Highlights

Substituting clinical features using synthetic medical phrases: Medical text data augmentation techniques

Mahdi Abdollahi,Xiaoying Gao,Yi Mei,Shameek Ghosh,Jinyan Li,Michael Narag

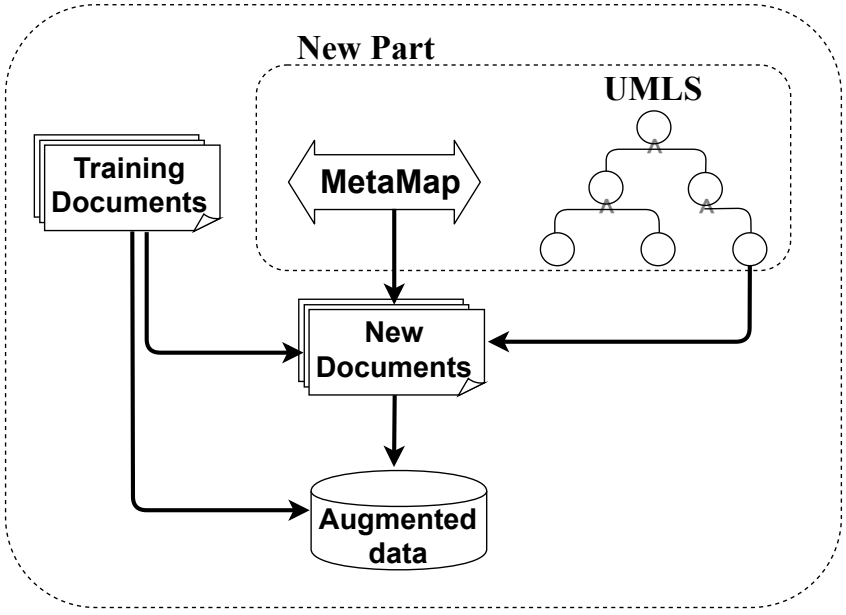
- This study presents two new approaches to fully automatic augmentation of medical discharge notes from original clinical notes based on an ontology and a dictionary.
- The first approach named $SciName$ is using Unified Medical Language System (UMLS) for data augmentation by replacing expression in the documents with their scientific names. (This method doubles the train set size.)
- The second combined approach named $SynName + SciName$ produces documents by using $SciName$ method plus using WordNet dictionary to replace phrases in the discharge notes with their synonyms. (This method triples the train set size.)
- The proposed augmentation approaches improved the performance of the Convolutional Neural Networks (CNN), Recurrent neural networks (RNN) and Hierarchical Attention Network (HAN) models by providing more instances in the training stage.

Graphical Abstract

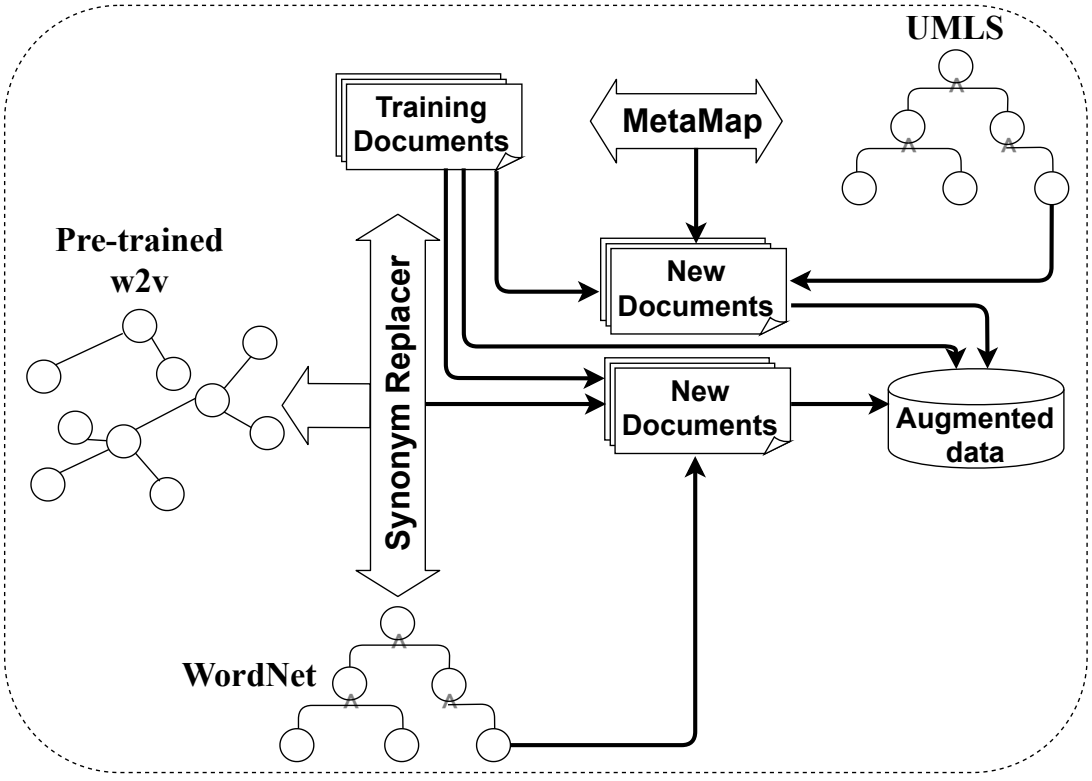
Substituting clinical features using synthetic medical phrases: Medical text data augmentation techniques

