term frequency-inverse document frequency (tf-idf) for extracting textual features. However, they tend to treat each feature separately, and ignore the dependencies between features. Ontology-based classification methods is introduced in Dollah and Aono ((2011)). They use ontologies such as Medical Subject Headings(MeSH), Systematized Nomenclature of Medicine(SNOMED) and Unified Medical Language System(UMLS) to improve classification performance Buchan et al. ((2017)).

Clinical documents has been used in tasks such as finding risk factors for diabetic patients, assessing Framingham risk score(FRF) for candidate population, distinguishing heart disease risk factors, and finding the risk of heart disease Shivade et al. ((2015)). In this research, we use ontology as a feature extraction technique for document classification to identify Coronary Artery Disease (CAD).

### 2.3 Feature selection in medical document classification

In medical document classification, choosing a more efficient feature selection method that works with small sets of features from a high dimensional set of features is necessary. In some research, traditional feature selection methods, such as information gain, are generally employed Gaizauskas et al. ((2014)). And then, after selecting a small set of features, learning algorithms such as Support Vector Machine (SVM) are used to learn classifiers. One of the promising methods in feature selection is PSO.

PSO has been used to predict and analyze different diseases in medical field. For example, Eberhart and Hu Eberhart and Hu ((1999)) utilized PSO to check human tremor. PSO is used to improve a neural network that makes a distinction between normal people and those have tremor.Fong et al. Fong et al. ((2014)) employed PSO to find optimal feature subsets.

## 3 Our three-stage method

In this section, the developed three-stage algorithm and the employed tools for extracting concepts of phrases and constructing new features are described in detail. Fig. 1 presents the flowchart of the proposed three-stage method.

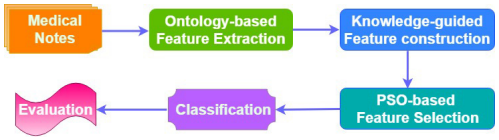


**Fig. 1.** The proposed three-stage method

The input of the proposed method is a set of medical discharge notes. Firstly, the method detects all of the meaningful phrases in the discharge notes by utilizing the MetaMap tool Aronson and Lang ((2010)) to extract their concepts from the United Medical Language System (UMLS). After eliminating unrelated features in the first stage, all the possible pairs of extracted expressions are created as the constructed features. Then, Particle Swarm Optimisation (PSO) is applied to select a feature subset from all of the extracted features in the first stage and the constructed features in the second stage. The classifier is learned along with the PSO feature selection.

It is expected that the proposed algorithm extracts meaningful features and selects more informative subset of the constructed features and maintains or enhances the classification accuracy.

### 3.1 Feature extraction method

UMLS is a dictionary in the biomedical area. An ontology structure of clinical vocabulary concepts is provided by UMLS. In this work our medical documents are the inputs of UMLS and the detected meaningful expressions are the outputs. In the first stage, the MetaMap tool is utilized to send all of the discharge documents to UMLS to extract the concepts of the detected meaningful expressions. Then, the classification task and the target label of the candidate problem is considered in the concept selection step. As the class label of the problem is the name of a disease and diseases have symptoms, all of the phrases whose concepts belong to "Disease or Syndrome" or "Sign or Symptom" are selected as a feature subset and the rest of the concepts are deleted. Fig. 2 shows the outline of the feature extraction and feature construction method.

A paragraph is given below as an example to describe how MetaMap works on the input discharge notes and what output it provides in classification process.

> *"Hyperlipidemia: The patient's Lipitor was increased to 80 mg q.d. A progress note in the patient's chart from her assisted living facility indicates that the patient has had shortness of breath for one day. The patient is a 63-year-old female with a three-year history of occasional weakness. Increasing large right-sided pulmonary edema."*

Fig. 3 presents the extracted concepts from MetaMap for the detected meaningful expressions in the paragraph. Table 1 shows the detected phrases based on their concepts. Some of the phrases such as "hyperlipidemia" and "shortest of breath" belong to more than one concept. As this research targets "[Disease or Syndrome]" and "[Sign or Symptom]" concepts, the scientific names of "hyperlipidaemia", "shortness of breath", "weakness" and "pulmonary oedema" are

