Improving Medical Document Classification via Feature Engineering

by

Mahdi Abdollahi

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the
requirements for the degree of
Doctor of Philosophy
in Computer Science.

Victoria University of Wellington
2021

Abstract

Document classification (DC) is the task of assigning the pre-defined labels to unseen documents by utilizing the model trained on the available labeled documents. DC has recently attracted much attention in the medical field because many issues can be formulated as classification problems. For example, categorizing clinical risk factors, automatic disease classification, and electronic health records classification are some applications of text classification. DC is critical for medical document management and analysis. Medical DC can assist doctors in decision making and correct decisions can reduce medical expenses. Medical documents have special attributes that distinguish them from other texts and make them difficult to analyze. For example, many acronyms and abbreviations, and short expressions make it more challenging to extract knowledge. The current classification performance on medical documents is not satisfactory. Furthermore, the source of data is not sufficient due to patients' privacy. This thesis aims to enhance the input feature sets of the medical DC methods to improve their classification performance. Additionally, it develops new data augmentation methods to deal with the shortage of data. To approach these goals, this work has developed new feature manipulation methods (such as future extraction, feature selection, and feature construction) in supervised learning systems to extract new meaningful feature sets. Moreover, it develops ontology and dictionary-based data augmentation approaches to create new synthetic documents. This thesis utilizes Evolutionary Computation (EC) techniques such as Particle Swarm Optimisation (PSO) and other deep learning methods such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Hierarchical Attention Network (HAN) to achieve its objectives.

The main goal of this thesis is to develop new feature engineering approaches to medical document classification by using domain-specific knowledge of the problem which automatically extracts prominent features, constructing new high-level features, selects informative features, and augments new synthetic documents from the original documents. These methods can improve medical document classification performance by enriching the quality of the input data.

This thesis develops three feature engineering approaches including domain-specific feature extraction, two-stage and three-stage PSO-based methods to automatically extract, construct, and select new high-level features for classification. The results demonstrate that two-stage and three-stage approaches outperformed the compared related works.

This thesis proposes two novel ontology-based data augmentation approaches to make new synthetic documents from the original training data sets for medical document classification. These approaches can make new synthetic documents from the original documents by employing a domain-specific ontology and a general dictionary to double/triple the size of the training data set and improve the performance of medical document classification. The results show that these approaches successfully improved medical document classification performance.

This thesis develops two dictionary-based data oversampling approaches to make new synthetic documents from the original training data sets for medical document classification problems. The proposed approach can make new synthetic documents with high variety compared to similar methods. The proposed approaches make an imbalanced data set balanced and improve the classification performance too. The results show better classification performance.

Acknowledgments

I would like to express my gratitude to those who gave me the assistance and support I needed to complete this thesis. Above all, I thank Allah for the blessings, wellness, and abilities, that He has given me during this work and in all my life.

First and foremost, I would like to express my very great appreciation to my supervisors Associate Professor Xiaoying Gao, and Dr. Yi Mei. They have spent many dedicated hours and efforts to train my research skills and provide encouraging and challenging feedbacks to improve my research work. They patiently and supportively paved my way from far away to reach this moment of my research journey. I am profoundly indebted to Associate Professor Xiaoying Gao, who, besides being a brilliant academic supervisor, is a caring human being. Her academic guidance was a cornerstone of my research. Her positivity, even when I am not doing well, kept me hopeful in very difficult times. I am completely grateful to Dr. Yi Mei for his constant support during my research. He had been generously giving me his unlimited knowledge and extensive expertise. I am deeply beholden to Dr. Shameek Gosh, Prof. Jinyan Li, and Dr. Michael Narag for their inexhaustible efforts in shaping my research abilities. They had been providing valuable and constructive feedback to improve my research work. To all my supervisors, you have been more than just academic supervisors to me.

I sincerely thank my dear friends Mashall Aryan, Ghassem Narimani, Baligh Al-Helali, Cao Truong Tran, and Mazhar Ansari Ardeh for being

good friends, and for sharing their knowledge with me. I would like to offer my thanks to the members of the Evolutionary Computation Research Group (ECRG) for creating an active and interesting research environment. I also would like to thank my officemates in the Cotton and Maru buildings for putting up with my annoyance. Special thanks go to my friends, Harisu Abdullahi Shehu, Qurrat Ul Ain, Joao Costa, Hiroshika Hinduramage, Duleeka Munasinghe, Fangfang Zhang, Shabbir Abbasi, Ying Bi, Muru Raj Odiathevar, Junaid Haseeb, Kirita-Rose Escott, and so many unlisted, who had been of great help to ease my life during stressful times.

I would like to express my thanks to Victoria University of Wellington for their invaluable funding. I also would like to acknowledge the efforts of all staff members of the School of Engineering and Computer Science. I would like to thank Dr. Harith Al-Sahaf, Prof Mengjie Zhang, Prof Bing Xue, A/Prof Hui Ma, Dr. Qi Chen, Dr. Mohammad Nekooei, and Dr. Bach Nguyen, whose discussions and works helped me in different stages of my research. I also want to show my gratitude to those who have taught me throughout my study journey. Especially, to Dr. Davoud Abdollahi, who has been always my example, and without him, I would never reach this moment.

Last, but most importantly, I would like to dedicate this work to my family. To my parents, Esmaeil and Delbar for working so hard to bring up their son. I would like also to thank my brother (Davoud) who is the rest of my senses and my sister (Saeedeh) who is my two eyes. Finally, this work is dedicated to the soul of my late brother Saeed. I know you wanted to be proud of your younger brother and I hope I can see this when we meet, hereafter.

Mahdi Abdollahi
12 *December* 2021
Wellington, New Zealand

List of Publications

- **Mahdi Abdollahi,** Xiaoying Gao, Yi Mei, Shameek Ghosh, and Jinyan Li. "Uncovering Discriminative Knowledge-Guided Medical Concepts for Classifying Coronary Artery Disease Notes." In Australasian Joint Conference on Artificial Intelligence, Wellington, New Zealand. pp. 104-110, Springer, 2018.

- **Mahdi Abdollahi,** Xiaoying Gao, Yi Mei, Shameek Ghosh, and Jinyan Li. "An Ontology-based Two-Stage Approach to Medical Text Classification with Feature Selection by Particle Swarm Optimisation." In 2019 IEEE Congress on Evolutionary Computation, Wellington, New Zealand. pp. 119-126, IEEE, 2019.

- **Mahdi Abdollahi,** Xiaoying Gao, Yi Mei, Shameek Ghosh, and Jinyan Li. "Stratifying Risk of Coronary Artery Disease Using Discriminative Knowledge-Guided Medical Concept Pairings from Clinical Notes." In: Nayak A., Sharma A. (eds) PRICAI 2019: Trends in Artificial Intelligence. PRICAI 2019. Lecture Notes in Computer Science, vol 11672. pp. 457-473, Springer, Cham., 2019.

- **Mahdi Abdollahi,** Xiaoying Gao, Yi Mei, Shameek Ghosh, and Jinyan Li. "Ontology-Guided Data Augmentation for Medical Document Classification." In: International Conference on Artificial Intelligence in Medicine. AIME 2020. pp. 78-88, Springer, Cham., 2020.
  (Invited paper to extend for "Artificial Intelligence In Medicine (AIIM) Journal")

- **Mahdi Abdollahi,** Xiaoying Gao, Yi Mei, Shameek Ghosh, and Jinyan Li. "A Dictionary-based Oversampling Approach to Clinical Document Classification on Small and Imbalanced Dataset." In: The 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology. WI-IAT'20. pp. 1-8, IEEE, 2020.

- **Mahdi Abdollahi,** Xiaoying Gao, Yi Mei, Shameek Ghosh, Jinyan Li and Michael Narag. "Substituting clinical features using synthetic medical phrases: Medical text data augmentation techniques." Artificial Intelligence in Medicine. pp. 1-14, Elsevier,120 (2021): 102167.

- **Mahdi Abdollahi,** Xiaoying Gao, Yi Mei, Shameek Ghosh, Jinyan Li. "Stratifying clinical diseases using UMLS concepts: Medical text feature engineering techniques." Pending to submit to: Journal of Biomedical Informatics. pp. 1-20, Elsevier, 2021.

Contents

Chapter 1

Introduction

1.1 Problem Statement

In recent years, much information is produced in digital form. Over 80% of the created information appears in text [90] such as newspapers, articles, and magazines, and around 90% of the generated text is unstructured [99]. Doctors have also generated many clinical texts to record the disease and symptoms of patients. Based on a report by the World Health Organization report [126], 56.9 million people died in 2016 and 54% of this number is due to ten different kinds of diseases such as Ischaemic Heart Disease, Stroke, Lower Respiratory Infections, Diabetes Mellitus and etc. There are many documents in the medical domain, but only a small number of those are available due to the privacy issues. Studying about these diseases (such as Coronary Artery Disease (CAD), Asthma, Clinical trials) using these documents is crucial. Hence, medical documents are a good resource to extract information related to a patient health condition to use in decision making and treatment. Furthermore, artificial Intelligence-driven decision making can assist doctors on medical text classification [146].

Clinical notes are one of the rich sources of data which should be analyzed to detect and extract useful information such as events (e.g. disease names and symptoms) and temporal times (e.g. date, time, duration, and

frequency). Comprehension of clinical timeline from patient's discharge note is necessary in deciding about a patient's problem [160]. The correct diagnosis can lead to the right treatment to solve the patient's issue and decrease the treatment costs. However, analyzing medical text is one of the challenging tasks from natural language processing aspect. Because these notes contain many acronyms, abbreviations, short expressions, and medical phrases which make it more complex to analyze in comparison with other texts. Text mining is a technology has been used to overcome the text analyzing difficulties.

Text mining is one of the important research areas in data mining which analyzes various unstructured and semi-structured texts to pull out useful information and knowledge. Text mining is closely related to natural language processing and has many practical application areas such as information retrieval (IR), document clustering, web mining, information extraction (IE), document classification [117], document summarization and semantic analysis.

Document classification is one of the broadly investigated natural language processing tasks. The goal of document classification is to learn a model from available training data set with predefined classes to predict the classes of unseen data. For instance, filtering spam emails, labeling client queries and tagging patient reports are a number of the document classification applications. There is a pipeline in text mining to classify the unlabeled documents which includes preprocessing, representing text, weighting features, selecting features, training, testing and evaluating. Since the data in text classification appears as the raw data such as medical discharge notes, hence, extracting meaningful information to use as the features in document classification is a substantial function.

Information extraction (IE) task targets to extract structured information from the unstructured and semi-structured texts. 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. Most of the existing methods are extracting all of the features in IE step which can be irrelevant. 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. Some other works are using the rule-based methods [167, 155, 192] to filter unnecessary information, however, these methods need expert people to define rules for each problem separately which can be costly, and the defined methods are not usable to another problem [133]. Feature selection can improve the performance of classification by selecting meaningful features and at the same time reducing the number of noisy features.

Evolutionary Computation (EC) has been used broadly in feature selection in classification problems. Evolutionary Computation [49] is a research area of Artificial Intelligence (AI) which is inspired from biological world and includes stochastic population-based search approaches, that mimics animals behaviour and evolution. Particle Swarm Optimisation (PSO) [85] is a robust EC algorithm that has been used to solve an extensive variety of NP-hard issues effectively, due to their capacity to discover good solutions in a reasonable time. The experimental results of applying EC algorithms on the classification problem has shown promising future [157, 186]. PSO has been widely utilized to feature selection [92] and feature construction [53], respectively. Although these methods have been applied to many issues, little work has been done in feature manipulation in the text classification domain, especially in the medical area.

This thesis will use a domain-specific knowledge of the problem for extracting and constructing meaningful features, and developing methods for feature selection in medical document classification problems. Fur-

thermore, two different approaches will be introduced to augment new documents by using a domain-specific ontology and general dictionary to deal with shortage of data in the medical field.

1.2 Motivations

Ambiguity is one of the leading challenges in analyzing text because of the complexity of natural language itself. Furthermore, Clinical texts, in contrast to edited published articles, are often not grammatically correct, may use locally used abbreviations and have misspellings [66]. This poses huge challenges for transferring tools that have been developed for general text analysis in the literature to the clinical context. Even well-formed medical texts have different characteristics from general domain texts, and require tailored solutions.

Another main challenge in document classification problems is classifying the documents with a high accuracy. Normally, texts are the raw data which need preprocessing to use in the training step of the classification task. However, the space of the extracted features in the preprocessing step can be large and doing text classification with a huge number of features can decrease the performance. Furthermore, the features can contain noisy and irrelevant data that can have a negative effect on the performance of categorization [60]. A large number of unknown words, non-words and poor grammatical sentences made up the noise in the clinical corpus. Unknown words are usually complex medical vocabularies, misspellings, acronyms and abbreviations where unknown non-words are generally the clinical patterns including scores and measures. These kinds of noise are very common and potentially affect the overall information extraction performance but they were not carefully investigated in most presented health informatics systems. Hence, feature selection is needed to overcome the mentioned issues.

One of the important preprocessing tasks in document classification is

$$\begin{array}{c|cccc} & T_1 & T_2 & \cdots & T_m \\ D_1 & w_{11} & w_{12} & \cdots & w_{1m} \\ D_2 & w_{21} & w_{22} & \cdots & w_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ D_n & w_{n1} & w_{n2} & \cdots & w_{nm} \end{array}$$

Figure 1.1: A vector matrix to represent documents.

to represent documents which reduces the complexity of the input documents and makes them simpler to be used in the next steps of preprocessing. In this step, all of the documents should be transformed from the text to a vector. Bag of Words (BoW) representation model is a common method to represent documents as weights of words. To address the input complexity, feature weighting approaches are used to allocate suitable weights to the feature. An example in Fig. 1.1 shows a vector matrix in which rows show the documents and the columns indicate the terms which are the extracted features from the documents. For example, W_{nm} demonstrates the weight of the feature m in the document n. Normally, not every word presents in every document. Hence, there are different ways such as Boolean weighting, word frequency weighting, term frequency and inverse document frequency (tf-idf), and entropy model to determine the weight of features. However, the main disadvantage of these models is that the output of them is a huge sparse matrix, which leads to high dimensionality problems.

Feature selection and feature construction [11] is a NP-hard problem. By supposing $X = \{X_1, \ldots, X_n\}$ as a feature set with size n, the set contains 2^n possible feature subsets which is a huge number to analyze. It can get worse by increasing the number of features, too. Feature selection is a task to choose a minimal subset of the original features that are meaningful and related to the classification problem. Moreover, it can decrease the dimension of features significantly. Feature selection methods

can be divided into three groups with respect to the applied feature evaluation method: filter methods (Chi-square test, Euclidean distance, Correlation criteria, Information Gain, Mutual Information, Correlation based feature selection (CFS)), wrapper methods (Sequential Selection Algorithms, Heuristic Search Algorithms) and embedded [165, 99]. These methods are different in evaluating the features. Filter methods assess feature subsets apart from classification approaches. They are fast in computation, but they do not consider dependency between features. In contrast, wrapper methods utilize classification approaches to evaluate feature subsets. They consider dependencies between features, however, they are slow in computation. As an alternative, embedded methods incorporate feature selection tasks into the training step of categorizer. They are faster than wrapper approaches, but they make actions which depend on the classifier and this may not act with other classifiers [65]. These feature selection methods have been used broadly to document classification problems, but there is very limited research in the medical text classification field.

Another key challenge in medical document classification problems is to create synthetic documents that can be used beside the original documents to deal with lack of medical notes. This is particularly important to train data-hungary models such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and Hierarchical Attention Networks (HAN). However, it is not easy to achieve this goal. There are a variety of methods applied for classification problems. The introduced approaches focused on the different aspects of classification such as character-level [198], sentence-level [76] and document-level [100]. These models have high flexibility in modeling the complex relationships and show good performance. Convolutional Neural Networks (CNN) is a deep learning model which is used broadly in image classification tasks, but in text classification the situation is a little different. The location of words in the sentence is very important and pooling and convolution operations can not keep the order of words [26]. It is really hard to fit en-

tity extraction and POS tagging applications into the classic Convolutional Neural Networks (CNN) architecture [22]. On the other hand, Recurrent Neural Networks (RNN) by utilizing long short term memory (LSTM) are able to remember the order of words and this ability makes them an appropriate candidate for text analysis. Hierarchical Attention Network (HAN) [56] is another useful method for classifying medical documents by analyzing discharge notes in word level and sentence level. However, all of the mentioned methods are data-hungry models and need to be fed with more data in the training step. It is very important in the medical field to construct meaningful synthetic medical documents.

1.2.1 Challenges of Medical Document Classification

- Clinical text is complicated because often these kinds of documents are unstructured, short and noisy. Moreover, high dimensionality, sparseness, nonlinear relationships among data elements, and complicated dependencies between variables are other challenges of medical documents [188]. These documents appear in different styles with many technical terms which need domain knowledge to help.

- Many existing acronyms and abbreviations, and short expressions in medical notes make it more challenging to extract meaningful information.

- Classification accuracy is not satisfactory because of the complexity of medical texts such as grammatical mistakes, locally used abbreviations, and misspellings.

- Because of privacy issues, there is limited available data in the medical area and collecting discharge notes from patients by hospitals and other health centers takes a long time.

- In practical settings, real life patient cases are unavailable to feed

data-hungry models (a rare disease is an example where there are not enough cases for training).

- Most of the developed methods are rule-based and need expert people to set rules for each specific task. It limits the suggested approach to a specific task and it is not appliable to other tasks and it needs to reset new rules for a new task.
- Most of the available data augmentation approaches are developed for general text rather than domain-specific text such as medical documents.

1.3 Research Goals

The overall goal of this thesis is to achieve higher accuracy solutions by using domain-specific knowledge of the problem for extracting and constructing meaningful features, and developing technologies for feature selection and data augmentation in medical document classification problems. The particular research objectives of this work can be organized as follows.

1. Improving the accuracy of medical document classification by increasing the quality of input features

   Information extraction (IE) aims to extract structured information from the unstructured and semi-structured texts [159]. Fig. 1.2 shows the ontology of the areas related to IE. From Fig. 1.2, it is clear that information extraction and document classification tasks are different branches of the text mining field. IE task is targeted to extract existing entities and relations between the extracted entities. On the other hand, the document classification task assigns the available labels to the unseen documents by using the created model from the labeled documents. Many works have been done in these two tasks

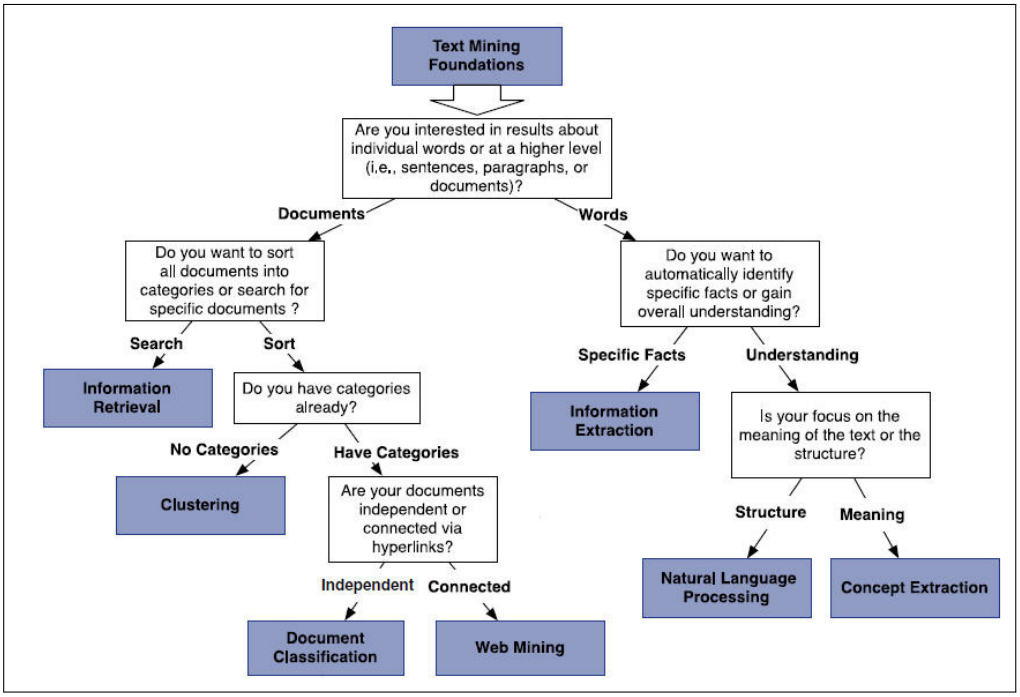

Figure 1.2: A decision tree for finding the right mining practice area [117].

separately, however, to the best of our knowledge, there is not much work that attempts to combine these tasks together. Our goal is to increase the classification accuracy by utilizing the extracted entities and relations from information extraction tasks as input features for document classification. As clinical texts are complex for analyzing, our expectation is that information extraction can extract meaningful features which can be used as input in document classification and this can lead to higher accuracy.

(a) Enriching input data by using domain knowledge

The Unified Medical Language System (UMLS) [1] was introduced by the US National Library of Medicine (NLM) for modeling the language of health and biomedicine. UMLS is a source of knowledge which improves the performance of information systems in the biomedical area. It provides three main resources:

the Metathesaurus, the Semantic Network and the SPECIALIST Lexicon. The largest component of UMLS is the Metathesaurus. It provides services such as finding biomedical concepts of phrases and relationships between concepts (e.g. SNOMED-CT, Mesh, etc.).

Our goal is to introduce a method that applies ontology by referring to Unified Medical Language System (UMLS) for entity recognition and then aggregates frequent entities to create features to improve classification accuracy. Most of the existing works are considering all of the information which some of them can be redundant. The main difference of the proposed method with previous works is that we are focused on the extracted information that is useful for the classification problem domain rather than considering all of the extracted information.

(b) Utilizing EC-based feature selection method for improving classification accuracy in 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 to improve the quality of features by removing irrelevant and noisy features. For instance, feature selection can reduce the dimensionality of the features by choosing a subset of relevant features. The objective is utilizing EC-based feature manipulation to enhance the quality of input features which is expected to improve the classification accuracy for the medical domain.

(c) Constructing new high level features by using domain knowledge in order to increase the quality of input features

In text classification, different terms of documents are consid-

ered as candidate features for training the model. These terms can be phrases such as "shortness of breath" which consists of three words. This level of feature is better than stand alone words as features. However, considering this level of features can increase features dimension extremely. Hence, we propose a method which applies ontology by referring to Unified Medical Language System (UMLS) for entity recognition, and then aggregates frequent entities to construct new levels of features to improve classification accuracy.

2. Developing ontology-based data augmentation methods for improving classification accuracy in medical domain

   One of the important factors which have an effect on the classification accuracy is the size of the data set for training the model. Generally, there is a lack of adequate data in the medical area. When the training data set is not big enough, the trained classification model does not had sufficient instances to learn. Hence, the prediction of the classifier will not be satisfactory. This issue can be worse when the data set does not have enough text inside of the documents such as a data set with only document abstracts. One possible solution to address the issue is to augment data for training the model. Most of the existing methods are developed for general text which it is possible to ignore some important domain-specific information. The available biomedical methods show good performance if the labels are not more sensitive to the order of words [79]. Hence, it is necessary to develop a domain-specific data augmentation method for biomedical document classification tasks.

   (a) Developing an ontology-guided data augmentation method to improve medical document classification accuracy

   Data augmentation is a methodology that empowers experts to build the assorted variety of data accessible for training mod-

els, without really gathering new data. Data augmentation has many applications in image classification, sound and speech classification. But there is not much work for text. It is not appropriate to augment the text by utilizing signal transformations as commonly used in image or speech classification. Because the order of words in text is important and may have semantic meaning, hence, the best approach for doing data augmentation is to paraphrase the sentences in the documents by human. But this is very expensive. A common method is to use normal dictionaries for augmentation but some domain specific terms or acronyms do not have synonyms in normal dictionaries. This issue can be addressed by using a domain-specific ontology to augment new medical documents by targeting concepts of words and expressions in the documents. This method will replace all of the words and phrases with their scientific names if they belong to a concept in the medical field.

(b) Developing a combined ontology and dictionary-based approach to improve medical document classification accuracy

As there are general vocabulary and acronyms in medical discharge notes, using only ontology-based methods may not be enough to augment medical notes with high variety. Hence, a combined data augmentation approach will be developed which combines an ontology-based method by targeting concepts of words and expressions in the documents to replace all of the words and phrases with their scientific names if they belong to a concept in medical field, and a synonym-based method by replacing all of the words and phrases with their highest similarity synonym. As this approach considers words and expressions' scientific names and synonyms, it will construct new words and phrases in documents by increasing the variety of the produced documents. It is expected that the new produced

documents with new features in them will improve the classification performance on the applied medical tasks.

3. Developing dictionary-based data augmentation methods for balancing data set and improving classification accuracy in medical domain

   Clinical discharge note classification is different from the generally document classification in terms of text vocabulary and date size. Normally due to privacy reasons, it takes a long time to collect a set of documents for a special disease. Furthermore, this type of data which are collected from real patients, often appear as imbalanced data where the patients with a particular disease are often the minority. As a result, the shortage of data and being extremely imbalanced are two common issues in medical discharge notes. In cases where the classifier is a data-hungry model and needs to be fed with a large amount of data, this kind of data will not be enough. Consequently, this can make the learning difficult for the candidate classifier [96]. In these scenarios, the learned model will be biased to the majority class. Hence, more investigation is needed to address this issue.

   (a) Developing a dictionary-based oversampling approach to clinical document classification on small and imbalanced data set

   Synonym based data augmentation can provide new documents while preserving the overall meaning of the original documents. We want to select better synonyms based on similarity. However, if a simple similarity function is used as the heuristic, the same synonym is often selected, and this reduces the variety when we need to generate multiple documents from the same original document. Medical documents can be severely imbalanced, hence, we do need to create multiple new documents to make the data set balanced. So the variety of the documents is important. If we create very similar new documents based

on the same original document, this can lead to overfitting in training the candidate classifier. Thus, an automatic dictionary-based method is introduced for oversampling the minority class documents by selecting more suitable synonyms and simultaneously increasing the diversity of the new produced documents.

(b) Developing an incremental dictionary-based oversampling approach to clinical document classification on small and imbalanced data set

Medical discharge notes are collected from real patients and they are imbalanced. Moreover, these data sets are not enough for data-hungry models (specially in rare disease cases). Both of these issues can lead to poor classification performance. Hence, the proposed dictionary-based data augmentation method is applied to all classes instead of just the minority class. The approach will balance the data set and increase the size of documents in all of the classes. It is expected the proposed approach will help the model in the training step to avoid biasing on the majority class and improve the classification performance.

1.4 Main Contributions

This section provides the following major contributions of this thesis.

1. This thesis presents a medical document classification approach to explore feature manipulation by using domain-specific knowledge. Three different approaches including feature extraction, feature construction and feature selection are proposed. Experimental results show that the accuracy of all of the classifiers in the test data set are increased significantly after applying the proposed methods. Using domain-specific ontology helps to detect meaningful words and phrases in documents and use them as features in training the can-

didate classifiers to get better classification performance. In the feature extraction approach, a medical-specific dictionary is used to extract the meaningful phrases by considering disease or symptom concepts. Using the extracted domain-specific features in training the candidate classifiers improved the classification performance. This thesis also proposes two multi-stage PSO approaches (two wrapper approaches) where prominent features are extracted in the first stage and informative features are selected by PSO in the second stage. In another method, in the first stage, important features are detected and in the second stage new high level features are constructed from the detected features. Finally, redundant features are filtered by PSO in the third stage. Experiment results of the wrapper approaches outperform the baseline machine learning approaches.

Parts of this contribution have been published in:

Mahdi Abdollahi, Xiaoying Gao, Yi Mei, Shameek Ghosh, and Jinyan Li. "Uncovering discriminative knowledge-guided medical concepts for classifying coronary artery disease notes." In Australasian Joint Conference on Artificial Intelligence, pp. 104-110. Springer, Cham, 2018.

Mahdi Abdollahi, Xiaoying Gao, Yi Mei, Shameek Ghosh, and Jinyan Li. "An ontology-based two-stage approach to medical text classification with feature selection by particle swarm optimisation." In 2019 IEEE Congress on Evolutionary Computation (CEC), pp. 119-126. IEEE, 2019.

Mahdi Abdollahi, Xiaoying Gao, Yi Mei, Shameek Ghosh, and Jinyan Li. "Stratifying risk of coronary artery disease using discriminative knowledge-guided medical concept pairings from clinical notes." In Pacific Rim International Conference on Artificial Intelligence, pp. 457-473. Springer, Cham, 2019.

2. This thesis presents how data augmentation can improve medical

document classification performance. This contribution suggests two data augmentation methods for oversampling on all of the classes to make the size of the data set bigger. The first method uses a domain-specific ontology to target concepts of words and expressions and replace them with their scientific names in documents. This approach doubles the size of the training data set. The second approach proposes a combined data augmentation method for oversampling on all of the training documents. In this approach a general dictionary is used beside a domain-specific ontology to construct new instances for the training data set. As this approach uses two ways to create new instances, it triples the size of the training data set. The obtained experimental results show applicability of the suggested approaches on real-world data sets. In comparison with existing methods, the proposed approaches achieved better classification performance in some of the tasks.

Parts of this contribution have been published in:

Mahdi Abdollahi, Xiaoying Gao, Yi Mei, Shameek Ghosh, and Jinyan Li. "Ontology-Guided Data Augmentation for Medical Document Classification." In International Conference on Artificial Intelligence in Medicine, pp. 78-88. Springer, Cham, 2020.

Mahdi Abdollahi, Xiaoying Gao, Yi Mei, Shameek Ghosh, Jinyan Li and Michael Narag. "Substituting clinical features using synthetic medical phrases: Medical text data augmentation techniques." Submitted to: Artificial Intelligence in Medicine (AIIM) Journal. pp. 1-14, Elsevier, 2021.

3. This thesis presents how oversampling can improve medical document classification performance. This contribution suggests two data augmentation methods for oversampling on the minority classes to make the imbalanced data set to be balanced. The first method uses a general dictionary to target synonyms of words and expres-

sions to replace with them in documents. The suggested approach can easily create new documents with high variety by using the extracted synonyms from the WordNet dictionary with awareness of synonyms' similarities with the original word. The second approach proposes an incremental data augmentation method for oversampling on the all of the training data set to make imbalanced data set to be balanced and at the same time increase the size of the data set. In this approach the same policy is used to generate new instances. In this approach we investigate how it works if we increase the size of all of documents in all of the classes. This approach is augmenting all of the classes by increasing their size equal to the double size of the large class. The suggested methods improve the performance in F1-measure by utilizing WordNet dictionary and three different English word embedding pre-trained models. The results show that the proposed approaches achieved better classification performance in some of the tasks in comparison with existing methods.

Parts of this contribution have been published in:

Mahdi Abdollahi, Xiaoying Gao, Yi Mei, Shameek Ghosh, and Jinyan Li. "A Dictionary-based Oversampling Approach to Clinical Document Classification on Small and Imbalanced Dataset." In: The 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology. WI-IAT'20. pp. 1-8, IEEE, 2020.

1.5 Organization of the thesis

The remainder of this thesis is organized as follows: Chapter 2 provides some fundamental background concepts and performs a literature review covering a range of works in this field. Chapters 3-5 present and discuss this thesis's contributions. Figure 1.3 shows the overall structure of Chapter 3-5 on the contributions of the thesis. Finally, the conclusion of the

thesis is presented in chapter 6.

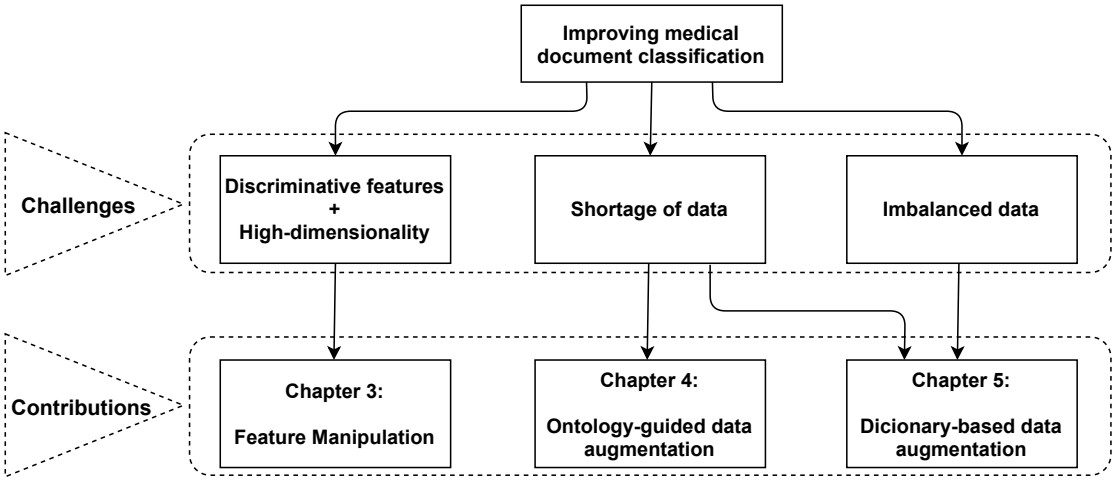

Figure 1.3: The overall structure of the contributions

Chapter 2 provides fundamental concepts and related background of medical document classification, machine learning, evolutionary computation and Particle Swarm Optimisation (PSO), feature manipulation (including feature extraction, feature selection and feature construction), data augmentation. Then, it presents an overview of the related works for medical document classification.

Chapter 3 presents three feature manipulation approaches including a ontology-based feature extraction, PSO-based feature selection and a discriminative knowledge-guided medical concept pairings from clinical notes for feature construction. This contribution targets concepts of the appeared words and expressions in documents by focusing on two specific concepts (diseases and symptoms) to extract features and construct new features. Finally, PSO is applied to reduce the dimensionality of features' space by eliminating redundant features.

Chapter 4 describes two data augmentation approaches to deal with shortage of data in medical field. The first approach is using an ontology-guided method to construct meaningful documents for training data set.

The second approach is combining two different methods to increase the size of all of the classes in training data set.

Chapter 5 proposes two oversampling approaches to make an imbalanced data to be balanced. In the first approach a synonym-based method which is using a WordNet dictionary to replace words with their best fit synonyms by measuring the similarity of synonyms with the original words.The second approach is an incremental method to increase the training set documents as much as necessary.

Chapter 6 summarises the thesis and draws the overall conclusions of the thesis. It provides main contributions and key research points of each chapter of the thesis. It then suggests some possible future research directions.

Chapter 2

Literature Review

This chapter introduces the basic concepts of using machine learning techniques, feature engineering methods, evolutionary computation, and related work on medical document classification. A brief overview of machine learning and its different approaches is described, followed by a review of Evolutionary Computation (EC) techniques such as Particle Swarm Optimization. Then, a number of steps involved in medical document classification and different kinds of features used in existing work are also described. Moreover, a description of Information Extraction (IE), document classification, feature manipulation, and data augmentation are provided. This chapter discusses recently proposed ideas in document classification headlining their advantages and limitations for the related work.

2.1 Background

2.1.1 Machine Learning

Machine learning is a broad research field that explores the study and construction of algorithms that can learn from and make predictions on data [115]. It is a sub-field of computer science, evolved from the study of pattern recognition and computational learning theory in artificial intelli-

gence [115]. The key factors of machine learning are representation and generalization [147]. While the former is concerned with the representation of data and various functions evaluated over this data, the latter represents the ability of the system or the model to handle unseen data based on the knowledge gained from the seen examples. Generally, algorithms of machine learning can be divided into the five following groups [10]: supervised learning; semi-supervised learning; unsupervised learning; reinforcement learning; and (5) transfer learning.

- **Supervised Learning:** A supervised learning algorithm aims to define a generalized function capable of predicting an output for unseen data relying on the information of the previously seen data. Classification and regression represent the most well-known example applications of supervised learning [118, 93]. In the case of classification, the system takes the available list of features or description of the inputs (training instances) and predicts the class label for each of them. The desired (known) outputs are used to guide the system during the training phase. Typical methods of this approach include Artificial Neural Networks (ANNs) [91], Decision Trees [156], and Naive Bayes (NB) [82].

- **Unsupervised Learning:** For unsupervised learning, the labels of the data set instances are not available. The model is trained by exploring the data to detect hidden patterns and structures [59]. Unsupervised machine learning is suitable when it is difficult to annotate the data or agree on what the right labels should be. It is usually used to detect groups of instances that are similar or relevant in some manner. Clustering and association rule learning are common tasks of unsupervised learning.

- **Semi-supervised Machine Learning:** The semi-supervised learning methods combine the schemes of both supervised and unsupervised learning methods, in which unlabelled data along with labeled data

are used to train a model. Generally, the number of labeled instances is smaller than the number of unlabelled instances. A typical example of semi-supervised learning methods is a transductive support vector machine (TSVM) [74, 105].

2.1.2 Classification Methods

The past decades witnessed a significant development of machine learning algorithms including classification algorithms. Many classifiers have been proposed. The rest of the thesis presents some commonly used classifiers including k-Nearest Neighbour (KNN), Decision Tree (DT), Support Vector Machines (SVM), Naive Bayes (NB), and Neural Networks.