Mahdi Abdollahi,Xiaoying Gao,Yi Mei,Shameek Ghosh,Jinyan Li,Michael Narag

SciName Approach (UMLS-based)



SynName+SciName Approach (Combined Dictionary and UMLS-based)



Substituting clinical features using synthetic medical phrases: Medical text data augmentation techniques

Mahdi Abdollahi[a,*], Xiaoying Gao[a], Yi Mei[a], Shameek Ghosh[b], Jinyan Li[c] and Michael Narag[b]

[a]Victoria University of Wellington, Wellington, New Zealand
[b]Medius Health, Sydney, Australia
[c]University of Technology Sydney, Sydney, Australia

ARTICLE INFO

Keywords:
Unified Medical Language System
Natural Language Processing
Machine Learning
Data Augmentation
Medical Document Classification

ABSTRACT

Biomedical natural language processing (NLP) has an important role in extracting consequential information in medical discharge notes. Detecting meaningful features from unstructured notes is a challenging task in medical document classification. The domain specific phrases and different synonyms within the medical documents make it hard to analyse them. Analyzing clinical notes becomes more challenging for short documents like abstract texts. All of these can result in poor classification performance, especially when there is a shortage of the clinical data in real life. Two new approaches (an ontology-guided approach and a combined ontology-based with dictionary-based approach) are suggested for augmenting medical data to enrich training data. Three different deep learning approaches are used to evaluate the classification performance of the proposed methods. The obtained results show that the proposed methods improved the classification accuracy in clinical notes classification.

1. Introduction

Clinical text classification is different from general text classification regarding the text vocabulary. In our clinical text classification task, each discharge note contains a set of clinical events for providing an accurate and comprehensive description about the patient's health history. Generally, this kind of text has domain-specific terminologies and synonyms which makes analyzing clinical notes remarkably different from general purpose text classification. Additionally, various orders of domain-specific medical events in clinical documents can illustrate a person's health condition absolutely in a different way. Deriving meaningful information to check medical notes is extremely essential.

The size of data set with training documents to feed to a learning model is one of the prominent points which has an effect on classification performance. Commonly, there is lack of sufficient data in the clinical area [23]. When the size of the training data set is not large enough, the trained classifier will not have enough samples to learn. Consequently, the classifier performance will not be convincing. This problem is often worse when the training set has notes with short content, like abstract texts. Data augmentation is one of the options to deal with this issue in training the classifier.

Data augmentation is a technique that enables researchers to make numerous kinds of data to use in training their models, without spending time to collect new data. Data augmentation is used in plenty of areas such as sound, speech and image classification [21]. However, to the best of our knowledge, there is not much research in the text area. It is not suitable to make new samples by employing signal transformation methods as generally utilized in speech and image classification, because these methods have not kept the word orders in text. The new notes produced by these methods may change the semantic meaning of the original text. Thus, paraphrasing all of the existing sentences within the documents by an individual can be the best way for augmenting new documents. However, paraphrasing can be very expensive due to being time consuming. Substituting words and phrases with their synonyms is an acceptable approach for augmenting new notes [34]. If we employ general dictionaries to build new documents and it is possible some domain-specific vocabulary do not have any synonyms in general dictionaries.