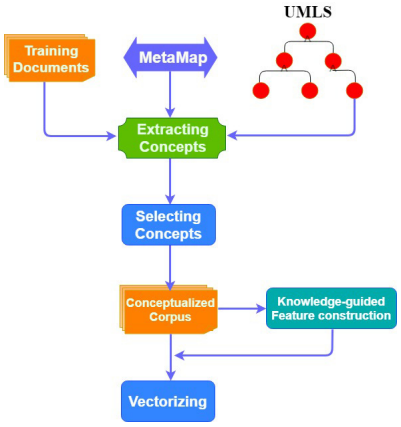


**Fig. 2.** Feature extraction method

```
1  Phrase: hyperlipidemia .
2  >>>>> Phrase
3  hyperlipidemia
4  <<<<< Phrase
5  >>>>> Mappings
6  Meta Mapping (1000):
7    1000   Hyperlipidaemia, NOS (Hyperlipidemia) [Disease or Syndrome]
8  Meta Mapping (1000):
9    1000   Hyperlipidemia (Serum lipids high (finding)) [Finding]
10 <<<<< Mappings
11 Processing 00000000.tx.7: MEDICATIONS ON ADMISSION : Lipitor , Flexeril ,
12 hydrochlorothiazide and Norvasc .
13 Phrase: shortness of breath
14 >>>>> Phrase
15 shortness of breath
16 <<<<< Phrase
17 >>>>> Mappings
18 Meta Mapping (1000):
19   1000   SHORTNESS OF BREATH (Dyspnea) [Sign or Symptom]
20 Meta Mapping (1000):
21   1000   Shortness of breath (Shortness of breath:-:Point in time:^Patient:-) [Clinical Attribute]
22 Meta Mapping (1000):
23   1000   Shortness of breath (How Often Shortness of Breath) [Intellectual Product]
24 <<<<< Mappings
25 Phrase: occasional weakness
26 >>>>> Phrase
27 occasional weakness
28 <<<<< Phrase
29 >>>>> Mappings
30 Meta Mapping (644):
31   569   Occasional (Infrequent) [Temporal Concept]
32   569   WEAKNESS (Weakness) [Sign or Symptom]
33<<<<< Mappings
34 Phrase: pulmonary edema
35 >>>>> Phrase
36 pulmonary edema
37 <<<<< Phrase
38 >>>>> Mappings
39 Meta Mapping (1000):
40   1000   PULMONARY OEDEMA (Pulmonary Edema) [Disease or Syndrome]
43 <<<<< Mappings
```

**Fig. 3.** A segment of returned results of extracted concepts using MetaMap

selected as a feature subset and the rest of the concepts are deleted. The scientific names of the expressions "Hyperlipidemia", "Dyspnea", "Weakness" and "Pulmonary Edema" are shown in lines 7, 19, 32 and 40 of Fig. 3, respectively.

### 3.2 Feature construction method

After the feature extraction, the obtained features are used to construct new features. To consider the relationship between the extracted diseases and symptoms, all of the possible pairs of (disease, disease), (disease, symptom) and (symptom, symptom) are constructed for each document and added to the extracted fea-

**Table 1.** The extracted concepts of example sentences using MetaMap.

| Sentences | Detected Phrases | Extracted Concepts | Selected |
|---|---|---|---|
| First Sentence | hyperlipidaemia | **[Disease or Syndrome]** | ✓ |
| | | [Finding] | × |
| | patient | [Patient or Disabled group] | × |
| | Lipitor | [Organic Chemical, Pharmacologic Substance] | × |
| | 80% | [Quantitative Concept] | × |
| | mg++ increased | [Finding] | × |
| Second Sentence | progress note | [Clinical Attribute] | × |
| | | [Intellectual Product] | × |
| | patient chart | [Manufactured Object] | × |
| | assisted living facility | [Healthcare Related Organization, Manufactured Object] | × |
| | patient | [Patient or Disabled group] | × |
| | shortness of breath | **[Sign or Symptom]** | ✓ |
| | | [Clinical Attribute] | × |
| | | [Intellectual Product] | × |
| | one day | [Temporal Concept] | × |
| Third Sentence | occasional | [Temporal Concept] | × |
| | weakness | **[Sign or Symptom]** | ✓ |
| Fourth Sentence | pulmonary oedema | **[Disease or Syndrome]** | ✓ |

tures. Table 2 shows the constructed features for the extracted features from the sample sentences.