- **K-Nearest Neighbors (K-NN):** k-Nearest Neighbour (KNN) [39, 81, 110] is a type of instance-based learning which does not perform any explicit induction or learning process. It compares new instances with instances seen in the training set. In KNN for classification, all the distances from the test instance to each instance of the training set are calculated to determine k nearest neighbours, where k is a positive and user predefined number. Then, the test instance is classified by a majority vote among these k nearest neighbours. Different proximity measures can be used to calculate the distance between instances such as Euclidean distance for continuous values and Hamming distance for discrete values. KNN is simple but works well in practice. It is also a nonparametric classifier which means no assumption about the probability distribution of the underlying data is needed. However, it is time and memory-consuming due to the need for distance calculation from the query instance to all training instances. Furthermore, because the classification decisions are made locally, it is quite susceptible to noise especially when k is small. Appropriate proximity measures and preprocessing steps have to be chosen for KNN to produce good predictions.

- **Decision tree:** Decision tree (DT) is another nonparametric classifier that is used for nominal, numeric, or mixture data [118]. A classifier in this method is presented as a tree in which each inner node is a decision stump or a split point and each leaf node is a class label. A query instance will traverse through the tree by testing its input values against the decision stumps and finally reaches its class label. To build a DT, a learning algorithm carries out a heuristic-based search using information gain as a criterion to choose the best split feature for each node. The best feature is the one that can split instances into separate groups that are as high homogeneous as possible. Different impurity measures were used to implement different DT algorithms such as C4.5 with information gain, CART with Ghini index, and CHAID with Chi-squared test [63, 44]. DT learning is computationally inexpensive and robust to noise. Moreover, the learned classifier can be translated into comprehensible rules. However, since only one feature is tested in a node, DT does not perform well for cases where boundaries between classes are not parallel with coordinate axes [103]. These cases usually happen in problems that have two-way or multi-way relationships among features.

- **Support Vector Machine (SVM):** Support vector machine (SVM) is a classification technique for continuous data that is rooted in statistical learning theory. SVM learns to find optimal hyperplanes in a higher-dimensional space, which have maximal margins to their nearest instances of different classes [68]. The original SVM method can only work with binary problems. Different versions of SVM have been proposed for multi-class problems such as SVM one versus one, one versus rest methods [98]. SVM has shown promising empirical results in many practical applications. However, users must provide the type of kernel function to use. Another drawback of SVM is its high computation time especially when the number of dimensions is high.

- **Bayesian:** A Bayesian classifier is an approach for modelling probabilistic relationships between the feature set and the class variable. Naive Bayes (NB) classifier is a simple implementation of a Bayesian classifier where class-conditional probability is estimated by assuming that features or attributes are conditionally independent given the class label [118]. Although NB has been shown to be competitive with DT [158], it is not feasible to real-world problems where feature conditionally independent assumption can not be held.

- **Neural Network:** A neural network classifier is a network of units, where the input units usually represent terms, the output unit(s) represents the category. For classifying a test document, its term weights are assigned to the input units; the activation of these units is propagated forward through the network, and the value that the output unit(s) takes up as a consequence determines the categorization decision. Some of the researchers use the single-layer perceptron, due to its simplicity of implementing [40]. The multi-layer perceptron which is more sophisticated is also widely implemented for classification tasks [139]. Models using back-propagation neural network (BPNN) and modified back-propagation neural network (MBPNN) are proposed in [104] for documents classification. An efficient feature selection method [123] is used to reduce the dimensionality as well as improve the performance. Trappey and et. al. are developed a document classification and search methodology based on neural network technology [164], which is helpful for companies to manage patent documents more effectively. New Neural network-based document classification methods are utilized in [56] such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Hierarchical Attention Network (HAN). They developed these networks to classify medical discharge notes by targeting classification performance.

2.1.3 Evolutionary Computation

Evolutionary computation (EC), inspired by the theory of natural selection and genetic inheritance, is a sub-field of artificial intelligence and refers to the family of algorithms for global optimization inspired by biological evolution [161]. They are a family of population-based trial and error problem solvers with a stochastic optimization character [115]. The increasingly active field of EC provides valuable tools, for problem-solving, machine learning, and optimization [28]. In particular, industrially relevant fields, such as signal and image processing, computer vision, pattern recognition, industrial control, scheduling and timetabling, telecommunication, and aerospace engineering, are using EC techniques to solve complex problems [28]. Under the EC umbrella, there are evolutionary algorithms (EAs), swarm intelligence (SI) algorithms, and other optimization techniques.

Evolutionary Algorithms (EAs) [127] is a sub-field of EC, which are population-based optimisation algorithms. EAs are based on mechanisms inspired by biological evolution including genetic operators like selection, reproduction, mutation, and crossover. In EAs, each candidate solution is represented as an individual in the population. The fitness or evaluation measure determines the goodness of each individual. The evolution of the population then takes place after the repeated application of the mentioned genetic operators. Examples of EAs include genetic algorithms, genetic programming, evolution strategy, and evolutionary programming.

Another area of EC is Swarm Intelligence (SI) algorithms which are inspired by the collective intelligence of social insects. A swarm is defined as a population of interacting individuals that is capable of optimizing global objectives through collaborative search. Here, intelligence lies in the networks of interactions among individuals, between individuals, and the environment [83]. There is a general stochastic tendency in a swarm for individuals to move towards a center of mass in the population, which results in convergence on an optimal solution [83]. The most common

optimization techniques in SI are Particle Swarm Optimisation (PSO) [46, 163] and Ant Colony Optimisation (ACO) [37, 119].

EC approaches are very powerful methods in solving optimization problems. They can deal very well with problems that have a huge search space such as feature selection and feature construction. Genetic Algorithms (GA) and Particle Swarm Intelligence Optimisation (PSO) are suitable for feature selection and Genetic Programming (GP) is appropriate for feature construction problems. Hence, we will describe GA, PSO, and GP in more detail.

Genetic Algorithms (GAs)

Genetic algorithms (GAs) provide an approach to learning that is based on biological evolution [118]. Candidate solutions are often encoded as bit strings whose interpretation depends on the application. The search for an appropriate solution begins with a randomly generated population, or collection, of initial solutions. Members of the current population give rise to the next generation population by applying operations such as mutation and crossover, which are modelled after processes in biological evolution [118]. At each generation, the solutions in the current population are evaluated based on a measure of fitness, with the most fit solutions selected probabilistically as parents for evolving the next generation.

GAs have been explored widely and applied successfully to a variety of learning and optimization problems [174]. For example, they have been used to learn collections of rules for robot control [113] and to optimize the topology and learning parameters for ANNs. They can search spaces of solutions containing complex interacting parts, where the impact of each part on the fitness of the overall solution may be difficult to predict [118]. However, GAs tend to be computationally expensive but they can be easily parallelized taking advantage of powerful computer hardware, hence, resulting in decreased costs [118].

Particle Swarm Optimization (PSO)

Particle swarm optimisation (PSO) is a population-based stochastic optimization technique inspired by the social behavior of birds flocking or fish schooling [84]. In PSO, each candidate solution is encoded as a particle moving in the search space according to a simple mathematical formula to update particles' position and velocity. Each particle remembers its local best-known position. Hence this collection of particles known as swarm searches for the optimal solution by updating the position of each particle based on the local best-known position of its own and its neighboring particles [84]. PSO is a simple but powerful search technique. It is a metaheuristic as it makes few or no assumptions about the problem being optimized and hence, can search very large spaces of candidate solutions. However, metaheuristics do not guarantee that an optimal solution is always found [83].

Genetic Programming (GP)

GP is an EC technique that evolves solutions in the form of computer programs. A GP evolved individual (program) is usually expressed as a syntax tree. This individual is made up of a root node, a number of internal nodes, and some leaf nodes. The internal nodes consist of functions that are usually arithmetic operations (e.g. $+$, $\times$ and $\div$). The terminal nodes (leaves) are variables or constants. Similar to other EC algorithms, a population of individuals are randomly generated in the initialization step in GP. In order to generate the individuals of the initial population, Koza [134] specified three different techniques which are growing, full and ramped-half-and-half. Ingrow method, nodes are randomly selected from the whole primitive set (i.e., functions and terminals) until a terminal is selected or a predefined maximum depth is reached. When the predefined maximum depth is reached, only terminals may be chosen. In the full method, nodes are randomly taken from the function set until the

predefined maximum depth is reached, and by reaching the maximum depth, only terminals can be chosen. ramped-half-and-half is a combination of growing and full which can provide individuals vary in shape and size [95]. In order to evaluate the goodness of each individual (program), an evaluation measure, called the fitness function is used. Since the performance of the evolved program is assessed by the fitness function, determining an appropriate fitness function is critical. GP system considers a selection method to choose individuals to produce a new generation. In this procedure, better individuals are more likely to be chosen than inferior individuals [134]. There are different selection methods such as fitness-proportional selection, truncation selection, and tournament selection [18]. To produce a new population, some genetic operators are employed to create new individuals (children) from the current individuals (parents). There are three genetic operators: reproduction (elitism), crossover, and mutation.

2.1.4 Information Extraction

The automatic extraction of information [50] from unstructured sources has opened up new avenues for querying, organizing, and analyzing data by drawing upon the clean semantics of structured databases and the abundance of unstructured data. The field of information extraction has its genesis in the natural language processing community where the primary impetus came from competitions centered around the recognition of named entities like people names and organizations from news articles. As society became more data-oriented with easy online access to both structured and unstructured data, new applications of structure extraction came around. Consequently, there are many different communities of researchers bringing in techniques from machine learning, databases, information retrieval, and computational linguistics for various aspects of the information extraction problem in a different field such as the medical domain.

Information extraction (IE) task targets to extract structured information from the unstructured and semi-structured documents [159]. IE is a very challenging task and it is very complex in medical documents because there is more domain knowledge in the medical documents. Information extraction by considering the domain of a candidate problem can provide relative entities which are meaningful. There are some works that tackled the problem in different ways. The available approaches are rule-based methods, dictionary-based methods, and machine learning methods. However, these approaches have some limitations. The rule-based methods should be designed by experts in the medical area and a suggested method that is suitable for the specific problem may not have generality to be used for other problems. One of the issues with the dictionary-based methods is that they are not complete. Because of the advancement of medical science and new discoveries over time, there are also new specialized terms and phrases that the dictionaries do not usually have. Hence, it is often possible to have an entity that is not in the dictionary. The machine learning methods show better performance in comparison to the rule-based and dictionary-based methods. Consequently, the machine learning approaches can be better choices by solving these mentioned problems. Furthermore, utilizing the latest updated dictionary with machine learning methods as a hybridization method to extract meaningful entities can be useful.

2.1.5 Feature Manipulation

Feature Weighting

Feature weighting [124] aims to assign a weight to each feature based on its degree of relevance to the target concept. Feature weighting gives high weights to the relevant features and low weights to the irrelevant features. Relief [101] is an example of feature weighting methods that use distance measures to evaluate the degree of feature relevance.

Feature Selection

In the literature, there are many definitions for feature selection based on different criteria, but most of them follow a similar intuition and/or content [41]. Feature selection is a process that aims to find a minimal subset of features to achieve similar or better performance than using all the original features by eliminating noisy and irrelevant features. The process of selecting informative and relevant features not only reduces the dimensionality, which can make the learning method faster but also improves the performance of the method. In addition, it is easier to interpret the constructed method by a smaller number of features [121].

A typical feature selection process has five major steps [41]: the initialization procedure, candidate feature subset generation, feature subset evaluation, stopping criteria, and a validation procedure.

1. The initialization procedure is the first step of a feature selection algorithm and the number of original features is taken as the dimensionality of the search space.

2. The candidate feature subset generation is known as the search procedure [102], which can start with no features, all features, or a random subset of features. In this step, many search techniques such as EC techniques, and conventional methods, can be employed to explore the best feature subset.

3. The generated candidates are evaluated based on a criterion which is called a fitness function. The fitness function will measure the goodness of each candidate, so it has an important role in guiding the algorithm to find an optimal solution.

4. The feature selection algorithm will be stopped when the stopping criterion is satisfied. The generation procedure and evaluation function can be used to determine the stopping criterion. For example, the algorithm can stop when a predetermined maximum number

of iterations have been reached, or a predefined number of features have been selected.

5. The validation procedure aims to check whether the subset of features can achieve good performance. Even though this part does not belong to the feature selection process, it is necessary to validate the selection method.

Feature Construction

Feature construction is a process that combines the original features to construct new high-level features [200]. Feature construction aims to improve the quality of representation by transforming the original representation space, i.e., features, into a new one in which the capability of a learning algorithm can be improved [122]. Constructed features are mathematical expressions of the original features. In order to enhance the performance of a method, the original feature set can be augmented or replaced by the newly constructed features [97]. A typical feature construction includes the following four steps.

1. Feature construction: New features are constructed by combining selected features using mathematical operators. The key point is to select appropriate features and operators, so the newly constructed features will have a higher discriminating ability than the original ones.

2. Feature evaluation: To guide the search algorithm, the constructed features are evaluated by means of a fitness function similar to feature selection methods.

3. Stopping criterion: Similar to feature selection methods, when a stopping criterion is met, the best-constructed features will be returned.

4. Validation procedure: This step is similar to feature selection methods. The constructed new features are required to be checked whether they can achieve a good performance.

Feature Selection and Feature Construction Approaches

There are three different types of approaches for feature selection, and feature construction: filter, wrapper and embedded methods.

Filter: Filter methods evaluate feature subsets based on the intrinsic characteristics of the training data rather than the feedback of a learning algorithm [144]. Filter methods can use different types of measures such as distance measure, information measure, dependence measure, and consistency measure [42]. Filter approaches have low computational cost and they are fast due to the avoidance of the inductive algorithms. However, evaluating the subsets in the search process is a challenging issue without relying on inductive algorithms. They are often not optimized to be used with a learning algorithm for specific tasks. They usually have lower classification performance than other methods (e.g. wrapper) on a particular learning algorithm, as the prediction performance of the selected features on a learning algorithm is not considered in the filter methods [144].

Wrapper: Unlike filter methods, wrapper methods employ an inductive algorithm to evaluate the goodness of the selected features [162, 51]. As compared to filter methods, wrapper methods result in more effective feature subsets, but the computational complexity is usually higher. This approach is claimed to be less general than the filter methods since the selected feature subsets have mainly relied on the predetermined learning algorithm. For example, the best feature subset evaluated by one learning algorithm(e.g. Decision Tree (DT)), may not improve the prediction performance of another learning algorithm (e.g. Support Vector Machine (SVM)) [35].

Embedded: Unlike wrapper methods, embedded methods make an interaction between the learning algorithms and the feature selection or construction approaches. Embedded methods determine the features and the learning algorithm (e.g. classifier) simultaneously during the training process [21]. For example, in DT, the tree is built by partitioning the data according to the importance of the features to the classification accuracy. Another example is GP which has an intrinsic capability of selecting or constructing features, and it improves the performance of the method [132]. The selected features by embedded methods are more effective than those generated by filter methods. Moreover, they have less computational cost than wrapper methods [181].

2.1.6 Text Mining and Document Classification

Text mining [199, 94] is one of the important topics in artificial intelligence which deals with analyzing different types of unstructured text to extract useful knowledge. There are many tasks in text mining such as document classification, document clustering, entity extraction, document summarization and semantic analysis. Document classification is one of the extensively studied natural language processing tasks. In document classification, the goal is to automatically classify text documents into one or more predefined classes. For example, detecting spam and non-spam emails, automatically tagging client queries and categorizing news articles are some applications of document classification. The main steps of document classification consist of preprocessing, text representation, feature weighting, feature selection, training, testing and interpretation.

Document classification [9] is the task of assigning label l_i to document d_j, where $l_i \in L = \{l_1, \cdots, l_{|L|}\}$ and $d_j \in D = \{d_1, \cdots, d_{|D|}\}$, using a function F:

$$F : D \rightarrow L \tag{2.1}$$

In formula (2.1), function F is a classifier which gets documents (D) as input and allocates labels (L) as output to each of the input documents.

During the past decades, machine learning algorithms like classification has been developed and many classifiers such as K-Nearest Neighbor (KNN) [16], Decision Tree (DT) [36], Support Vector Machines (SVM) [67], Naive Bayes (NB) [173], and Neural Networks [142] have been proposed. Statistical algorithms and artificial intelligence techniques have been used to automatically classify documents [8, 87, 145]. In document classification tasks, the frequency of words or the letter combination is considered an important attribute to construct basic features. In general, the number of letter combinations is huge, and the frequency of each combination is not high. In the stage of pre-processing, some texts will be filtered, such as "and". In the process of feature extraction, it is popular to use n-gram techniques [8, 30]. If we only consider the word-specific combination, the bag-of-words model can be considered as a special case in the n-gram model. The number of occurrences of each word is a general and basic feature in a bag-of-words model. Considering the frequency of each word is not high and the set of words is large in a document classification task, we need effective feature selection techniques to select small sets of features from a high dimensional and sparse set of basic features. Traditional feature selection methods, such as information gain [191], are generally employed [8]. After a small set of features is selected, learning algorithms, such as Support Vector Machine (SVM) [73], are used to build classifiers.

There are a variety of methods applied for classification problems. For example, neural network methods have been used a lot in researches for document classification problems. The introduced approaches focused on the different aspect of classification such as character-level [198], sentence-level [76] and document-level [100]. These models have high flexibility in modeling the complex relationships and show good performance, how-

ever, they are not fast in training and test phases [76]. Hence, by increasing the number of documents, the speed of learning will decrease and it will be time-consuming. Additionally, by increasing the number of hidden nodes, the required parameters for Neural Networks will increase which can lead to overfitting of the data [99]. Deep learning (DL) which is an artificial neural network (ANN) model, needs more data in the training phase to have better performance, but high volume data is not always available, especially in the medical field. Convolutional Neural Networks (CNN) is a deep learning model which is used broadly in image classification tasks, but in document classification, the situation is a little different. It is really hard to fit entity extraction and POS tagging applications into the classic Convolutional Neural Networks (CNN) architecture [22]. On the other hand, Recurrent Neural Networks (RNN) by utilizing long short term memory (LSTM) are able to remember the order of words and this ability makes them an appropriate candidate for text analysis. Extreme learning machine (ELM) is another useful method for cases that the dimension of features is high [78]. Generally, all of the document classification approaches for general text are similar to medical document classification. One of the limitation of these methods is that they are not using domain-specific clinical resources such as Unified Medical Language System (UMLS) to be able to serve additional feature [133].

2.1.7 Data Augmentation

Data augmentation is a technique that enables researchers to make a suitable variety of instances available for training models, without really collecting new instances. Data augmentation has numerous applications in image classification, sound and speech classification [138]. With regard to text, it is not proper to increase the text by employing the signal transformations as generally utilized in image or speech classification because the words in the text are valuable and can have semantic meaning. There-

fore, the best method for augmenting new instances is to paraphrase the sentences in the documents by humans. However, paraphrasing is very time-consuming due to the large size of existing sentences in documents. Another intuitive way to produce new discriminative data set is shuffling words or sentences. However, as the order of words and sentences can have semantic meaning in medical discharge notes, shuffling can change the explanation of a patient's health condition. Another method is randomly choosing n words from a sentence and replacing them with one of their synonyms randomly [171]. Since the synonyms might have different levels of similarity to the target word, random selection is not optimal [198].

Data augmentation is more important for imbalanced data sets. A data set with a different number (unequal distribution) of instances for each class defines as an imbalanced data set. This type of data can make learning difficult for the candidate classifier [96]. In these scenarios, the learned model will be biased on the majority class. Hence, more investigation is needed to address this issue. There are three main approaches to imbalanced data set: data-level methods, algorithm-level methods, and hybrid methods [96]:

- **Data-level:** Data-level methods aim to make a balanced distribution of classes by generating new instances and/or deleting some instances.
- **Algorithm-level:** Algorithm-level methods try to modify learning approaches to reduce the bias towards majority classes and tune them to learn data with biased distributions.
- **Hybrid:** Hybrid methods integrate the data-level and algorithm-level methods to get the advantages of both of the methods.

In this thesis, we target data-level approaches to deal with imbalanced data, which is appropriate for data-hungry deep learning methods. Data-level provides more information for the candidate model in the learning

step and at the same time solves the imbalanced issue. As the imbalance issue can be solved by a data-level approach, it is not necessary to tune algorithms to learn the data with biased distribution (Algorithm-level). Consequently, it is not necessary to use both of the approaches at the same time (Hybrid).

2.2 Related Works

The organisation of the section follows the targeted objectives order. First, medical document classification is discussed as the main goal, then feature extraction, feature selection, and feature construction are discussed as main tasks to do for improving document classification accuracy, and finally, data augmentation approaches are reviewed.

2.2.1 Document Classification in Medical Domain

Make a prediction model by classification algorithms in medicine, especially in genomics, and forecasting outcomes were studied in [20]. This paper focused on classification algorithms such as Decision Tree, Decision Rule, Logistic Regression, Neural Network, Naive Bayes, Bayesian Network, Support Vector Machine, and K-nearest Neighbor. This study tries to provide a comprehensive framework to organize the application of data mining in medicine. Applications of data mining algorithms in healthcare and biomedicine are proposed in Yoo et al. [196]. In this paper, three data mining tasks are selected and then the application of each task in medicine is reviewed. For this purpose, a brief discussion about each task and their advantage/disadvantage is presented. The main medical aspects that mentioned are: predicting health costs, prognosis, and diagnosis, extracting hidden knowledge from biomedicine data, discovering relationship among diseases and among drugs. This paper summarizes three problems of data mining in medicine: setting and calibrating param-

eters of algorithms, the accuracy of data mining is not reliable yet and, lack of data mining package for the medical domain. Wagholikar, Sundararajan, and Deshpande [170] review applications of more than eight modeling techniques in diagnosis. In this study, more than ten diseases were selected. Authors conclude that the application of these techniques in gastroenterology, oncology, and cardiovascular are more than the other diseases. Feature selection methods using Gini Index were employed along with models like Bayesian networks and decision trees to improve Medline document classification [130]. Furthermore, Parlak and Uysal [129] were studied the feature selection impact on medical document classification. They analyzed the performance of two classifiers including Bayesian networks and C4.5 decision trees on OHSUMED and MEDLINE data sets. The experimental results showed that utilizing Bayesian networks with distinguishing feature selectors achieved better classification performance.

In 1988, a multi-layer perceptron (MLP) network was used to diagnose low back pain and sciatic. Then the system performance was compared with three groups of doctors and the other computer program [25]. Another useful example of MLP is one that was established to support the diagnosis of heart diseases [189]. In a similar case [148] the authors applied the MLP method to diagnose tumors based on chromatography analysis of urinary nucleoside as a practical pattern recognition instrument to differentiate cancer patients from healthy persons. Authors in [19] were evaluated the performance of machine learning algorithms on biomedical document classification and found that the future goal would be improving the classifiers such as deep learning approaches that are more adaptive to large data size. Hashemzadeh et al. [64] developed a semi-supervised transfer learning method for the cases there are not enough labeled medical documents. The limitation of this work is a shortage of the labeled hospital documents which has a negative effect on their results.

Extreme learning machine (ELM) proposed by Huang et al. [70], provides a simple and efficient learning algorithm for single hidden layer

feedforward neural networks (SFLNs). It has an extremely fast learning speed. The hidden nodes are randomly started and then they are fixed without iteratively tuning [69]. Therefore, ELM has many advantages to be used in medical or biomedical data which have high dimensional features. There have been different studies that applied ELM such as thyroid disease diagnosis [106] which used ELM to assist the tasks using the most discriminative new feature set and the optimal parameters. The ELM approach is also used for blood indexes to predict overweight statuses consisting of 251 healthy subjects and 225 overweight (297 females and 179 males). The results showed the differences in blood and biomedical indexes. Lately, Anthimopoulos et al. [14] proposed and evaluated a conventional neural network (CNN) method for the classification of interstitial lung disease (ILD) with a classification performance of 85.5%.

A CNN-based approach was developed by Hughes et al. [71] to classify automatically medical sentences which are extracted from PubMed data set. They used different representations such as Doc2vec and Word2Vec for training their model. The developed CNN method by Word2Vec representation showed better performance in comparison with the state-the-art approaches. However, still, the achieved accuracy is poor and needs to be improved. In [135], authors have developed a new hierarchical neural network approach which is using an attention mechanism to analyze the medical notes in the word level and sentence level to deal with data sparsity issues. The proposed model used a Convolutional Neural Network (CNN) to extract features from sentences and a bidirectional gated recurrent unit (BIGRU) to remember the appeared features before and after each feature. The approach has trained on two types of data sets (medical notes and clinical literature) separately. The obtained results showed their method outperformed baseline methods, however, the method is expensive to implement. Li et al. [108] were suggested a combined three-stage approach by using a bi-directional long short-memory which utilizing an attention-based approach to classify medical documents. The approach

extracted features to pass to the construction step which is based on the regular expression rule. The method has been designed to provide interpretable results to assist humans in modifying their decisions. However, the implantation of the approach is expensive.

SVMs techniques have also been successfully used in the medical field. Wen et al. [172] used the support vector machine (SVM) approach which originates from the root of the Kernel method. Another useful research used the SVM tool to predict head and neck patients' criticalities for adaptive tomotherapy treatments [62]. In addition, an SVM classifier was used to recognize breast cancer for different conditions (normal and malignant) of the breast. In [55] the authors developed new methods to analyze microarray expression data of cancer tissue samples by using SVMs which consist of classification of the tissue samples and an exploration of the data for mislabeled or doubtful tissue results. Moreover, Acharya et al. [7] aimed to assess the possibility of utilizing thermal imaging as a possible instrument for recognizing breast cancer by utilizing an SVM classifier for automatic classification of normal and malignant breast circumstances. Authors in [48] were developed a supervision system to detect patients' documents if their discharge notes include hospital-related infections. They applied support vector machines and gradient tree boosting classifiers on Swedish sick discharge notes to classify them based on hospital infections. The best results were obtained by the gradient tree boosting classifier using prepossessing methods and tuning classifier parameters. Although the suggested approach by the authors outperformed the pipeline methods, the achieved results are still poor and need to improve. Additionally, the distribution of the used data sets is not like real-life data set.

Researchers used other classifiers for medical document classification too [13]. Sako and Palimote [140] applied a Naive Bayesian classifier for classifying heart disease documents. They employed WEKA [169] tool for preprocessing and extracting features to train the classifier. Generally, all

of the document classification approaches for general text are similar to medical document classification. One of the limitations of these methods is that they are not using domain-specific clinical resources such as Unified Medical Language System (UMLS) to be able to serve additional feature [133]. Furthermore, the accuracy of medical document classification methods still is poor and needs more investigation.

2.2.2 Information Extraction in Medical document Classification

The extraction and selection of features for document classification problems have received a lot of interest in the past. Typically, a lot of these algorithms rank features using statistics from the distribution of features in the given corpus [149, 145]. Existing methods have employed metrics associated with word frequency, information gain, mutual information, term frequency-inverse document frequency (TF-IDF) for extracting textual features [88]. However, the aforementioned techniques tend to treat each feature separately, i.e they ignore the dependencies between features.

In medical 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 texts include heavy use of domain-specific terminology and the frequent inclusion of acronyms, which makes medical text analysis very different from standard text mining. Especially, a discriminative combination of domain-specific medical events reported within a clinical note can be highly indicative of a patient's condition.

There has been research that applies document classification to medical document. Previously, Pratt et al [194] employed words, medical phrases, and their combinations as features for medical document classification. Multi-label classification performance based on an associative classifier is examined on medical articles [137]. In another study, Hid-

den Markov models were used for classification [195]. In a recent study, an approach using support vector machines and latent semantic indexing was applied to some data sets including the ones consisting of medical abstracts [166]. The performances of classifiers on medical document classification were analyzed for two cases where stemming was applied and not applied [128]. The impact of different text representations of biomedical documents on the performance of classification was analyzed [193]. In another work [80], the impact of using different ontologies (such as iCSV and SEMCON) was investigated in medical document classification performance.

Besides, there exist a number of studies in the literature where ontology-based classification approaches have been applied [29, 45]. The use of ontologies like Unified Medical Language System (UMLS), Systematized Nomenclature of Medicine (SNOMED), and Medical Subject Headings (MeSH) have proved very useful for improving classification performance [57, 149, 27, 154]. Recently, Shanvas et al. [150] used an ontology-based approach to make rich graphs of concepts for clinical document classification.

In addition, some work has used clinical records for prominent tasks such as finding risk factors for diabetic patients [86], extracting Framingham risk score (FRF) for target population [75], using rule-based and dictionary-based methods to identify heart disease risk factors [190], and applying a rule-based method by combining with a regular expression and UMLS to spot risk of heart disease [152]. Our approach applies an ontology as a feature selection method for document classification.

By analyzing the previous work, it is noticeable that the majority of disease-targeted systems have tended to develop static rule-based systems to extract meaningful information which requires human interventions every time the model is updated with new features. Such systems are not scalable for practical machine learning purposes. Our expectation is that the extracted meaningful information in the IE step can be utilized as fea-

tures to train classifiers to achieve a better model which can predict classes of test data with high accuracy.

2.2.3 Feature Selection in Medical Field

Computational intelligence approaches (CI) techniques are classified based on single and hybrid methods, where single methods refer to those studies which use only one of the machine learning techniques (i.e. genetic programming (GP), particle swarm optimization (PSO), artificial immune system (AIS) and artificial neural network (ANN)) as the main method and the other classification refers to those studies that used hybridization of each two (or more than two) methods like Neuro-Fuzzy (NF) and Fuzzy Support Vector Machine (FSVM).

Healthcare is one of the domains that used the aid of PSO to predict or analyze diseases and it is also a useful approach for the medical field. For instance, Eberhart and Hu [47] presented an approach for the examination of human tremor using PSO which addressed two forms of human tremor as important tremor and Parkinson's disease. PSO is used to improve a neural network that differentiates between normal subjects and those with tremor. Another useful example of PSO which also had the same topic based on previous researches is for the prediction of Parkinson's disease tremor. This method uses a Radial Basis Function Neural Network (RBFNN) based on PSO to detect that Parkinsonian tremors are happening from Local Field Potential (LFP) signals [176]. Furthermore, Fong et al. [54] used PSO to search for optimal feature subsets, together with three classifiers, i.e., pattern network, decision tree, and Navies Bayes. The approach was shown to achieve high accuracy in classification in two empirical biomedical data sets, i.e., the Arrhythmia data set and the Micro Mass data set.

GA and PSO were proposed for gene selection and were tested on three benchmark gene expression datasets including breast cancer, colon, and

leukemia. The effectiveness of this hybrid method shows that it is able to decrease the dimensionality of the dataset and improve the most informative gene subset and enhance classification accuracy [107]. In another example, PSO and GA were used to optimize feature selection using facial and clothing information for gender classification. The PSO-GA was used to choose the best significant features set which more evidently demonstrates the gender and therefore, the data size dimensions can be decreased [120].

SVM-PSO approach is widely applied in the medical field and it has shown significant performances. For example, PSO and SVM were combined as a hybrid method for gene selection and tumor classification where the PSO was used to choose a gene, whereas SVM was used as the classifier. Then, the suggested method was applied to microarray data and it had a great prediction performance for accuracy [151]. In another study, Jiang et al. [72] used a new approach for identifying liver cancer that works based upon the PSO-SVM method. PSO was applied to select the parameters automatically for SVM. This is an essential advantage that makes it possible to select parameters more objectively and it keeps away the subjectivity in the old SVM model. In another research, SVM-PSO was proposed as a feature selection for enhancing medical diagnosis validity by applying machine learning ensembles in order to obtain higher accuracy of ensembles [112]. Liu and Fu [109] introduced a novel method that hybridizes SVM, PSO, and Cuckoo Search (CS). This technique uses 2 steps: (i) developed a method of Cuckoo search for improving the SVM parameters to identify the better initial parameter of a kernel function, (ii) PSO is applied to the remainder of the training of SVM and identifies the best parameter in SVM.

Many ideas have been proposed to improve the performance of PSO-based feature selection algorithms. The ideas include modifications in the initialization strategy, representation, fitness function, or search mechanisms. In [185], Xue et al. proposed three new initialisation mechanisms,

which mimic the sequential feature selection approach. While the small initialisation used about 10% of original features to initialize the particles, particles in the large initialisation were constructed based on 50% of original features. These two initialisation mechanisms were combined in the mixed initialisation, which used the small initialisation for most of the particles and the large initialisation for the rest. In addition, three new updating mechanisms for $pbest$ and $gbest$ were proposed in the paper. The experimental results showed that the new initialisation and updating mechanisms led to smaller feature subsets with better classification performance than the standard PSO and two-stage binary PSO algorithm [183, 184].

In order to solve feature selection problems, PSO has been used with different filter measures, for example, rough set theory [32, 31], fuzzy set, and information theory [34]. The goal of attribute reduction using a rough set is to find the smallest feature subset, which still preserves the classification quality as the original feature set. Therefore, Cervante et al. [32] proposed a new fitness function, which used probabilistic rough set theory to minimize the number of equivalence classes and maximize the number of instances in each equivalence class. The reported results illustrated that the new fitness function could guide PSO to search for a small subset, which had better classification performance than the subsets evolved by two other PSO-based feature selection approaches using a rough set.

Feature selection problem can be seen as a multi-objective problem because its two objectives usually conflict with each other. However, most wrapper feature selection approaches used only classification performance as the fitness function or combined the classification accuracy and the number of selected features into a single fitness function [183]. In [179], two wrapper multi-objective PSO-based feature selection approaches were proposed, which applied the non-dominated sorting idea (NSPSOFS) and the crowding, mutation, and dominance concept (CMDPSOFS) to evolve a set of non-dominated feature subsets. The performance of the proposed algorithms was better than three other multi-objective algorithms, including

NSGAAII, SPEA2, and PAES. In addition, Xue et al. [178] also conducted a comparison between binary and continuous PSO for multi-objective feature selection problems, which indicated that generally continuous PSO achieved better performance than binary PSO.

Multi-objective PSO was also combined with filter measures to form multi-objective feature selection approaches. Xue et al. [178] proposed two multi-objective PSO-based feature selection algorithms, which simultaneously minimized the number of selected features and maximized the relevance of the selected feature subset. In these algorithms, the relevant measure was calculated by applying either pair-wise mutual information or information gain. The results illustrated that the proposed multi-objective algorithms outperformed single-objective algorithms. More multi-objective PSO-based filter feature selection approaches could be found in [180, 33, 131]

Hybridized algorithms were proposed to solve the feature selection problem. Meenachi and Ramakrishnan [114] proposed two meta-heuristic approaches to feature selection in the medical area by targeting cancer disease. Firstly, they hybridized Ant Colony Optimisation (ACO) and Tabu search (TS) algorithm by using Fuzzy Rough set to find optimal feature set. Secondly, they hybridized the Genetic Algorithm (GA) and Tabu search algorithm by using the Fuzzy Rough set to find the optimal feature set. Ant Colony Optimisation and Genetic Algorithm utilized as a global search to find the best feature set. Then, Tabu search is used as a local search to find the best feature. Their techniques are addressing global search approaches issue in finding the best local optimal and local search approaches issue in finding best global feature by hybridizing both of the approaches. These methods achieved better performance in comparison to other related works by selecting lesser features. However, Ant Colony Optimization-based approach is slower than the Genetic Algorithm-based method. In another work, Situmorang et al. [153] developed a hybrid PSO and GA feature selection approach by combining Logistic Regression and

SVM algorithms to classify coronary heart disease. They improved the results of classifiers by applying hybrid PSO and GA to remove redundant features.

2.2.4 Data Augmentation in Classification

Data augmentation in general classification

Data augmentation is a method to deal with data shortage issues in training models for distinct tasks such as classification. Data augmentation approaches [187, 103] have been suggested to deal with data shortage for training deep learning methods for time series or sequences. Yadav et al. used ordinary differential equations (ODEs) for generating time series to improve the performance of recurrent neural networks (RNNs) for anomaly detection [187]. Guennec et al. have recently suggested data augmentation by applying warping and window slicing in Convolutional Neural Networks (CNNs) for time series classification (TSC) [103]. Malhotra et al. have used a pre-trained encoder network on various time series with different lengths in a deep recurrent neural network for time series classification [111]. Wong et al. have investigated the benefit of data augmentation with artificially produced instances during training a classifier. They introduced data warping and synthetic over-sampling methods for making new samples by applying to handwritten digit dataset [175]. Salamon et al. have studied the effect of different data augmentation approaches on the performance of deep convolutional neural networks for environmental audio classification [141].

There are some usual methods such as augmenting by dictation errors, generating new documents by translating to French language and back into English language [197]. Other work has utilized adding textual noise [177] and using language models to predict synonym replacement [89]. These methods are not used often due to the high cost of implementation [171].