Since clinical documents contain domain-specific acronyms and terminologies, replacing words and expressions with their synonyms is insignificant and needs to consider domain related knowledge. General dictionaries such as WordNet do not contain all of domain-specific vocabulary and acronyms. 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 dictionary which can provide more information regarding the detected medical phrases in documents. In this work, an ontology-based approach is proposed for augmenting medical documents by considering concepts of terms and phrases in the discharge notes. This approach will substitute the vocabulary and expressions with their scientific names if they relate to a concept in the clinical area. Furthermore, an incremental approach is proposed which considers ontology-based and dictionary-based methods at the same time to augment new medical documents.

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 pa-

*Corresponding author
mahdi.abdollahi@ecs.vuw.ac.nz (M. Abdollahi); xiaoying.gao@ecs.vuw.ac.nz (X. Gao); yi.mei@ecs.vuw.ac.nz (Y. Mei); shameek.ghosh@mediushealth.org (S. Ghosh); Jinyan.Li@uts.edu.au (J. Li); michael.narag@mediushealth.org (M. Narag)
ORCID(s): 0000-0002-6115-1342 (M. Abdollahi); 0000-0002-6326-7947 (X. Gao); 0000-0003-0682-1363 (Y. Mei); 0000-0003-1833-7413 (J. Li); 0000-0002-6061-2653 (M. Narag)

per for medical document classification problems. Different from most suggested 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. In this paper 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 $SynName$ (an existed synonym-based) method by analyzing their effectiveness for medical document augmentation.
- Combine the two approaches ($SynName+SciName$) to enhance the classification performance by increasing the size of training data sets.
- Compare the classification performance of proposed data augmentation approaches with other non-data augmentation approach

The rest of the paper is organized as follows. Section 2 presents the related work. Section 3 describes the proposed ontology-guided data augmentation approach to construct new documents from the original documents. Section 4 presents the proposed combined ontology and dictionary based approach. Section 5 describes the experiment design, classification methods, data sets and parameter settings for comparison. Section 6 presents the results and discussion. Section 7 provides further analysis and clinical assessment on the obtained best cases compared with some existing works. The achievements of the two approaches, and their limitations are summarised in section 8.

2. Related Work

2.1. Document classification in medical domain

Many machine leaning algorithms are used in medical domain including K-nearest neighbor, decision trees, decision rule, logistic regression, naive Bayes, support vector machine Bayesian network and neural network. In [6], authors used classifier models for clinical research prediction. Yoo et al. studied the advantages and disadvantages of utilizing data mining approaches in the biomedical area [33]. Their suggested medical features involve prognosis and diagnosis, hidden information, and prediction health costs from biomedical data sets. They derive relation between drugs and between diseases, and then the extracted information is employed for prediction. Furthermore in this work, three different issues related to data mining methods in clinical areas are considered such as how to determine and tune parameters of algorithms, precision of data mining and data mining package shortage for the medical domain. In [28], researchers have utilized more than ten approaches to analyze more than ten diseases. In this investigation, these approaches show better efficiency for some diseases such as cardiovascular, oncology and gastroenterology. Authors in [1] suggested an ontology-guided approach by employing the Unified Medical Language System (UMLS) to extract the concepts of available meaningful phrases inside documents. They considered diseases or symptoms as features to classify coronary artery disease notes. Furthermore, they applied particle swarm optimisation (PSO) on the extracted concepts as features to find the best feature subset to increase the accuracy of the classification [2]. Additionally, they constructed different pairs of the derived concepts to use as new features to enrich the feature set to improve the medical classification accuracy [3].

2.2. Data augmentation in classification

Data augmentation is an approach to solve data shortage issues for training different models for various tasks like classification. The mostly used methods to augment text are spelling words with errors and adding to the text, using regular expressions and syntax trees to paraphrase sentences, adding noise to the text and substituting with synonyms.In text data augmentation, synonym substituting is a common methods among the above approaches.