**Table 2.** The constructed features for the extracted features from the sample sentences

| Cases | Pairs | Constructed Features |
|---|---|---|
| Case 1 | (Disease, Disease) | (Hyperlipidemia, Pulmonary Edema) |
| Case 2 | (Disease, Symptom) | (Hyperlipidemia, Dyspnea), (Hyperlipidemia, Weakness) (Pulmonary Edema, Dyspnea), (Pulmonary Edema, Weakness) |
| Case 3 | (Symptom, Symptom) | (Dyspnea, Weakness) |
| Case 4 | Case 1 + Case 2 | (Hyperlipidemia, Pulmonary Edema), (Hyperlipidemia, Dyspnea), (Hyperlipidemia, Weakness), (Pulmonary Edema, Dyspnea), (Pulmonary Edema, Weakness) |
| Case 5 | Case 1 + Case 3 | (Hyperlipidemia, Pulmonary Edema), (Dyspnea, Weakness) |
| Case 6 | Case 2 + Case 3 | (Hyperlipidemia, Dyspnea), (Hyperlipidemia, Weakness) (Pulmonary Edema, Dyspnea), (Pulmonary Edema, Weakness), (Dyspnea, Weakness) |
| Case 7 | Case 1 + Case 2 + Case 3 | (Hyperlipidemia, Pulmonary Edema), (Hyperlipidemia, Dyspnea), (Hyperlipidemia, Weakness), (Pulmonary Edema, Dyspnea), (Pulmonary Edema, Weakness), (Dyspnea, Weakness) |

After the feature construction step, all of the created pairs are added to the obtained feature set in the concept selection step. In Table 2, the last column presents the total feature size for each case. The obtained output will be used instead of the original documents in the binary classification problem. The first stage keeps the informative features and the second stage enrich the feature set. For giving weights to the extracted phrases of the documents, TF-IDF is utilized in the vectorization phase and each document is represented as a vector of weights based on the TF-IDF function.

**Table 3.** Possible Pairs and the Number of Features

| Cases | Pairs | Number of Original Features (100%) | Number of UMLS Features (10.33%) | Number of Features (UMLS + Pairs)(%) |
|---|---|---|---|---|
| Case 1 | (Disease, Disease) | 7554 | 780 | 10107(133.80) |
| Case 2 | (Disease, Symptom) | 7554 | 780 | 11261(149.07) |
| Case 3 | (Symptom, Symptom) | 7554 | 780 | 4199(55.59) |
| Case 4 | (Disease, Disease) + (Disease, Symptom) | 7554 | 780 | 20578(272.41) |
| Case 5 | (Disease, Disease) + (Symptom, Symptom) | 7554 | 780 | 13518(178.95) |
| Case 6 | (Disease, Symptom) + (Symptom, Symptom) | 7554 | 780 | 14670(194.20) |
| Case 7 | (Disease, Disease) + (Disease, Symptom) + (Symptom, Symptom) | 7554 | 780 | 24074(318.69) |

### 3.3 PSO-based algorithm for feature selection

In the second step, different pairs are made from disease and symptoms. As the pairs are constructed using all the extracted features, there might be redundant features among the obtained feature set. Hence, it is necessary to do feature selection. In this stage, PSO is applied to remove the irrelevant and unnecessary features from the extracted and constructed features in the first and second stage. The value for each particle is initialized randomly between [-1, 1]. Each particle in PSO indicates a feature subset and is represented as a vector. For instance, a negative value indicates the feature is not selected and a positive value means the feature is selected. The dimension of each vector is $d$ and each vector includes real numbers. The dimension of the search space is represented by $d$ which is equal to the size of the obtained features by the first and second steps. The position and velocity of each particle is initialized randomly. Then, particles moves by updating their *gbest* (the best position) and *pbest* (best position has found so far). At the end of the method, *gbest* is found using the fitness values of particles and also the obtained best particle is used to form the selected feature set. Algorithm 1 shows the pseudocode for PSO for feature selection in the third stage. The fitness value for each particle is calculated by the classification accuracy(see line 5).

The method used in this work is a wrapper approach. Hence, a classifier is utilized to run with PSO to calculate the value of fitness function.