Zhang et al. [198] used data augmentation in Convolutional Neural Network (CNN) for classifying document by employing English thesaurus gained from WordNet. They replaced the words and phrases with their most similar synonyms in the context to generate new document. Targeting high similar synonyms to generate new documents can decrease the variety of the produced instances. Rosario [138] suggested an approach to data augmentation for short documents classification by making similar short documents for each original short document to produce a longer text by considering a semantic space. Quijas has studied the effect of data augmentation in training CNNs and RNNs for document classification [136]. Kobayashi has proposed the "contextual augmentation" method which generates matches of words by utilizing a bidirectional language model and replaces words with their matches in sentences [89]. Coulombe in [38] has suggested other textual data augmentation by using different approaches such as paraphrase generation. The approaches were examined on different neural network architectures. Jungiewicz [77] has introduced a method to textual data augmentation for training CNN models for sentence classification. The researcher converted sentences by maintaining their lengths the same as their original lengths. The researcher has used a thesaurus that belongs to Princeton University's WordNet. However, these methods are expensive to implement.

Data augmentation in medical classification

Wei and Zou proposed four different approaches including Synonym Replacement (SR), Random Insertion (RI), Random Swap (RS), and Random Deletion (RD). SR is utilizing synonyms of words by randomly choosing n words from a sentence and replacing them with one of their synonyms randomly [171]. RI method chooses a random synonym for a random word in each sentence and puts that synonym in a random place in the sentence. RS approach chooses randomly two words in a sentence and swaps their places. RD method removes a word in the sentence with probability

p. As these methods target only some words of each document to replace, delete or change, they may not produce new instances with enough variety. SR method is similar to the random synonym-selection method and, as the order of words and sentences can have semantic meaning in medical discharge notes, RI, RS, and RD methods can change the explanation of a patient's health condition. In this paper, we extract synonyms from WordNet dictionary and employ similarities of the extracted synonyms with their corresponding words as a feature extraction approach to detect meaningful synonyms for oversampling the minority class to produce new documents with high diversity.

Borrajo et al. [24] investigated the effects of oversampling and subsampling by using dictionaries on imbalanced scientific document classification problems in the bio-medicine domain. They employed three different classifiers (K-nearest neighbors (KNN), support vector machine (SVM), and Naive-Bayes) to check the performance of classification. Precision, Recall, F-measure, and Utility metrics are utilized for evaluation. The SVM classifier is achieved the best results by using the sub-sampling approach and considering NLPBA, Protein dictionaries.

Ollagnier and Williams [125] proposed two approaches for augmenting synthetic documents for clinical case coding tasks (CLEF eHealth 2020) to deal with the class-imbalanced issues. They suggested a word-level text transformation and a text generation approach to augment new labeled data. A Convolutional Neural Network and Long Short-Term Memory Network (CNN-LSTM) classifier are used for training the suggested approaches. Multilingual BERT (M-BERT) is used to represent the cases. The data augmentation approach is employed a WordNet dictionary [171] to replace randomly selected 10% of each document's words with their synonyms. In the text generation method, a pre-trained model is made by considering n-gram distribution probabilities by using the input corpus. Then the pre-trained language model is used to generate new synthetic data. 30% of each document is selected to replace with the new augmented

data. Experimental results showed the data augmentation model achieved the best f1-score results in comparison to the other proposed approaches.

Kang et al. [79] proposed a UMLS-based approach to augment new instances on biomedical research literature data sets for improving deep learning models' performance. They developed an easy data augmentation (EDA) technique for named entity recognition (NER) and sentence classification tasks. They replaced synonyms with extracted concepts from UMLS ontology. The performance of the two approaches is compared in this work. Firstly, the base model of EDA by using biLSTM-CRF from BlueBERT is used to check the effectiveness of transfer learning. Secondly, the base model of EDA is used with the proposed data augmentation (UMLS-EDA) approach to deal with NER and sentence classification tasks. The UMLS-EDA achieved the best results in comparison with the transfer learning approach. Furthermore, they combined these two approaches to achieve better results. However, it decreased the results performance. They concluded that it is not a good idea to use data augmentation besides transfer learning as the used EDA approach makes noise for the used transfer learning method. Additionally, data augmentation is a better choice to have in lack of high-level infrastructure.

Guan et al. [61] developed a Generative Adversarial Network (GAN) model to generate synthetic electronic clinical notes. The model is trained by the reinforcement technique. The input of their model is tags of diseases and the output is generated medical document. Their goal is to protect the privacy of patients by making more real synthetic notes with diversity. Two deep learning models including Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) used as generative models separately. A binary classification task is designed to discriminate the real electronic medical document from the constructed synthetic document. One of the disadvantages of the method is that the relationship dependency for the long-term is complicated if document length is too extensive.

2.3 Chapter Summary

Some of the limitations in the existing work are illustrated here which become the motivation of this thesis.

- The extraction of text-based and document features for document classification problems has received a lot of interest in the past. Typically, a lot of these algorithms rank features using statistics from the distribution of features in the given corpus [149, 145]. Existing methods have employed metrics associated with word frequency, information gain, mutual information, term frequency-inverse document frequency (TF-IDF) for extracting textual features [88]. However, most of the existing techniques tend to treat each feature separately, i.e they ignore the dependencies between features. Developing a domain-specific approach to map free text to concepts from an ontology that encodes semantic relationships between concepts and improves the classification performance.

- The majority of previous researches on document classification utilizes only one method to carry out feature selection and has problems because of the extremely large search space [17]. Specifically, given an extremely large feature set, a single feature selection method such as PSO can still result in a large number of selected features, which limits the effectiveness of the feature selection. To overcome this drawback and improve the effectiveness of feature selection in a very high dimensional feature space, developing a two-stage method to extract and select meaningful features for document classification can improve the performance of classification.

- In medical document classification, there are thousands of features and often there are redundant and irrelevant features that 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 [17] and feature construction [182] 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 large number of redundant features [17]. To solve this issue and improve the performance of document classification, a three-stage method by using discriminative knowledge-guided medical concept pairings from clinical notes for constructing new high-level features can improve the classification performance.

- One of the important factors which have an effect on the classification accuracy is the size of the data set for training the model. Generally, there is a lack of adequate data in the medical area [143]. One possible solution to address the issue is to augment data for training the model. Replacing words and expressions with their synonyms is a common approach in data augmentation [198]. However, these methods are using normal dictionaries for augmentation and some domain-specific terms or acronyms do not have synonyms in normal dictionaries. Hence, developing an ontology-based method to augment new instances by targeting concepts of words and expressions in the documents can improve classification performance.

- Normally due to privacy reasons, it takes a long time to collect a set of documents for a special disease. This type of data which is collected from real patients often appears as imbalanced data where the patients with a particular disease are often the minority. As a result, the shortage of data and being extremely imbalanced are two common issues in medical discharge notes. In cases where the classifier is a data-hungry model and needs to be fed with a large amount of data, this kind of data will not be enough. Consequently, this can make the learning difficult for the candidate classifier [96]. In these scenarios, the learned model will be biased on the majority class.

Most of the existing work replaced the words and phrases with their most similar synonyms in the context to generate new document. Targeting high similar synonyms to generate new documents can decrease the variety of the produced instances. Furthermore, some methods target only some words of each document to replace, delete or change which may not produce new instances with enough variety. Having a new probabilistic dictionary-based data augmentation approach can address these issues by oversampling on the minority class and improve classification performance.

Chapter 3

Feature Manipulation for Medical Document Classification

3.1 Introduction

In medical document classification, there are a large number of features and often there are redundant and irrelevant features that 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 [17] and feature construction to improve the quality of features by removing irrelevant and noisy features.

There has been research that applies text classification to clinical text. There exist a number of studies in the literature where ontology-based classification approaches have been applied [29, 45]. The use of ontologies like Unified Medical Language System (UMLS), Systematized Nomenclature of Medicine (SNOMED), and Medical Subject Headings (MeSH) have been proved to be very useful for improving classification performance [57, 149, 27, 150]. Most of the existing methods are extracting all of the features in IE step which can be irrelevant. 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. In addition, some work has used clinical records for prominent tasks such as finding risk factors for diabetic patients [86], extracting Framingham risk score (FRF) for target population [75], using rule-based and dictionary-based methods to identify heart disease risk factors [190], and applying a rule-based method by combining with a regular expression and UMLS to spot risk of heart disease [152]. Furthermore, other works are using the rule-based methods [167, 155, 192] to filter unnecessary information, however, these methods need expert people to define rules for each problem separately which can be costly, and the defined methods are not usable to another problem [133].

By analyzing the previous work, it is noticeable that the majority of disease-targeted systems have tended to develop static rule-based systems which require human interventions every time the model is updated with new features [150]. Such systems are not scalable for practical machine learning purposes. Our systems allow an easier and flexible selection of different types of medical concepts to enable automatic extraction of features and the generation of a prediction model. This is an easier way to investigate domain concepts and determine which concepts are most discriminative to a classification problem. Furthermore, our systems provide the ability to allow a domain expert to interactively change the concepts and auto-build machine learning models for different diseases investigation.

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 precisely and comprehensively 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. More-

over, various combinations of domain-specific medical events in a clinical report can describe a patient's conditions totally differently. Generally, all of the text classification approaches for general text are similar to medical text classification. One of the limitations of these methods is that they are not using domain-specific clinical resources such as Unified Medical Language System (UMLS) to be able to serve additional features [133]. Furthermore, the accuracy of medical document classification methods still is poor and needs more investigation. Hence, an ontology-based approach for feature extraction, a two-stage wrapper approach for feature selection, and an ontology-based approach for feature construction will be explored. This chapter will introduce the three different classification approaches for medical document classification.

3.1.1 Chapter Objectives

This chapter develops a new ontology-based method that applies ontology by referring to Unified Medical Language System(UMLS) for entity recognition. This method focuses on concepts of words and expressions for extracting meaningful features. The method targets two concepts "Disease or Syndrome" and "Sign or Symptom" instead of extracting all of the concepts can increase the feature space dimension. Furthermore, a two-stage method is developed to extract and select meaningful features for medical document classification. This method is utilizing PSO for the feature selection step. Additionally, a discriminative knowledge-guided medical concept pairings (diseases and symptoms) approach is developed for feature construction from the original set of features for classifying clinical discharge notes. This chapter explores the following research objectives to improve medical document classification accuracy:

- Design a new ontology-based method to extract meaningful features from raw data by considering the concepts of words and expressions regarding data set labels.

- Design a two-stage method to extract and select meaningful features by utilizing PSO for feature selection for medical document classification.
- Design a new discriminative knowledge-guided medical concept pairings approach for feature construction from the original set of features by considering diseases and symptoms concepts.

3.1.2 Chapter Organization

The rest of the chapter is organized as follows: Section 3.2 introduces the proposed method to discover discriminative knowledge-guided medical concepts for classifying medical documents. Section 3.3 presents an Ontology-based two-Stage approach to medical document classification with feature selection by Particle Swarm Optimisation (PSO). Section 3.4 describes an ontology-based feature construction approach for document classification by using discriminative knowledge-guided medical concept pairings from clinical notes. Section 3.5 provides the experiment design, classification methods, data sets, and parameter settings for comparison. Section 3.6 presents the results and discussion. The achievements of the three approaches and their limitations are summarised in section 3.7.

3.2 Discriminative knowledge-guided concepts

In this section, the proposed method for extracting concepts of phrases is described in detail. Additionally, our way of labeling the candidate data set is introduced.

3.2.1 Overview

One of the important issues in text classification problems is to investigate the domain of documents that should be classified and the domain

of classes that documents should be labeled with. This can help to select only the related features of the documents to the domain for the training phase and improve the accuracy of prediction for unseen documents. In the clinic text classification task, the documents are discharge notes of patients in the medical domain. The candidate class is whether a disease such as Coronary Artery Disease (CAD) is present (1) or not (0). Our goal is to select features that have relations with the disease. In this case, the performance of the learned model can be improved.

To achieve the above goal, the proposed algorithm employs the knowledge in candidate data sets and the UMLS library by targeting two specific concepts (diseases and symptoms). For this purpose, the MetaMap tool is used to extract all the concepts of existing phrases for each document using the UMLS. As shown in Fig. 3.1, the concept extraction step is employed on both the training and the test documents. Then, by considering the medical domain, the concept selection step is performed on the obtained concepts. As the first step, two concepts are selected among all the concepts: ”Disease or Syndrome” and ”Sign or Symptom”. By following this way of concept selection, the meaningful concepts will be selected which will assist the training phase to learn better in order to increase the accuracy of classification.

3.2.2 Unified Medical Language System (UMLS)

The Unified Medical Language System (UMLS) [1] was introduced for modeling the language of health and biomedicine. UMLS is a source of knowledge that improves the performance of information systems in the biomedical area. It provides three main resources: the Metathesaurus, the Semantic Network, and the SPECIALIST Lexicon. The largest component of UMLS is the Metathesaurus. It gives services such as finding biomedical concepts and relationships between concepts (e.g. SNOMED-CT, Mesh, etc.). The Semantic Network includes a collection of extensive

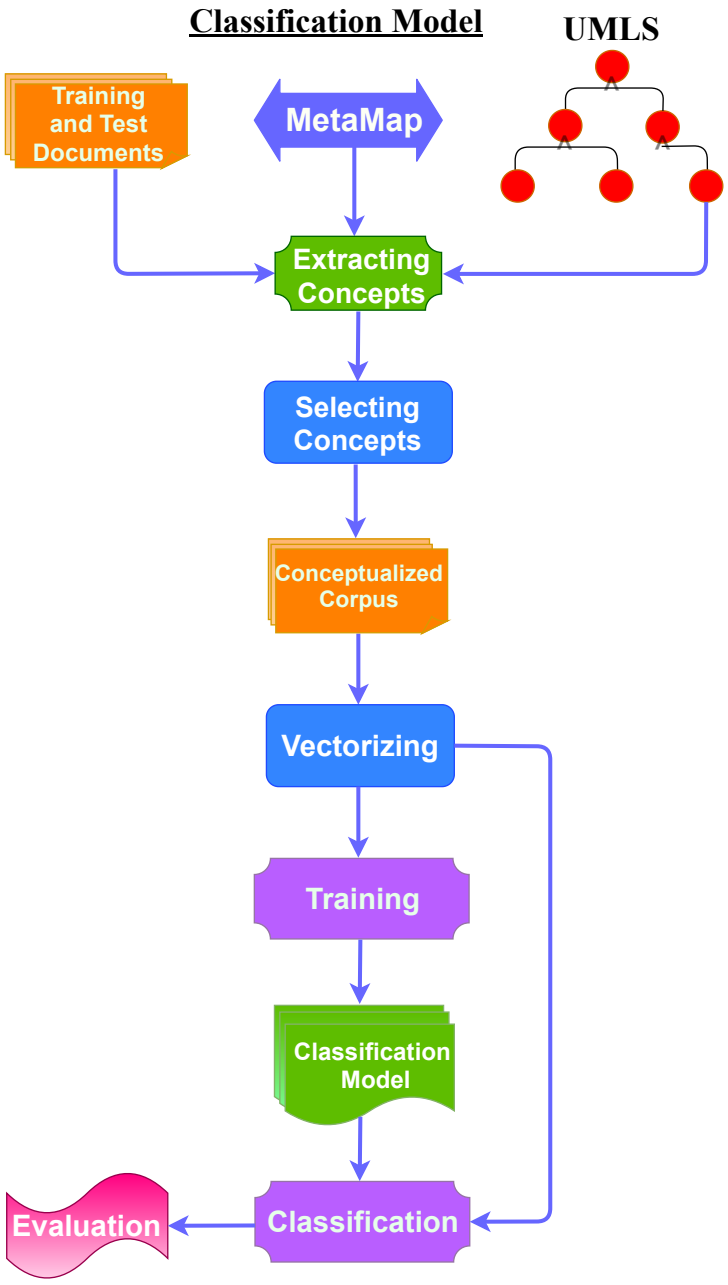

Figure 3.1: The flowchart of the architecture of using MetaMap and UMLS for text classification.

topic classes, and different types of Semantics, which cover a matchable classification of concepts provided in the UMLS Metathesaurus, and a category of relationships and Semantic Relations between Semantic Types. The SPECIALIST lexicon includes a specialised English vocabulary of biomedical words.

3.2.3 MetaMap Tool

MetaMap [15] is a configurable program created by Dr. Alan (Lan) Aronson at the National Library of Medicine (NLM). It maps text to the UMLS Metathesaurus to find Metathesaurus concepts in the text. MetaMap is a knowledge concentrated approach that utilizes computational-linguistic, natural-language processing (NLP), and symbolic methods. MetaMap is applied broadly in Information Retrieval (IR), and data mining applications. Furthermore, it is utilized for automatically biomedical literature indexing in the U.S. National Library of Medicine (NLM). It allows the mapping between text content and related concepts in UMLS. To achieve this goal, MetaMap breaks the content into expressions and after that, for each expression, it selects the mapping alternatives based on the ranking of mapping quality.

3.2.4 Conceptualization

Two sentences are given below as a sample to show how MetaMap works on the input notes and what output it provides in the 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."*

Fig. 3.2 shows a segment of the returned results from MetaMap. Table 3.1 summarizes the extracted concepts of detected meaningful phrases from the sample sentences using MetaMap. As can be observed, the phrase

"hyperlipidemia" belongs to "[Disease or Syndrome]" and "[Finding]" concepts. The phrase "shortest of breath" is allocated to the "[Sign or Symptom]", "[Clinical Attribute]", and "[Intellectual Product]" concepts. Considering the medical domain and the type of the classes in the selected data set, we choose concepts that appear in the "[Disease or Syndrome]" or "[Sign or Symptom]" categories. First, we identify these two categories which are in square brackets, then the phrase that is within the round parentheses at the same line will be extracted as the main phrase. For example, the phrase "Dyspnea" is extracted in line 19 of Fig. 3.2 for the phrase "shortness of breath". After finishing the concept selection step, the obtained phrases will be used instead of the original documents in the binary classification problem. In order to give weights to the extracted terms of the documents, TF-IDF is applied in the vectorization step and each document is represented as a vector of weights based on the TF-IDF function. Then, the weighted term can be used in training and test phases.

```
   ---------------------------------------------------------------------------------------------------
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
   ---------------------------------------------------------------------------------------------------
```

Figure 3.2: A segment of returned results of extracted concepts using MetaMap.

Table 3.1: 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]	×

3.3 Feature Selection by PSO

In this section, the proposed two-stage algorithm for extracting concepts of phrases is described in detail. Fig. 3.3 shows the flowchart of the proposed two-stage method.

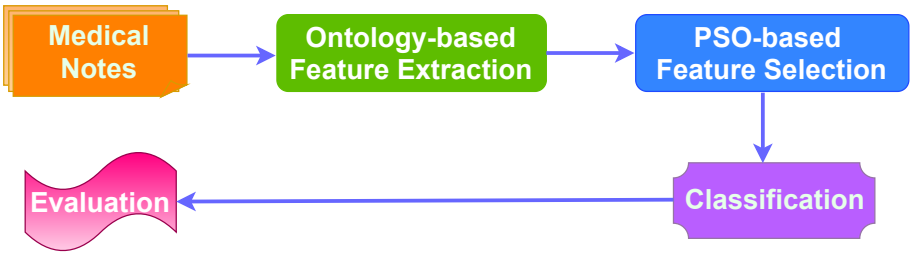

Figure 3.3: The proposed two-stage method

The input of the proposed approach is a set of clinic texts. Firstly, the approach detects all of the meaningful expressions in the documents and then applies the MetaMap tool to extract their concepts from the UMLS.

After deleting redundant features in the first step, PSO is employed to select a feature subset from the features extracted in the first stage. The output is a classifier along with the selected features that predict the label of a text. The first stage reduces the size of the search space for PSO and assists it to better search. It is expected that the suggested method extracts meaningful features and selects a more informative subset of them and maintains or enhances the classification performance.

The main idea for introducing the mentioned two-stage algorithm is to take the advantage of information extraction approaches to extract domain-specific features, and then to select the features by considering feature interactions and related features to the problem domain. In the first stage, all of the features are extracted as described in section 3.2.

3.3.1 PSO-based Algorithm in the Second Stage

In this stage, PSO is employed to further eliminate the irrelevant and unnecessary features from the extracted features in the first stage. The value of particles is initialized randomly by numbers in [-1, 1]. Each particle in PSO corresponds to a feature subset and is coded as a vector. For example, a positive number indicates the corresponding feature is selected and a negative number means the feature is not selected. The dimension of a vector is d and the vector consists of real numbers. In other words, d represents the dimension of the search space which is equal to the number of the primary features obtained by the first step. A random value is initialized for the position and velocity of each particle. Next, PSO moves particles by updating their $pbest$ (best position has been found so far) and $gbest$ (the best position). Toward the end of the process, $gbest$ is obtained based on particles' fitness value and the gained best particle will be investigated to achieve the selected feature subset. Algorithm 3.1 presents the pseudo-code for PSO for feature selection in the second stage. During the algorithm (line 5), the fitness value of each particle is evaluated

based on the medical document classification F1-measure. Our approach is a wrapped-based method. Hence, a classifier is employed to run with the selected features to evaluate the value of the fitness function.

Algorithm 3.1: PSO to select best feature subset

Input : Training instances
Output: The best feature subset ($gbest$)

```
Keep only the features that are extracted in the first stage;
Randomly initialize the position and velocity of particles;
gen ← 0
while gen < maxGen do
    Evaluation: Evaluate fitness of particles based on classification accuracy on
      the validation set;
    for i = 1 to |Particle| do
        Update pbest and gbest for particle i;
    for i = 1 to |Particle| do
        for d = 1 to dimension do
            Update the velocity of particle i
            Update the position of particle i
    gen ← gen + 1
return the position of gbest;
```

Fig. 3.4 shows the flowchart of how we calculate the fitness function value for each particle. All the training data is entered as input for PSO to do the feature selection. 10-fold cross-validation is used to compute a particle's fitness value. The training data is divided into 10 subsets. Nine training subsets are used as input for PSO and one validation subset is used for calculating the fitness of each particle. The average of calculated 10 classification accuracies will be the fitness value of a particle. Note that the test data set is not used in this PSO feature selection process. The test set is only used in the final evaluation where the final classification accuracy is calculated for the selected best feature subsets. The used approach is a wrapper method that utilizing five different classifiers in-

cluding Logistic Regression (LR), Linear Support Vector Machine (LSVM), Naive Bayes (NB), Decision Tree (DT), and K-Nearest Neighbor (KNN) separately. Regarding the time efficiency of the algorithm, PSO takes more time to select the best feature subset in the suggested method, but it takes the same amount of time in the testing step.

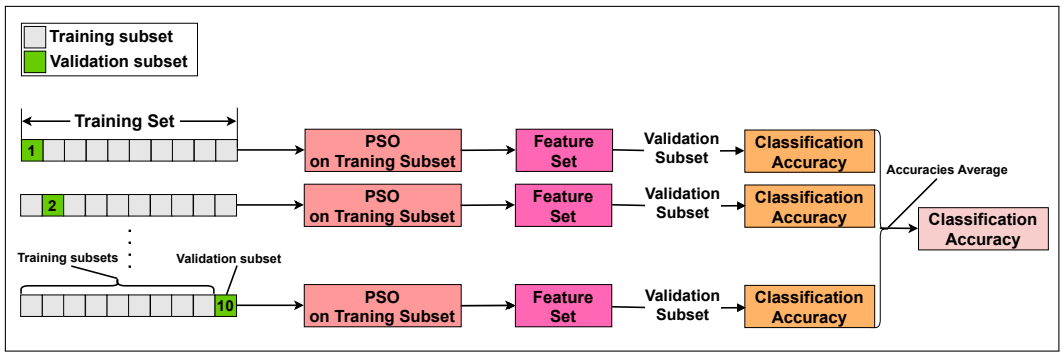

Figure 3.4: PSO for feature selection using 10 fold cross validation

3.4 knowledge-guided medical concept pairings

In this section, the developed three-stage algorithm for extracting concepts of phrases and constructing new features is described in detail. Fig. 3.5 presents the flowchart of the proposed three-stage method. The method will construct new features by pairing medical concepts.

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 [15] 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.

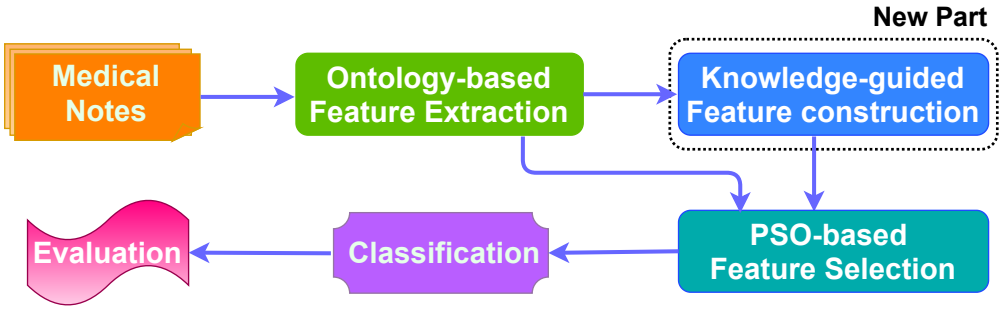

Figure 3.5: The proposed three-stage method

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

3.4.1 Feature construction method

All of the features are extracted based on the described approach in section 3.2. A paragraph is given below as an example to describe how MetaMap works on the input discharge notes and what output it provides in the classification process. Below is an example of raw clinical notes. Table 3.2 shows the detected phrases based on their concepts.

> *"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."*

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 features. Table 3.3 shows the constructed features for the extracted features from the sample sentences.

Table 3.2: The extracted concepts of the example notes 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]**	✓

Table 3.3: 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 3.4, 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 enriches 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.

3.5 Experimental design

3.5.1 Dataset and preprocessing

The performance of the proposed approaches is evaluated on two data sets, the 2010 Informatics for Integrating Biology and the Bedside (i2b2) and the 2008 Informatics for Integrating Biology and the Bedside (i2b2) [58] data sets. The labels of the 2010 i2b2 [168] data set are CAD (Coronary Artery Disease) and non-CAD that form a binary classification problem. The data set includes 426 documents which are 170 documents for training and 256 documents for testing. This is the first time that the data set is used for a text classification problem. This approach focuses on binary classification, so all the documents are labeled based on whether or not Coronary Artery Disease (CAD) is present. Each document in the original data set has three files consisting of "Concepts.con", "Relations.rel", and "Assertions.ast" which were provided by the i2b2 organization for Relations Challenge. We used the content of "Assertions.ast" file of each document to determine its label. As shown in Fig. 3.6, there are a number of problem names inside each Assertion file. To label all of the documents, at the first step, all the lines of the file are searched for the "Coronary Artery Disease" phrase. If the phrase is found by the search, the second step will

Table 3.4: 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)

be checking whether the disease is present or not. If the name of illness appears with the phrase "present" in the same line, we will consider that the document is in the CAD class. The first line of Fig. 3.6 indicates a sample of both the phrases "Coronary Artery Disease" and "present" occurring in the same line. By following this rule, all of the labels of 170 training documents and 256 test documents are extracted.

```
---------------------------------------------------------------
Assertions

c="coronary artery disease" 24:8 24:10||t="problem"||a="present"
c="myocardial infarction" 24:19 24:20||t="problem"||a="absent"
c="cyanosis" 44:6 44:6||t="problem"||a="absent"
c="hypertension" 24:0 24:0||t="problem"||a="present"
        .
        .
        .
c="chest pain" 47:11 47:12||t="problem"||a="present"
---------------------------------------------------------------
```

Figure 3.6: A subpart of the Assertions file.

Table 3.5 presents the labels of i2b2(2008) [58]: Present, Absent and Questionable. The data set has sixteen tasks, in which six of the tasks have not Questionable labels and the size of instances for Questionable label is very small for the other tasks. Hence, we filtered this label.

In the first approach, 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, the extracted features are used in training the classifiers. In the second approach, PSO is applied to the extracted features from the first approach to eliminate redundant features. In the third approach, all of the possible pairs of obtained features from the first approach are constructed for the output of each document separately and added to the extracted features of each document. Next, the following preprocessing steps are applied to the obtained results of the feature extraction step:

- Hold only words and delete punctuation, numbers, etc. Convert all

Table 3.5: Distribution of Intuitive Judgments into Training and Test Sets

Classes	Present		Absent		Questionable		Total	
Diseases	Training	Test	Training	Test	Training	Test	Training	Test
Asthma	93	68	596	403	0	0	682	471
CAD	391	272	265	185	5	1	661	458
CHF	308	205	318	229	1	4	627	438
Depression	142	105	555	372	0	0	697	477
Diabetes	473	333	205	146	5	0	683	479
Gallstones	101	80	609	411	0	0	710	491
GERD	144	93	447	331	1	2	592	426
Gout	94	61	616	439	2	0	712	500
Hypercholesterolemia	315	242	287	189	1	0	603	431
Hypertension	511	358	127	88	0	0	638	446
Hypertriglyceridemia	37	25	665	461	0	0	702	486
Obesity	285	192	379	255	1	0	665	447
OSA	99	66	606	427	8	2	713	495
OA	117	91	554	367	1	4	672	462
PVD	110	65	556	399	1	1	667	465
Venous Insufficiency	54	29	577	398	0	0	631	427
Total	3267	2285	7362	5100	26	14	10655	7399

CAD = coronary artery disease; CHF = congestive heart failure; DM = diabetes mellitus; GERD = gastroesophageal reflux disease; HTN = hypertension; OSA = obstructive sleep apnea; OA = osteo arthritis; PVD = peripheral vascular disease.

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.

The 2010 i2b2 data set [168] includes 426 documents with 7554 unique terms. Table 3.4 shows the total number of features for each case after applying the first and second stages (check the last column). Table 3.8 shows the total number of features for each task of the 2008 i2b2 data set [58] after applying the first and second stages (check the fourth column).

3.5.2 Parameter Settings

The size of used datasets is small due to privacy issues in the medical area. Hence, 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 separately. These methods can be a good choice when the dataset is not large. The classification F1-measure is calculated on the testing documents to evaluate the performance of the classifiers.

Some of the classifiers' parameters are tuned to get better results. The inverse of regularization strength ("C") is adjusted to 10 in the Logistic Regression. The number of the neighbors is set to the value 28 in KNN. The maximum depth of the tree and the random number generator is adjusted to values 14 and 11 in the Decision Tree classifier, respectively. Furthermore, an early stopping rule is chosen to avoid overfitting in training SVM and Logistic Regression classifiers. In the SVM classifier, the strength of the regularization ("C") is adjusted to 5e5. The Kernel value is set to 'rbf'. The kernel coefficient ("gamma") and the tolerance for stopping criterion

Table 3.6: PSO parameter setting

PSO Parameters	**Value**
Population Size	30
Maximum Number of Iteration	100
Dimension of Original+PSO [3]	7554
Dimension of UMLS+PSO [3]	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 (θ)	0
Acceleration Coefficients	2.0
Run Times	30

("tol") parameters are set to 1e-06 and 1e-05, respectively. The rest of the classifiers' parameters are kept the same as default values.

Table 3.6 presents the set parameters of PSO which are proposed in [17]. The values for particles are initialised using numbers in [-1, 1], and the threshold (θ) is set to zero, hence, about 50% of the features will be selected. Some documents will disappear if less than 50% of features are selected.

3.5.3 Evaluation criteria

As the used data sets in this chapter are imbalanced, the macro F1-measure metric is used as it is common for analyzing imbalanced data sets. The performance of each classifier is calculated by evaluating the macro F1-

measure metric for all of the used methods:

$$Precision = \frac{TP}{(TP + FP)} \tag{3.1}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{3.2}$$

$$F1\ measure = \frac{1}{N} \sum_{i=1}^{N} 2 * \frac{(precision * recall)}{(precision + recall)} \tag{3.3}$$

where N presents the number of classes, TP (True Positive) is the number of correctly identified documents, FP (False Positive) is the number of incorrectly identified documents and FN (False Negative) is the number of incorrectly rejected documents. Our approach is a wrapped-based method. Hence, a classifier is employed to run with PSO to evaluate the value of fitness function parameters (TP, FP, and FN).

The evaluation in the medical document classification context is the same as the general evaluation criteria. If the suggested features manipulation approaches improve the classification F1-measures, it means the extracted, selected and, constructed features are meaningful and lead to better classification performance.

3.6 Results, discussion and further analysis

3.6.1 Results

Five different classifiers are employed to assess the proposed approaches on the i2b2 (2010) data set [168]. As i2b2 (2008) data set [58] is bigger and there are 16 tasks, the best feature construction method is selected to apply to the i2b2 (2008) data set [58]. The suggested approaches are applied to the training set using 30 independent PSO runs.

Discussion based on feature size

Our three-stage approach has seven cases (case1 to case7) and they use different pair combinations shown in Table 3.4. All of the cases of different features are shown in Table 3.7. It 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 the MetaMap tool. The "$UMLS + Pairs$" method is utilizing the detected features from UMLS and the constructed pairs of features. "$Original + PSO$" [3], "$UMLS + PSO$" [3] and "$UMLS + Pairs + PSO$" are stochastic methods by applying PSO to select a feature subset. The smallest feature subset belongs to the "$UMLS + PSO$" method which contains only 10.33% of the original features. Among seven cases of different pairing, 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 the feature size of the stochastic method ($UMLS+PSO$) is approximately 50% smaller than the deterministic method.

Table 3.8 shows the statistical results related to the number of features of each task in the i2b2 (2008) data set [58] for the methods. The lowest numbers of the selected features belong to the two-stage approach which is approximately 3.5% of the original number of features. The highest numbers of the features are selected by the three-stage approach which is almost 90% of the original number of features for each task.

Table 3.7: Number of Selected Features for the i2b2(2010)

#	Classifiers / Cases	NB	LSVM	KNN	DT	LR
		Ave±Std				
1	Original (100%) [2]	7554	7554	7554	7554	7554
2	UMLS [2]	780	780	780	780	780
3	Original+PSO [3]	3779.35±38.01	3768.75±48.22	3774.13±39.36	3775.25±43.04	3767.65±32.77
4	UMLS+PSO [3]	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

Table 3.8: Number of Selected Features for the i2b2 (2008) data set

Methods / Diseases	Original Deterministic	UMLS Deterministic	UMLS+Pairs Deterministic	Original+PSO Stochastic	Two-Stage Stochastic	Three-Stage Stochastic
				Ave(%)±Std		
Asthma	21140 (100%)	1456 (6.89%)	38942 (184.21%)	10538(49.85%)±67.48	701(3.32%)±21.06	19427(91.90%)±87.25
CAD	20803 (100%)	1453 (6.98%)	37948 (182.42%)	10389(49.94%)±68.07	723(3.48%)±17.82	18950(91.09%)±91.51
CHF	20339 (100%)	1430 (7.03%)	35959 (176.80%)	10142(49.86%)±47.24	712(3.50%)±15.47	17981(88.41%)±65.83
Depression	21629 (100%)	1476 (6.82%)	39415 (182.23%)	10808(49.97%)±77.47	703(3.25%)±16.56	19718(91.16%)±97.77
Diabetes	21274 (100%)	1463 (6.88%)	38698 (181.90%)	10639(50.01%)±83.64	721(3.39%)±18.70	19347(90.94%)±122.03
Gallstones	21927 (100%)	1480 (6.75%)	39345 (179.44%)	10964(50.00%)±69.20	725(3.31%)±18.39	19666(89.69%)±90.39
GERD	19344 (100%)	1377 (7.12%)	33743 (174.44%)	9660(49.94%)±56.76	684(3.54%)±18.55	16861(87.16%)±99.74
Gout	21884 (100%)	1485 (6.79%)	39862 (182.15%)	10910(49.85%)±81.70	726(3.32%)±15.79	19899(90.93%)±109.15
Hypercholesterolemia	19756 (100%)	1421 (7.19%)	35995 (182.20%)	9848(49.85%)±55.22	695(3.52%)±19.55	17989(91.06%)±86.89
Hypertension	20610 (100%)	1436 (6.97%)	37317 (181.06%)	10293(49.94%)±75.85	697(3.38%)±19.27	18641(90.45%)±92.38
Hypertriglyceridemia	21736 (100%)	1482 (6.82%)	39846 (183.32%)	10864(49.98%)±74.48	731(3.36%)±19.41	19883(91.47%)±83.87
Obesity	20941 (100%)	1444 (6.90%)	37831 (180.66%)	10456(49.93%)±86.10	701(3.35%)±18.55	18854(90.03%)±108.13
OSA	21838 (100%)	1473 (6.75%)	39715 (181.86%)	10889(49.86%)±63.58	724(3.32%)±15.24	19845(90.87%)±94.89
OA	21096 (100%)	1452 (6.88%)	38065 (180.44%)	10519(49.86%)±84.88	721(3.42%)±15.52	19021(90.16%)±118.57
PVD	21033 (100%)	1442 (6.86%)	38247 (181.84%)	10504(49.94%)±66.84	711(3.38%)±15.13	19093(90.78%)±86.39
Venous Insufficiency	20223 (100%)	1426 (7.05%)	36072 (178.37%)	10100(49.94%)±61.93	707(3.50%)±16.97	18022(89.12%)±114.88

Discussion based on classification performance

Table 3.9 compares the statistical results of the deterministic and stochastic versions of the proposed feature construction approaches with the pairs. The best results are highlighted and the three-stage method (with PSO) shows better performance than the three-stage method (without PSO) [3] in Naive Bayes, SVM, KNN, and Logistic Regression classifiers. A Friedman test is applied to rank the seven feature construction methods' performance based on classifiers. Therefore, 35 different combinations of methods (seven) and classifiers (five) are fed to the Friedman test. Case 2 with the SVM classifier is selected as the best combination. Hence, the case 2 feature construction approach by using the SVM classifier is applied on i2b2 (2008) [58] for comparison with other approaches.