In the rest of this subsection, some of the approaches introduced for image, speech, time series and text classification are discussed. Data augmentation approaches [31, 18] 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 performance of recurrent neural networks (RNNs) for anomaly detection [31]. Guennec et al. have recently suggested data augmentation by applying warping and window slicing in Convolutional Neural Networks (CNNs) for time series classification (TSC) [18]. Malhotra et al. have employed a pre-trained encoder network on different series of time with various lengths in deep recurrent neural network for classifying series of time [19]. Wong et al. have investigated the benefit of data augmentation with artificially produced instances during training a classifier. They introduced data warping and synthetic oversampling methods for making new samples by applying on handwritten digit data set [29]. Salamon et al. have studied the effect of different data augmentation approaches on the performance of deep convolutional neural networks for environmental audio classification [22].

Zhang et al. [34] employed augmenting data in Convolutional Neural Network (CNN) for classifying text by using English thesaurus extracted from WordNet dictionary. They substituted the terms and phrases with their corresponding synonyms in the notes to build new notes by considering the original data set. Rosario [21] presented an approach for augmenting data for classifying short notes by making similar terms for each short note to build a note with a long length based on the main note's semantic space. Quijas has researched the result of augmenting data for training CNNs and RNNs models for classifying text [20]. Kobayashi has proposed "contextual augmentation" technique which makes correspondents of terms by employing a bidirectional lan-

guage model and substitutes terms with their correspondents in sentences. They tested their approach on various data sets and presented increasments in the results [17]. Coulombe in [9] has presented another text-based augmenting data by employing various approaches containing paraphrasing, spelling errors, making text noise, back-translation technique and synonyms substitution. The suggested approaches were examined on various neural network structures. Jungiewicz [16] has introduced a method for augmenting text data for training CNN models by classifying sentences. The author changed sentences by preserving their lengths the same as their main lengths. The researcher has used a thesaurus which was extracted from Princeton University's WordNet dictionary. Nonetheless, these approaches are employing general dictionaries for augmenting new notes and one limitation is that some domain-specific acronyms or words do not have any synonyms in general dictionaries.

2.3. Feature extraction in medical document classification

Shah and Patel [24] have utilized statistical methods by considering distribution of features in text classification for ranking features. The proposed approaches applied information gain (IG), mutual information, word frequency and term frequency-inverse document frequency (tf-idf) metrics for extracting text features. However, these approaches consider all of the features independently without taking into account the existing relationship between features. Ontology-based classification approach is proposed in [10]. Dollah and Aono have presented ontology-based approaches for classifying biomedical short documents [10]. Researchers in [8, 1, 2, 3] use various ontologies like Unified Medical Language System (UMLS), Systematized Nomenclature of Medicine (SNOMED) and Medical Subject Headings (MeSH) to improve document classification performance.

Clinical notes have been used in various tasks like investigating Framingham risk score (FRF), evaluating factors of risk for those patients suffering from diabetic, diagnosing factors of heart disease risk, and discovering factors of risk for patients with heart disease [25]. In this work, we use ontology and a dictionary for extracting important features to detect meaningful terms and phrases for building new documents.

2.4. Feature selection in medical document classification

In medical text classification, choosing an appropriate feature selection method which can discover the best existing features from features set with a high dimension is important. Information gain (IG) is a commonly utilized traditional method for selecting features [13]. Machine learning (ML) methods like Support Vector Machine (SVM) employed the selected features by information gain to feed to the classifier to learn. Particle swarm optimisation (PSO) is another feature selection approach which has impressive capacity in discovering the best feature subset from the feature set.

PSO has been used for analyzing and diagnosing different types of diseases in the biomedical area. For instance, [11] utilized PSO for analyzing tremors in individuals. They checked two various types of tremors: Parkinson's disease and important tremor. PSO is applied for evolving with neural network weights to enhance the neural network performance in diagnosing individuals infected with tumors from individuals without tumors. In another work, a PSO-based method which uses a Radial Basis Function Neural Network (RBFNN) employed to find Parkinsonian tremors [30]. [12] utilized PSO to discover best subsets of features. They applied PSO with three various classifiers: decision tree, pattern network, and navies bayes. This work showed a high performance of classification in two various experimental medical data sets: the Arrhythmia and the Micro Mass data sets.

3. The proposed Ontology-guided data augmentation approach

This section introduces the suggested novel data augmentation approach and the employed tools for finding concepts of terms and phrases for building new notes. The proposed method detects terms and phrases which belong to a concept to substitute them with their scientific names. Fig. 1 presents the flowchart of the suggested ontology-based method for augmenting documents.