**Algorithm 1:** Pseudo-code of PSO to select the best feature subset

```
   Input  : Training instances
   Output: The best feature subset (gbest)
 1: Keep only the features that are extracted in the first and second stages;
 2: Randomly initialize the position and velocity of particles;
 3: iter ← 0
 4: while iter < maxIter do
 5:     Evaluation: Evaluate fitness of particles based on classification accuracy on the training set;
 6:     for i = 1 to |Particle| do
 7:         Update pbest and gbest for particle i;
 8:     end
 9:     for i = 1 to |Particle| do
10:         for d = 1 to dimension do
11:             Update the velocity of particle i
12:             Update the position of particle i
13:         end
14:     end
15:     iter ← iter + 1
16: end
17: return the position of gbest;
```

The process of calculating the fitness function value for each particle is presented in Fig. 4. All of the training documents are feeded as input to PSO for selecting features. Fitness value of a particle is computed by 10-fold cross validation. The training document set is separated into 10 subsets. One training subset is used for evaluating the particle's fitness value and the nine remained training subsets are utilized as input to PSO for training a classifier. The fitness value of a particle is the average of computed ten classification accuracies. In this stage, only the training set is considered to train the candidate classifier and the test set is only utilized after the training to evaluate the classification accuracy of the selected best feature subsets.

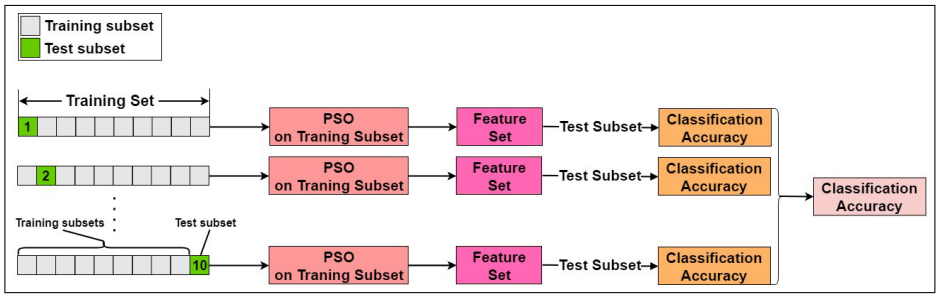


**Fig. 4.** PSO for feature selection using 10 fold cross validation

# 4 Experimental design

## 4.1 Dataset and preprocessing

The performance of the proposed three-stage method is evaluated on the 2010 Informatics for Integrating Biology and the Bedside (i2b2) data set. The labels of the 2010 i2b2 data set are CAD (Coronary Artery Disease) and non-CAD that form a binary classification problem. The data set includes 426 documents which 170 documents for training and 256 documents for testing. All of the features are extracted by considering two specific concepts ("Disease or Syndrome" and "Sign or Symptom") by employing the MetaMap tool and utilizing the UMLS. Then, all of the possible pairs of obtained features are constructed for the output of each document separately. Next, the following preprocessing steps are applied on the obtained results of the feature extraction step:

- Hold only words and delete punctuation, numbers, etc. Convert all words to lowercase.
- Delete words which are less than 3 letters long. For example, removing "am" but keeping "are".
- Remove the 524 SMART stopwords.
- Extract stems of the remained words.

## 4.2 Parameter Settings

The 2010 i2b2 data set includes 426 documents with 7554 various terms. Table 3 shows the total number of attributes for each case after applying the first and second stages (check the last column). Five different classifiers (Logistic Regression (LR), Linear Support Vector Machine (LSVM), Naive Bayes (NB), Decision Tree (DT) and K-Nearest Neighbor (KNN)) are employed for the experimental comparison. The classification accuracy is calculated on the testing documents to evaluate the performance of the classifiers. Table 4 presents the set parameters of PSO which are proposed in Bai et al. ((2018)). The values for particles are initialised using numbers in [-1, 1], and zero is set to the threshold ($\theta$), hence, about 50% of the features is selected. Some documents will disappear if less than 50% of features are selected.

**Table 4.** PSO parameter setting

| PSO Parameters | Value |
|---|---|
| Population Size | 30 |
| Maximum Number of Iteration | 100 |
| Dimension of All+PSO | 7554 |
| Dimension of UMLS+PSO | 780 |
| Dimension of case 1 | 10107 |
| Dimension of case 2 | 11261 |
| Dimension of case 3 | 4199 |
| Dimension of case 4 | 20578 |
| Dimension of case 5 | 13518 |
| Dimension of case 6 | 14670 |
| Dimension of case 7 | 24074 |
| Velocity | [-3, 3] |
| Threshold ($\theta$) | 0 |
| Acceleration Coefficients | 2.0 |
| Run Times | 40 |

Some of the classifiers' parameters are tuned to get better results. The inverse of regularization strength ("C") is adjusted to the value "1e1" in the Logistic Regression. The number of the neighbors ("n_neighbors") is set to the value 28 in KNN. The maximum depth of the tree ("max_depth") and the random number generator ("random_state") are adjusted to values 14 and 11 in Decision Tree classifier, respectively. Furthermore, early stopping rule is chosen to avoid overfitting in training Linear SVM and Logistic Regression classifiers. The rest of the classifiers' parameters are kept the same as default values.