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 F1-measures of the 30 selected feature subsets. Table 3.10 compares the statistical results for six approaches. The standard deviation and average of F1-measures are calculated for all of the classifiers and the Wilcoxon signed ranks test [43] with a significance level of 0.05 is used to test whether the suggested approach has made a significant difference in classification accuracy. In Table 3.10, the "T" column presents the significance test of the proposed approach against the other approaches in the previous column, 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.

From Table 3.10, it can be concluded that the proposed three-stage approach has achieved considerably higher classification F1-measure than other methods for Naive Bayes, SVM, and Logistic Regression classifiers. The UMLS method [2] shows better performance with the KNN classifier by using only 780 features which is 10.33% of the original features' number. The three-Stage approach gains significantly better classification

Table 3.9: F1-measure of Classifiers on i2b2 (2010) data set [168] for the Seven Cases without PSO and with PSO

Classifiers	NB		SVM		KNN		DT		LR	
Cases	F1-measure (%) Ave±Std									
	Without PSO	With PSO	Without PSO	With PSO	Without PSO	With PSO	Without PSO	With PSO	Without PSO	With PSO
Case 1	44.76±0.00	**72.32±0.04**	86.57±0.00	**89.81±0.01**	84.84±0.00	**89.76±0.01**	**84.87±0.00**	82.40±0.03	86.90±0.00	**88.65±0.01**
Case 2	46.90±0.00	**72.11±0.04**	86.04±0.00	**90.02±0.01**	88.20±0.00	**89.49±0.01**	**87.14±0.00**	83.42±0.03	85.15±0.00	**89.08±0.01**
Case 3	46.10±0.00	**71.97±0.04**	87.15±0.00	**89.76±0.01**	87.15±0.00	**90.02±0.01**	**85.57±0.00**	82.52±0.03	86.04±0.00	**88.50±0.01**
Case 4	44.76±0.00	**71.21±0.04**	86.90±0.00	**89.80±0.01**	87.75±0.00	**89.57±0.01**	**86.17±0.00**	82.18±0.03	84.25±0.00	**88.90±0.01**
Case 5	44.76±0.00	**72.01±0.04**	87.75±0.00	**89.48±0.01**	83.33±0.00	**89.77±0.01**	**80.40±0.00**	83.53±0.03	84.25±0.00	**88.23±0.01**
Case 6	46.90±0.00	**71.44±0.03**	86.04±0.00	**89.72±0.01**	86.04±0.00	**90.02±0.01**	**85.56±0.00**	81.64±0.04	82.38±0.00	**88.59±0.01**
Case 7	44.76±0.00	**71.85±0.03**	86.04±0.00	**89.96±0.01**	87.98±0.00	**89.71±0.01**	**79.99±0.00**	83.19±0.03	82.97±0.00	**88.63±0.01**

Table 3.10: Comparison of classification F1-measure and standard deviation averages using 30 independent runs on the i2b2 (2010) data set [168]. The highlighted entries are significantly better (Wilcoxon Test, $\alpha = 0.05$)

Methods	Original [2] Deterministic	UMLS [2] Deterministic		UMLS+Pairs [4] Deterministic		Original+PSO [3] Stochastic			Two-Stage [3] Stochastic			Three-Stage [4] Stochastic		
Classifiers	F1-measure	F1-measure	T	F1-measure	T	F1-measure Ave±Std	F1-measure Best(Lowest)	T	F1-measure Ave±Std	F1-measure Best(Lowest)	T	F1-measure Ave±Std	F1-measure Best(Lowest)	T
NB	49.78	60.86	+	46.90	- -	49.22±0.218	54.41(45.41)	= - +	60.66±0.063	79.61(50.99)	+ = + +	**72.11±0.039**	78.27(62.03)	+ + + + +
SVM	74.14	89.99	+	86.04	+ -	75.06±0.013	76.77(71.99)	+ - -	81.92±0.110	90.79(50.13)	+ - - +	**90.02±0.005**	91.57(89.37)	+ = + + +
KNN	65.13	**91.71**	+	86.04	+ -	74.00±0.042	80.91(62.44)	+ - -	76.83±0.175	91.85(44.44)	+ - - +	89.49±0.014	91.85(85.37)	+ - + + +
DT	74.84	82.57	+	**85.56**	+ +	82.91±0.040	88.25(72.88)	+ = -	81.96±0.034	89.30(74.18)	+ - - -	83.42±0.030	89.73(72.77)	+ + - + +
LR	72.36	87.26	+	82.38	+ -	74.88±0.016	78.32(71.33)	+ - -	85.95±0.044	90.62(75.30)	+ - + +	**89.08±0.013**	91.42(86.03)	+ + + + +

F1-measure in most of the cases.

The statistical results of i2b2 (2008) [58] are presented in Table 3.11 for the proposed approaches. Wilcoxon test is used to compare the approaches with $\alpha = 0.05$ in Table 3.11. The highlighted entries are significantly better with $\alpha < 0.05$. The $Three - Stage$ approach is outperformed the other methods in five tasks (Asthma, CHF, Gallstones, Obesity, and OA). The $Two - Stage$ approach showed better performance in four tasks (GERD, Hypertension, OSA, and PVD). The best results for CAD, Depression, Hypertriglyceridemia, and Venous Insufficiency diseases are achieved by the $Original + PSO$ method. The $Three - Stage$ approach is showed better performance as the number of the achieved best results is more than the other approaches.

3.6.2 Further analysis

For further analyzing three-stage, two-stage, and UMLS approaches, we compared the obtained experimental results by the proposed approaches with five different methods. The compared methods are divided into two groups. Kappa [167], Solt [155] and Yao [192] used rule-based approaches to classify i2b2 (2008) data set [58]. Meanwhile, Ambert [12] and Garla [58] used automatic feature engineering methods to solve the i2b2 (2008) challenge [58]. Our methods three-stage, two-stage, and UMLS utilizes an automatic system to enrich the data set.

Table 3.12 compares the statistical results of the three-stage, two-stage, original+PSO, UMLS+Pairs, UMLS and Original with the other five methods. The suggested three-stage approach shows better performance in five tasks (CHF, GERD, OSA, OA, and PVD). The Superlin [58] method shows better performance in four tasks including Depression, Gout, Hypertriglyceridemia, and Venous Insufficiency. Yao's method [192] showed better F1-measure in three tasks (Asthma, Diabetes, and Obesity). The proposed three-stage approach showed better performance by improving five

Table 3.11: Comparison of classification F1-measure and standard deviation averages using 30 independent runs on the i2b2 (2008) data set [58]. The highlighted entries are significantly better (Wilcoxon Test, $\alpha = 0.05$)

Methods	Original Deterministic	UMLS Deterministic		UMLS+Pairs Deterministic		Original+PSO Stochastic			Two-Stage Stochastic			Three-Stage Stochastic		
Diseases	F1-measure	F1-measure	T	F1-measure	T	F1-measure Ave±Std	F1-measure Best(Lowest)	T	F1-measure Ave±Std	F1-measure Best(Lowest)	T	F1-measure Ave±Std	F1-measure Best(Lowest)	T
Asthma	61.75	84.05	+	55.82	- -	71.69±0.013	74.83(67.71)	+ - +	92.40±0.011	94.50(90.31)	+ + + +	**92.42±0.007**	**94.00(90.99)**	+ + + + =
CAD	84.99	80.37	+	78.68	- -	**85.20±0.011**	**87.37(83.38)**	+ + +	78.38±0.010	80.06(76.20)	- - - -	78.65±0.009	80.69(76.38)	- - = - =
CHF	84.04	87.76	+	83.36	- -	84.66±0.009	86.32(83.11)	+ - +	88.68±0.009	90.05(86.38)	+ + + +	**88.86±0.011**	**90.54(90.54)**	+ + + + =
Depression	47.69	43.75	-	43.82	- =	**56.04±0.021**	**60.49(52.05)**	+ + +	44.22±0.008	47.20(43.55)	- + + -	43.96±0.006	45.46(43.48)	- = = - -
Diabetes	78.02	86.03	+	85.75	+ -	85.07±0.014	88.31(82.89)	+ - -	87.98±0.011	89.88(84.36)	+ + + +	87.97±0.007	89.66(86.60)	+ + + + =
Gallstones	45.57	51.57	+	46.85	+ -	48.02±0.007	49.33(46.85)	+ - +	55.84±0.013	58.02(53.82)	+ + + +	**56.05±0.013**	**60.00(53.82)**	+ + + + +
GERD	52.23	81.60	+	66.03	+ -	60.67±0.021	64.32(56.68)	+ - -	**84.21±0.008**	**86.04(82.63)**	+ + + +	84.19±0.010	86.04(82.11)	+ + + + =
Gout	56.01	76.71	+	59.82	+ -	79.01±0.016	82.39(76.21)	+ + +	87.56±0.010	90.06(85.56)	+ + + +	87.78±0.010	90.06(85.56)	+ + + + =
Hypercholesterolemia	73.01	74.46	+	72.51	- -	75.39±0.015	78.52(71.79)	+ + +	80.97±0.015	84.18(77.64)	+ + + +	80.99±0.013	83.24(78.17)	+ + + + =
Hypertension	49.10	53.31	+	49.10	= -	53.02±0.021	57.36(50.06)	+ = +	**64.85±0.025**	**72.99(60.59)**	+ + + +	64.62±0.024	68.83(59.16)	+ + + + =
Hypertriglyceridemia	48.68	48.68	=	48.68	= =	**59.03±0.028**	**62.68(52.58)**	+ + +	55.52±0.029	59.55(48.68)	+ + + -	54.18±0.034	59.55(48.68)	+ + + - -
Obesity	73.25	84.66	+	77.58	+ -	77.00±0.012	79.64(74.62)	+ - -	93.84±0.006	95.15(92.55)	+ + + +	**94.12±0.006**	**95.61(92.33)**	+ + + + +
OSA	69.11	94.31	+	91.63	+ -	78.23±0.018	81.08(73.63)	+ - -	**96.45±0.009**	**97.77(93.54)**	+ + + +	96.32±0.008	98.20(94.47)	+ + + + =
OA	60.13	83.07	+	58.39	- -	61.12±0.013	64.29(58.92)	+ - +	85.08±0.006	86.23(83.91)	+ + + +	**85.33±0.007**	**87.13(84.39)**	+ + + + =
PVD	65.51	89.69	+	72.73	+ -	74.95±0.022	78.71(67.96)	+ - +	**92.59±0.007**	**93.73(91.00)**	+ + + +	92.38±0.012	94.26(89.85)	+ + + + =
Venous Insufficiency	51.63	48.24	-	48.24	- =	**57.58±0.015**	**60.60(54.81)**	+ + +	50.96±0.031	57.79(48.18)	- + + -	49.80±0.020	54.81(48.24)	- + + - -

tasks of i2b2 (2008) [58] which this number is more than the best-achieved results by the other methods. This performance is impressive considering that our method is fully automatic without using rules.

3.7 Chapter Summary

This chapter has proposed three new approaches to medical document classification by focusing on feature extraction, feature construction, and feature selection methods. The first approach proposes a medical ontology-driven feature engineering method to reduce the number of features as well as persist with meaningful features. In conjunction with the MetaMap tool, we map meaningful phrases in the medical text to specific UMLS medical concepts. The related concepts to the problem domain are selected as features. The number of features is reduced significantly by selecting "Disease or Syndrome" and "Sign or Symptom" concepts, which are the most important in the domain of clinical notes. Experimental results show that the suggested approach can accomplish significantly better classification F1-measure than the common pipeline method. The second approach introduces a two-stage approach to investigate domain concepts and determine which concepts are discriminative to a classification problem. It is able to extract meaningful features from the document set and reduce the number of features. Moreover, the two-stage approach improves the classification performance in the majority of the candidate classifiers by using a small-size feature subset. Experimental results illustrate that the proposed method can achieve significantly better classification performance than the first approach. The third approach introduces a three-stage method to utilize 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 constructing new features and increase the classification performance in the majority of the targeted classifiers. From the experimental examinations, it can be seen that the sug-

Table 3.12: Macro-averaged F1 on I2B2 2008 test set [58]. Best scores are highlighted in bold.

Methods / Diseases	Kappa [167]	Solt [155]	Yao [192]	Ambert [12]	Superlin [58]	Original Deterministic	UMLS Deterministic	UMLS+Pairs Deterministic	Original+PSO Stochastic	Two-Stage Stochastic	Three-Stage Stochastic
Asthma	76.00	**97.84**	**97.84**	97.00	97.00	61.75	84.05	55.82	74.83	94.50	94.00
CAD	81.00	60.22	62.33	63.00	61.80	84.99	80.37	78.68	**87.37**	80.06	80.69
CHF	74.00	62.36	62.36	61.20	61.20	84.04	87.76	83.36	86.32	90.05	**90.54**
Depression	86.00	93.46	96.02	93.50	**97.90**	47.69	43.75	43.82	60.49	47.20	45.46
Diabetes	87.00	96.82	**97.31**	91.50	96.00	78.02	86.03	85.75	88.31	89.88	89.66
Gallstones	90.00	**97.29**	96.89	96.10	95.00	45.57	51.57	46.85	49.33	58.02	60.00
GERD	59.00	57.68	57.68	57.90	57.90	52.23	81.60	66.03	64.32	**86.04**	**86.04**
Gout	92.00	97.71	97.71	98.10	**98.20**	56.01	76.71	59.82	82.39	90.06	90.06
Hypercholesterolemia	68.00	90.53	91.13	**91.20**	90.80	73.01	74.46	72.51	78.52	84.18	83.24
Hypertension	67.00	88.51	92.40	89.90	92.90	49.10	53.31	49.10	53.02	72.99	68.83
Hypertriglyceridemia	72.00	79.81	70.92	87.60	**92.80**	48.68	48.68	48.68	59.03	59.55	59.55
Obesity	86.00	97.24	**97.47**	97.30	97.20	73.25	84.66	77.58	77.00	95.15	95.61
OSA	92.00	88.05	88.05	65.30	65.60	69.11	94.31	91.63	78.23	97.77	**98.20**
OA	76.00	62.86	63.07	63.10	60.40	60.13	83.07	58.39	61.12	86.23	**87.13**
PVD	73.00	63.48	63.14	62.30	60.60	65.51	89.69	72.73	74.95	93.73	**94.26**
Venous Insufficiency	44.00	80.83	80.83	72.50	**81.60**	51.63	48.24	48.24	60.60	57.79	54.81

gested approach can achieve significantly better classification F1-measure than the suggested state of the art method for i2b2 (2008) data set [58].

This chapter presents the potential of utilizing UMLS for feature extraction, construction and selection method in clinical document classification, however, it still needs more research to improve the classification performance.

Chapter 4

Ontology-guided data augmentation for medical document classification

4.1 Introduction

Medical document classification is different from common document classification in terms of text terminology. In our medical document classification, the content explains a set of medical events in a discharge note, with the objective of providing a clarification as accurately and comprehensively as conceivable when explaining the health condition of a patient. Mainly, such text massively uses domain-specific vocabulary and acronyms, making medical note analysis significantly different from commonly considered document classification. In addition, different combinations of domain-specific clinical events in a medical discharge note can explain a patient's health status completely differently. Hence, extracting important information to analyze clinical documents is exceptionally imperative.

One of the important factors which has effect on the classification accuracy is the size of the data set for training the model. Generally, there is

a lack of adequate data in medical area [143]. When the training data set is not big enough, the trained classification model does not have sufficient instances to learn. Hence, the prediction of the classifier will not be satisfactory. This issue can be worse when the data set contains documents with not enough text, such as document abstracts. One possible solution to address the issue is to augment data for training the model.

Data augmentation is a methodology that empowers experts to build the assorted variety of data accessible for training models, without really gathering new data. Data augmentation has many applications in image classification, sound and speech classification [138]. But there is not much work for text. It is not appropriate to augment the text by utilizing signal transformations as commonly used in image or speech classification. Because the order of words in text is important and may have semantic meaning. Hence, the best approach for doing data augmentation is to paraphrase the sentences in the documents by human. But this is very expensive due to lack of data. Replacing words and expressions with their synonyms can be a reasonable choice in data augmentation [198]. However, these methods are using normal dictionaries for augmentation and some domain specific terms or acronyms do not have synonyms in normal dictionaries.

As there are domain-specific vocabulary and acronyms in medical discharge notes, finding synonyms is not trivial and this requires domain knowledge. General dictionaries such as WordNet do not contain all of domain-specific vocabulary and acronyms. Hence, there is not any guarantee they will detect all of the domain-specific vocabulary. They provide a set of synonyms for the detected words and expressions in documents which can not be enough knowledge to use for data augmentation. It can be more promising if we use a domain-specific ontology which can provide more information regarding the detected phrases in documents. In this chapter, an ontology-based method is introduced for data augmentation by targeting concepts of words and expressions in the documents.

This method will replace all of the words and phrases with their scientific names if they belong to a concept in the medical field.

4.1.1 Chapter Objectives

By considering the domain of the targeted classification tasks, two data augmentation methods (a domain-specific approach and a hybrid approach) are introduced in this chapter for medical document classification problems. Different from most existing approaches, the proposed methods aim at generating new documents using a domain specific ontology (Unified Medical Language System) and a general dictionary (WordNet) to construct discriminative and more informative new documents. The overall goal of this chapter is to propose effective augmentation methods which can improve the medical document classification performance. In this chapter the following research objectives are addressed:

- Design a new ontology-based data augmentation approach ($SciName$) for constructing new documents by considering the concept of words and expressions.

- Compare the $SciName$ method with $SynName$ [171] (an existed synonym-based for general texts) method by analyzing their effectiveness for medical document augmentation.

- Combine the two methods ($SynName$ and $SciName$) to enhance the classification performance by increasing the size of training data sets.

- Compare the classification performance of proposed data augmentation approaches ($SciName$ and $SynName + SciName$) with other developed approaches for the i2b2 (2008) challenge [58].

4.1.2 Chapter Organization

The rest of the chapter is organized as follows. Section 4.2 describes the proposed ontology-guided data augmentation approach to construct new documents from the original documents. Section 4.3 presents the proposed combined ontology and dictionary based approach for medical document classification problems. Section 4.4 describes the experiment design, classification methods, data sets and parameter settings for comparison. Section 4.5 presents the results and discussion. Section 4.6 provides further analysis on the obtained best cases compared with some existing works. The achievements of the two approaches, and their limitations are summarised in section 4.7.

4.2 The proposed Ontology-guided data augmentation approach

In this section, we describe the proposed new data augmentation method and the utilized tools for extracting concepts of words and expressions for producing new documents. The proposed approach targets on concepts of words and expressions to replace them with their scientific names. Fig. 4.1 shows the flowchart of the proposed ontology-based approach for data augmentation.

The input of the proposed approach is a set of clinical documents. Firstly, the method parses each document and tokenize the content based on sentences. Then, MetaMap tool [15] is employed to detect the meaningful phrases and their concepts in each sentence from the Unified Medical Language System (UMLS) [23]. After finding the phrases with a concept, the scientific name of the detected words or expressions are used to replace their corresponding phrases in the sentence. All of the new documents are created by applying the method. Next, all of the features are extracted from the original data set and the newly created data set. Then, a classifi-

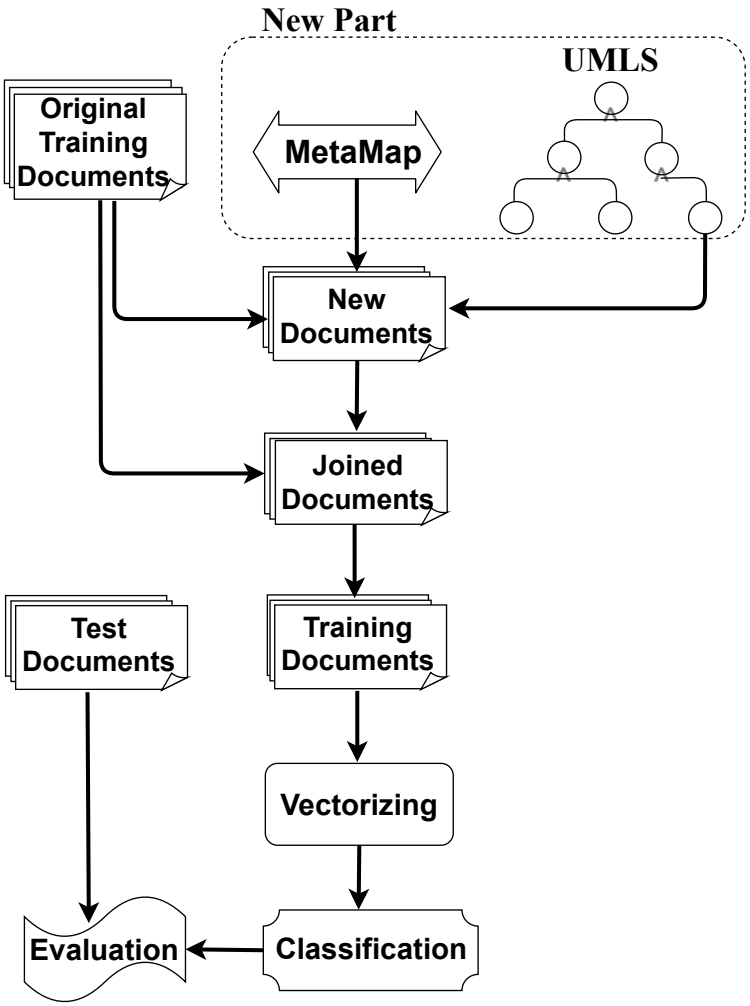

Figure 4.1: The proposed data augmentation for medical document classification

cation method is used to predict the test set labels such as Convolutional Neural Network (CNN) [56], Recurrent Neural Network (RNN) [56] and Hierarchical Attention Network (HAN) [56] methods. The output predicts the label of a document.

It is expected that the proposed approach which produces meaningful documents and keeps their class label based on the original documents,

can enhance the classification accuracy.

4.2.1 Data augmentation based on UMLS

There are many domain-specific words and expressions in medical document and data augmentation requires domain knowledge. In this section, an ontology-guided approach is introduced for short text augmentation as a preprocessing stage.

UMLS is a domain-specific dictionary in the biomedical field. It provides an ontology structure of medical terminology concepts. In the proposed approach ("$SciName$") [6], each document in the data set (D) is analyzed independently. Firstly, the xth document (D_x) is tokenized to sentences (S). Then, the ith sentence (S_i) is sent to the UMLS by using the MetaMap tool. MetaMap extracts all of the concepts of the detected meaningful expressions in S_i from the UMLS. Next, all of the detected phrases are replaced with their extracted scientific names from the UMLS. Finally, S_i is updated in D_x. This process is repeated on all of the sentences of documents to make new documents. Algorithm 4.1 shows the pseudo code of the proposed ontology-guided data augmentation. All of the expressions belong to a concept in UMLS and have a unique scientific name in UMLS called meaningful phrases in the algorithm.

A document segment is given below to illustrate how MetaMap works on the input medical documents and what output it returns in the data augmentation process. The following is a sample of a clinical note.

> *"An 80-year-old female with known rheumatic heart disease, with mitral regurgitation, paroxysmal atrial fibrillation, who underwent an AV ablation and post ablation requiring a permanent pacemaker, who now complains about progressive shortness of breath, orthopnea, and ankle edema, and referred for surgery."*

Figure 4.2 shows the output of the MetaMap for the sample document. Table 4.1 presents the detected expressions with their concepts and scien-

```
----------------------------------------------------------------------------------------------------
Phrase: An 80-year-old female with known rheumatic heart disease,
>>>>> Mappings
Meta Mapping (717):
   739   female (Woman) [Population Group]
<<<<< Mappings
----------------------------------------------------------------------------------------------------
Phrase: with mitral regurgitation,
>>>>> Mappings
Meta Mapping (1000):
  1000   MITRAL REGURGITATION (Mitral Valve Insufficiency) [Pathologic Function]
<<<<< Mappings
----------------------------------------------------------------------------------------------------
Phrase: paroxysmal atrial fibrillation,
>>>>> Mappings
Meta Mapping (1000):
  1000   PAROXYSMAL ATRIAL FIBRILLATION (Paroxysmal Atrial Fibrillation by ECG Finding)
                                                                        [Disease or Syndrome]
<<<<< Mappings
----------------------------------------------------------------------------------------------------
Phrase: an AV ablation
>>>>> Mappings
Meta Mapping (861):
   861   ablation (Destructive procedure (surgical)) [Therapeutic or Preventive Procedure]
<<<<< Mappings
----------------------------------------------------------------------------------------------------
Phrase: now
>>>>> Mappings
Meta Mapping (1000):
  1000   Now (Now (temporal qualifier)) [Temporal Concept]
<<<<< Mappings
----------------------------------------------------------------------------------------------------
Phrase: complains about progressive shortness of breath,
>>>>> Mappings
Meta Mapping (799):
   833   SHORTNESS OF BREATH (Dyspnea) [Sign or Symptom]
<<<<< Mappings
----------------------------------------------------------------------------------------------------
Phrase: ankle edema,
>>>>> Mappings
Meta Mapping (1000):
  1000   Ankle oedema (Ankle edema (finding)) [Pathologic Function]
<<<<< Mappings
----------------------------------------------------------------------------------------------------
Phrase: referred for surgery.
>>>>> Mappings
Meta Mapping (746):
   790   referred (Referring) [Functional Concept]
   790   Surgery (Surgery specialty) [Biomedical Occupation or Discipline]
<<<<< Mappings
----------------------------------------------------------------------------------------------------
```

Figure 4.2: A segment of returned results of extracted concepts using MetaMap

Algorithm 4.1: Ontology-guided data augmentation

Input : Set of medical documents (D)

Output: Set of new medical documents

$x \leftarrow 0$

$maxIter \leftarrow |D|$

while $x < maxIter$ **do**

  Tokenization: Tokenize D_x to sentences (S);

  for *i = 1 to* $|S|$ **do**

    Detect all of the meaningful phrases which belong to a concept in UMLS by using MetaMap for S_i

    Replace the detected phrases with their extracted unique scientific names from the UMLS ($S_i^{updated}$)

    Replace S_i with $S_i^{updated}$ in D_x

  $x \leftarrow x + 1$

return set of the new produced medical documents;

tific names for each phrase of the sample document. The concepts and scientific names of each detected phrase in the table is extracted by analyzing the lines 4, 9, 14, 19, 24, 29, 34, 39 and 40 shown in Figure 4.2 for the example document segment. Firstly, the phrase appearing in square brackets is extracted as a concept of the detected expression in the sentence. Then, the phrase appeared within the round parentheses at the same line is extracted as a scientific name of the detected expression. Finally, the extracted scientific name is used to replace the original expression in the sentence. This process is applied on all of the three sentences of the sample note. Below is the final output of the proposed method for the example clinical note.

> *"An 80-year-old woman with known rheumatic heart disease, with mitral valve insufficiency, paroxysmal atrial fibrillation by ECG finding, who underwent an AV destructive procedure (surgical) and post destructive procedure (surgical) requiring a permanent pacemaker,*

Table 4.1: The detected phrases of the example notes using MetaMap.

#	Detected Phrases	Extracted Concepts	Replaced Phrases
1	female	[Population Group]	woman
2	mitral regurgitation	[Pathologic Function]	mitral valve insufficiency
3	paroxysmal atrial fibrillation	[Disease or Syndrome]	paroxysmal atrial fibrillation by ECG finding
4	ablation	[Therapeutic or Preventive Procedure]	destructive procedure (surgical)
5	now	[Temporal Concept]	now (temporal qualifier)
6	shortness of breath	[Sign or Symptom]	dyspnea
7	ankle edema	[Pathologic Function]	ankle edema (finding)
8	referred	[Functional Concept]	referring
9	surgery	[Biomedical Occupation or Discipline]	surgery specialty

who now (temporal qualifier) complains about progressive dyspnea, orthopnea, and ankle edema (finding), and referring for surgery specialty."

The proposed data augmentation approach is able to provide more particular knowledge from UMLS. Hence, the length of the output is longer in comparison with the length of the original input. For instance, the phrase *"ablation"* is exchanged to *"destructive procedure (surgical)"* and the expression *"mitral regurgitation"* is transformed to *"mitral valve insufficiency"*. The suggested approach is able to provide more meaningful information by utilizing UMLS. At the end, the newly constructed notes are utilized to feed to the candidate classifier in the training step together with the original discharge notes to increase the clinical document classification performance.

The classification performance is used to evaluate the resulting documents. If a trained classification model by the original medical document plus the resulting documents can improve the medical document classification performance, it can be concluded that the resulting documents helped the classifier to learn better in the training step.

4.3 The proposed combined ontology and dictionary-based approach

In this section, we combine the suggested ontology-guided approach with the synonym-based method to provide more data for the training model. Figure 4.3 demonstrates the introduced method overall. In the synonym-based method (SynName [171]), WordNet dictionary is used to find the synonyms of words and then a 100-dimensional pretrained GloVe model [1] which is trained on Wikipedia data is employed to find the highest similarity synonym of each word to replace it in the document. The produced documents from the SynName method and the constructed documents from the SciName method are added to the target set. The output of the combined approach increases the size of the training data three times. Then, the tripled data feeds to the model for training. Figure 4.4 describes the oversampling on the classes in detail. It is expected that the newly produced documents with new features in them will improve the classification performance on the applied medical tasks.

4.4 Experiment design

The aim of evaluation is to figure out if the proposed augmentation approaches can improve the medical document classification performance and solve the shortage of data issues in the medical area.

4.4.1 Classification methods

To get an advantage from deep learning methods, it is better to have more data to feed these data-hungry models. Hence, in the ontology-based and the combined data augmentation approaches, the new documents are made and mixed with the original documents to use for classification. The

[1] From the GloVe website http://nlp.stanford.edu/data/glove.6B.zip

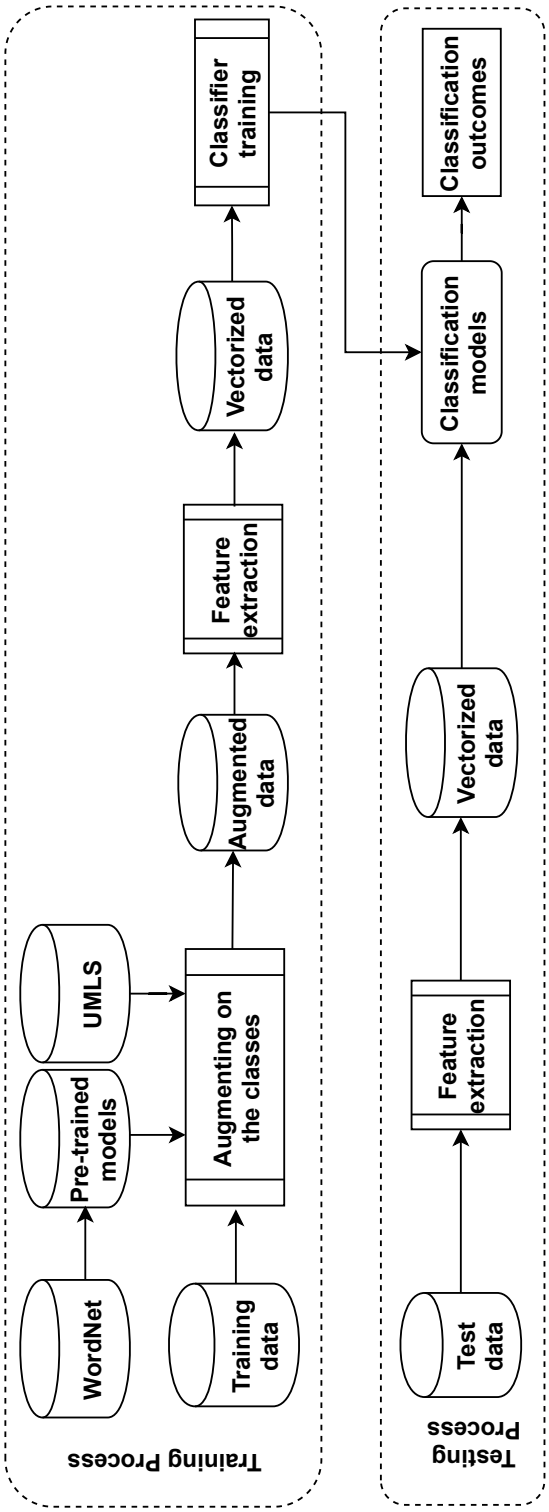

Figure 4.3: The combination of SynName and SciName oversampling methods for medical document classification

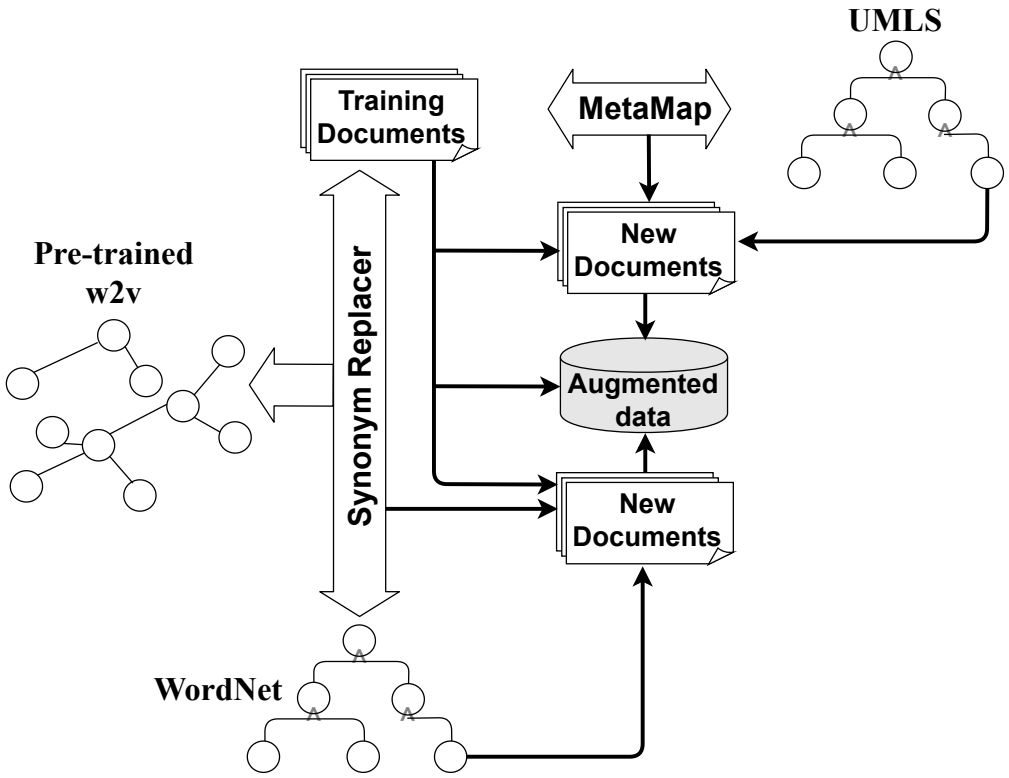

Figure 4.4: The oversampling on the classes by using SynName and SciName methods for medical document classification in detail

baseline approach is the "Original" method. We use three deep learning (DL) models, including a convolutional neural network (CNN) [56], a recurrent neural network (RNN) [56], and a hierarchical attention network (HAN) [56]. The performance is calculated by evaluating the macro F1-measure metric for all of the used ML methods.

Word2Vec is a presentation approach to convert words to numbers. Word2Vec word embedding is used to represent word tokens into numerical vectors. Word embedding represents the semantic meaning of each word in a numerical vector form. Word2Vec makes word embedding by utilizing a feed-forward neural network to anticipate the context words for an input word. The word embedding is trained on all documents lo-

cally in the training and test data sets and transformed each word to its corresponding embedding. Then, the learned word embedding is used to generate the input for CNN [56], RNN [56] and HAN [56]. The size of word embedding is 350.

4.4.2 Data set and preprocessing

The performance of the suggested ontology-guided data augmentation is evaluated on the 2010 Informatics for Integrating Biology and the Bedside (i2b2 2010) [168], the data set of 2008 Informatics for Integrating Biology and the Bedside (i2b2) [58] and the PubMed data set [56]. All of the details about i2b2(2010) [168] and i2b2(2008) [58] data sets are provided in Chapter 3.

The labels of the PubMed data set [56] are metabolism, physiology, genetics, chemistry, pathology, surgery, psychology, and diagnosis. The data set includes 8000 documents, each class has 1000 documents with 70% of documents for training and 30% for testing. The PubMed data set [56] has 30178 various terms. The size of the training sets of all data sets will double/triple by adding the new produced documents to the original ones. For example, the number of the train documents in $SynName$ and $SciName$ is increased to (5600$\times$2=) 11200 in PubMed dataset [56] and (426$\times$2=) 852 in i2b2(2010) [168] by adding the new augmented documents in $SynName$ and $SciName$ approaches. The 852 input documents from i2b2 (2010) contain 14920 various terms. The 11200 input documents from PubMed [56] include 59151 different terms.