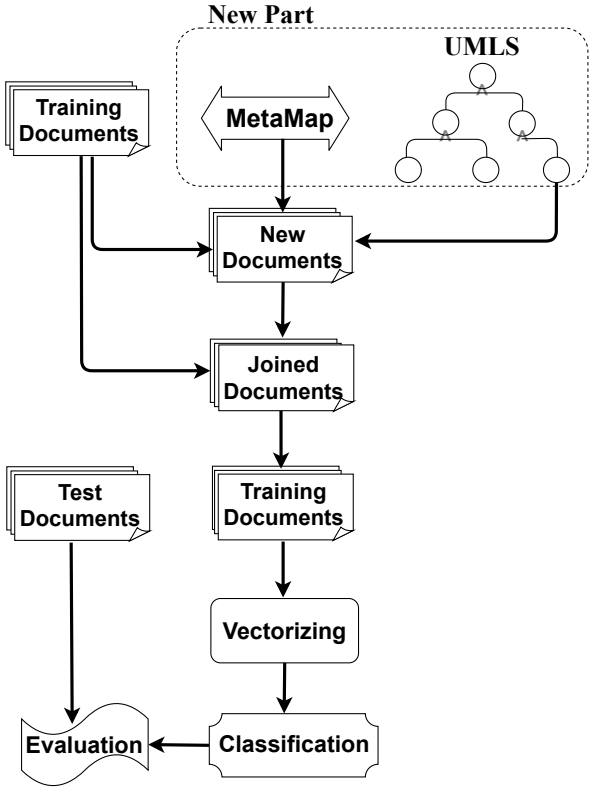


Figure 1: The proposed data augmentation for medical document classification

The input of the suggested method is a number of medical discharge notes. First of all, the approach analyzes each

note and tokenizes the note's content according to sentences. Next, the MetaMap tool [5] is utilized to find the important expressions and their related concepts in any sentence from the Unified Medical Language System (UMLS) [7]. After detecting the expressions with a concept, the scientific name of the found terms or phrases are utilized to substitute their corresponding expressions in the sentence. Whole of the new notes are made by using the approach. Then, all of the features are extracted from the original set of documents and the newly generated documents. Next, a classification approach such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and Hierarchical Attention Network (HAN) is employed to predict the labels of test documents.

It is anticipated that the proposed method which builds meaningful new notes and preserves their class label according to the main notes, can improve the performance of classification.

3.1. Data augmentation method

There are plenty of domain-specific terms and phrases in clinical discharge notes and augmenting new notes needs domain knowledge. In this section, an ontology-guided method is proposed for augmenting short documents as a preprocessing step.

The newly developed data augmentation method is based on UMLS. UMLS is a set of files and applications which group a wide range of terms and expressions as a domain-specific dictionary in the biomedical area. It presents an ontology-based structure of clinical vocabulary concepts. In the proposed method ("$SciName$"), all of the documents in the training set (D) are analyzed separately. First of all, the xth document (D_x) is tokenized to sentences (S). Next, the ith sentence (S_i) is transferred to UMLS by utilizing the MetaMap application. MetaMap finds the concepts of all of the found meaningful phrases in S_i from UMLS. Then, all of the found expressions are substituted with their found scientific names from the UMLS. At the end, S_i is renewed in the xth document (D_x). These steps are applied to the sentences of all of the documents to construct new documents. Algorithm 1 presents the pseudo code of the suggested ontology-guided data augmentation approach.

A piece of a document is provided to demonstrate what is MetaMap's output for the input clinical discharge notes and what is the content of the returned output in the data augmentation process. The sample of a medical discharge note is presented below.

> *"Early resistance to pathogens requires a swift response from nk cells. In largeint giorgio trinchieri identified an nk growth factor and activator later called interleukin 12 il 12. This discovery helped reveal the regulatory link between innate and adaptive immunity."*

Figure 2 presents the MetaMap's output for the provided piece of the document. Table 1 shows the found phrases with their corresponding concepts and scientific names for each expression of the example note. The concepts and scientific names of each found expression in the table is extracted by parsing the lines seven, fifteen, sixteen, seventeen displayed in Figure 2 for the first sentence, lines twenty five, thirty three, thirty four for the second sentence and lines forty two, fifty and fifty one for the third sentence. First of all, the expression between square brackets is detected as a concept of the found phrase in the sentence. Next, the appeared expression in the round parentheses at the same line is detected as a scientific name of the found phrase. At the end, the found scientific name is utilized to substitute the main phrase in the sentence. This process is used on all of the three sentences of the example text. The final output of the suggested data augmentation approach for the sample medical note is provided below.