## 5 Results and further analysis

### 5.1 Results

Five different classifier are employed to assess the proposed approach, and the results are shown in figure 5-9 for each classifier respectively. Our three stage approach has six cases (case1 to case6) and they use different pair combinations shown in Table 3. The six methods are compared with four other methods: "All_Deter" which uses all unique term features; "UMLs_Deter" which uses UMLS concepts as features; "All+PSO" which uses PSO to select features from all terms; and "UMLS+PSO" which uses PSO to select from UMLS concepts. The efficiency of the classifiers are assessed based on classification accuracy. From figures 5 to 9 it is obvious that the proposed technique with three stages (case 1 to case 6) is significantly better than the other compared methods.

### 5.2 Further analysis

**Number of selected features:** Table 5 shows the average (and standard deviation values for stochastic methods) of the selected features by different approaches. "Original", "UMLS" and "UMLS+Pairs" methods are deterministic and use all of the features without any feature selection. "Original" is using all unique terms in the original documents. "UMLS" approach is using the extracted features from UMLS by applying MetaMap tool. "UMLS+Pairs" method is utilizing the detected features from UMLS and the constructed pairs of features. "All+PSO", "UMLS+PSO" and "UMLS+Pairs+PSO" are stochastic methods by applying PSO to select a feature subset. The smallest feature subset belongs

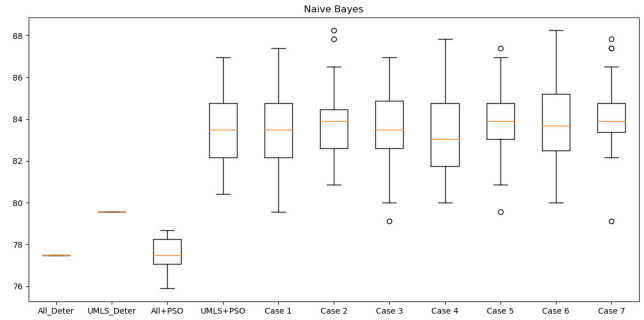


**Fig. 5.** Comparison of Naive Bates classifier accuracy

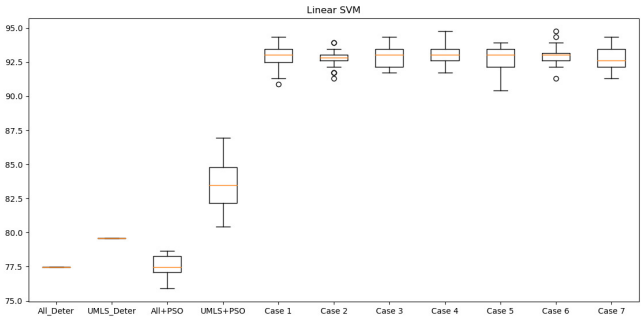


**Fig. 6.** Comparison of Linear SVM Classifier accuracy

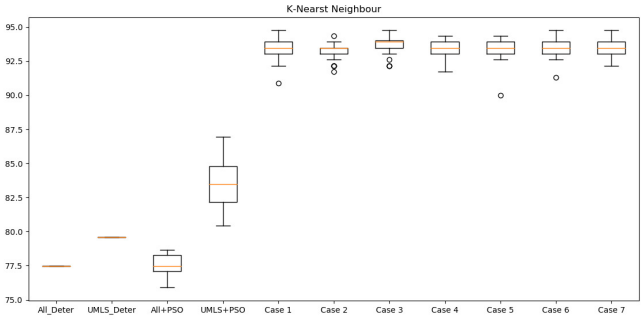


**Fig. 7.** Comparison of K-Nearest Neighbor classifiers accuracy

to "UMLS+PSO" method which contains only 10.33% of the original features. The smallest number of features is allocated for case 3 in "UMLS+Pairs" and "UMLS+Pairs+PSO" with 55.59% and 27.02%, respectively. By comparing the number of the selected features for the deterministic and stochastic versions of the proposed approach, it can be concluded that case 3 has the smallest size of the features in both methods which is smaller than "Original" method's feature