4.4.3 Parameter Settings

Three different ML methods are used to evaluate the proposed approach. Ten percent of the original training set in all of the tasks are considered as a validation set. The suggested parameters in [56] are used for the employed neural network approaches. Early stopping by considering the validation

F1-measure (three epochs without any improvement) is used to terminate the training step.

The used CNN [56] architecture has three separate parallel convolutional layers with 100 filters for each one. The input documents are fed to each CNN layer at the same time. One CNN has a kernel (window) size of 3, the other has a size of 4 and the third one has a size of 5. The output of each CNN layers goes through a separate max pooling operation and the results (3 vectors) are concatenated into one vector which is then sent to the fully connected layer. The output of this architecture for each input document is 300 channels $\times$ number of words. The applied dropout rate is 50% [56].

HAN [56] is a deep learning model developed for document classification. It contains two hierarchies. The lower hierarchy analyzes a line in word level and it feeds with a word embedding. Then, it uses a bidirectional GRU (Gated Recurrent Unit) to apply an attention mechanism to find more important words. The output is a line embedding which is feed to the upper hierarchy to analyze a document in line level. A dropout with 50% value is applied on the produced document embedding and finally, a softmax function is employed to predict a label of each document.

The RNN architecture uses an attention mechanism (which is similar to a single hierarchy of HAN method). A bidirectional GRU with attention and 200 number of hidden cells is utilized with dropout and softmax. The used optimizer is Adam with learning rate of 0.0002. The applied dropout rate is 50%.

4.4.4 Evaluation criteria

Macro F1-measure metric is used as it is common for analyzing imbalanced data sets. As the PubMed [56] is a balanced data set, accuracy metric is applied on it too. Hence, the performance of each classifier is calculated by evaluating macro F1-measure/accuracy metrics for all of the used

methods based on the candidate data set:

$$\begin{aligned} Accuracy = & \frac{(TN + TP)}{(TN + TP + FN + FP)} \\ & = \frac{Number\ of\ correct\ assessments}{Number\ of\ all\ assessments} \end{aligned} \tag{4.1}$$

where TP (True Positive) is the number of correctly identified documents, FP (False Positive) is the number of incorrectly identified documents, TN (True Negative) is the number of correctly rejected documents and FN (False Negative) is the number of incorrectly rejected documents. Our approach is a wrapped-based method. Hence, a classifier is employed to run with PSO to evaluate value of fitness function parameters (TP, FP, TN and FN).

4.5 Results and discussion

Three different approaches are applied on the original documents in each data set. In first approach ($SynName$), we used WordNet dictionary to extract all of the synonyms of the every word appeared inside of a document. Then, the most similar synonym is found by using the GloVe pre-trained model [2] and used to replace the original word to generate new documents. The used GloVe model provides 100-dimensional vector which is trained on Wikipedia data with 6 billion tokens (expressions) and a 400,000 word vocabulary. In the second approach (our proposed $SciName$), UMLS is employed to find scientific names of the appeared phrases in the documents based on their concepts to replace with the original phrase in the document. In the third approach, we added together the augmented document from both the $SynName$ and $SciName$ approaches with the original data set (denoted as SynName+SciName). ALL the augmentation methods are applied on the training set only. Then, experimental results are calculated using 30 independent runs on the original test set.

[2]From the GloVe website http://nlp.stanford.edu/data/glove.6B.zip

4.5.1 Significant test

Tables 4.2 to 4.7 compare the statistical results of the four methods, i.e., the original one without augmentation, $SynName$, $SciName$ and $SynName + SciName$ methods. The average, best and standard deviation of accuracies and F1-measure are provided for each ML method and the statistical significance test is done on the experiment results of the 30 runs to compare the three approaches. The paired Wilcoxon signed rank test with significance level of 0.05 is used to assess whether the proposed approaches ($SciName$ and $SynName + SciName$) have made significant difference in the classification performance. In tables 4.2 to 4.7, the "T" column indicates the significance test of each approach against the previous columns (methods), where "+" indicates the suggested method is significantly better, "=" no significant difference, and "-" significantly less accurate. The best results are highlighted in the tables. For example, table 4.2 presents the significant test of the $SynName + SciName$ approach for PDV disease with "+ - =" result. Here, "+" means $SynName + SciName$ is significantly better than $Original$, "-" means $SynName + SciName$ is significantly less accurate than $SynName$ and "=" means there is no significant difference between $SynName + SciName$ and $SciName$.

4.5.2 Discussion based on classifiers (i2b2 2008 dataset)

Tables 4.2, 4.3 and 4.4 present the obtained results for $SynName+SciName$, $SciName$, $SynName$ and $Original$ approaches. In table 4.2, $SciName$ approach has improved the classification performance in ten tasks out of the sixteen tasks by using CNN in comparison to the $Original$ and $SynName$ methods. The method shows big improvement in eight tasks (Asthma, Depression, Gallstones, GERD, Gout, Hypercholesterolemia, Hypertension and OA). Table 4.3 provides the achieved results by RNN. $SciName$ has enhanced F1-measure in comparison to the $Original$ and $SynName$ methods in eight tasks in which the improvements in six of them (Asthma, CHF,

Depression, Gallstones, Gout and Hypertriglyceridemia) are substantial. Comparison of HAN classification performance for the $SciName$ method is shown in table 4.4. The results demonstrate that the improvement in six tasks which the classification performance is noteworthy in three of them (CHF, Depression and Gallstones). Overally, the $SciName$ has achieved the highest F1-measure (89.66%) for Gout task by using RNN in comparison to the $original$ and $SynName$ methods.

Tables 4.2, 4.3 and 4.4 present the F1-measure obtained results for the combined method in comparison with other three methods ($Original$, $SynName$ and $SciName$) on i2b2(2008) data set [58]. The combined method with CNN model improved in ten tasks. The improvement is remarkable in eight tasks (Asthma, CAD, Gallstones, GERD, Gout, Hypercholesterolemia, Hypertension and OA). The highest enhancement achieved in CAD task with 92.09% F1-measure. Table 4.3 demonstrates the performance of the suggested approach by using RNN model. It outperformed the other methods in thirteen tasks by showing a big improvement in ten of them (Asthma, Depression, Diabetes, Gallstones, GERD, Gout, Hypercholesterolemia, OA, PVD and Venous Insufficiency). The improvement by combined method and RNN model is noticeable in Hypercholesterolemia, OA and Venous Insufficiency tasks with 84.31%, 88.30% and 74.20% F1-measures, respectively. From table 4.4, it is obvious that the combined method by using the HAN model upgraded the F1-measure values in majority of the tasks except the CHF task. While HAN improved fourteen tasks in comparison with the other approaches, the obtained F1-measures by RNN model for the combined method is higher almost in all of the tasks.

4.5.3 Discussion based on classifiers (PubMed dataset [56])

The $SciName$ approach is tested on two other data sets (PubMed [56] and i2b2(2010) [168]) too. By analyzing tables 4.5, 4.6, and 4.7, it is clear that the neural network methods are improved in accuracy and F1-measure by

using the domain-specific ontology to augment the training documents in comparison to the $Original$ and $SynName$ methods which provides more information in learning process to the used model. The $SciName$ method presents better performance in comparison with $SynName$ in PubMed [56] and i2b2(2010) [168] data sets. RNN shows significant improvement in both of the data sets. It achieved 85.57% accuracy, 85.37% F1-measure for the PubMed data set [56] and 94.66% F1-measure for the i2b2(2010) data set [168].

By analyzing tables 4.5 and 4.6, it is clear that the neural network methods are improved in accuracy and F1-measure by using combination of the original data set with the obtained augmented data sets from the two introduced augmentation methods for PubMed data set [56]. The highest accuracy and F1-measure in tables 4.5 and 4.6 belong to RNN with values 90.80% and 90.91%, respectively. Table 4.7 provides the statistical results for i2b2(2010) data set [168]. In the table, CNN, RNN and HAN show high F1-measure in the combination approach ($SynName$+$SciName$). The highest F1-measure belongs to RNN with 96.43%.

4.5.4 Conclusion

In tables 4.2 to 4.7, the suggested combined ($SynName + SciName$) approach shows better performance in comparison with the $SciName$ and the $SynName$ methods [198].

Interestingly, although some of the resulting documents look gibberish to human readers, the generated documents improve the classifier performance. This might be because our approach works at word level which can tolerate non-sense sentences as long as they contain meaningful words. Our method not only works well on clinical notes (such as i2b2(2008) [58] and i2b2(2010) [168]), but also it shows promising results on related biomedical notes (PubMed set [56]).

Table 4.2: Comparison of CNN [56] classification performance (F1-measure) and standard deviation averages using 30 independent runs for i2b2 2008 data set [58]. The significant test is for the suggested approach against the original data set (Wilcoxon Test, $\alpha = 0.05$)

Methods	Original	SynName		SciName		SynName+SciName	
Diseases	Ave±Std (Best)	Ave±Std (Best)	T	Ave±Std (Best)	T	Ave±Std (Best)	T
Asthma	57.16±4.091 (65.64)	63.30±5.493 (71.92)	+	79.69±8.178 (87.17)	++	**85.71±3.904 (89.78)**	+++
CAD	90.44±1.396 (92.93)	90.53±2.790 (93.13)	=	90.80±0.840 (92.19)	++	**92.09±0.690 (93.64)**	+++
CHF	90.63±1.551 (93.07)	90.04±2.645 (92.16)	=	**91.00±1.703 (92.85)**	=+	90.27±2.038 (93.30)	== -
Depression	45.34±1.875 (52.38)	46.51±3.683 (58.59)	+	**63.59±7.505 (74.17)**	++	53.96±7.321 (72.57)	++ -
Diabetes	90.01±3.329 (93.022)	**90.46±3.221 (95.04)**	=	88.66±4.199 (93.83)	- -	90.19±3.517 (94.54)	==+
Gallstones	49.63±3.173 (55.95)	68.90±11.398 (83.39)	+	74.30±11.820 (85.81)	++	**84.57±1.473 (87.75)**	+++
GERD	53.28±4.528 (61.52)	59.50±6.493 (72.43)	+	68.09±1.078 (70.29)	++	**80.26±1.501 (82.11)**	+++
Gout	52.06±5.049 (61.75)	67.62±9.645 (82.72)	+	72.72±12.457 (89.11)	++	**86.90±4.650 (91.65)**	+++
Hypercholesterolemia	66.17±2.686 (69.65)	75.96±5.131 (82.45)	+	76.07±4.593 (80.93)	+=	**81.33±2.239 (84.84)**	+++
Hypertension	54.25±5.153 (66.68)	60.39±6.810 (72.75)	+	62.01±5.498 (72.62)	++	**69.97±7.334 (81.44)**	+++
Hypertriglyceridemia	48.68±1.421e-14 (48.68)	48.68±1.421 (48.68)	=	48.68±0.010 (48.68)	==	48.67±0.016 (48.68)	===
Obesity	84.70±5.914 (91.39)	90.04±4.736 (92.56)	+	89.61±4.981 (91.84)	+=	**90.39±2.736 (92.08)**	+=+
OSA	92.51±3.121 (95.32)	92.41±1.568 (94.54)	=	92.13±2.635 (94.89)	==	**92.61±2.555 (95.32)**	===
OA	49.23±4.064 (59.92)	49.21±2.903 (56.43)	=	67.09±4.159 (74.73)	++	**73.54±5.632 (82.38)**	+++
PVD	87.60±3.723 (91.56)	**90.18±1.970 (92.12)**	+	89.54±3.139 (92.34)	+=	89.16±2.241 (92.00)	+ - =
Venous Insufficiency	**52.98±3.530 (60.18)**	50.56±2.633 (57.79)	-	50.20±2.433 (59.78)	- =	52.43±3.862 (62.76)	=++

Table 4.3: Comparison of RNN classification performance (F1-measure) and standard deviation averages using 30 independent runs for i2b2 2008 data set [58]. The significant test is for the suggested approach against the original data set (Wilcoxon Test, $\alpha = 0.05$)

Methods	Original	SynName		SciName		SynName+SciName	
Diseases	Ave±Std (Best)	Ave±Std (Best)	T	Ave±Std (Best)	T	Ave±Std (Best)	T
Asthma	79.44±15.278 (89.50)	87.21±2.260 (91.52)	+	87.34±2.254 (90.05)	+=	**88.44±1.354 (90.44)**	+++
CAD	95.52±1.128 (97.28)	95.27±0.64 (96.40)	=	**95.61±0.44 (96.62)**	==	95.25±0.53 (96.37)	===
CHF	90.27±1.280 (93.29)	91.57±0.833 (92.83)	+	**92.35±1.112 (94.92)**	++	92.07±0.832 (93.26)	++=
Depression	49.42±3.555 (56.03)	63.15±4.898 (73.59)	+	73.34±2.011 (79.10)	++	**76.01±3.096 (82.00)**	+++
Diabetes	93.16±2.002 (94.93)	94.32±0.646 (95.35)	+	93.41±0.746 (94.76)	+ -	**95.26±0.790 (97.00)**	+++
Gallstones	69.99±10.597 (84.44)	84.60±1.397 (87.23)	+	85.13±1.348 (87.90)	++	**86.25±0.986 (88.34)**	+++
GERD	69.23±7.923 (80.42)	80.4±1.827 (83.82)	+	79.52±1.416 (82.49)	+ -	**81.41±1.264 (83.67)**	+++
Gout	79.96±1.588 (83.78)	84.64±3.165 (91.6)	+	89.66±2.552 (93.71)	++	**92.08±1.766 (95.83)**	+++
Hypercholesterolemia	70.63±3.489 (78.98)	81.79±1.877 (85.29)	+	81.24±2.200 (85.32)	+=	**84.31±1.240 (86.65)**	+++
Hypertension	77.63±3.176 (83.16)	84.0±2.181 (87.37)	+	80.76±2.817 (85.26)	+ -	**84.34±2.934 (88.47)**	+=+
Hypertriglyceridemia	48.68±1.421e-14 (48.68)	48.97±1.120 (52.19)	=	50.17±1.949 (53.85)	++	**50.82±2.460 (56.73)**	++=
Obesity	89.85±1.858 (92.09)	91.26±0.877 (92.32)	+	90.92±0.850 (92.39)	+ -	**92.02±0.818 (94.47)**	+++
OSA	93.98±2.788 (97.80)	97.18±1.189 (98.70)	+	96.15±1.390 (98.20)	+ -	**97.25±0.740 (98.27)**	+=+
OA	69.26±3.326 (74.51)	86.08±2.003 (89.48)	+	84.86±1.439 (87.42)	+ -	**88.30±1.149 (90.70)**	+++
PVD	88.21±1.192 (91.13)	89.02±1.012 (91.51)	+	89.32±1.374 (92.23)	+=	**90.11±1.193 (92.23)**	+++
Venous Insufficiency	65.58±3.803 (73.79)	71.15±3.853 (77.93)	+	68.45±6.065 (76.04)	+ -	**74.20±3.616 (81.63)**	+++

Table 4.4: Comparison of HAN classification performance (F1-measure) and standard deviation averages using 30 independent runs for i2b2 2008 data set [58]. The significant test is for the suggested approach against the original data set (Wilcoxon Test, $\alpha = 0.05$)

Methods	Original	SynName		SciName		SynName+SciName	
Diseases	Ave±Std (Best)	Ave±Std (Best)	T	Ave±Std (Best)	T	Ave±Std (Best)	T
Asthma	79.73±14.047 (90.45)	87.12±2.790 (91.53)	+	86.10±2.472 (90.45)	+ -	**88.01±1.359 (90.58)**	+++
CAD	95.16±1.525 (97.29)	94.16±1.74 (96.61)	-	95.47±0.34 (96.14)	=+	**95.60±0.51 (96.61)**	=+=
CHF	89.47±1.583 (91.91)	90.81±1.537 (93.52)	+	**92.03±0.776 (93.31)**	++	91.98±0.755 (93.54)	++=
Depression	51.69±4.984 (66.92)	57.16±6.356 (68.93)	+	73.78±1.850 (76.92)	++	**81.47±5.831 (90.94)**	+++
Diabetes	93.36±1.405 (95.04)	94.42±0.651 (96.07)	+	93.54±0.582 (95.31)	= -	**95.58±0.808 (97.26)**	+++
Gallstones	58.43±7.694 (80.18)	78.58±13.12 (87.20)	+	84.31±1.547 (86.17)	++	**85.38±3.259 (88.20)**	+++
GERD	60.34±7.734 (75.92)	79.52±1.948 (83.50)	+	78.83±1.944 (82.28)	+ -	**80.58±1.703 (83.65)**	+++
Gout	62.24±13.106 (80.38)	83.33±2.323 (91.48)	+	80.63±6.058 (89.11)	+ -	**90.48±2.192 (95.27)**	+++
Hypercholesterolemia	67.00±3.220 (74.24)	80.18±1.576 (84.16)	+	79.56±4.084 (86.15)	+ -	**82.13±1.866 (85.11)**	+++
Hypertension	72.88±8.741 (81.21)	81.61±2.946 (85.99)	+	78.50±3.763 (84.42)	+ -	**82.75±2.307 (87.38)**	+++
Hypertriglyceridemia	48.68±1.421e-14 (48.68)	48.76±0.410 (50.96)	=	49.58±1.882 (55.92)	++	**49.59±1.480 (52.38)**	++=
Obesity	89.65±1.922 (91.91)	91.47±0.592 (92.59)	+	91.01±1.011 (93.27)	+=	**92.00±0.803 (93.51)**	+++
OSA	95.35±1.682 (97.41)	96.52±1.371 (98.68)	+	95.45±1.801 (97.83)	= -	**97.24±1.120 (99.14)**	+++
OA	66.87±4.959 (74.51)	87.86±1.774 (91.01)	+	83.67±3.382 (88.24)	+ -	**88.65±1.728 (91.32)**	+++
PVD	88.52±1.077 (90.69)	89.30±1.577 (91.81)	=	**89.89±1.128 (91.92)**	+=	89.68±1.264 (92.34)	+==
Venous Insufficiency	61.57±3.179 (67.83)	57.67±4.733 (70.54)	-	59.38±7.091 (71.86)	- +	**71.06±5.635 (78.22)**	+++

Table 4.5: Comparison of classification accuracy and standard deviation averages using 30 independent runs for PubMed data set [56]. The significant test is for the combined approach against others(Wilcoxon Test, $\alpha = 0.05$)

Methods	Original	SynName		SciName		SynName+SciName	
Classifiers	Accuracy	Accuracy		Accuracy		Accuracy	
	Ave±Std (Best)	Ave±Std (Best)	T	Ave±Std (Best)	T	Ave±Std (Best)	T
CNN [56]	71.64±0.55 (72.67)	80.64±0.63 (81.84)	+	80.80±0.51 (82.08)	+=	**84.16±0.66 (85.42)**	+++
RNN [56]	71.53±1.03 (73.50)	84.42±0.90 (86.38)	+	85.57±0.62 (86.75)	++	**90.80±0.45 (91.63)**	+++
HAN [56]	71.29±0.69 (72.88)	84.29±0.85 (85.38)	+	85.00±0.94 (86.92)	++	**90.75±0.49 (91.79)**	+++

Table 4.6: Comparison of classification F1-measure and standard deviation averages using 30 independent runs for PubMed data set [56]. The significant test is for the combined approach against others(Wilcoxon Test, $\alpha = 0.05$)

Methods	Original	SynName		SciName		SynName+SciName	
Classifiers	F1-measure	F1-measure		F1-measure		F1-measure	
	Ave±Std (Best)	Ave±Std (Best)	T	Ave±Std (Best)	T	Ave±Std (Best)	T
CNN [56]	71.54±0.76 (72.64)	80.42±0.69 (81.42)	+	80.48±0.71 (81.81)	+=	**84.34±0.54 (85.36)**	+++
RNN [56]	71.62±0.73 (72.97)	84.07±0.88 (85.64)	+	85.37±0.90 (87.00)	++	**90.91±0.58 (91.98)**	+++
HAN [56]	70.96±0.82 (72.21)	84.21±1.23 (86.16)	+	84.95±0.73 (86.36)	+=	**90.85±0.79 (91.99)**	+++

Table 4.7: Comparison of classification F1-measure and standard deviation averages using 30 independent runs for i2b2 2010 data set [168]. The significant test is for the combined approach against others(Wilcoxon Test, $\alpha = 0.05$)

Methods	Original	SynName		SciName		SynName+SciName	
Classifiers	F1-measure	F1-measure		F1-measure		F1-measure	
	Ave±Std (Best)	Ave±Std (Best)	T	Ave±Std (Best)	T	Ave±Std (Best)	T
CNN [56]	77.66±13.16 (90.22)	92.40±1.95 (95.62)	+	92.72±0.90 (95.10)	+=	**94.15±1.12 (97.44)**	+++
RNN [56]	85.51±8.00 (91.37)	91.30±2.74 (97.51)	+	94.66±1.32 (96.92)	++	**96.43±0.68 (98.12)**	+++
HAN [56]	56.11±18.78 (90.35)	86.90±7.23 (97.48)	+	93.75±2.55 (96.87)	++	**96.27±0.73 (97.51)**	+++

4.6 Further analysis

For further analyzing the proposed $SciName$ and $SynName + SciName$ methods, we compared the obtained experimental results by three differ-

ent ML models with five different existing methods. The compared methods are divided into two groups. Kappa [167], Solt [155] and Yao [192] used rule-based approaches to classify i2b2 (2008) data set [58]. Meanwhile, Ambert [12] and Garla [58] used automatic feature engineering methods to solve i2b2 (2008) challenge [58]. $SciName$ and $SynName + SciName$ methods utilize an automatic system to enrich the data set.

Table 4.8 compares the statistical results of the $SynName$, $SciName$ and $SynName+SciName$ with the other five methods. The best results are highlighted in bold. The second best results of each task is bolded too. The dictionary-based method ($SynName$) improved the F1-measure values in three tasks (CAD, GERD and PVD). Similarly, the ontology-based method ($SciName$) increased the F1-measure of three different tasks (CAD, CHF and PVD). The combined approach ($SynName + SciName$) outperformed the other methods in seven tasks out of sixteen tasks (CAD, Diabetes, GERD, OSA, OA, PVD and Venous Insufficiency). This performance is impressive considering that our method is fully automatic without using rules.

4.6.1 Clinical Assessment

In this chapter, two methods are developed for augmenting clinical discharge notes:

- $SciName$: The first method using UMLS for data augmentation by replacing expression in the documents with their scientific names. (This method doubles the train set size.)

- $SynName + SciName$: The combined method which produces documents by using $SciName$ method idea and plus using WordNet dictionary to replace expressions in the documents with their synonyms. (This method triples the train set size.)

Table 4.8: Macro-averaged F1 on I2B2 2008 test set. Best scores from our study indicated in bold.

Methods						SynName			SciName			SynName+SciName		
Disease	Kappa [167]	Solt [155]	Yao [192]	Ambert [12]	Superlin [58]	CNN	RNN	HAN	CNN	RNN	HAN	CNN	RNN	HAN
Asthma	76.00	**97.84**	**97.84**	97.00	97.00	71.92	91.52	91.53	87.17	90.05	90.45	89.78	90.44	90.58
CAD	81.00	61.22	62.33	63.00	61.80	93.13	96.40	**96.61**	92.19	**96.62**	96.14	93.64	96.37	**96.61**
CHF	74.00	62.36	62.36	61.20	61.20	92.16	92.83	93.52	92.85	**94.92**	93.31	93.30	93.26	93.54
Depression	86.00	93.46	96.02	93.50	**97.90**	58.59	73.59	68.93	74.17	79.10	76.92	72.57	82.00	90.94
Diabetes	87.00	96.82	**97.31**	91.50	96.00	95.04	95.35	96.07	93.83	94.76	95.31	94.54	97.00	**97.26**
Gallstones	90.00	**97.29**	96.89	96.10	95.00	83.39	87.23	87.20	85.81	87.90	86.17	87.75	88.34	88.20
GERD	59.00	57.68	57.68	57.90	57.90	72.43	**83.82**	**83.50**	70.29	82.49	82.28	82.11	**83.67**	**83.65**
Gout	92.00	97.71	97.71	98.10	**98.20**	82.72	91.60	91.48	89.11	93.71	89.11	91.65	95.83	95.27
Hypercholesterolemia	68.00	90.53	91.13	**91.20**	90.80	82.45	85.29	84.16	80.93	85.32	86.15	84.84	86.65	85.11
Hypertension	67.00	88.51	92.40	89.90	**92.90**	72.75	87.37	85.99	72.62	85.26	84.42	81.44	88.47	87.38
Hypertriglyceridemia	72.00	79.81	70.92	87.60	**92.80**	48.68	52.19	50.96	48.68	53.85	55.92	48.68	56.73	52.38
Obesity	86.00	97.24	**97.47**	97.30	97.20	92.56	92.32	92.59	91.84	92.39	93.27	92.08	94.47	93.51
OSA	92.00	88.05	88.05	65.30	65.60	94.54	98.70	98.68	94.89	98.20	97.83	95.32	98.27	**99.14**
OA	76.00	62.86	63.07	63.10	60.40	56.43	89.48	91.01	74.73	87.42	88.24	82.38	90.70	**91.32**
PVD	73.00	63.48	63.14	62.30	60.60	**92.12**	91.51	91.81	**92.34**	92.23	91.92	**92.00**	**92.23**	**92.34**
Venous Insufficiency	44.00	80.83	80.83	72.50	81.60	57.79	77.93	70.54	59.78	76.04	71.86	62.76	**81.63**	78.22

As the second method ($SynName + SciName$) presents better performance, we found those documents which this method predicted correctly but the other methods could not predict correctly. This case happened in the four tasks (Gout, Hypertension, Obesity and OA (osteoarthritis)).

The second method ($SynName+SciName$) presents better performance, and is able to predict correctly four conditions; namely, gout, hypertension, obesity and osteoarthritis which other methods are not able to predict correctly.

Two cases (a positive patient and a negative patient) from the original training data sets and their corresponding documents from the new augmented data sets of the mentioned diseases in this section are selected for better understanding the reason for the better performance. On a clinical perspective, the primary reasons may be due to the following specificity of certain terms/words to a condition. This can be categorised into three relatively important medical concepts.

1. Medications/drugs that are specifically used for the condition.
2. Medical terms or phrases that enable to identify a diagnosis.
3. Medical diagnosis from the past and current situation.

Medications or drugs that are mainly used for a specific condition warranted an increase in prediction to certain condition such as gout and hypertension. For both gout positive cases, an anti-gout treatments (e.g. allopurinol and colchicine) are in the documents. Both anti-gout treatments are only utilised in patients with gout. This increases the prediction accuracy because both $SynName$ and $SciName$ narrows it to just the condition gout.

Similarly, drugs commonly used for (e.g. metoprolol, felodipine, hydralazine) treating hypertension are also evident in the documents with labeled as positive for hypertension despite no mention of the diagnosis of hypertension.

In the second concept where medical terms or phrases enable identification of a condition, this clearly strengthens the $SciName$ as it is based on UMLS. However, $SynName$ can also point to specific conditions that it may be synonymous with. Such a good example is the phrase "elevated blood pressure", which both $SciName$ and $SynName$ will definitely map it to hypertension. Another is the term "arthritis", which probably increased the performance in the Osteoarthritis cases. Even though arthritis is a collective term for inflammation of the joints, it provides common arthritic conditions such as Rheumatoid arthritis and Osteoarthritis.

In the obesity cases, both diagnose obesity in the past and current situation as indicated in both documents.

In the cases analyzed, there is clear indication that $SciName$ would perform better than $SynName$. This is because UMLS has a wide-range of concepts where it can identify and map it to a diagnosis. Combining this with $SynName$ enables it to bring it a notch higher in performance.

However, what we noticed is that $SynName$ may also result in decreasing the prediction due to a concept which is translated differently enabling a different wording or concept to come about.

What we would recommend is to focus more on the $SciName$ rather than the $SynName$. A percentage of what to be used should be higher on $SciName$ and lesser on $SynName$.

Additionally, another database specifically for medications or diagnostics should also be use as it may increase the performance as they are more specific compared to UMLS.

As mentioned in early this chapter, in numerous practical works on modeling, data augmentation is extremely important. This is a situation that we encounter when in practical settings, real life patient cases are unavailable to feed data-hungry models (a rare disease is an example where available cases are few). In fact, synthetic data synthesis and augmentation has strong advantages with respect to advancing healthcare models research by protecting patient confidentiality, and is a promising tool for

situations where real world data is difficult to obtain or unnecessary. At that time, in combination with data augmentation we can also perform simulations to generate new training cases.

4.7 Chapter Summary

This chapter has developed a domain-specific method and a combined method to augment clinical data for solving binary and multi-class medical document classification. The methods aim to produce new documents from original documents by replacing meaningful expressions with their scientific names from UMLS and synonyms from WordNet dictionary to deal with the data shortage issue in medical document classification. Medical domain knowledge is borrowed from UMLS to find scientific names of expression based on their concepts. The meaningful synonyms are extracted from the WordNet dictionary to construct new documents. The introduced approaches are able to improve the precision of classification in the neural network models. Experimental results of accuracy and F1-measure show that the proposed method can improve the performance of the CNN [56], RNN [56], and HAN [56] models by using the suggested ontology-based approach ($SciName$) and combination approach of $SynName$ and $SciName$ ($SynName + SciName$) to provide more samples in the training phase. Providing more samples for data-hungry methods such as CNN [56], RNN [56], and HAN [56] helps these models learn better the sample dataset and improve the classification accuracy. From the obtained results for the candidate classifiers, it can be concluded that the RNN model demonstrates better results in comparison with CNN [56] and HAN [56].

The proposed methods in this chapter can address the shortage of data in medical field by doubling and tripling the size of documents. These approaches can apply in cases there is not access to real cases to collect data. Though these methods have shown promise in utilizing an ontology-

guided and a combined data augmentation approach in medical document classification, they produce one new document for each original document. They are not able to make different versions from a single document as they have only one option for each detected expression in a document to replace in the new document. Consequently, they are not able to solve the existing imbalance issue in the dataset. Hence, to tackle this issue, the next chapter will develop a suitable dictionary-based method where multiple set of documents can be constructed for each document to improve classification performance in medical document classification. The dictionary-based approach can deal with imbalanced data issues. This approach is converting imbalanced data to balanced data by increasing the size of small classes' documents to make it equal to the size of the largest class in the dataset.

Chapter 5

A Dictionary-based Oversampling Approach to Clinical Document Classification on Small and Imbalanced Dataset

5.1 Introduction

Synonym based data augmentation can provide new documents while preserving the overall meaning of the original documents. We want to select better synonyms based on similarity. However, if a simple similarity function is used as the heuristic, the same synonym is often selected, and this reduces the variety when we need to generate multiple documents from the same original document. Medical documents can be severely imbalanced, hence, we do need to create multiple new documents to make the data set balanced. So the variety of the documents are important. If we create very similar new documents based on the same original document, this can lead to overfitting in training the candidate classifier. Thus, an automatic dictionary-based method is introduced for oversampling the

minority class documents by selecting more suitable synonyms and simultaneously increasing the diversity of the new produced documents. Then, we will propose an incremental approach which using the same applied method on the minority class for all of the classes. It will make the data set balanced and at the same time make the size of all of the classes double size of the majority class.

5.1.1 Chapter Objectives

This chapter develops a new dictionary-based approach for dealing with imbalanced issues in medical discharge notes. Different from most available approaches, the proposed approach ($SynNameSim$) targets a dictionary-based approach by employing different pre-trained models to measure the similarity between an appeared word in a document and its extracted synonyms from a dictionary. Furthermore, the proposed approach is applied for replacing a suitable synonym among the related synonyms with the detected word in the document in all of the classes by making the classes balanced and increasing their sizes at the same time ($IncSynNameSim$ method). Two other existing approaches ($HighSimSynName$ [171] and $RandomSynName$ [116]) are used in comparison step. The HighSimSynName [171] approach has used the highest similarity synonym to replace a word in generating new instances, and the $RandomSynName$ [116] approach has randomly selected a synonym from synonyms set to replace a word in constructing new documents. This chapter investigates the following research objectives:

- Design a new dictionary-based method that can generate new discriminative documents for the minority class from the original document set. This method can address the limitation of the suggested approaches in the previous chapter in generating more than one document from the original document.
- Compare the $SynNameSim$ method with $HighSimSynName$ [171]

and $RandomSynName$ [116] (existing synonym-based) methods by analyzing their effectiveness for medical document augmentation.

- Compare the classification performance of proposed data augmentation approaches ($SynNameSim$ and $IncSynNameSim$) with other non-data augmentation approaches.

- Further analyze the effectiveness of the proposed approaches in increasing the size of all of the classes in the targeted medical document classification tasks.

5.1.2 Chapter Organization

The rest of the chapter is organized as follows. Section 6.2 presents the proposed dictionary-based data augmentation approach to construct new documents from the original documents. Section 6.3 describes the proposed augmentation approach for increasing the size of all of the classes in the targeted data sets. Section 6.4 presents the experiment design, classification methods, data sets and parameter settings for comparison. Section 6.5 describes the results and discussion. Section 6.6 provides further analysis on the obtained best cases compared with some existing works. The achievements of the two approaches, and their limitations are summarized in section 6.7.

5.2 The new dictionary-based oversampling

In this section, we propose a new dictionary-based oversampling method. We describe the used tools for extracting synonyms of words and similarities of the extracted synonyms with the words for producing new documents for the minority class. The proposed approach employs the synonyms of a word to replace one of them with the targeted word. This

approach is using a WordNet dictionary to extract all of the existing synonyms of each word and three pretrained models to measure the similarity between the synonyms and the corresponding word.

The input of the proposed method is a set of medical documents. Firstly, the method checks for the minority class to target to do oversampling. Next, the method tokenizes each document of the minority class based on sentences and then words. After that, for each word, the proposed method finds a suitable synonym to replace the word in the main sentence. We create multiple new documents from the same original document using this method to make the data set balanced.

It is expected that the proposed method which produces meaningful documents for the minority class and retains their class label based on the original documents, can increase the classification performance.

5.2.1 Probabilistic Oversampling Method

Fig. 5.1 demonstrates the proposed dictionary-based oversampling method in details.

First Phase: The extracted words are fed one by one to the system to find a proper synonym to replace the input word in the original sentence. The first phase extracts all of the available synonyms of the input word from WordNet dictionary. Then, the second phase gets the main word and its corresponding synonyms vector as an input.

$$Synonyms = [syn_1,\ syn_2,\ syn_3,\ \ldots,\ syn_n]$$

Second Phase: The second phase utilizes three different pre-trained word embeddings to extract the similarity of each synonym with the input word. The used three pre-trained models are: (1) English word embeddings pre-trained on biomedical texts from PubMed and PMC with five billion words[1], (2) English word embeddings pre-trained on biomedi-

[1]From the website http://evexdb.org/pmresources/vec-space-models/wikipedia-pubmed-and-PMC-w2v.bin

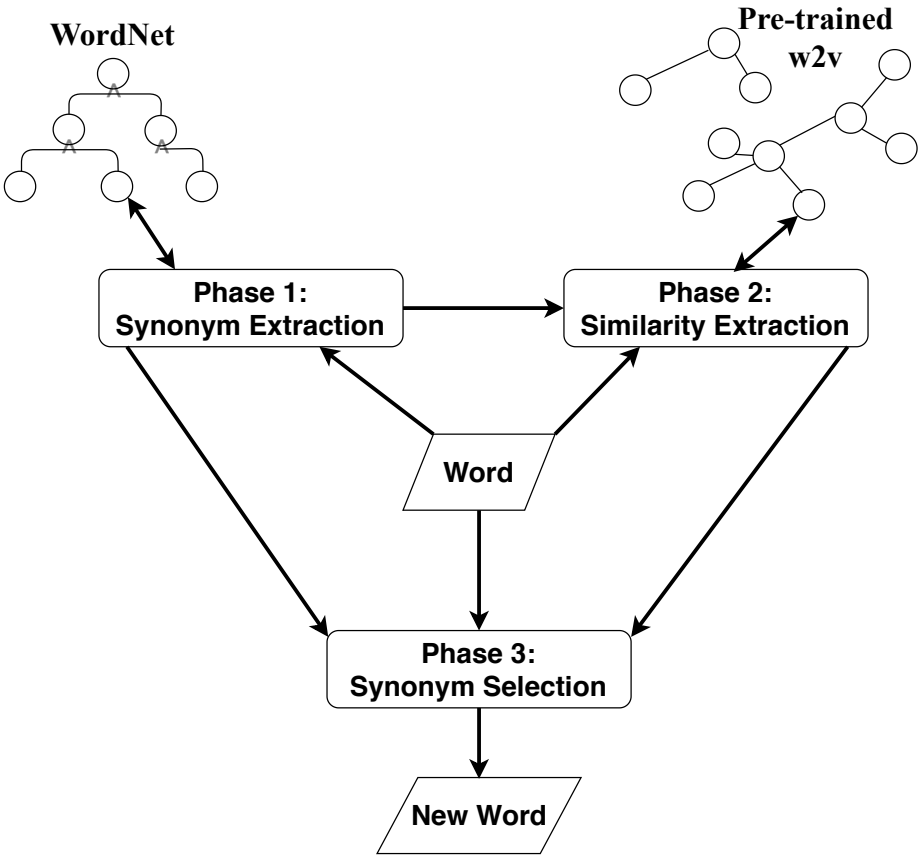

Figure 5.1: The proposed dictionary-based oversampling method for data augmentation

cal texts from MEDLINE®/PubMed with 2,665,547 types distinct words[2], and (3) English word embeddings pre-trained on Wikipedia data with six billion tokens and a 400,000 word vocabulary[3].