> *"Early resistance to pathogenic organisms requires a family apodidae response process from natural killer cells. In largeint giorgio trinchieri identified an natural killer cells growth factor and activator later decisioned interleukin 12 illinois (geographic location) 12. This discovery assisted (qualifier value) reveal the regulator links list between innate and adaptive immunity."*

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 acronym *"nk"* is exchanged to *"natural killer"* and the acronym *"il"* is transformed to *"illinois (geographic location)"*. 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.

Algorithm 1: Pseudo-code of ontology-guided data augmentation approach

```
Input  : Set of clinical notes (D)
Output: Set of new clinical notes
x ← 0
while x < |D| do
    Tokenization: Tokenize D_x to sentences (S);
    for i = 1 to |S| do
        Detect all of the meaningful expressions by
          using MetaMap tool for S_i
        Substitute the detected expressions with their
          extracted scientific names from the UMLS
    end
    x ← x + 1
end
return set of the new made clinical notes;
```

```
   ----------------------------------------------------------------------
1  Phrase: Early resistance to pathogens
2  >>>>> Phrase
3  early resistance to pathogens
4  <<<<< Phrase
5  >>>>> Mappings
6  Meta Mapping (834):
7     570   Pathogen (Pathogenic organism) [Organism]
8  <<<<< Mappings
   ----------------------------------------------------------------------
9  Phrase: a swift response from nk cells.
10 >>>>> Phrase
11 a swift response from nk cells
12 <<<<< Phrase
13 >>>>> Mappings
14 Meta Mapping (719):
15    586   Swift (Family Apodidae) [Bird]
16    753   Response (Response process) [Organism Attribute]
17    623   NK Cells (Natural Killer Cells) [Cell]
18 <<<<< Mappings
   ----------------------------------------------------------------------
19 Phrase: an nk growth factor
20 >>>>> Phrase
21 nk growth factor
22 <<<<< Phrase
23 >>>>> Mappings
24 Meta Mapping (901):
25    660   NK (Natural Killer Cells) [Cell]
26 <<<<< Mappings
   ----------------------------------------------------------------------
27 Phrase: activator later called interleukin 12 il 12.
28 >>>>> Phrase
29 activator later called interleukin 12 il 12
30 <<<<< Phrase
31 >>>>> Mappings
32 Meta Mapping (745):
33    595   Call (Decision) [Mental Process]
34    795   IL (Illinois (geographic location)) [Geographic Area]
35 <<<<< Mappings
   ----------------------------------------------------------------------
36 Phrase: helped
37 >>>>> Phrase
38 helped
39 <<<<< Phrase
40 >>>>> Mappings
41 Meta Mapping (966):
42    966   Help (Assisted (qualifier value)) [Qualitative Concept]
43 <<<<< Mappings
   ----------------------------------------------------------------------
44 Phrase: the regulatory link between innate
45 >>>>> Phrase
46 the regulatory link between innate
47 <<<<< Phrase
48 >>>>> Mappings
49 Meta Mapping (695):
50    593   regulatory [Regulation or Law]
51    760   Link (Links List) [Intellectual Product]
52 <<<<< Mappings
   ----------------------------------------------------------------------
```

Figure 2: A piece of the returned output of the extracted concepts utilizing MetaMap tool

4. 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 3 demonstrates the introduced method overall. In the synonym-based method ($SynName$) WordNet dictionary is used to find the synonyms of words and then a 100-dimensional pretrained GloVe model 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 describes the oversampling on the classes in detail. It is expected that the new produced discharge notes with new

Table 1
The detected phrases of the example notes using MetaMap.