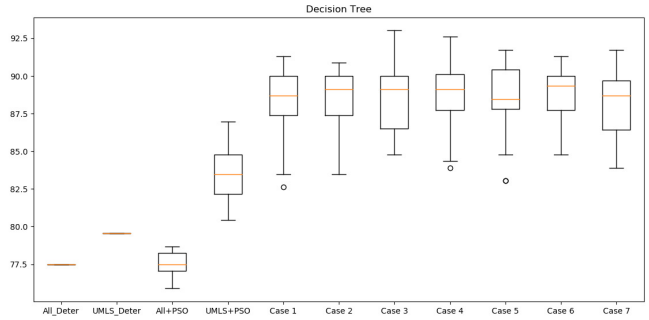


**Fig. 8.** Comparison of Decision Tree classifier accuracy

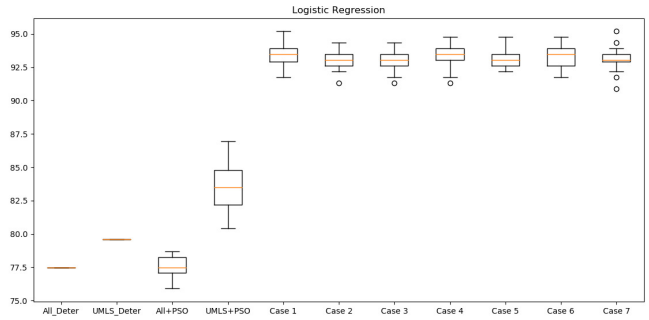


**Fig. 9.** Comparison of Logistic Regression classifier accuracy

size and the feature size of stochastic method is approximately 50% smaller than the deterministic method.

**Table 5.** Number of Selected Features

| # | Classifiers / Cases | NB | LSVM | KNN | DT | LR |
|---|---|---|---|---|---|---|
| | | Ave±Std | | | | |
| 1 | Original (100%) [ Abdollahi et al. ((2018))] | 7554 | 7554 | 7554 | 7554 | 7554 |
| 2 | UMLS [ Abdollahi et al. ((2018))] | 780 | 780 | 780 | 780 | 780 |
| 3 | All+PSO | 3779.35±38.01 | 3768.75±48.22 | 3774.13±39.36 | 3775.25±43.04 | 3767.65±32.77 |
| 4 | UMLS+PSO | 387.20±14.61 | 386.08±14.79 | 394.35±10.68 | 388.60±15.14 | 388.25±12.31 |
| 5 | UMLS+Pairs (Case 1) | 10107 | 10107 | 10107 | 10107 | 10107 |
| 6 | UMLS+Pairs (Case 2) | 11261 | 11261 | 11261 | 11261 | 11261 |
| 7 | UMLS+Pairs (Case 3) | 4199 | 4199 | 4199 | 4199 | 4199 |
| 8 | UMLS+Pairs (Case 4) | 20578 | 20578 | 20578 | 20578 | 20578 |
| 9 | UMLS+Pairs (Case 5) | 13518 | 13518 | 13518 | 13518 | 13518 |
| 10 | UMLS+Pairs (Case 6) | 14670 | 14670 | 14670 | 14670 | 14670 |
| 11 | UMLS+Pairs (Case 7) | 24074 | 24074 | 24074 | 24074 | 24074 |
| 12 | UMLS+Pairs+PSO (Case 1) | 5051.68±56.22 | 5055.95±51.53 | 5048.78±52.02 | 5049.85±55.25 | 5041.68±53.57 |
| 13 | UMLS+Pairs+PSO (Case 2) | 5630.18±56.41 | 5625.6±53.50 | 5616.0±44.85 | 5625.1±51.37 | 5630.55±54.53 |
| 14 | UMLS+Pairs+PSO (Case 3) | 2097.25±34.79 | 2090.85±34.84 | 2100.0±35.59 | 2089.93±33.19 | 2103.33±29.34 |
| 15 | UMLS+Pairs+PSO (Case 4) | 10276.4±81.09 | 10292.38±81.56 | 10275.6±83.59 | 10288.23±67.09 | 10274.93±80.68 |
| 16 | UMLS+Pairs+PSO (Case 5) | 6756.98±71.62 | 6747.9±59.01 | 6762.73±47.58 | 6763.4±63.10 | 6752.05±56.78 |
| 17 | UMLS+Pairs+PSO (Case 6) | 7310.73±53.22 | 7329.95±55.86 | 7343.43±59.93 | 7343.9±68.94 | 7329.78±59.80 |
| 18 | UMLS+Pairs+PSO (Case 7) | 12038.48±75.39 | 12042.25±79.63 | 12037.60±62.95 | 12035.55±77.71 | 12026.95±69.64 |

**With or without PSO:** Table 6 compares the statistical results of the deterministic and stochastic versions of the proposed approach with the pairs. The best results are highlighted and three-stage method (with PSO) shows better

performance than two-stage method (without PSO) in Naive Bayes, Linear SVM, KNN and Logistic Regression classifiers.