In this step, the system checks for the similarity of each synonym with the input word in the first word embedding (1). If there is a similarity value between them in the word embedding, the value is extracted. If there is no similarity, the system checks the second word embedding (2)

[2]From the website https://archive.org/download/pubmed2018_w2v_400D.tar/pubmed2018_w2v_400D.tar.gz

[3]From the GloVe website http://nlp.stanford.edu/data/glove.6B.zip

to measure the similarity. If there is not any similarity, the system checks for the third word embedding (3). At the end, if there is not any similarity, the synonym is removed from the synonym vector. If the system finds any similarity in one of the word embeddings, it does not check for other word embeddings and moves to check the next synonym's similarity. These steps applies on all of the synonyms to extract the similarities vector.

$$Similarities = [sim_1,\ sim_2,\ sim_3,\ \ldots,\ sim_n]$$

Third Phase: The third phase gets the main word, the extracted synonyms vector from the first phase and the extracted similarities vector from the second phase. The system selects one of the synonyms to replace the input word by considering the similarities vector. As we need to augment several different versions of each document, a good selection method is needed to select various synonyms to increase the variety of the new produced documents to avoid overfitting in the training step. To address this issue, the synonym is selected by producing a random vector $Rand\_Vec$ with uniformly distributed elements between 0 and 1 ($U(0,1)$) in the same size with the similarities vector:

$$Rand\_Vec = [r_1, r_2, r_3, \ldots, r_n]\ \ r_1, r_2, r_3, \ldots, r_n\ \sim U(0,1)$$

Then, we have vector D by

$$D = Similarities - Rand\_Vec$$

$$D = [sim_1 - r_1, sim_2 - r_2, sim_3 - r_3, \ldots, sim_n - r_n]$$

Finally, the maximum elements of D will be found and its corresponding index's synonym in $Synonyms$ will be selected as the synonym to replace with the main word in the sentence.

5.2.2 An Example

A document piece is given below to describe how the proposed approach ($SynNameSim$) [5] works on the input medical discharge notes and what

output it returns in the oversampling process. The following is an example of a clinical document.

> *"History of present illness: this is a 33-year-old female with a history of postpartum cardiomyopathy (ef 15% to 20%), status post pacer/icd placement, status post mitral valve replacement/tricuspid valve replacement, who presents with abdominal pain, decreased appetite, nausea, vomiting, and occasional chest pain."*

Fig. 5.2 shows the extracted synonyms vectors from WordNet dictionary for each detected word from the example sentence in the first phase.

Next, the second phase extracts all of the available similarities. Then, the third phase selects proper synonyms to replace with the main words in the original sentence. Table. 5.1 demonstrates the selected synonyms for each word by apply the selection process.

Table 5.1: The detected words and the selected synonyms for each word.

Word	Replaced Synonym		Word	Replaced Synonym
history	chronicle		pain	bother
present	existing		appetite	appetence
illness	malady		nausea	sickness
cardiomyopathy	myocardiopathy		vomiting	vomit
status	condition		occasional	casual
placement	positioning		chest	thorax
status	position		pain	painfulness

The following is the final output after applying the three mentioned phases.

> *"Chronicle of existing malady: this is a 33-year-old female with a chronicle of postpartum myocardiopathy (ef 15% to 20%), condition post pacer/icd positioning, position post mitral valve replacement/tricuspid valve replenishment, who presents with abdominal*

```
'history'= ['history', 'story', 'account', 'chronicle']
'present'= ['lay_out', 'portray', 'face', 'pose', 'award', 'show', 'gift', 'represent', 'existing',
            'deliver', 'exhibit', 'salute', 'demo', 'demonstrate', 'submit', 'present', 'give', 'nowadays',
            'confront', 'introduce', 'present_tense', 'acquaint', 'stage']
'illness'= ['sickness', 'unwellness', 'illness', 'malady']
'female'= ['female_person', 'female']
'cardiomyopathy'= ['cardiomyopathy', 'myocardiopathy']
'status= ['position', 'condition', 'status']
'placement'= ['arrangement', 'position', 'positioning', 'emplacement', 'locating', 'placement', 'location']
'status'= ['position', 'condition', 'status']
'replacement'= ['surrogate', 'replenishment', 'substitution', 'alternate', 'replacement', 'replacing',
                'permutation', 'switch', 'substitute', 'refilling', 'transposition', 'renewal', 'successor']
'abdominal'= ['ab', 'abdominal', 'abdominal_muscle']
'pain'= ['pain_in_the_ass', 'annoyance', 'botheration', 'pain', 'pain_sensation', 'painfulness',
         'pain_in_the_neck', 'infliction', 'hurting', 'nuisance', 'bother', 'painful_sensation']
'appetite'= ['appetence', 'appetite', 'appetency']
'nausea'= ['sickness', 'nausea']
'vomiting'= ['disgorgement', 'puking', 'regurgitation', 'vomit', 'emesis', 'vomiting']
'occasional'= ['occasional', 'periodic', 'episodic', 'casual']
'chest'= ['chest', 'thorax', 'pectus', 'dresser', 'breast', 'bureau', 'chest_of_drawers']
```

Figure 5.2: Extracted synonyms of each detected word from the example sentence

bother, decreased appetence, sickness, vomit, and casual thorax painfulness."

The proposed approach can easily create new documents with high variety by using the extracted synonyms from WordNet with awareness of synonyms' similarities with the original word. Finally, the new created documents are utilized for the training stage together with the original documents to enhance the performance of medical document classification.

5.3 Incremental Data Augmentation

In this section, the proposed approach is applied on all of the classes' documents to make the classes balanced and increase the size of the classes too. $IncSynNameSim$ approach is proposed for augmenting all of the classes by increasing their size equal to the double size of the large class.

In the proposed approach ("$IncSynNameSim$"), each document in the data set (D) is analyzed independently. The data set is grouped based on the documents' classes ($D \leftarrow \{C_1, C_2, ..., C_n\}$). For example, C_1 is set of all of the documents with the same label. The number of the documents in each class increases until the class size is equal to the double size of the biggest class in the classes. As it is possible to produce more than one instance from each document. Hence, the approach checks to make sure all of the existing documents in the class will be considered in the new documents' construction process. Firstly, the document dth in class i (C_d^i) is tokenized to sentences (S). Then, the jth sentence (S_j) is tokenized to words (W). Next, the kth word (W_k) is sent to the WordNet dictionary. WordNet extracts all of the synonyms of the kth word. After that, three different pre-trained models are employed to find the similarities of the extracted synonyms with kth word (W_k). Next, the proposed selection approach is applied to select a suitable synonym by considering its similarity value to

W_k. Finally, the selected synonym is replaced with W_k. This process is repeated on all of the words of tokenized sentences in documents to make new documents. The number of documents in each class will increase to double size of the biggest class size in each data set. Algorithm 5.1 shows the pseudo code of the proposed incremental data augmentation.

5.4 EXPERIMENTAL DESIGN

5.4.1 Baseline methods

In this chapter three main baseline methods are used in comparison with the proposed approaches. The $Original$ method is using the main data set without any new augmented documents. The $RandomSynName$ [116] approach is choosing randomly a synonym to swap with the main word without considering the similarity between the selected synonym and the main word. The $HighSimSynName$ [171] approach is selecting a synonym with the highest similarity to the main word to swap with the main word.

5.4.2 Classification methods

The performance of both of the ideas ($SynNameSim$ and $IncSynNameSim$) are tested by using three deep learning (DL) models independently, including a convolutional neural network (CNN), a recurrent neural network (RNN), and a hierarchical attention network (HAN) [56] is used. The performance of each classifier is calculated by evaluating the macro F1-measure metric for all of the used methods. Word2Vec word embedding is utilized to represent word tokens into numerical vectors. The size of word embedding is 350, which is the same as Chapter 4.

Algorithm 5.1: $IncSynNameSim$ data augmentation

Input : Set of medical documents based on classes ($D \leftarrow \{C_1, C_2, ..., C_n\}$)

Output: A set of new medical documents

$maxIter \leftarrow |max(D)| * 2$
for *i = 1 to* $|D|$ **do**
  $counter \leftarrow |C_i|$
  $d \leftarrow 0$
  while $counter < maxIter$ **do**
    if $d == |C_i| - 1$ **then**
      $d \leftarrow 0$
    Tokenization: Tokenize document d in class i (C_i^d) to sentences (S);
    for *j = 1 to* $|S|$ **do**
      Tokenization: Tokenize sentence j (S_j) in document d to words (W);
      for *k = 1 to* $|W|$ **do**
        Phase 1: Extract synonyms of word W_k by using WordNet dictionary
        Phase 2: Find similarities between each extracted synonym with W_k
        Phase 3: Select a suitable synonym by considering its similarity value to W_k
        Replace the selected synonym with W_k
    $counter \leftarrow counter + 1$
    $d \leftarrow d + 1$
return set of the new produced medical documents;

5.4.3 Data set and preprocessing

The performance of the proposed dictionary-based oversampling approaches are assessed on the data set of 2008 Informatics for Integrating Biology and

the Bedside (i2b2) [58] and the data set of 2010 Informatics for Integrating Biology and the Bedside (i2b2) [168].

The proposed data augmentation approaches are applied on the data sets and the size of the training sets of all minority classes will increase to their corresponding majority class in each task by adding the new produced documents to the original ones ($SynNameSim$) and $IncSynNameSim$ approach will increase the size of all of the classes in each training set to the double size of their corresponding majority class in each task by adding the new produced documents to the original ones.

5.4.4 Parameter Settings

Three different deep learning methods are utilized to assess the proposed approaches (the same as Chapter 5). All of the used parameters are the same as Chapter 4.

5.5 Results and discussions

The performance of the proposed approaches are evaluated based on macro F1-measure for the 2008 Integrating Informatics with Biology and the Bedside (I2B2) obesity challenge and the i2b2(2010) [168] data sets.

We compared our proposed approaches with three different approaches and to evaluate the performance of each one, three different machine learning models (namely, CNN, RNN, HAN) are used separately. In the first approach ($Original$), the imbalanced data set without any changes in the data set is used. In the second approach ($HighSimSynName$ [171]), the synonyms with the highest similarity to the main word are used to replace the main word to augment the minority class instances. In the third approach ($RandomSynName$ [116]), one of the synonyms of each word is selected randomly to augment the new instances. In the suggested dictionary-based approach ($SynNameSim$), we utilized WordNet dictio-

nary to extract all of the synonyms of the main word that appeared inside of a document from the minority class. Then, three different English word embedding pre-trained models are used to find the similarities of the extracted synonyms with the main word. Next, the selection method (the proposed probabilistic approach) is used to generate new documents for the minority class to make the train data set of each task balanced. Finally, experimental results using 30 independent runs are measured on the original test set.

5.5.1 Significant test

Tables 5.2–5.9 compare the statistical results for seven approaches in each table. The average and standard deviation of F1-measure is presented for each machine learning method and the significance test is done using the experiment results of the 30 runs to compare the two approaches. The Wilcoxon signed rank test with significance level of 0.05 is utilized to evaluate whether the proposed idea has made a significant difference in classification performance. The highlighted entries are significantly better with $alpha < 0.05$. In tables 5.2–5.9, "T" column shows the significance test of the each method against the other previous methods (columns) in each table, where "+" indicates the proposed approach is significantly better, "=" indicates no significant difference, and "-" indicates significantly less accurate. The best results are highlighted boldly in the tables.

5.5.2 Discussion based on the proposed SynNameSim method

By analyzing tables 5.2–5.5, it is clear that the dictionary-based oversampling method ($SynNameSim$) improves the performance in F1-measure by utilizing WordNet dictionary and three different English word embedding pre-trained models (one general and two domain-specific models). It achieved the highest F1-measure in the majority of the tasks in all of the employed machine learning models. $SynNameSim$ provides more docu-

Table 5.2: Comparison of CNN classification performance (F1-measure) and standard deviation averages using 30 independent runs for i2b2 2008 dataset [58]. The significant test is for the suggested approach against the previous columns (Wilcoxon Test, $\alpha = 0.05$)

Methods	Original	HighSimSynName		RandomSynName		SynNameSim	
Diseases	Ave±Std (Best)	Ave±Std (Best)	T	Ave±Std (Best)	T	Ave±Std (Best)	T
Asthma	57.16±4.091 (65.64)	82.57±6.731 (88.37)	+	64.12±18.881 (82.49)	+ -	**87.15±0.772 (88.37)**	+ + +
CAD	90.44±1.396 (92.93)	90.57±1.244 (92.28)	=	89.59±2.103 (92.67)	- -	**91.21±0.595 (92.99)**	+ + +
CHF	90.63±1.551 (93.07)	89.62±1.796 (92.61)	-	89.66±2.254 (92.84)	- =	**91.33±0.434 (92.61)**	+ + +
Depression	**45.34±1.875 (52.38)**	44.07±0.590 (46.47)	-	44.26±0.657 (46.62)	- =	44.62±0.925 (48.09)	- = =
Diabetes	90.01±3.329 (93.022)	89.68±3.240 (93.71)	-	89.12±4.247 (94.52)	- =	**92.87±0.926 (94.80)**	+ + +
Gallstones	49.63±3.173 (55.95)	58.81±11.465 (79.77)	+	65.23±14.853 (82.22)	+ +	**67.40±5.173 (76.85)**	+ + +
GERD	53.28±4.528 (61.52)	51.92±6.387 (67.84)	-	52.76±5.569 (63.16)	- +	**58.45±4.307 (65.43)**	+ + +
Gout	52.06±5.049 (61.75)	57.27±22.626 (86.28)	+	41.85±20.892 (77.21)	- -	**72.33±7.292 (83.12)**	+ + +
Hypercholesterolemia	66.17±2.686 (69.65)	65.96±3.108 (71.67)	-	65.46±2.547 (69.56)	- =	**68.03±1.063 (70.76)**	+ + +
Hypertension	54.25±5.153 (66.68)	53.84±4.626 (61.85)	-	54.22±5.111 (63.08)	= +	**60.99±3.512 (70.27)**	+ + +
Hypertriglyceridemia	48.68±1.421e-14 (48.68)	38.02±18.338 (53.62)	-	25.60±19.670 (51.84)	- -	**50.34±2.451 (57.09)**	+ + +
Obesity	84.70±5.914 (91.39)	90.90±1.093 (91.85)	+	90.64±0.745 (91.60)	+ =	**91.50±0.151 (91.87)**	+ + +
OSA	**92.51±3.121 (95.32)**	58.74±27.907 (92.03)	-	49.13±25.798 (85.53)	- -	77.21±8.843 (90.76)	- + +
OA	49.23±4.064 (59.92)	50.88±6.418 (67.73)	+	48.49±3.563 (55.19)	- -	**52.37±4.205 (64.33)**	+ + +
PVD	**87.60±3.723 (91.56)**	71.70±10.524 (88.51)	-	61.18±14.006 (83.83)	- -	76.24±5.936 (85.22)	- + +
Venous Insufficiency	52.98±3.530 (60.18)	54.52±12.924 (71.25)	+	48.94±18.937 (69.60)	- -	**64.72±3.788 (71.32)**	+ + +

Table 5.3: Comparison of RNN classification performance (F1-measure) and standard deviation averages using 30 independent runs for i2b2 2008 dataset [58]. The significant test is for the suggested approach against the previous columns (Wilcoxon Test, $\alpha = 0.05$)

Methods	Original	HighSimSynName		RandomSynName		SynNameSim	
Diseases	Ave±Std (Best)	Ave±Std (Best)	T	Ave±Std (Best)	T	Ave±Std (Best)	T
Asthma	79.44±15.278 (89.50)	86.41±3.445 (89.34)	+	85.05±10.250 (89.64)	+ -	**89.24±0.641 (90.99)**	+ + +
CAD	**95.52±1.128 (97.28)**	94.77±1.096 (96.62)	**-**	94.50±1.108 (96.61)	- =	**95.24±0.498 (96.85)**	= + +
CHF	90.27±1.280 (93.29)	90.12±1.220 (92.82)	**=**	90.18±1.232 (92.36)	= =	**91.17±0.672 (92.84)**	+ + +
Depression	49.42±3.555 (56.03)	44.98±3.028 (54.04)	-	45.13±3.850 (61.15)	- +	**50.70±6.358 (71.94)**	+ + +
Diabetes	**93.16±2.002 (94.93)**	92.42±1.400 (94.52)	-	91.51±1.619 (94.04)	- -	**93.27±0.616 (94.35)**	= + +
Gallstones	69.99±10.597 (84.44)	83.05±7.009 (86.70)	+	81.30±9.688 (86.87)	+ -	**85.87±0.794 (87.75)**	+ + +
GERD	69.23±7.923 (80.42)	79.97±1.428 (83.19)	+	79.77±1.589 (82.85)	+ =	**81.38±0.800 (83.36)**	+ + +
Gout	79.96±1.588 (83.78)	**89.44±2.196 (90.95)**	+	86.15±7.360 (90.95)	+ -	**89.92±0.587 (91.37)**	+ = +
Hypercholesterolemia	70.63±3.489 (78.98)	71.29±4.405 (79.02)	+	71.92±4.939 (79.15)	+ =	**74.16±2.580 (79.97)**	+ + +
Hypertension	77.63±3.176 (83.16)	73.76±9.123 (83.24)	-	75.99±5.349 (83.16)	- +	**81.27±1.492 (83.72)**	+ + +
Hypertriglyceridemia	48.68±1.421e-14 (48.68)	49.00±0.995 (52.57)	+	48.61±0.108 (48.68)	= -	**49.17±1.037 (52.58)**	+ = +
Obesity	89.85±1.858 (92.09)	**91.32±0.207 (91.84)**	+	**91.27±0.357 (91.87)**	+ =	**91.46±0.190 (92.08)**	+ = =
OSA	93.98±2.788 (97.80)	95.03±2.296 (97.83)	+	89.65±12.40 (97.38)	- -	**96.28±1.042 (98.25)**	+ + +
OA	69.26±3.326 (74.51)	81.93±7.420 (88.04)	+	81.64±6.732 (88.94)	+ =	**85.10±1.105 (87.72)**	+ + +
PVD	88.21±1.192 (91.13)	88.86±1.512 (91.26)	=	87.19±8.573 (91.56)	- -	**90.14±0.825 (92.03)**	+ + +
Venous Insufficiency	65.58±3.803 (73.79)	66.78±5.652 (70.34)	+	63.53±5.747 (70.34)	- -	**68.46±2.028 (71.79)**	+ + +

Table 5.4: Comparison of HAN classification performance (F1-measure) and standard deviation averages using 30 independent runs for i2b2 2008 dataset [58]. The significant test is for the suggested approach against the previous columns (Wilcoxon Test, $\alpha = 0.05$)

Methods	Original	HighSimSynName		RandomSynName		SynNameSim	
Diseases	Ave±Std (Best)	Ave±Std (Best)	T	Ave±Std (Best)	T	Ave±Std (Best)	T
Asthma	79.73±14.047 (90.45)	86.25±2.233 (89.60)	+	85.79±2.125 (89.78)	+ -	**88.90±0.338 (89.90)**	+ + +
CAD	**95.16±1.525 (97.29)**	94.85±1.181 (96.85)	**-**	94.82±1.142 (97.29)	- =	**95.60±0.536 (97.06)**	= + +
CHF	89.47±1.583 (91.91)	89.76±1.354 (92.37)	**=**	89.89±1.178 (91.91)	= =	**90.75±1.045 (93.98)**	+ + +
Depression	51.69±4.984 (66.92)	46.92±4.676 (57.37)	-	48.38±5.081 (57.12)	- +	**53.37±1.428 (57.87)**	+ + +
Diabetes	**93.36±1.405 (95.04)**	92.43±1.210 (94.20)	-	92.74±1.207 (95.34)	- =	**93.54±0.670 (95.36)**	= + +
Gallstones	58.43±7.694 (80.18)	82.77±7.062 (87.75)	+	78.64±11.463 (86.17)	+ -	**85.15±1.275 (87.23)**	+ + +
GERD	60.34±7.734 (75.92)	60.31±11.574 (80.26)	=	61.98±11.164 (80.16)	+ +	**71.33±7.219 (81.48)**	+ + +
Gout	62.24±13.106 (80.38)	89.74±0.497 (90.81)	+	85.00±9.377 (90.52)	+ -	**90.01±0.389 (91.52)**	+ + +
Hypercholesterolemia	67.00±3.220 (74.24)	66.48±2.623 (73.45)	-	67.40±3.425 (74.22)	= +	**68.37±1.677 (73.33)**	+ + +
Hypertension	**72.88±8.741 (81.21)**	59.10±12.21218 (79.65)	-	63.35±12.319 (80.67)	- +	**72.56±7.047 (82.87)**	= + +
Hypertriglyceridemia	48.68±1.421e-14 (48.68)	48.68±1.421e-14 (48.68)	=	48.80±0.701 (52.58)	= =	**49.06±1.150 (52.58)**	+ + +
Obesity	89.65±1.922 (91.91)	**91.11±0.647 (91.824)**	+	**91.33±0.260 (92.06)**	+ =	**91.43±0.173 (92.17)**	+ = =
OSA	**95.35±1.682 (97.41)**	92.77±3.540 (97.44)	-	93.17±2.958 (97.34)	- +	94.19±1.396 (96.88)	- + +
OA	66.87±4.959 (74.51)	82.44±4.228 (88.84)	+	79.12±7.086 (86.76)	+ -	**84.83±1.472 (88.34)**	+ + +
PVD	88.52±1.077 (90.69)	88.26±1.524 (90.71)	=	88.80±0.993 (90.57)	= =	**89.97±0.821 (91.56)**	+ + +
Venous Insufficiency	61.57±3.179 (67.83)	60.35±5.051 (66.41)	-	60.16±4.869 (68.10)	- =	**64.86±2.457 (69.70)**	+ + +

Table 5.5: Comparison of classification F1-measure and standard deviation averages using 30 independent runs for i2b2 2010 data set [168]. The significant test is for each approach against the previous columns (Wilcoxon Test, $\alpha = 0.05$)

Methods	Original	HighSimSynName		RandomSynName		SynNameSim	
Classifiers	Ave±Std (Best)	Ave±Std (Best)	T	Ave±Std (Best)	T	Ave±Std (Best)	T
CNN	77.66±13.16 (90.22)	**87.61±5.73 (93.97)**	+	85.50±9.31 (93.32)	+ -	**87.36±2.77 (92.96)**	+ = +
RNN	85.51±8.00 (91.37)	**85.94±7.25 (95.68)**	+	77.04±14.62 (95.10)	- -	**85.81±7.29 (94.87)**	+ = +
HAN	56.11±18.78 (90.35)	83.97±11.43 (93.88)	+	80.08±13.35 (92.76)	+ -	**88.80±3.65 (95.10)**	+ + +

ments by considering the similarity values of synonyms to make the imbalanced candidate data sets balanced. It provides some new features in the new documents and simultaneously it helps CNN, RNN and HAN methods to avoid overfitting. Our data argumentation method assists the mentioned machine learning methods to learn better in the training step and predict the labels of the test set with higher accuracy. In table 5.5, the proposed approach ($SynNameSim$) shows comparable performance as the $HighSimSynName$ [171] approach with CNN and RNN classifiers, but it achieves better performance with HAN classifier.

5.5.3 Discussion based on the proposed IncSynNameSim method

Furthermore, we investigated whether increasing the size of all of the classes with the proposed approach can help to improve the classification performance or not. Hence, $IncSynNameSim$ approach is suggested to increase the size of all of the classes and at the same time make the data set to be balanced. The idea is applied to two other approaches which are called $IncHighSimSynName$ and $IncRandomSynName$. Both of the methods are similar to $HighSimSynName$ [171] and $RandomSynName$ [116] approaches. The main difference is in the size of the classes in outputs which are bigger. Then the proposed approach ($IncSynNameSim$) is compared with three different approaches (Original, IncHighSimSynName and In-

Table 5.6: Comparison of CNN classification performance (F1-measure) and standard deviation averages using 30 independent runs for i2b2 2008 data set [58]. The significant test is for each approach against the previous columns (Wilcoxon Test, $\alpha = 0.05$)

Methods	Original	IncHighSimSynName		IncRandomSynName		IncSynNameSim	
Diseases	Ave±Std (Best)	Ave±Std (Best)	T	Ave±Std (Best)	T	Ave±Std (Best)	T
Asthma	57.16±4.091 (65.64)	87.43±4.02 (88.78)	+	87.42±1.27 (89.34)	+ =	**88.71±0.15 (88.78)**	+ + +
CAD	90.44±1.396 (92.93)	90.63±2.38 (94.60)	=	90.80±1.79 (93.11)	= =	**91.10±0.69 (92.45)**	+ + +
CHF	90.63±1.551 (93.07)	90.51±1.63 (92.62)	=	90.09±2.90 (93.28)	= =	**91.45±1.06 (93.08)**	+ + +
Depression	45.34±1.875 (52.38)	**53.06±7.28 (71.00)**	+	44.68±1.73 (51.38)	- -	45.79±2.78 (53.65)	+ - +
Diabetes	**90.01±3.329 (93.022)**	86.25±6.12 (93.55)	-	86.28±4.03 (92.46)	- -	88.01±4.35 (93.52)	- + +
Gallstones	49.63±3.173 (55.95)	**82.47±0.72 (83.61)**	+	80.66±4.70 (82.93)	+ -	**82.59±0.37 (83.27)**	+ = +
GERD	53.28±4.528 (61.52)	72.33±4.68 (80.36)	+	**73.83±2.57 (80.42)**	+ +	72.90±2.90 (80.16)	+ = -
Gout	52.06±5.049 (61.75)	**89.33±1.19 (90.37)**	+	87.63±2.32 (90.81)	+ -	87.14±1.97 (90.22)	+ - =
Hypercholesterolemia	66.17±2.686 (69.65)	**77.02±4.16 (84.03)**	+	73.24±5.28 (82.55)	+ -	73.91±3.73 (81.25)	+ - =
Hypertension	54.25±5.153 (66.68)	53.33±7.95 (75.13)	-	51.24±5.86 (66.30)	- -	**58.16±8.64 (76.79)**	+ + +
Hypertriglyceridemia	48.68±1.421e-14 (48.68)	**56.76±1.45 (62.68)**	+	55.06±3.62 (64.45)	+ -	54.62±2.81 (62.17)	+ - -
Obesity	84.70±5.914 (91.39)	86.66±10.89 (91.91)	+	88.77±6.91 (92.31)	+ +	**90.73±1.47 (92.06)**	+ + +
OSA	**92.51±3.121 (95.32)**	89.62±4.72 (93.36)	-	90.64±3.40 (92.83)	- +	91.42±0.84 (92.30)	- + +
OA	49.23±4.064 (59.92)	**82.29±2.65 (85.85)**	+	62.33±12.19 (82.05)	+ -	**82.42±2.73 (85.89)**	+ = +
PVD	87.60±3.723 (91.56)	88.61±2.05 (84.02)	+	87.13±2.75 (91.00)	= -	**89.13±1.34 (91.13)**	+ + +
Venous Insufficiency	52.98±3.530 (60.18)	**74.81±4.42 (83.76)**	+	**74.68±4.36 (82.34)**	+ =	**74.85±3.20 (82.34)**	+ = =

Table 5.7: Comparison of RNN classification performance (F1-measure) and standard deviation averages using 30 independent runs for i2b2 2008 data set [58]. The significant test is for each approach against the previous columns (Wilcoxon Test, $\alpha = 0.05$)

Methods	Original	IncHighSimSynName		IncRandomSynName		IncSynNameSim	
Diseases	Ave±Std (Best)	Ave±Std (Best)	T	Ave±Std (Best)	T	Ave±Std (Best)	T
Asthma	79.44±15.278 (89.50)	**88.29±1.56 (89.75)**	+	**88.46±1.11 (90.71)**	+ =	**88.86±0.38 (89.75)**	+ = =
CAD	**95.52±1.128 (97.28)**	**95.41±1.35 (97.95)**	=	94.96±0.64 (96.37)	- -	**95.43±0.69 (97.04)**	= = +
CHF	90.27±1.280 (93.29)	**92.05±0.98 (93.98)**	+	91.45±1.19 (93.76)	+ -	91.82±0.81 (93.98)	+ - =
Depression	**49.42±3.555 (56.03)**	48.65±9.74 (83.84)	-	44.36±1.10 (47.49)	- -	44.41±1.25 (49.30)	- - =
Diabetes	93.16±2.002 (94.93)	**94.36±0.76 (95.72)**	+	93.74±1.27 (95.77)	= -	**94.59±0.74 (96.33)**	+ = +
Gallstones	69.99±10.597 (84.44)	**77.35±12.43 (86.98)**	+	71.30±16.70 (85.83)	+ -	75.37±14.46 (87.04)	+ - +
GERD	69.23±7.923 (80.42)	71.63±11.80 (82.58)	+	70.90±12.22 (83.50)	+ -	**73.60±11.21 (83.48)**	+ + +
Gout	79.96±1.588 (83.78)	87.11±7.87 (92.20)	+	77.44±17.15 (92.08)	- -	**89.70±1.22 (91.52)**	+ + +
Hypercholesterolemia	70.63±3.489 (78.98)	81.76±2.37 (88.21)	+	79.17±4.22 (85.17)	+ -	**82.16±2.21 (86.43)**	+ + +
Hypertension	**77.63±3.176 (83.16)**	52.17±12.84 (84.44)	-	51.63±9.74 (44.97)	- -	49.65±7.41 (79.98)	- - -
Hypertriglyceridemia	48.68±1.421e-14 (48.68)	**57.26±8.47 (76.12)**	+	56.72±9.85 (86.85)	+ -	**57.23±9.49 (78.36)**	+ = +
Obesity	89.85±1.858 (92.09)	90.68±1.03 (92.09)	+	**91.69±0.63 (90.17)**	+ +	**91.22±0.52 (92.12)**	+ + =
OSA	93.98±2.788 (97.80)	**97.28±0.94 (99.13)**	+	**97.22±0.87 (99.14)**	+ =	**97.11±0.76 (99.13)**	+ = =
OA	69.26±3.326 (74.51)	79.94±2.07 (88.86)	+	77.89±14.69 (90.40)	+ -	**86.90±1.88 (90.63)**	+ + +
PVD	88.21±1.192 (91.13)	88.85±3.19 (91.62)	=	88.38±3.75 (92.34)	= =	**89.80±1.27 (92.34)**	+ + +
Venous Insufficiency	65.58±3.803 (73.79)	70.73±4.27 (76.74)	+	70.32±6.97 (77.02)	+ =	**71.27±3.30 (79.97)**	+ + +

Table 5.8: Comparison of HAN classification performance (F1-measure) and standard deviation averages using 30 independent runs for i2b2 2008 data set [58]. The significant test is for each approach against the previous columns (Wilcoxon Test, $\alpha = 0.05$)

Methods	Original	IncHighSimSynName		IncRandomSynName		IncSynNameSim	
Diseases	Ave±Std (Best)	Ave±Std (Best)	T	Ave±Std (Best)	T	Ave±Std (Best)	T
Asthma	79.73±14.047 (90.45)	**88.18±0.94 (89.60)**	+	87.30±2.13 (89.60)	+ -	***88.86±0.67 (89.90)***	+ = +
CAD	**95.16±1.525 (97.29)**	94.84±1.08 (96.84)	**-**	94.91±1.09 (96.83)	- =	**95.38±0.55 (96.39)**	= + +
CHF	89.47±1.583 (91.91)	**91.61±1.03 (93.50)**	**+**	**91.08±1.44 (93.53)**	+ =	**91.44±0.83 (93.31)**	+ = =
Depression	**51.69±4.984 (66.92)**	43.85±0.18 (44.81)	-	43.97±0.65 (47.37)	- =	43.94±0.53 (46.66)	- = =
Diabetes	**93.36±1.405 (95.04)**	**93.63±1.25 (95.57)**	=	**93.52±1.24 (96.56)**	= =	**93.85±0.96 (95.29)**	= = =
Gallstones	58.43±7.694 (80.18)	68.29±17.29 (83.49)	+	64.17±17.51 (82.70)	+ -	**72.95±14.32 (83.96)**	+ + +
GERD	60.34±7.734 (75.92)	56.35±12.52 (80.76)	-	**64.32±11.48 (80.66)**	+ +	61.99±12.35 (83.13)	+ + -
Gout	62.24±13.106 (80.38)	**89.45±1.28 (90.95)**	+	86.78±8.61 (92.20)	+ -	**89.16±1.02 (91.52)**	+ = +
Hypercholesterolemia	67.00±3.220 (74.24)	**82.72±2.94 (87.55)**	+	79.17±3.71 (85.32)	+ -	**82.32±2.40 (86.88)**	+ = +
Hypertension	**72.88±8.741 (81.21)**	48.15±7.02 (79.52)	-	51.02±6.96 (78.01)	- +	52.11±10.51 (82.43)	- + +
Hypertriglyceridemia	48.68±1.421e-14 (48.68)	**52.69±2.69 (56.19)**	+	51.66±5.01 (68.05)	+ -	51.37±2.40 (59.55)	+ - =
Obesity	89.65±1.922 (91.91)	**91.10±0.52 (91.82)**	+	**91.11±1.60 (92.36)**	+ =	**91.37±0.27 (92.11)**	+ = =
OSA	95.35±1.682 (97.41)	**96.50±1.55 (99.13)**	+	**96.98±1.70 (98.70)**	+ =	**96.94±1.08 (99.57)**	+ = =
OA	66.87±4.959 (74.51)	74.79±12.70 (89.33)	+	66.02±17.24 (87.83)	= -	**80.99±8.94 (88.64)**	+ + +
PVD	88.52±1.077 (90.69)	86.70±8.15 (91.10)	-	**89.58±1.57 (92.25)**	+ +	**89.83±1.23 (92.77)**	+ + =
Venous Insufficiency	61.57±3.179 (67.83)	65.54±4.66 (74.73)	+	**68.18±4.25 (76.39)**	+ +	67.55±3.00 (74.73)	+ + -

Table 5.9: Comparison of classification F1-measure and standard deviation averages using 30 independent runs for i2b2 2010 data set [168]. The significant test is for each approach against the previous columns (Wilcoxon Test, $\alpha = 0.05$)

Methods	Original	IncHighSimSynName		IncRandomSynName		IncSynNameSim	
Classifiers	Ave±Std (Best)	Ave±Std (Best)	T	Ave±Std (Best)	T	Ave±Std (Best)	T
CNN	77.66±13.16 (90.22)	89.15±2.38 (92.42)	+	88.49±4.93 (92.42)	+ -	**90.05+1.67 (91.90)**	+ + +
RNN	85.51±8.00 (91.37)	90.22±8.89 (96.27)	+	91.01±9.26 (96.92)	+ +	**93.02±1.39 (90.86)**	+ + +
HAN	56.11±18.78 (90.35)	90.06±4.93 (95.68)	+	**91.50±3.57 (97.51)**	+ +	**91.95±2.26 (96.21)**	+ + =

cRandomSynName).

Tables 5.6 to 5.9 provide the statistical experiments for $IncSynNameSime$, $IncHighSimSynName$ and $IncRandomSynName$ approaches. The proposed approached ($IncSynNameSime$) improves the classification performance in most of the diseases such as Asthma, CAD, CHF, Gallstones, GERD, Gout, Hypertension, Hypercholesterolemia, Obesity, OA, PVD and Venous Insufficiency. By checking tables 5.6 to 5.8, it is clear that all of the classifiers have improved the classification of Asthma, Obesity, OA and PVD diseases. Furthermore, the proposed approach enhances the F1-measure accuracy in comparison with other three approaches in all of the classifiers.

5.5.4 Comparison on all of the proposed methods in this chapter

Tables 5.10 to 5.12 present comparison of CNN, RNN and HAN classifiers performance (F1-measure) and standard deviation averages using 30 independent runs for i2b2 2008 data set [58]. The significance test is for each approach against their corresponding incremental approaches. In table 5.10, the proposed incremental approach improved approximately twelve tasks in each method by using CNN classifier. In table 5.11, the proposed incremental approach increased the performance of RNN classifiers in approximately eight tasks in each method. In table 5.12, $IncRandomSynName$

presented better performance by improving the F1-measure in ten tasks by using the HAN classifier in comparison to the $RandomSynName$ [116] method. In general, it can be concluded that incremental methods demonstrated better performance in the majority of the tasks by using CNN and RNN.