**Table 6.** Accuracy of Classifiers for the Seven Cases without PSO and with PSO

| Classifiers | NB | | LSVM | | KNN | | DT | | LR | |
|---|---|---|---|---|---|---|---|---|---|---|
| Cases | Accuracy (%) Ave±Std | | | | | | | | | |
| | Without PSO | With PSO | Without PSO | With PSO | Without PSO | With PSO | Without PSO | With PSO | Without PSO | With PSO |
| Case 1 | 81.05±0.00 | **83.47**±0.018 | 92.49±0.00 | **92.75**±0.007 | 92.09±0.00 | **93.32**±0.007 | **91.30±0.00** | 88.50±0.022 | 92.89±0.00 | **93.32**±0.007 |
| Case 2 | 81.42±0.00 | **83.85**±0.016 | 92.49±0.00 | **92.80**±0.006 | **93.28±0.00** | 93.22±0.006 | **92.09±0.00** | 88.49±0.019 | 92.09±0.00 | **93.03**±0.006 |
| Case 3 | 79.45±0.00 | **83.66**±0.017 | 92.89±0.00 | **92.98**±0.007 | 92.89±0.00 | **92.98**±0.007 | **90.12±0.00** | 88.42±0.020 | 92.49±0.00 | **92.91**±0.007 |
| Case 4 | 81.03±0.00 | **83.39**±0.019 | 92.89±0.00 | **92.97**±0.007 | 93.28±0.00 | **93.37**±0.006 | **91.70±0.00** | 88.85±0.021 | 91.70±0.00 | **93.35**±0.008 |
| Case 5 | 81.03±0.00 | **83.84**±0.017 | 93.28±0.00 | **92.77**±0.008 | 91.30±0.00 | **93.33**±0.007 | 88.14±0.00 | **88.60**±0.020 | 91.70±0.00 | **92.97**±0.006 |
| Case 6 | 81.42±0.00 | **83.79**±0.021 | 92.49±0.00 | **92.99**±0.007 | 92.49±0.00 | **93.52**±0.007 | **90.91±0.00** | 88.92±0.016 | 90.91±0.00 | **93.34**±0.008 |
| Case 7 | 81.03±0.00 | **84.05**±0.016 | 92.49±0.00 | **92.90**±0.008 | 93.28±0.00 | **93.47**±0.005 | 86.56±0.00 | **88.10**±0.022 | 91.30±0.00 | **93.15**±0.007 |

**Significance test** The suggested three-stage approach is applied on the training set using 40 independent PSO runs. Next, the quality of the selected feature subsets is evaluated on the test set by using the gained best feature subsets from each run. The experimental results are computed by considering the classification accuracies of the 40 selected feature subsets. Table 8 compares the statistical results for six approaches. The standard deviation and average of accuracies are calculated for all of the classifiers and the Wilcoxon signed ranks test with significance level of 0.05 is used to test whether the suggested approach has made significant difference in classification accuracy. In Table 7, "T" column presents the significance test of the proposed approach against the other five approaches, where "+" means the suggested three-stage method is significantly more accurate, "=" means no significant difference, and "-" means significantly less accurate. The best results are highlighted in the table.

**Table 7.** Comparison of classification accuracy and standard deviation averages using 40 independent runs. The highlighted entries are significantly better(Wilcoxon Test, $\alpha = 0.05$)

| Methods | Three-Stage Stochastic | | All [ Abdollahi et al. ((2018))] Deterministic | | UMLS [ Abdollahi et al. ((2018))] Deterministic | | UMLS+Pairs Deterministic | | All+PSO Stochastic | | | Two-Stage Stochastic | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Classifiers | Accuracy Ave±Std | Accuracy Best(Lowest) | Accuracy | T | Accuracy | T | Accuracy | T | Accuracy Ave±Std | Accuracy Best(Lowest) | T | Accuracy Ave±Std | Accuracy Best(Lowest) | T |
| NB | **83.79**±0.021 | **88.26**(80.00) | 77.47 | + | 79.57 | + | 81.42 | + | 77.58±0.007 | 78.66(75.89) | + | 83.50±0.018 | 86.96(80.43) | + |
| LSVM | **92.99±0.007** | **94.78(91.30)** | 87.35 | + | 92.61 | + | 92.49 | + | 87.22±0.008 | 88.93(84.98) | + | 92.87±0.007 | 93.91(91.30) | + |
| KNN | 93.52±0.007 | **94.78**(91.30) | 84.98 | + | 94.78 | + | 93.28 | + | 86.80±0.014 | 89.33(82.21) | + | 93.61±0.005 | 94.78(92.61) | = |
| DT | 88.92±0.016 | 91.30(84.78) | 85.77 | + | 87.39 | + | 92.09 | - | 90.09±0.011 | **92.25(86.96)** | - | 88.71±0.021 | 91.30(82.61) | = |
| LR | **93.34**±0.008 | **94.78(91.74)** | 86.96 | + | 92.61 | + | 92.09 | + | 87.62±0.008 | 89.33(86.17) | + | 93.27±0.007 | 94.35(91.74) | + |

## 6 Conclusions and Future Work

This work introduces a three-stage method to utilise domain concepts and their relations to enrich the input data for a classification problem. The proposed approach is able to improve the quality of the input data set by construction new features and increase the classification accuracy in the majority of the targeted classifiers. From the experimental and statistical examinations it can be seen that the suggested approach can achieve significantly better classification accuracy.

This work shows promise in using a third-stage feature extraction, construction and selection method in clinical document classification, however, it still needs more research to improve the classification performance. We will study other ways to construct features for the second stage by analyzing the distance of the detected features in the document to guide our feature construction method in making pairs. In the meantime, we will consider different fitness functions to enhance the PSO method.

# Bibliography


M. Abdollahi, X. Gao, Y. Mei, S. Ghosh, and J. Li. Uncovering discriminative knowledge-guided medical concepts for classifying coronary artery disease notes. In *Australasian Joint Conference on Artificial Intelligence*, pages 104–110. Springer, 2018.

A. R. Aronson and F.-M. Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

X. Bai, X. Gao, and B. Xue. Particle swarm optimization based two-stage feature selection in text mining. In *2018 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE, 2018.

R. Bellazzi and B. Zupan. Predictive data mining in clinical medicine: current issues and guidelines. *International journal of medical informatics*, 77(2):81–97, 2008.

K. Buchan, M. Filannino, and Ö. Uzuner. Automatic prediction of coronary artery disease from clinical narratives. *Journal of biomedical informatics*, 72: 23–32, 2017.

R. B. Dollah and M. Aono. Ontology based approach for classifying biomedical text abstracts. *Int. J. Data Eng*, 2(1):1–15, 2011.

R. C. Eberhart and X. Hu. Human tremor analysis using particle swarm optimization. In *Proceedings of the 1999 Congress on Evolutionary Computation-CEC99 (Cat. No. 99TH8406)*, volume 3, pages 1927–1930. IEEE, 1999.

S. Fong, S. Deb, X.-S. Yang, and J. Li. Feature selection in life science classification: metaheuristic swarm search. *IT Professional*, 16(4):24–29, 2014.

R. Gaizauskas, E. Barker, M. L. Paramita, and A. Aker. Assigning terms to domains by document classification. In *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*, pages 11–21, 2014.

F. P. Shah and V. Patel. A review on feature selection and feature extraction for text classification. In *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pages 2264–2268. IEEE, 2016.

C. Shivade, P. Malewadkar, E. Fosler-Lussier, and A. M. Lai. Comparison of umls terminologies to identify risk of heart disease using clinical notes. *Journal of biomedical informatics*, 58:S103–S110, 2015.

W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, and G. Wang. Data processing and text mining technologies on electronic medical records: a review. *Journal of healthcare engineering*, 2018, 2018.

K. B. Wagholikar, V. Sundararajan, and A. W. Deshpande. Modeling paradigms for medical diagnostic decision support: a survey and future directions. *Journal of medical systems*, 36(5):3029–3049, 2012.

I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, and L. Hua. Data mining in healthcare and biomedicine: a survey of the literature. *Journal of medical systems*, 36(4):2431–2448, 2012.