5.6 Further analysis

For further analyzing $SynNameSim$ and $IncSynNameSim$ approaches, we compared the obtained experimental results by three different machine learning models with five different methods. The compared methods are divided into two groups. Kappa [167], Solt [155] and Yao [192] used rule-based approaches to classify i2b2 (2008) data set [58]. Meanwhile, Ambert [12] and Garla [58] used automatic feature engineering methods to solve the i2b2 (2008) challenge [58]. Our methods $SynNameSim$ and $IncSynNameSim$ utilizes an automatic system to enrich the data set.

Table 5.13 compares the statistical results of the $SynNameSim$ and $IncSynNameSim$ with the other five methods. Our methods with three DL models show better performance in seven tasks (CAD, CHF, GERD, OSA, OA, PVD and Venous Insufficiency). RNN and HAN achieved higher F1-measure than other employed machine learning models. This performance is impressive considering that our method is fully automatic without using rules or feature engineering. CNN classifier improved the classification performance from 81.60 to 82.34 for Venous Insufficiency disease. Furthermore, all of the classifiers achieved better F1-measure values for six diseases (CAD, CHF, GERD, OSA, OA and PVD) in comparison with other approaches.

Table 5.14 presents a comparison on a test set of i2b2 (2008) and i2b2 (2010) [168] data sets for all of the proposed approaches in this thesis. The best results are highlighted in bold. By analyzing the obtained results, it can be concluded that the $SynName + SciName$ approach with RNN

Table 5.10: Comparison of CNN classification performance (F1-measure) and standard deviation averages using 30 independent runs for i2b2 2008 dataset [58]. The significant test is for the all of the methods against their corresponding approaches (Wilcoxon Test, $\alpha = 0.05$)

Methods	HighSimSynName Balanced	IncHighSimSynName Balanced Doubled		RandomSynName Balanced	IncRandomSynName Balanced Doubled		SynNameSim Balanced	IncSynNameSim Balanced Doubled	
Diseases	Ave±Std (Best)	Ave±Std (Best)	T	Ave±Std (Best)	Ave±Std (Best)	T	Ave±Std (Best)	Ave±Std (Best)	T
Asthma	82.57±6.731 (88.37)	**87.43±4.02 (88.78)**	+	64.12±18.881 (82.49)	**87.42±1.27 (89.34)**	+	87.15±0.772 (88.37)	**88.71±0.15 (88.78)**	+
CAD	90.57±1.244 (92.28)	90.63±2.38 (94.60)	=	89.59±2.103 (92.67)	**90.80±1.79 (93.11)**	**+**	91.21±0.595 (92.99)	91.10±0.69 (92.45)	=
CHF	89.62±1.796 (92.61)	**90.51±1.63 (92.62)**	+	89.66±2.254 (92.84)	**90.09±2.90 (93.28)**	**+**	91.33±0.434 (92.61)	91.45±1.06 (93.08)	=
Depression	44.07±0.590 (46.47)	**53.06±7.28 (71.00)**	+	44.26±0.657 (46.62)	44.68±1.73 (51.38)	=	44.62±0.925 (48.09)	**45.79±2.78 (53.65)**	+
Diabetes	**89.68±3.240 (93.71)**	86.25±6.12 (93.55)	-	**89.12±4.247 (94.52)**	86.28±4.03 (92.46)	-	**92.87±0.926 (94.80)**	88.01±4.35 (93.52)	-
Gallstones	58.81±11.465 (79.77)	**82.47±0.72 (83.61)**	+	65.23±14.853 (82.22)	**80.66±4.70 (82.93)**	+	67.40±5.173 (76.85)	**82.59±0.37 (83.27)**	+
GERD	51.92±6.387 (67.84)	**72.33±4.68 (80.36)**	+	52.76±5.569 (63.16)	**73.83±2.57 (80.42)**	**+**	58.45±4.307 (65.43)	**72.90±2.90 (80.16)**	+
Gout	57.27±22.626 (86.28)	**89.33±1.19 (90.37)**	+	41.85±20.892 (77.21)	**87.63±2.32 (90.81)**	+	72.33±7.292 (83.12)	**87.14±1.97 (90.22)**	+
Hypercholesterolemia	65.96±3.108 (71.67)	**77.02±4.16 (84.03)**	**+**	65.46±2.547 (69.56)	**73.24±5.28 (82.55)**	+	68.03±1.063 (70.76)	**73.91±3.73 (81.25)**	+
Hypertension	53.84±4.626 (61.85)	53.33±7.95 (75.13)	=	**54.22±5.111 (63.08)**	51.24±5.86 (66.30)	-	**60.99±3.512 (70.27)**	58.16±8.64 (76.79)	-
Hypertriglyceridemia	38.02±18.338 (53.62)	**56.76±1.45 (62.68)**	+	25.60±19.670 (51.84)	**55.06±3.62 (64.45)**	+	50.34±2.451 (57.09)	**54.62±2.81 (62.17)**	+
Obesity	**90.90±1.093 (91.85)**	86.66±10.89 (91.91)	-	**90.64±0.745 (91.60)**	88.77±6.91 (92.31)	-	**91.50±0.151 (91.87)**	90.73±1.47 (92.06)	-
OSA	58.74±27.907 (92.03)	**89.62±4.72 (93.36)**	+	49.13±25.798 (85.53)	**90.64±3.40 (92.83)**	**+**	77.21±8.843 (90.76)	**91.42±0.84 (92.30)**	+
OA	50.88±6.418 (67.73)	**82.29±2.65 (85.85)**	+	48.49±3.563 (55.19)	**62.33±12.19 (82.05)**	**+**	52.37±4.205 (64.33)	**82.42±2.73 (85.89)**	+
PVD	71.70±10.524 (88.51)	**88.61±2.05 (84.02)**	+	61.18±14.006 (83.83)	**87.13±2.75 (91.00)**	**+**	76.24±5.936 (85.22)	**89.13±1.34 (91.13)**	+
Venous Insufficiency	54.52±12.924 (71.25)	**74.81±4.42 (83.76)**	+	48.94±18.937 (69.60)	**74.68±4.36 (82.34)**	+	64.72±3.788 (71.32)	**74.85±3.20 (82.34)**	+

Table 5.11: Comparison of RNN classification performance (F1-measure) and standard deviation averages using 30 independent runs for i2b2 2008 dataset [58]. The significant test is for the all of the methods against their corresponding Incremental approaches (Wilcoxon Test, $\alpha = 0.05$)

Methods	HighSimSynName Balanced	IncHighSimSynName Balanced Doubled		RandomSynName Balanced	IncRandomSynName Balanced Doubled		SynNameSim Balanced	IncSynNameSim Balanced Doubled	
Diseases	Ave±Std (Best)	Ave±Std (Best)	T	Ave±Std (Best)	Ave±Std (Best)	T	Ave±Std (Best)	Ave±Std (Best)	T
Asthma	86.41±3.445 (89.34)	**88.29±1.56 (89.75)**	+	85.05±10.250 (89.64)	**88.46±1.11 (90.71)**	+	**89.24±0.641 (90.99)**	88.86±0.38 (89.75)	-
CAD	94.77±1.096 (96.62)	**95.41±1.35 (97.95)**	+	94.50±1.108 (96.61)	94.96±0.64 (96.37)	=	95.24±0.498 (96.85)	95.43±0.69 (97.04)	=
CHF	90.12±1.220 (92.82)	**92.05±0.98 (93.98)**	+	90.18±1.232 (92.36)	**91.45±1.19 (93.76)**	+	91.17±0.672 (92.84)	91.82±0.81 (93.98)	=
Depression	44.98±3.028 (54.04)	**48.65±9.74 (83.84)**	+	**45.13±3.850 (61.15)**	44.36±1.10 (47.49)	-	**50.70±6.358 (71.94)**	44.41±1.25 (49.30)	-
Diabetes	92.42±1.400 (94.52)	**94.36±0.76 (95.72)**	+	91.51±1.619 (94.04)	**93.74±1.27 (95.77)**	+	93.27±0.616 (94.35)	**94.59±0.74 (96.33)**	+
Gallstones	**83.05±7.009 (86.70)**	77.35±12.43 (86.98)	-	**81.30±9.688 (86.87)**	71.30±16.70 (85.83)	-	**85.87±0.794 (87.75)**	75.37±14.46 (87.04)	-
GERD	**79.97±1.428 (83.19)**	71.63±11.80 (82.58)	-	**79.77±1.589 (82.85)**	70.90±12.22 (83.50)	-	**81.38±0.800 (83.36)**	73.60±11.21 (83.48)	-
Gout	**89.44±2.196 (90.95)**	87.11±7.87 (92.20)	-	**86.15±7.360 (90.95)**	77.44±17.15 (92.08)	-	89.92±0.587 (91.37)	89.70±1.22 (91.52)	=
Hypercholesterolemia	71.29±4.405 (79.02)	**81.76±2.37 (88.21)**	+	71.92±4.939 (79.15)	**79.17±4.22 (85.17)**	+	74.16±2.580 (79.97)	**82.16±2.21 (86.43)**	+
Hypertension	**73.76±9.123 (83.24)**	52.17±12.84 (84.44)	-	**75.99±5.349 (83.16)**	51.63±9.74 (44.97)	-	**81.27±1.492 (83.72)**	49.65±7.41 (79.98)	-
Hypertriglyceridemia	49.00±0.995 (52.57)	**57.26±8.47 (76.12)**	+	48.61±0.108 (48.68)	**56.72±9.85 (86.85)**	+	49.17±1.037 (52.58)	**57.23±9.49 (78.36)**	+
Obesity	**91.32±0.207 (91.84)**	90.68±1.03 (92.09)	-	91.27±0.357 (91.87)	91.69±0.63 (90.17)	=	91.46±0.190 (92.08)	91.22±0.52 (92.12)	=
OSA	95.03±2.296 (97.83)	**97.28±0.94 (99.13)**	+	89.65±12.40 (97.38)	**97.22±0.87 (99.14)**	+	96.28±1.042 (98.25)	**97.11±0.76 (99.13)**	+
OA	**81.93±7.420 (88.04)**	79.94±2.07 (88.86)	-	**81.64±6.732 (88.94)**	77.89±14.69 (90.40)	-	85.10±1.105 (87.72)	**86.90±1.88 (90.63)**	+
PVD	88.86±1.512 (91.26)	88.85±3.19 (91.62)	=	87.19±8.573 (91.56)	**88.38±3.75 (92.34)**	+	**90.14±0.825 (92.03)**	89.80±1.27 (92.34)	-
Venous Insufficiency	66.78±5.652 (70.34)	**70.73±4.27 (76.74)**	+	63.53±5.747 (70.34)	**70.32±6.97 (77.02)**	+	68.46±2.028 (71.79)	**71.27±3.30 (79.97)**	+

Table 5.12: Comparison of HAN classification performance (F1-measure) and standard deviation averages using 30 independent runs for i2b2 2008 dataset [58]. The significant test is for the all of the methods against their corresponding approaches (Wilcoxon Test, $\alpha = 0.05$)

Methods	HighSimSynName Balanced	IncHighSimSynName Balanced Doubled		RandomSynName Balanced	IncRandomSynName Balanced Doubled		SynNameSim Balanced	IncSynNameSim Balanced Doubled	
Diseases	Ave±Std (Best)	Ave±Std (Best)	T	Ave±Std (Best)	Ave±Std (Best)	T	Ave±Std (Best)	Ave±Std (Best)	T
Asthma	86.25±2.233 (89.60)	**88.18±0.94 (89.60)**	+	85.79±2.125 (89.78)	**87.30±2.13 (89.60)**	+	88.90±0.338 (89.90)	88.86±0.67 (89.90)	=
CAD	94.85±1.181 (96.85)	94.84±1.08 (96.84)	=	94.82±1.142 (97.29)	94.91±1.09 (96.83)	=	95.60±0.536 (97.06)	95.38±0.55 (96.39)	=
CHF	89.76±1.354 (92.37)	**91.61±1.03 (93.50)**	+	89.89±1.178 (91.91)	**91.08±1.44 (93.53)**	+	90.75±1.045 (93.98)	**91.44±0.83 (93.31)**	+
Depression	**46.92±4.676 (57.37)**	43.85±0.18 (44.81)	-	**48.38±5.081 (57.12)**	43.97±0.65 (47.37)	-	**53.37±1.428 (57.87)**	43.94±0.53 (46.66)	-
Diabetes	92.43±1.210 (94.20)	**93.63±1.25 (95.57)**	+	92.74±1.207 (95.34)	**93.52±1.24 (96.56)**	+	93.54±0.670 (95.36)	93.85±0.96 (95.29)	=
Gallstones	**82.77±7.062 (87.75)**	68.29±17.29 (83.49)	-	**78.64±11.463 (86.17)**	64.17±17.51 (82.70)	-	**85.15±1.275 (87.23)**	72.95±14.32 (83.96)	-
GERD	**60.31±11.574 (80.26)**	56.35±12.52 (80.76)	-	61.98±11.164 (80.16)	**64.32±11.48 (80.66)**	+	**71.33±7.219 (81.48)**	61.99±12.35 (83.13)	-
Gout	89.74±0.497 (90.81)	89.45±1.28 (90.95)	=	85.00±9.377 (90.52)	**86.78±8.61 (92.20)**	+	**90.01±0.389 (91.52)**	89.16±1.02 (91.52)	-
Hypercholesterolemia	66.48±2.623 (73.45)	**82.72±2.94 (87.55)**	+	67.40±3.425 (74.22)	**79.17±3.71 (85.32)**	+	68.37±1.677 (73.33)	**82.32±2.40 (86.88)**	+
Hypertension	**59.10±12.21218 (79.65)**	48.15±7.02 (79.52)	-	**63.35±12.319 (80.67)**	51.02±6.96 (78.01)	-	**72.56±7.047 (82.87)**	52.11±10.51 (82.43)	-
Hypertriglyceridemia	48.68±1.421e-14 (48.68)	**52.69±2.69 (56.19)**	+	48.80±0.701 (52.58)	**51.66±5.01 (68.05)**	+	49.06±1.150 (52.58)	**51.37±2.40 (59.55)**	+
Obesity	91.11±0.647 (91.824)	91.10±0.52 (91.82)	=	91.33±0.260 (92.06)	91.11±1.60 (92.36)	=	91.43±0.173 (92.17)	91.37±0.27 (92.11)	=
OSA	92.77±3.540 (97.44)	**96.50±1.55 (99.13)**	+	93.17±2.958 (97.34)	**96.98±1.70 (98.70)**	+	94.19±1.396 (96.88)	**96.94±1.08 (99.57)**	+
OA	**82.44±4.228 (88.84)**	74.79±12.70 (89.33)	-	**79.12±7.086 (86.76)**	66.02±17.24 (87.83)	-	**84.83±1.472 (88.34)**	80.99±8.94 (88.64)	-
PVD	**88.26±1.524 (90.71)**	86.70±8.15 (91.10)	-	88.80±0.993 (90.57)	**89.58±1.57 (92.25)**	+	89.97±0.821 (91.56)	89.83±1.23 (92.77)	=
Venous Insufficiency	60.35±5.051 (66.41)	**65.54±4.66 (74.73)**	+	60.16±4.869 (68.10)	**68.18±4.25 (76.39)**	+	64.86±2.457 (69.70)	**67.55±3.00 (74.73)**	+

Table 5.13: Macro-averaged F1 on I2B2 2008 test set. Best scores from our study indicated in bold.

Methods / Diseases	Kappa [167]	Solt [155]	Yao [192]	Ambert [12]	Superlin [58]	SynNameSim (CNN) Balanced	SynNameSim (RNN) Balanced	SynNameSim (HAN) Balanced	IncSynNameSim (CNN) Balanced Doubled	IncSynNameSim (RNN) Balanced Doubled	IncSynNameSim (HAN) Balanced Doubled
Asthma	76.00	**97.84**	**97.84**	97.00	97.00	88.37	90.99	89.90	88.78	89.75	89.90
CAD	81.00	60.22	62.33	63.00	61.80	**92.99**	**96.85**	**97.06**	**92.45**	**97.04**	**96.39**
CHF	74.00	62.36	62.36	61.20	61.20	**92.61**	**92.84**	**93.98**	**93.08**	**93.98**	**93.31**
Depression	86.00	93.46	96.02	93.50	**97.90**	48.09	71.94	57.87	53.65	49.30	46.66
Diabetes	87.00	96.82	**97.31**	91.50	96.00	94.80	94.35	95.36	93.52	96.33	95.29
Gallstones	90.00	**97.29**	96.89	96.10	95.00	76.85	87.75	87.23	83.27	87.04	83.96
GERD	59.00	57.68	57.68	57.90	57.90	**65.43**	**83.36**	**81.48**	**80.16**	**83.48**	**83.13**
Gout	92.00	97.71	97.71	98.10	**98.20**	83.12	91.37	91.52	90.22	91.52	91.52
Hypercholesterolemia	68.00	90.53	91.13	**91.20**	90.80	70.76	79.97	73.33	81.25	86.43	86.88
Hypertension	67.00	88.51	92.40	89.90	**92.90**	70.27	83.72	82.87	76.79	79.98	82.43
Hypertriglyceridemia	72.00	79.81	70.92	87.60	**92.80**	57.09	52.58	52.58	62.17	78.36	59.55
Obesity	86.00	97.24	**97.47**	97.30	97.20	91.87	92.08	92.17	92.06	92.12	92.11
OSA	92.00	88.05	88.05	65.30	65.60	90.76	**98.25**	**96.88**	**92.30**	**99.13**	**99.57**
OA	76.00	62.86	63.07	63.10	60.40	64.33	**87.72**	**88.34**	**85.89**	**90.63**	**88.64**
PVD	73.00	63.48	63.14	62.30	60.60	**85.22**	**92.08**	**91.56**	**91.13**	**92.34**	**92.77**
Venous Insufficiency	44.00	80.83	80.83	72.50	81.60	71.32	71.79	69.70	**82.34**	79.97	74.73

shows better performance by achieving better classification F1-measure in four tasks (Gallstones, Gout, Hypertension, and CAD-i2b2(2010) [168]). This method uses WordNet dictionary to provide synonyms for general vocabulary and UMLS ontology to provide scientific names for domain-specific terms to utilize for data augmentation. Utilizing UMLS and WordNet together enables this approach to show better classification performance.

5.7 Chapter Summary

Building practical ML models that fit a business case for various segments of users, and scale to maximize user needs is challenging, especially when existing enterprise datasets may just represent a microcosm of the actual population of customers. For example, a model trained on insurance customers in a specific age group/segment does not predict outcomes for a different age group/segment. Synthetic features, smart data augmentations, and event simulations are solutions that help address these issues, and provide the capacity to train models that are able to provide better coverage and scale. In this regard, algorithms for the generation of "synthetic phrases" to augment text datasets are developed in chapters 4 and 5. Furthermore, the suggested approaches in chapter 5 can deal with imbalance data issue in medical and general areas.

This chapter has proposed a new dictionary-based oversampling method and a new incremental method to augment clinical data for solving binary medical document classification. The methods aim to produce new documents from original documents by replacing words with their synonyms. The proposed approaches use the WordNet dictionary to extract synonyms of words. Then, three domain specific English embedding models are employed to measure similarities of the extracted synonyms to find a suitable synonym to replace a word in the main sentence to construct new documents and deal with the data shortage issue in medical docu-

Table 5.14: Macro-averaged F1 on test set for all of the proposed approaches in the thesis. Best scores are highlithed in bold.

Methods	Two-Stage	Three-Stage	SciName			SynName+SciName			SynNameSim			IncSynNameSim CNN		
Diseases	SVM		CNN	RNN	HAN	CNN	RNN	HAN	CNN	RNN	HAN	CNN	RNN	HAN
	Chapter 3 - Objective 1		Chapter 4 - Objective 2			Chapter 4 - Objective 2			Chapter 5 - Objective 3			Chapter 5 - Objective 3		
Asthma	**94.50**	94.00	87.17	90.05	90.45	89.78	90.44	90.58	88.37	90.99	89.90	88.78	89.75	89.90
CAD	80.06	80.69	92.19	96.62	96.14	93.64	96.37	96.61	92.99	96.85	**97.06**	92.45	97.04	96.39
CHF	90.05	90.54	92.85	**94.92**	93.31	93.30	93.26	93.54	92.61	92.84	93.98	93.08	93.98	93.31
Depression	47.20	45.46	74.17	79.10	76.92	72.57	82.00	90.94	48.09	71.94	57.87	53.65	49.30	46.66
Diabetes	89.88	89.66	93.83	94.76	95.31	94.54	97.00	97.26	94.80	94.35	95.36	93.52	96.33	95.29
Gallstones	58.02	60.00	85.81	87.90	86.17	87.75	**88.34**	88.20	76.85	87.75	87.23	83.27	87.04	83.96
GERD	**86.04**	**86.04**	70.29	82.49	82.28	82.11	83.67	83.65	65.43	83.36	81.48	80.16	83.48	83.13
Gout	90.06	90.06	89.11	93.71	89.11	91.65	**95.83**	95.27	83.12	91.37	91.52	90.22	91.52	91.52
Hypercholesterolemia	84.18	83.24	80.93	85.32	86.15	84.84	86.65	85.11	70.76	79.97	73.33	81.25	86.43	**86.88**
Hypertension	72.99	68.83	72.62	85.26	84.42	81.44	**88.47**	87.38	70.27	83.72	82.87	76.79	79.98	82.43
Hypertriglyceridemia	59.55	59.55	48.68	53.85	55.92	48.68	56.73	52.38	57.09	52.58	52.58	62.17	**78.36**	59.55
Obesity	95.15	**95.61**	91.84	92.39	93.27	92.08	94.47	93.51	91.87	92.08	92.17	92.06	92.12	92.11
OSA	97.77	98.20	94.89	98.20	97.83	95.32	98.27	99.14	90.76	98.25	96.88	92.30	99.13	**99.57**
OA	86.23	87.13	74.73	87.42	88.24	82.38	90.70	**91.32**	64.33	87.72	88.34	85.89	90.63	88.64
PVD	93.73	**94.26**	92.34	92.23	91.92	92.00	92.23	92.34	85.22	92.08	91.56	91.13	92.34	92.77
Venous Insufficiency	57.79	54.81	59.78	76.04	71.86	62.76	81.63	78.22	71.32	71.79	69.70	**82.34**	79.97	74.73
CAD - i2b2 (2010)	90.79	91.57	95.10	96.92	96.87	97.44	**98.12**	97.51	92.96	94.87	98.10	91.90	90.86	96.21

ment classification. The proposed approaches are able to increase the classification precision in the neural network models. Experimental results shown as for F1-measure indicate that the proposed dictionary-based approaches ($SynNameSim$ and $IncSynNameSim$) can improve the performance of the CNN, RNN and HAN models by considering the existing similarities among a word and its synonyms to generate more samples in the training phase for all of classes.

This chapter shows promise in using a dictionary-based data oversampling technique in medical document classification, however, there is still some room for more research to improve the classification performance. Different variations of dictionary-based data augmentation and oversampling for clinical discharge notes could be investigated. We can also take some additional criteria into account to reduce the ambiguity in the produced documents. Moreover, insightful detailed analysis could be conducted including different evaluation metrics.

Chapter 6

Conclusions

This chapter makes conclusions for the discussions of this thesis and highlights the proposed contributions and possible research ideas for the future work.

The overall goal of this thesis was to develop new feature engineering approaches to medical document classification by utilizing PSO for feature selection, a domain-specific ontology and a WordNet dictionary for feature extraction and data augmentation. The proposed methods are able to automatically extract features, construct meaningful features, select prominent features and make new synthetic instances from the original medical documents which can improve medical document classification performance. The goal has been achieved by proposing new feature engineering approaches which consider the concept of words and just two relevant concepts (disease names and disease symptoms) to select and extract new high-level features which help to discriminate documents with higher performance. The proposed approaches have been evaluated on three real-world medical document data sets and compared with related methods. The obtained experimental results show that the suggested approaches in this thesis have achieved better classification performance than the related approaches.

The rest of this chapter summarizes conclusions for the proposed con-

tributions and highlights the important findings from each objective and introduces some possible research ideas as a future work.

6.1 Achieved Objectives

The main goal of this thesis was to propose some new feature engineering approaches to medical document classification by automatically extracting meaningful features, selecting prominent features, constructing new high-level features and making new synthetic documents by targeting concepts of words and expressions which can improve the quality of training data set and increase the precision of medical classification notes. The following objectives have been achieved in this thesis:

- Develop three different feature engineering approaches including feature extraction, feature selection and feature construction for medical document classification. The first approach is an ontology-based feature extraction, whereas the second approach is a wrapper-based feature selection method and the third approach is a feature construction approach. The first method extracts important features from raw text (stage-1). Then, feature construction is applied on the extracted features from stage-1 (stage-2). Finally, feature selection is employed to eliminate redundant features from the extracted features in stage-1 and the constructed features in stage-2. The selected features are utilized to form a vector matrix which represent documents with the selected features. The matrix is given to a classifier for training a model to reuse in medical document classification. The three-stage approach achieves better classification performance than the two-stage approach. All of the proposed approaches can solve binary and multi-class classification problems. The suggested methods have shown better classification performance compared to using all features set.

- Proposes two novel ontology-based data augmentation approaches to make new synthetic documents from the original training data sets for medical document classification. The first method utilizes an ontology-based approach, whereas the second one uses a combined approach (from a dictionary-based method and the ontology-based method) and provides better results than the ontology-based approach. These approaches can make new synthetic documents from the original data set by employing a domain-specific ontology and a general dictionary to double/triple the size of the training data set and improve the performance of medical document classification. These approaches have successfully shown improved medical document classification performance and outperformed the related methods. From the experimental results, it can be concluded that the contribution of the produced documents by ontology-based method should be more than the dictionary-based method in the proposed combination method. It may help to improve those diseases the ontology-based showed better performance than combined approach.

- Develop two dictionary-based data augmentation approaches to make new synthetic documents from the original training data sets for medical document classification problems. The first approach is using a WordNet dictionary to find all of the synonyms of the words in documents. Then, three different pretrained models are employed to measure the similarities between the found synonyms and their corresponding word. A new synonym selection approach is proposed to select a synonym by considering its similarity value during the selection process. The proposed approach is able to make new synthetic documents with high variety compared to the similar methods. It makes an imbalanced data set balanced and improves the classification performance too. To analyze the potential of the proposed synonym selection approach, it is applied on all of the classes

of a data set to double size of the large class in the same data set. The experiment results demonstrate better classification performance in comparison to the related methods.

6.2 Main Conclusions

This thesis finds that feature engineering is a promising direction to address the issue of medical documents for classification by automatically extracting features from documents and selecting prominent features and constructing new and informative features. Most of the suggested approaches in this thesis showed better classification performance than the related works. All of the suggested methods in this thesis are targeted to improve medical document classification performance by using feature manipulation and enriching the quality of the datasets. As the labels of the targeted datasets are names of diseases, the trained models by using the proposed approaches predict disease of an unseen medical document and consequently assist doctors in making final decisions on a patient's health by considering the model prediction for the patient's document. The conclusions of the proposed contribution chapters (including from chapter 3 to 5) are drawn and discussed in this section.

6.2.1 Feature Engineering Approaches to Improving Medical Document Classification

Chapter 4 proposes three feature engineering approaches including feature extraction, feature selection and feature construction methods, where PSO-based feature selection method is a wrapper approach. The proposed approaches from three stages. In the first stage, the feature extraction method extracts features based on their concepts. In the second stage, different pairs of concepts are constructed to make new features. Finally, PSO is applied to select prominent and informative features to improve

the performance of medical document classification.

Knowledge-guided Feature Extraction

In this thesis a knowledge-guided feature extraction method is proposed to diminish the number of features and at the same time to keep the meaningful features to decline the irrelevant and noisy features. The MetaMap tool is utilized to map the meaningful phrases to their concepts in UMLS and then, the related concepts to the problem's domain are selected to be used as features instead of original documents. Conceptualization significantly reduced the number of features and selecting "Disease or Syndrome" and "Sign or Symptom" concepts, which cover the domain of the targeted notes, decreased the unrelated features and increased the precision of the classification. Experimental and statistical results exhibit that the proposed technique is more effective than utilizing classification based on all raw features in some tasks and shows better performance in comparing with existing feature selection/extraction methods. Moreover, in comparison to the routine method of classification, the suggested approach can accomplish better classification performance.

Two-Stage Feature Engineering

This thesis finds that a two-stage wrapper based feature extraction and feature selection method achieves better medical document classification performance than the traditional approach of feature selection. The suggested two-stage approach investigates domain concepts and determines which concepts are discriminative to a classification problem. It is able to extract meaningful features from the document set and reduce the number of the features. Moreover, the two-stage approach improves the classification accuracy in the majority of the candidate classifiers by using a small size of feature subset. Experimental and statistical results illustrate that the proposed method can achieve significantly better classification

performance for some diseases in comparing with existing feature selection/extraction methods.

Three-Stage Feature Engineering

This thesis proposes a three-stage wrapper based feature extraction, feature construction and feature selection method to utilise domain concepts and their relations to enrich the input data for medical document classification problems. Since there are redundant or irrelevant features, an ontology-based method is applied to extract more related features to the candidate problem by considering concepts of features. After eliminating some of the redundant or irrelevant features, new high-level features are constructed from the extracted feature from the first level. Finally, a wrapper PSO method is applied to remove redundant or irrelevant features from the extracted features from the fist stage and constructed features from the second stage. The proposed approach is able to improve the quality of the input data set by constructing new high-level features and increase the classification performance 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 performance for some diseases in comparison with existing feature selection/extraction methods.

6.2.2 Ontology-based Data Augmentation Approaches to Improve Medical Document Classification

Chapter 5 proposes two data augmentation approaches including an ontology-based and a combined approach. Both of the methods increase the size of the training data set by making new synthetic data to provide more data in training a classifier. The ontology-based approach is using domain-specific knowledge to make new instances and the combined method is a combination of an ontology-based and a dictionary-based method. These

methods help a classifier to learn the differences between classes of a targeted data set and improve medical document classification performance.

Ontology-guided Data Augmentation

Our ontology-based data augmentation found that employing MetaMap to extract more related features from UMLS in augmenting new data can improve the medical classification performance. The ontology-based approach is using UMLS to find scientific names of all of the words and expressions belonging to a concept and replacing the corresponding words and expressions with their scientific names in each document to make new synthetic documents. This method doubles the size of the training data set. The experimental results show that the proposed approach can increase the classification performance in comparison with the related approaches. This shows that domain-specific knowledge of the problem is helpful in augmenting new instances.

Combined Data Augmentation

A combined data augmentation approach to build new instances is proposed in this thesis. It is found that combining the ontology-based data augmentation with a dictionary-based data augmentation can enable the approach to produce more instances by targeting general aspects and domain-specific aspects of contents at the same time. This approach triples the size of the training data set which provides more data than the proposed ontology-based method. The experimental results show that the proposed combined method achieves better classification performance than the existing approaches for the i2b2 (2008) challenge. The reason is that this method provides more training data sets and targeting general and domain-specific aspects to provide new synthetic documents with high variety.

6.2.3 Dictionary-based Data Augmentation Approaches to Improve Medical Document Classification

Chapter 6 proposes two data augmentation approaches including a dictionary-based and an incremental approach. Both of the approaches increase the size of the training data set by using the extracted synonyms of words from a WordNet dictionary. The proposed methods in chapter 5 have limitations in increasing the size of the training data set. They can double/triple the size of the original data set. Hence, the incremental approach is proposed to increase the size of the data set as much as what is required. These methods make an imbalanced data set balanced, increase the size of it and improve medical document classification performance.

Dictionary-based Data Augmentation

A new dictionary-based oversampling method to balance an imbalanced data set is proposed in this thesis. It is found that employing three domain specific English embedding models to measure similarities to find a suitable synonym to replace a word in the main sentence can improve the medical document classification performance. This method is able to augment diverse instances from the original instance. Experimental results for F1-measure indicate that the proposed approach can improve the performance of the CNN, RNN and HAN models by utilizing the proposed dictionary-based approach to generate more samples in the training phase for the minority class.

Incremental Data Augmentation

In this thesis, the proposed dictionary-based approach is applied on all of the classes' documents to make the classes balanced and at the same time increase the size of the classes. The method is augmenting all of the classes by increasing their size equal to the double size of the large class. It found

that the proposed method is capable of providing more synthetic documents for all of the data in the classes by using the suggested synonym selection method. By analyzing the potential of the proposed approach in making new instances, it can be concluded that the approach improves the medical document classification performance by balancing and increasing the size of all of the classes' data.

6.3 Future Work

In this section, some possible ideas are presented as future work.

6.3.1 Relation-guided Feature Construction Approach for Medical Document Classification

In this thesis, some feature engineering approaches are proposed to extract and construct meaningful features by considering the concept of the words and expressions. In the construction approach, the relations between extracted concepts (diseases and symptoms) is not considered. The performance of classification may be raised by considering relations between diseases and symptoms and contributing to the ones that are interconnected as pairs [52]. Hence, there are better ways to construct features for the second stage (feature construction) by analyzing the distance of the detected features in the document to guide our feature construction method in making pairs.

6.3.2 Relation-guided Data Augmentation Approach for Medical Document Classification

In this thesis, an ontology-guided data augmentation is proposed by replacing the scientific names of those words and expressions which belong to a concept. The limitation of this method in augmenting new documents

is the number of the available scientific names to replace with words and expressions. This limitation can only produce one new synthetic document for each original document. It is better to develop a method that can produce more than one synthetic document from each original one.

To achieve this goal, extracting the parent and children of the corresponding scientific name to the word from UMLS can provide more options to replace with the original word in augmenting new synthetic documents. It is expected that the proposed approach will improve medical document classification performance by making more synthetic documents from each original document and providing more instances for training machine learning methods.

6.3.3 Rule-based Data Augmentation for Medical Document Classification

In this thesis, two different data augmentation (ontology-based and dictionary-based) are proposed for medical document classification. The created documents are useful to feed as an input to a machine learning method and they have improved the classification performance. However, some of the created sentences could have ambiguous in meaning. Hence, we can take some additional criteria into account to reduce the ambiguity in the produced documents.

6.3.4 Data Augmentation by Using Autoencoders

In this thesis some data augmentation approaches are proposed to deal with lack of data in the medical area. All of the proposed methods are augmenting data in data space which has its own limitations. These methods are limited to the data space and the suggested methods should be applicable for the targeted data space. For example, those methods suggested for documents are not easy to apply for image data and vice versa.

Hence, it is better to develop a generic augmentation method that does not rely on data type and can be applied on different data types.

As all types of input data in machine learning methods can be presented in numbers, developing an augmentation approach in number space can be usable for different types of data such as text and image. Autoencoders can help to achieve this goal. Autoencoders are able to produce similar presented numbers for the original data set by preserving the same semantic meaning of the original data. This approach will have potential to augment different versions of presentation for any text or image by keeping the same semantic meaning of the original data. Then the produced new synthetic presentations with the original presentation of the data will feed to the candidate classifier for training. It is expected the proposed approach will improve classification performance.

Bibliography

[1] Unified medical language system (umls®), http://www.nlm.nih.gov/research/umls/. last updated 20 apr 2016.

[2] ABDOLLAHI, M., GAO, X., MEI, Y., GHOSH, S., AND LI, J. Uncovering discriminative knowledge-guided medical concepts for classifying coronary artery disease notes. In *Australasian Joint Conference on Artificial Intelligence* (2018), Springer, pp. 104–110.

[3] ABDOLLAHI, M., GAO, X., MEI, Y., GHOSH, S., AND LI, J. An ontology-based two-stage approach to medical text classification with feature selection by particle swarm optimisation. In *2019 IEEE Congress on Evolutionary Computation(CEC)* (2019), IEEE, pp. 1–8.

[4] ABDOLLAHI, M., GAO, X., MEI, Y., GHOSH, S., AND LI, J. Stratifying risk of coronary artery disease using discriminative knowledge-guided medical concept pairings from clinical notes. In *Pacific Rim International Conference on Artificial Intelligence* (2019), Springer, pp. 457–473.

[5] ABDOLLAHI, M., GAO, X., MEI, Y., GHOSH, S., AND LI, J. A dictionary-based oversampling approach to clinical document classification on small and imbalanced dataset. In *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)* (2020), IEEE, pp. 357–364.

[6] ABDOLLAHI, M., GAO, X., MEI, Y., GHOSH, S., AND LI, J. Ontology-guided data augmentation for medical document classification. In *International Conference on Artificial Intelligence in Medicine* (2020), Springer, pp. 78–88.

[7] ACHARYA, U. R., NG, E. Y.-K., TAN, J.-H., AND SREE, S. V. Thermography based breast cancer detection using texture features and support vector machine. *Journal of medical systems 36*, 3 (2012), 1503–1510.

[8] AGARWAL, B., AND MITTAL, N. Text classification using machine learning methods-a survey. In *Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012* (2014), Springer, pp. 701–709.

[9] AGGARWAL, C. C., AND ZHAI, C. A survey of text classification algorithms. In *Mining text data*. Springer, 2012, pp. 163–222.

[10] ALPAYDIN, E. Introduction to machine learning. *MIT press* (2009), 1–673.

[11] AMALDI, E., AND KANN, V. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science 209*, 1-2 (1998), 237–260.

[12] AMBERT, K. H., AND COHEN, A. M. 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 16*, 4 (2009), 590–595.

[13] AMELIO, L., AND AMELIO, A. New frontiers in document classification with applications to medical context. In *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (2019), IEEE, pp. 1665–1670.

[14] ANTHIMOPOULOS, M., CHRISTODOULIDIS, S., EBNER, L., CHRISTE, A., AND MOUGIAKAKOU, S. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. *IEEE transactions on medical imaging 35*, 5 (2016), 1207–1216.

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

[16] ASSEGIE, T. A. An optimized k-nearest neighbor based breast cancer detection. *Journal of Robotics and Control (JRC) 2*, 3 (2021), 115–118.

[17] BAI, X., GAO, X., AND XUE, B. Particle swarm optimization based two-stage feature selection in text mining. In *2018 IEEE Congress on Evolutionary Computation (CEC)* (2018), IEEE, pp. 1–8.

[18] BANZHAF, W., NORDIN, P., KELLER, R. E., AND FRANCONE, F. D. *Genetic programming: an introduction: on the automatic evolution of computer programs and its applications*, vol. 1. Morgan Kaufmann San Francisco, 1998, pp.1–451.

[19] BEHERA, B., AND KUMARAVELAN, G. Performance evaluation of machine learning algorithms in biomedical document classification. *Performance evaluation 29*, 5 (2020), 5704–5716.

[20] BELLAZZI, R., AND ZUPAN, B. Predictive data mining in clinical medicine: current issues and guidelines. *International journal of medical informatics 77*, 2 (2008), 81–97.

[21] BENGIO, Y., DELALLEAU, O., ROUX, N., PAIEMENT, J.-F., VINCENT, P., AND OUIMET, M. Feature extraction: Foundations and applications, chapter spectral dimensionality reduction, 2003, pp. 1–765.

[22] BHANDARE, A., BHIDE, M., GOKHALE, P., AND CHANDAVARKAR, R. Applications of convolutional neural networks. *International Journal of Computer Science and Information Technologies* (2016), 2206–2215.

[23] BODENREIDER, O. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research 32*, suppl_1 (2004), D267–D270.

[24] BORRAJO, L., ROMERO, R., IGLESIAS, E. L., AND MAREY, C. R. Improving imbalanced scientific text classification using sampling strategies and dictionaries. *Journal of integrative bioinformatics 8*, 3 (2011), 90–104.

[25] BOUNDS, D. G., LLOYD, P. J., MATHEW, B., AND WADDELL, G. A multilayer perceptron network for the diagnosis of low back pain. In *Proc. IEEE Int. Conf. on Neural Networks* (1988), vol. 2, pp. 481–489.

[26] BRITZ, D. Understanding convolutional neural networks for nlp. *URL: http://www. wildml. com/2015/11/understanding-convolutional-neuralnetworks-for-nlp/(visited on 11/07/2015)* (2015, pp. 1-4).

[27] BUCHAN, K., FILANNINO, M., AND UZUNER, Ö. Automatic prediction of coronary artery disease from clinical narratives. *Journal of biomedical informatics 72* (2017), 23–32.

[28] CAGNONI, S., POLI, R., SMITH, G. D., CORNE, D., OATES, M., HART, E., LANZI, P. L., WILLEM, E. J., LI, Y., PAECHTER, B., ET AL. *Real-World Applications of Evolutionary Computing: EvoWorkshops 2000: EvoIASP, EvoSCONDI, EvoTel, EvoSTIM, EvoRob, and EvoFlight, Edinburgh, Scotland, UK, April 17, 2000 Proceedings*. Springer Science & Business Media, 2000, pp. 1–395.

[29] CAMOUS, F., BLOTT, S., AND SMEATON, A. F. Ontology-based medline document classification. In *Bioinformatics Research and Development*. Springer, 2007, pp. 439–452.

[30] CAVNAR, W. B., TRENKLE, J. M., ET AL. N-gram-based text categorization. *Ann arbor mi 48113*, 2 (1994), 161–175.

[31] CERVANTE, L., XUE, B., SHANG, L., AND ZHANG, M. A dimension reduction approach to classification based on particle swarm optimisation and rough set theory. In *Australasian Joint Conference on Artificial Intelligence* (2012), Springer, pp. 313–325.

[32] CERVANTE, L., XUE, B., SHANG, L., AND ZHANG, M. Binary particle swarm optimisation and rough set theory for dimension reduction in classification. In *Evolutionary Computation (CEC), 2013 IEEE Congress on* (2013), IEEE, pp. 2428–2435.

[33] CERVANTE, L., XUE, B., SHANG, L., AND ZHANG, M. A multi-objective feature selection approach based on binary pso and rough set theory. In *European Conference on Evolutionary Computation in Combinatorial Optimization* (2013), Springer, pp. 25–36.

[34] CERVANTE, L., XUE, B., ZHANG, M., AND SHANG, L. Binary particle swarm optimisation for feature selection: A filter based approach. In *Evolutionary Computation (CEC), 2012 IEEE Congress on* (2012), IEEE, pp. 1–8.

[35] CHAKRABORTY, B. Feature subset selection by particle swarm optimization with fuzzy fitness function. In *Intelligent System and Knowledge Engineering, 2008. ISKE 2008. 3rd International Conference on* (2008), vol. 1, IEEE, pp. 1038–1042.

[36] CHARBUTY, B., AND ABDULAZEEZ, A. Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends 2*, 01 (2021), 20–28.

[37] COLORNI, A., DORIGO, M., MANIEZZO, V., ET AL. Distributed optimization by ant colonies. In *Proceedings of the first European conference on artificial life* (1991), vol. 142, Paris, France, pp. 134–142.

[38] COULOMBE, C. Text data augmentation made simple by leveraging nlp cloud apis. *arXiv preprint arXiv:1812.04718* (2018, pp. 1–33).

[39] COVER, T., AND HART, P. Nearest neighbor pattern classification. *IEEE transactions on information theory 13*, 1 (1967), 21–27.

[40] DAGAN, I., KAROV, Y., AND ROTH, D. Mistake-driven learning in text categorization. *arXiv preprint cmp-lg/9706006* (1997, pp. 1–9).

[41] DASH, M., AND LIU, H. Feature selection for classification. *Intelligent data analysis 1*, 3 (1997), 131–156.

[42] DASH, M., AND LIU, H. Consistency-based search in feature selection. *Artificial intelligence 151*, 1-2 (2003), 155–176.

[43] DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research 7*, Jan (2006), 1–30.

[44] DEVASENAPATHY, K., AND DURAISAMY, S. Evaluating the performance of teaching assistant using decision tree id3 algorithm. *International Journal of Computer Applications 164*, 7 (2017), 23–27.

[45] DOLLAH, R. B., AND AONO, M. Ontology based approach for classifying biomedical text abstracts. *International Journal of Data Engineering (IJDE) 2*, 1 (2011), 1–15.

[46] EBERHART, R., AND KENNEDY, J. A new optimizer using particle swarm theory. In *Micro Machine and Human Science, 1995. MHS'95., Proceedings of the Sixth International Symposium on* (1995), IEEE, pp. 39–43.

[47] EBERHART, R. C., AND HU, X. Human tremor analysis using particle swarm optimization. In *Proceedings of the congress on evolutionary computation* (1999), IEEE Press Piscataway, NJ, pp. 1927–1930.

[48] EHRENTRAUT, C., EKHOLM, M., TANUSHI, H., TIEDEMANN, J., AND DALIANIS, H. Detecting hospital-acquired infections: a document classification approach using support vector machines and gradient tree boosting. *Health informatics journal 24*, 1 (2018), 24–42.

[49] EIBEN, A., AND SMITH, J. Introduction to evolutionary computing springer natural computing series, 2003, pp. 1–307.

[50] EIKVIL, L. Information extraction from world wide web-a survey. Tech. rep., Citeseer, 1999, pp. 1–40.

[51] EL ABOUDI, N., AND BENHLIMA, L. Review on wrapper feature selection approaches. In *2016 International Conference on Engineering & MIS (ICEMIS)* (2016), IEEE, pp. 1–5.

[52] ERNST, P., SIU, A., AND WEIKUM, G. Knowlife: a versatile approach for constructing a large knowledge graph for biomedical sciences. *BMC bioinformatics 16*, 1 (2015), pp. 1–13.

[53] ESPEJO, P. G., VENTURA, S., AND HERRERA, F. A survey on the application of genetic programming to classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 40*, 2 (2010), 121–144.

[54] FONG, S., DEB, S., YANG, X.-S., AND LI, J. Feature selection in life science classification: metaheuristic swarm search. *IT Professional*, 4 (2014), 24–29.

[55] FUREY, T. S., CRISTIANINI, N., DUFFY, N., BEDNARSKI, D. W., SCHUMMER, M., AND HAUSSLER, D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics 16*, 10 (2000), 906–914.

[56] GAO, S., YOUNG, M. T., QIU, J. X., YOON, H.-J., CHRISTIAN, J. B., FEARN, P. A., TOURASSI, G. D., AND RAMANTHAN, A. Hierarchical attention networks for information extraction from cancer

pathology reports. *Journal of the American Medical Informatics Association 25*, 3 (2017), 321–330.

[57] GARLA, V. N., AND BRANDT, C. Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification. *Journal of the American Medical Informatics Association 20*, 5 (2012), 882–886.

[58] GARLA, V. N., AND BRANDT, C. Ontology-guided feature engineering for clinical text classification. *Journal of biomedical informatics 45*, 5 (2012), 992–998.

[59] GENTLEMAN, R., AND CAREY, V. Unsupervised machine learning. In *Bioconductor Case Studies*. Springer, 2008, pp. 137–157.

[60] GHAREB, A. S., BAKAR, A. A., AND HAMDAN, A. R. Hybrid feature selection based on enhanced genetic algorithm for text categorization. *Expert Systems with Applications 49* (2016), 31–47.

[61] GUAN, J., LI, R., YU, S., AND ZHANG, X. A method for generating synthetic electronic medical record text. *IEEE/ACM transactions on computational biology and bioinformatics* (2019, pp. 173–182).

[62] GUIDI, G., MAFFEI, N., VECCHI, C., CIARMATORI, A., MISTRETTA, G. M., GOTTARDI, G., MEDURI, B., BALDAZZI, G., BERTONI, F., AND COSTI, T. A support vector machine tool for adaptive tomotherapy treatments: prediction of head and neck patients criticalities. *Physica Medica 31*, 5 (2015), 442–451.

[63] GULATI, P., SHARMA, A., AND GUPTA, M. Theoretical study of decision tree algorithms to identify pivotal factors for performance improvement: A review. *Int. J. Comput. Appl 141*, 14 (2016), 19–25.

[64] HASSANZADEH, H., KHOLGHI, M., NGUYEN, A., AND CHU, K. Clinical document classification using labeled and unlabeled data

across hospitals. In *AMIA annual symposium proceedings* (2018), vol. 2018, American Medical Informatics Association, pp. 545–554.

[65] HIRA, Z. M., AND GILLIES, D. F. A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics 2015* (2015, pp. 1–14).

[66] HOLZINGER, A., SCHANTL, J., SCHROETTNER, M., SEIFERT, C., AND VERSPOOR, K. Biomedical text mining: state-of-the-art, open problems and future challenges. In *Interactive knowledge discovery and data mining in biomedical informatics*. Springer, 2014, pp. 271–300.

[67] HOUSSEIN, E. H., ABDELMINAAM, D. S., HASSAN, H. N., AL-SAYED, M. M., AND NABIL, E. A hybrid barnacles mating optimizer algorithm with support vector machines for gene selection of microarray cancer classification. *IEEE Access 9* (2021), 64895–64905.

[68] HSU, C.-W., CHANG, C.-C., LIN, C.-J., ET AL. A practical guide to support vector classification, 2003, pp. 1–16.

[69] HUANG, G., HUANG, G.-B., SONG, S., AND YOU, K. Trends in extreme learning machines: A review. *Neural Networks 61* (2015), 32–48.

[70] HUANG, G.-B., ZHU, Q.-Y., AND SIEW, C.-K. Extreme learning machine: theory and applications. *Neurocomputing 70*, 1-3 (2006), 489–501.

[71] HUGHES, M., LI, I., KOTOULAS, S., AND SUZUMURA, T. Medical text classification using convolutional neural networks. In *Informatics for Health: Connected Citizen-Led Wellness and Population Health*. IOS Press, 2017, pp. 246–250.

[72] JIANG, H., TANG, F., AND ZHANG, X. Liver cancer identification based on pso-svm model. In *Control Automation Robotics &*

Vision (ICARCV), 2010 11th International Conference on (2010), IEEE, pp. 2519–2523.

[73] JOACHIMS, T. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning* (1998), Springer, pp. 137–142.

[74] JOACHIMS, T. Transductive inference for text classification using support vector machines. In *ICML* (1999), vol. 99, pp. 200–209.

[75] JONNAGADDALA, J., LIAW, S.-T., RAY, P., KUMAR, M., CHANG, N.-W., AND DAI, H.-J. Coronary artery disease risk assessment from unstructured electronic health records using text mining. *Journal of biomedical informatics 58* (2015), S203–S210.

[76] JOULIN, A., GRAVE, E., BOJANOWSKI, P., AND MIKOLOV, T. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759* (2016, pp. 1–5).

[77] JUNGIEWICZ, M., AND SMYWIŃSKI-POHL, A. Towards textual data augmentation for neural networks: synonyms and maximum loss. *Computer Science 20* (2019, pp. 57–83).

[78] KALANTARI, A., KAMSIN, A., SHAMSHIRBAND, S., GANI, A., ALINEJAD-ROKNY, H., AND CHRONOPOULOS, A. T. Computational intelligence approaches for classification of medical data: State-of-the-art, future challenges and research directions. *Neurocomputing 276* (2018), 2–22.

[79] KANG, T., PEROTTE, A., TANG, Y., TA, C., AND WENG, C. Umls-based data augmentation for natural language processing of clinical research literature. *Journal of the American Medical Informatics Association 28*, 4 (2021), 812–823.

[80] KASTRATI, Z., AND YAYILGAN, S. Y. Improving document classification effectiveness using knowledge exploited by ontologies. In *International Conference on Applications of Natural Language to Information Systems* (2017), Springer, pp. 435–438.

[81] KATARIA, A., AND SINGH, M. A review of data classification using k-nearest neighbour algorithm. *International Journal of Emerging Technology and Advanced Engineering 3*, 6 (2013), 354–360.

[82] KAUR, G., AND OBERAI, E. N. A review article on naive bayes classifier with various smoothing techniques. *International Journal of Computer Science and Mobile Computing 3*, 10 (2014), 864–868.

[83] KENNEDY, J. Swarm intelligence. In *Handbook of nature-inspired and innovative computing*. Springer, 2006, pp. 187–219.

[84] KENNEDY, J. Particle swarm optimization. In *Encyclopedia of machine learning*. Springer, 2011, pp. 760–766.

[85] KENNEDY, J., AND EBERHART, R. Particle swarm optimization in proceedings of ieee international conference on neural networks. *Piscataway December* (1995, pp. 1942–1948).

[86] KHALIFA, A., AND MEYSTRE, S. Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes. *Journal of biomedical informatics 58* (2015), S128–S132.

[87] KHAN, A., BAHARUDIN, B., LEE, L. H., KHAN, K., AND TRONOH, U. T. P. A review of machine learning algorithms for text-documents classification. In *Journal of Advances In Information Technology, VOL* (2010, pp. 4–20).

[88] KO, Y. A study of term weighting schemes using class information for text classification. In *Proceedings of the 35th international ACM*

SIGIR conference on Research and development in information retrieval (2012), ACM, pp. 1029–1030.

[89] KOBAYASHI, S. Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201* (2018, pp. 1–6).

[90] KORDE, V., AND MAHENDER, C. N. Text classification and classifiers: A survey. *International Journal of Artificial Intelligence & Applications 3*, 2 (2012), 85–99.

[91] KOSKIVAARA, E. Artificial neural networks in analytical review procedures. *Managerial Auditing Journal* (2004, pp. 191–223).

[92] KOTHARI, V., ANURADHA, J., SHAH, S., AND MITTAL, P. A survey on particle swarm optimization in feature selection. In *Global Trends in Information Systems and Software Applications*. Springer, 2012, pp. 192–201.

[93] KOTSIANTIS, S. B., ZAHARAKIS, I., PINTELAS, P., ET AL. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering 160*, 1 (2007), 3–24.

[94] KOWSARI, K., JAFARI MEIMANDI, K., HEIDARYSAFA, M., MENDU, S., BARNES, L., AND BROWN, D. Text classification algorithms: A survey. *Information 10*, 4 (2019), 1–68.

[95] KOZA, J. Genetic programming: on the programming of computers by means of natural selection. bradford, 1992, pp. 1–26.

[96] KRAWCZYK, B. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence 5*, 4 (2016), 221–232.

[97] KRAWIEC, K. Genetic programming-based construction of features for machine learning and knowledge discovery tasks. *Genetic Programming and Evolvable Machines 3*, 4 (2002), 329–343.

[98] KREBEL, U.-G. Pairwise classification and support vector machines. *Advances in kernel methods: support vector learning* (1999), 255–268.

[99] KUMBHAR, P., AND MALI, M. A survey on feature selection techniques and classification algorithms for efficient text classification. *International Journal of Science and Research 5*, 5 (2016), 1267–1275.

[100] LAI, S., XU, L., LIU, K., AND ZHAO, J. Recurrent convolutional neural networks for text classification. In *AAAI* (2015), vol. 333, pp. 2267–2273.

[101] LANE, M. C., XUE, B., LIU, I., AND ZHANG, M. Gaussian based particle swarm optimisation and statistical clustering for feature selection. In *European conference on evolutionary computation in combinatorial optimization* (2014), Springer, pp. 133–144.

[102] LANGLEY, P., ET AL. Selection of relevant features in machine learning. In *Proceedings of the AAAI Fall symposium on relevance* (1994), vol. 184, pp. 245–271.

[103] LE GUENNEC, A., MALINOWSKI, S., AND TAVENARD, R. Data augmentation for time series classification using convolutional neural networks. In *ECML/PKDD workshop on advanced analytics and learning on temporal data* (2016, pp. 1–9).

[104] LI, C. H., AND PARK, S. C. An efficient document classification model using an improved back propagation neural network and singular value decomposition. *Expert Systems with Applications 36*, 2 (2009), 3208–3215.

[105] LI, G., HOI, S. C., AND CHANG, K. Two-view transductive support vector machines. In *Proceedings of the 2010 SIAM International Conference on Data Mining* (2010), SIAM, pp. 235–244.

[106] LI, L.-N., OUYANG, J.-H., CHEN, H.-L., AND LIU, D.-Y. A computer aided diagnosis system for thyroid disease using extreme learning machine. *Journal of medical systems 36*, 5 (2012), 3327–3337.

[107] LI, S., WU, X., AND TAN, M. Gene selection using hybrid particle swarm optimization and genetic algorithm. *Soft Computing 12*, 11 (2008), 1039–1048.

[108] LI, X., CUI, M., LI, J., BAI, R., LU, Z., AND AICKELIN, U. A hybrid medical text classification framework: Integrating attentive rule construction and neural network. *Neurocomputing 443* (2021), 345–355.

[109] LIU, X., AND FU, H. Pso-based support vector machine with cuckoo search technique for clinical disease diagnoses. *The Scientific World Journal 2014* (2014, pp. 1–8).

[110] LUBIS, H., SIRAIT, P., AND HALIM, A. Knn method on credit risk classification with binary particle swarm optimization based feature selection. *INFOKUM 9*, 2, June (2021), 211–218.

[111] MALHOTRA, P., TV, V., VIG, L., AGARWAL, P., AND SHROFF, G. Timenet: Pre-trained deep recurrent neural network for time series classification. *arXiv preprint arXiv:1706.08838* (2017, pp. 1–10).

[112] MANDAL, I. Svm-pso based feature selection for improving medical diagnosis reliability using machine learning ensembles. *Computer Science & Information Technology (CS & IT) 2012* (2012), 267–276.

[113] MANIKAS, T. W., ASHENAYI, K., AND WAINWRIGHT, R. L. Genetic algorithms for autonomous robot navigation. *IEEE Instrumentation & Measurement Magazine 10*, 6 (2007, pp. 1–6).

[114] MEENACHI, L., AND RAMAKRISHNAN, S. Metaheuristic search based feature selection methods for classification of cancer. *Pattern Recognition 119* (2021), 1–14.

[115] MICHALSKI, R. S., CARBONELL, J. G., AND MITCHELL, T. M. *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013, pp. 1–572.

[116] MILLER, G. A. Wordnet: a lexical database for english. *Communications of the ACM 38*, 11 (1995), 39–41.

[117] MINER, G., ELDER IV, J., AND HILL, T. *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press, 2012, pp. 1–1047.

[118] MITCHELL, T. M., ET AL. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill 45*, 37 (1997), 870–877.

[119] MULLEN, R. J., MONEKOSSO, D., BARMAN, S., AND REMAGNINO, P. A review of ant algorithms. *Expert systems with Applications 36*, 6 (2009), 9608–9617.

[120] NAZIR, M., MAJID-MIRZA, A., AND ALI-KHAN, S. Pso-ga based optimized feature selection using facial and clothing information for gender classification. *Journal of applied research and technology 12*, 1 (2014), 145–152.

[121] NESHATIAN, K. Feature manipulation with genetic programming (2010, pp. 1–244).

[122] NESHATIAN, K., AND ZHANG, M. Genetic programming for performance improvement and dimensionality reduction of classification problems. In *Evolutionary Computation, 2008. CEC 2008.(IEEE World Congress on Computational Intelligence). IEEE Congress on* (2008), IEEE, pp. 2811–2818.

[123] NG, H. T., GOH, W. B., AND LOW, K. L. Feature selection, perceptron learning, and a usability case study for text categorization. In *ACM SIGIR Forum* (1997), vol. 31, ACM, pp. 67–73.

[124] NIÑO-ADAN, I., MANJARRES, D., LANDA-TORRES, I., AND PORTILLO, E. Feature weighting methods: A review. *Expert Systems with Applications* (2021), 115424, pp. 1–16.

[125] OLLAGNIER, A., AND WILLIAMS, H. T. Text augmentation techniques for clinical case classification. In *CLEF (Working Notes)* (2020, pp. 1–9).

[126] ORGANIZATION, W. H., AND OF SUBSTANCE ABUSE UNIT, W. H. O. M. *Global status report on alcohol and health, 2018*. World Health Organization, 2018, pp. 1–24.

[127] PAL, S. K., BANDYOPADHYAY, S., AND RAY, S. S. Evolutionary computation in bioinformatics: A review. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 36*, 5 (2006), 601–615.

[128] PARLAK, B., AND UYSAL, A. K. Classification of medical documents according to diseases. In *Signal Processing and Communications Applications Conference (SIU), 2015 23th* (2015), IEEE, pp. 1635–1638.

[129] PARLAK, B., AND UYSAL, A. K. The impact of feature selection on medical document classification. In *2016 11th Iberian conference on information systems and technologies (CISTI)* (2016), IEEE, pp. 1–5.

[130] PARLAK, B., AND UYSAL, A. K. On feature weighting and selection for medical document classification. In *Developments and Advances in Intelligent Systems and Applications*. Springer, 2018, pp. 269–282.

[131] PAUL, D., JAIN, A., SAHA, S., AND MATHEW, J. Multi-objective pso based online feature selection for multi-label classification. *Knowledge-Based Systems 222* (2021), 106966, pp. 1–14.

[132] PENG, Y., WU, Z., AND JIANG, J. A novel feature selection approach for biomedical data classification. *Journal of Biomedical Informatics 43*, 1 (2010), 15–23.

[133] PERCHA, B. Modern clinical text mining: A guide and review. *Annual Review of Biomedical Data Science 4* (2021, pp. 165–187).

[134] POLI, R., LANGDON, W. B., MCPHEE, N. F., AND KOZA, J. R. *A field guide to genetic programming*. Lulu. com, 2008, pp. 1–225.

[135] QING, L., LINHONG, W., AND XUEHAI, D. A novel neural network-based method for medical text classification. *Future Internet 11*, 12 (2019), 255, pp. 1–13.

[136] QUIJAS, J. K. *Analysing the effects of data augmentation and free parameters for text classification with recurrent convolutional neural networks*. The University of Texas at El Paso, 2017, pp. 1–54.

[137] RAK, R., KURGAN, L. A., AND REFORMAT, M. Multilabel associative classification categorization of medline articles into mesh keywords. *IEEE engineering in medicine and biology magazine 26*, 2 (2007), 47–55.

[138] ROSARIO, R. R. *A Data Augmentation Approach to Short Text Classification*. PhD thesis, UCLA, 2017, pp. 1–210.

[139] RUIZ, M. E., AND SRINIVASAN, P. Automatic text categorization using neural networks. In *Proceedings of the 8th ASIS SIG/CR Workshop on Classification Research* (1998), pp. 59–72.

[140] SAKO, D., AND PALIMOTE, J. A medical document classification system for heart disease diagnosis using naïve bayesian classifier. *Int. J. Appl. Sci. Math. Theory 4*, 1 (2018), 69–79.

[141] SALAMON, J., AND BELLO, J. P. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters 24*, 3 (2017), 279–283.

[142] SAMEK, W., MONTAVON, G., LAPUSCHKIN, S., ANDERS, C. J., AND MÜLLER, K.-R. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE 109*, 3 (2021), 247–278.

[143] SÁNCHEZ, D., BATET, M., AND VIEJO, A. Utility-preserving privacy protection of textual healthcare documents. *Journal of biomedical informatics 52* (2014), 189–198.

[144] SÁNCHEZ-MAROÑO, N., ALONSO-BETANZOS, A., AND TOMBILLA-SANROMÁN, M. Filter methods for feature selection–a comparative study. In *International Conference on Intelligent Data Engineering and Automated Learning* (2007), Springer, pp. 178–187.

[145] SEBASTIANI, F. Machine learning in automated text categorization. *ACM computing surveys (CSUR) 34*, 1 (2002), 1–47.

[146] SECINARO, S., CALANDRA, D., SECINARO, A., MUTHURANGU, V., AND BIANCONE, P. The role of artificial intelligence in healthcare: a structured literature review. *BMC Medical Informatics and Decision Making 21*, 1 (2021), 1–23.

[147] SEGARAN, T. *Programming collective intelligence: building smart web 2.0 applications*. " O'Reilly Media, Inc.", 2007, pp. 1–323.

[148] SEIDEL, P., SEIDEL, A., AND HERBARTH, O. Multilayer perceptron tumour diagnosis based on chromatography analysis of urinary nucleosides. *Neural Networks 20*, 5 (2007), 646–651.

[149] SHAH, F. P., AND PATEL, V. A review on feature selection and feature extraction for text classification. In *Wireless Communications, Sig-*

nal Processing and Networking (WiSPNET), International Conference on (2016), IEEE, pp. 2264–2268.

[150] SHANAVAS, N., WANG, H., LIN, Z., AND HAWE, G. Ontology-based enriched concept graphs for medical document classification. *Information Sciences 525* (2020), 172–181.

[151] SHEN, Q., SHI, W.-M., KONG, W., AND YE, B.-X. A combination of modified particle swarm optimization algorithm and support vector machine for gene selection and tumor classification. *Talanta 71*, 4 (2007), 1679–1683.

[152] SHIVADE, C., MALEWADKAR, P., FOSLER-LUSSIER, E., AND LAI, A. M. Comparison of umls terminologies to identify risk of heart disease using clinical notes. *Journal of biomedical informatics 58* (2015), S103–S110.

[153] SITUMORANG, S., SIRAIT, P., ET AL. Combination of logistic regression and svm algorithm with hybrid pso and ga based selection feature in coronary heart disease classification. *INFOKUM 9*, 2, June (2021), 204–210.

[154] ŠKORIĆ, M., AND DRAGONI, M. Medical domain document classification via extraction of taxonomy concepts from mesh ontology. *Infotheca* (2019, pp. 55–69).

[155] SOLT, I., TIKK, D., GÁL, V., AND KARDKOVÁCS, Z. T. Semantic classification of diseases in discharge summaries using a context-aware rule-based classifier. *Journal of the American Medical Informatics Association 16*, 4 (2009), 580–584.

[156] SOMVANSHI, M., CHAVAN, P., TAMBADE, S., AND SHINDE, S. A review of machine learning techniques using decision tree and support vector machine. In *2016 international conference on computing*

communication control and automation (ICCUBEA) (2016), IEEE, pp. 1–7.

[157] SOUSA, T., SILVA, A., AND NEVES, A. Particle swarm based data mining algorithms for classification tasks. *Parallel computing 30*, 5-6 (2004), 767–783.

[158] SPIEGELHALTER, D., TAYLOR, C., AND CAMPBELL, J. Machine learning, neural and statistical classification. *University of Strachclide* (1994, pp. 1–298).

[159] SUN, W., CAI, Z., LI, Y., LIU, F., FANG, S., AND WANG, G. Data processing and text mining technologies on electronic medical records: a review. *Journal of healthcare engineering 2018* (2018, pp. 1–10).

[160] SUN, W., RUMSHISKY, A., AND UZUNER, O. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association 20*, 5 (2013), 806–813.

[161] SVINGEN, B. Using genetic programming for document classification. In *FLAIRS Conference* (1998), pp. 63–67.

[162] TALAVERA, L. An evaluation of filter and wrapper methods for feature selection in categorical clustering. In *International Symposium on Intelligent Data Analysis* (2005), Springer, pp. 440–451.

[163] TRAN, B., XUE, B., AND ZHANG, M. Overview of particle swarm optimisation for feature selection in classification. In *Asia-Pacific conference on simulated evolution and learning* (2014), Springer, pp. 605–617.

[164] TRAPPEY, A. J., HSU, F.-C., TRAPPEY, C. V., AND LIN, C.-I. Development of a patent document classification and search platform

using a back-propagation network. *Expert Systems with Applications 31*, 4 (2006), 755–765.

[165] UYSAL, A. K., AND GUNAL, S. A novel probabilistic feature selection method for text classification. *Knowledge-Based Systems 36* (2012), 226–235.

[166] UYSAL, A. K., AND GUNAL, S. Text classification using genetic algorithm oriented latent semantic features. *Expert Systems with Applications 41*, 13 (2014), 5938–5947.

[167] UZUNER, Ö. Recognizing obesity and comorbidities in sparse data. *Journal of the American Medical Informatics Association 16*, 4 (2009), 561–570.

[168] UZUNER, Ö., SOUTH, B. R., SHEN, S., AND DUVALL, S. L. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association 18*, 5 (2011), 552–556.

[169] VANAM, M. K., JIWANI, B. A., SWATHI, A., AND MADHAVI, V. High performance machine learning and data science based implementation using weka. *Materials Today: Proceedings* (2021, pp. 1–7).

[170] WAGHOLIKAR, K. B., SUNDARARAJAN, V., AND DESHPANDE, A. W. Modeling paradigms for medical diagnostic decision support: a survey and future directions. *Journal of medical systems 36*, 5 (2012), 3029–3049.

[171] WEI, J., AND ZOU, K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196* (2019, pp. 1–9).

[172] WEN, C., WANG, Z., ZHANG, M., WANG, S., GENG, P., SUN, F., CHEN, M., LIN, G., HU, L., MA, J., ET AL. Metabolic changes in rat

urine after acute paraquat poisoning and discriminated by support vector machine. *Biomedical Chromatography 30*, 1 (2016), 75–80.

[173] WICKRAMASINGHE, I., AND KALUTARAGE, H. Naive bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. *Soft Computing 25*, 3 (2021), 2277–2293.

[174] WITTEN, I. H., FRANK, E., HALL, M. A., AND PAL, C. J. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016, pp. 1–128.

[175] WONG, S. C., GATT, A., STAMATESCU, V., AND MCDONNELL, M. D. Understanding data augmentation for classification: when to warp? In *2016 international conference on digital image computing: techniques and applications (DICTA)* (2016), IEEE, pp. 1–6.

[176] WU, D., WARWICK, K., MA, Z., GASSON, M. N., BURGESS, J. G., PAN, S., AND AZIZ, T. Z. Prediction of parkinson's disease tremor onset using a radial basis function neural network based on particle swarm optimization. *International journal of neural systems 20*, 02 (2010), 109–116.

[177] XIE, Z., WANG, S. I., LI, J., LÉVY, D., NIE, A., JURAFSKY, D., AND NG, A. Y. Data noising as smoothing in neural network language models. *arXiv preprint arXiv:1703.02573* (2017, pp. 1–12).

[178] XUE, B., CERVANTE, L., SHANG, L., BROWNE, W. N., AND ZHANG, M. A multi-objective particle swarm optimisation for filter-based feature selection in classification problems. *Connection Science 24*, 2-3 (2012), 91–116.

[179] XUE, B., CERVANTE, L., SHANG, L., BROWNE, W. N., AND ZHANG, M. Multi-objective evolutionary algorithms for filter based feature

selection in classification. *International Journal on Artificial Intelligence Tools 22*, 04 (2013), 1350024, pp. 2107–2115.

[180] XUE, B., CERVANTE, L., SHANG, L., BROWNE, W. N., AND ZHANG, M. Binary pso and rough set theory for feature selection: A multi-objective filter based approach. *International Journal of Computational Intelligence and Applications 13*, 02 (2014), 1450009, pp. 1–34.

[181] XUE, B., AND ZHANG, M. Evolutionary computation for feature manipulation: Key challenges and future directions. In *Evolutionary Computation (CEC), 2016 IEEE Congress on* (2016), IEEE, pp. 3061–3067.

[182] XUE, B., AND ZHANG, M. Evolutionary computation for feature selection and feature construction. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference Companion* (2020), pp. 1283–1312.

[183] XUE, B., ZHANG, M., AND BROWNE, W. N. New fitness functions in binary particle swarm optimisation for feature selection. In *Evolutionary Computation (CEC), 2012 IEEE Congress on* (2012), IEEE, pp. 1–8.

[184] XUE, B., ZHANG, M., AND BROWNE, W. N. Novel initialisation and updating mechanisms in pso for feature selection in classification. In *European Conference on the Applications of Evolutionary Computation* (2013), Springer, pp. 428–438.

[185] XUE, B., ZHANG, M., AND BROWNE, W. N. Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms. *Applied Soft Computing 18* (2014), 261–276.

[186] XUE, B., ZHANG, M., BROWNE, W. N., AND YAO, X. A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on Evolutionary Computation 20*, 4 (2016), 606–626.

[187] YADAV, M., MALHOTRA, P., VIG, L., SRIRAM, K., AND SHROFF, G. Ode-augmented training improves anomaly detection in sensor data from machines. *arXiv preprint arXiv:1605.01534* (2016, pp. 1–5).

[188] YADAV, P., STEINBACH, M., KUMAR, V., AND SIMON, G. Mining electronic health records (ehrs): a survey. *ACM Computing Surveys (CSUR) 50*, 6 (2018), 85, pp. 1–40.

[189] YAN, H., JIANG, Y., ZHENG, J., PENG, C., AND LI, Q. A multilayer perceptron-based medical decision support system for heart disease diagnosis. *Expert Systems with Applications 30*, 2 (2006), 272–281.

[190] YANG, H., AND GARIBALDI, J. M. A hybrid model for automatic identification of risk factors for heart disease. *Journal of biomedical informatics 58* (2015), S171–S182.

[191] YANG, Y., AND PEDERSEN, J. O. A comparative study on feature selection in text categorization. In *Icml* (1997), vol. 97, pp. 412–420.

[192] YAO, L., MAO, C., AND LUO, Y. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC medical informatics and decision making 19*, 3 (2019), 71, pp. 31–39.

[193] YEPES, A. J. J., PLAZA, L., CARRILLO-DE ALBORNOZ, J., MORK, J. G., AND ARONSON, A. R. Feature engineering for medline citation categorization with mesh. *BMC bioinformatics 16*, 1 (2015), 113, pp. 1–12.

[194] YETISGEN-YILDIZ, M., AND PRATT, W. The effect of feature representation on medline document classification. In *AMIA annual symposium proceedings* (2005), vol. 2005, American Medical Informatics Association, pp. 849–853.

[195] YI, K., AND BEHESHTI, J. A hidden markov model-based text classification of medical documents. *Journal of Information Science 35*, 1 (2009), 67–81.

[196] YOO, I., ALAFAIREET, P., MARINOV, M., PENA-HERNANDEZ, K., GOPIDI, R., CHANG, J.-F., AND HUA, L. Data mining in healthcare and biomedicine: a survey of the literature. *Journal of medical systems 36*, 4 (2012), 2431–2448.

[197] YU, A. W., DOHAN, D., LUONG, M.-T., ZHAO, R., CHEN, K., NOROUZI, M., AND LE, Q. V. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541* (2018, pp. 1–16).

[198] ZHANG, X., ZHAO, J., AND LECUN, Y. Character-level convolutional networks for text classification. In *Advances in neural information processing systems* (2015), pp. 649–657.

[199] ZHANG, Y., CHEN, M., AND LIU, L. A review on text mining. In *2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS)* (2015), IEEE, pp. 681–685.

[200] ZHAO, H., SINHA, A. P., AND GE, W. Effects of feature construction on classification performance: An empirical study in bank failure prediction. *Expert Systems with Applications 36*, 2 (2009), 2633–2644.