Sentences	Detected Phrases	Extracted Concepts	Replaced Phrases
First Sentence	pathogens	[Organism]	Pathogenic organism
	swift	[Bird]	Family Apodidae
	response	[Organism Attribute]	Response process
	nk cells	[Cell]	Natural Killer Cells
Second Sentence	nk	[Cell]	Natural Killer Cells
	call	[Mental Process]	Decision
	il	[Geographic Area]	Illinois (geographic location)
Third Sentence	help	[Qualitative Concept]	Assisted (qualifier value)
	regulatory	[Regulation or Law]	regulatory
	link	[Intellectual Product]	Links List

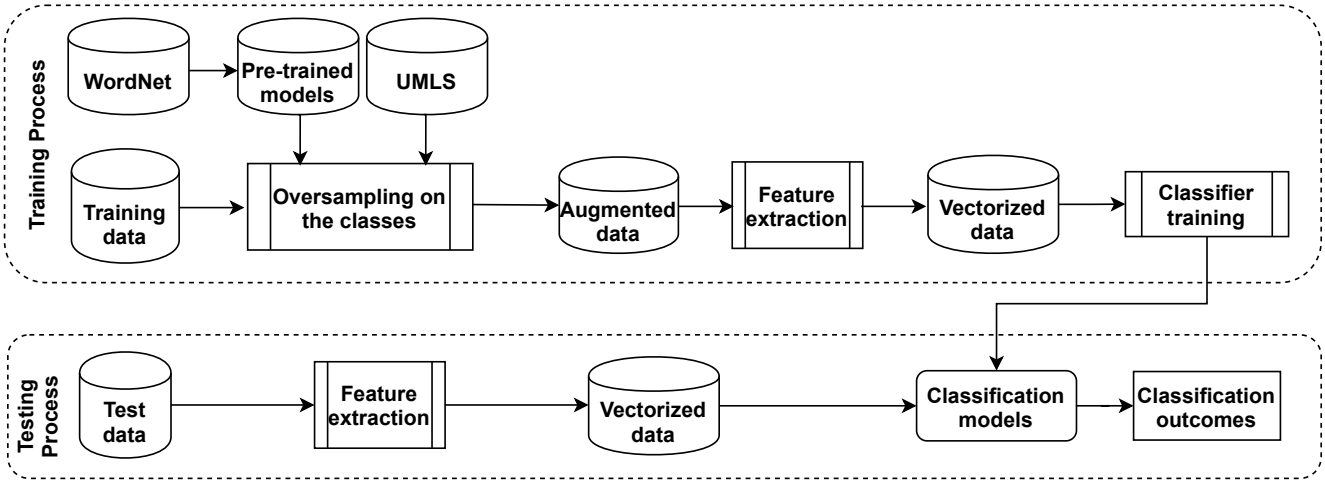


Figure 3: The combination of SynName and SciName oversampling methods for medical document classification

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

5. Experiment design

5.1. Classification approaches

In the proposed ontology-based and combined data augmentation approaches, the new medical discharge notes are constructed and mixed with the original discharge notes utilized for classification. Three various deep learning (DL) models, including a convolutional neural network (CNN), a recurrent neural network (RNN), and a hierarchical attention network (HAN) [14] are used as a classifier separately. The performance of the classifiers are computed by considering macro F1-measure metric for whole of the utilized machine learning approaches (see formula 1):

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

In Eg.(1), N is the size of classes. Word2Vec word embedding is utilized to present term tokens within numeral vectors. The semantic meaning of all of the words are represented by word embedding in a numeral vector format. Word embedding made by Word2Vec via using a feedforward neural network to predict the appeared words in the context for an input term. The used word embedding is trained based on the whole of the notes locally within the training and test documents and converted each term to its matching embedding. Next, the trained word embedding is utilized to make the input for the employed deep learning networks (CNN, RNN and HAN). The word embedding size is 350.

5.2. Data set and preprocessing

The performance of the proposed ontology-guided and combined data augmentation approaches are evaluated on the PubMed data set, the data set of 2008 Informatics for Integrating Biology and the Bedside (i2b2 2008) and the 2010 Informatics for Integrating Biology and the Bedside (i2b2 2010). CAD and non-CAD are i2b2(2010) data set labels which make an imbalanced binary classification task. The size of CAD documents for i2b2 (2010) training and testing documents are 25 and 48 abstract documents, respectively. The i2b2(2010) data set includes 7481 various words. It has overall 426 documents which the distribution is 170 for train-

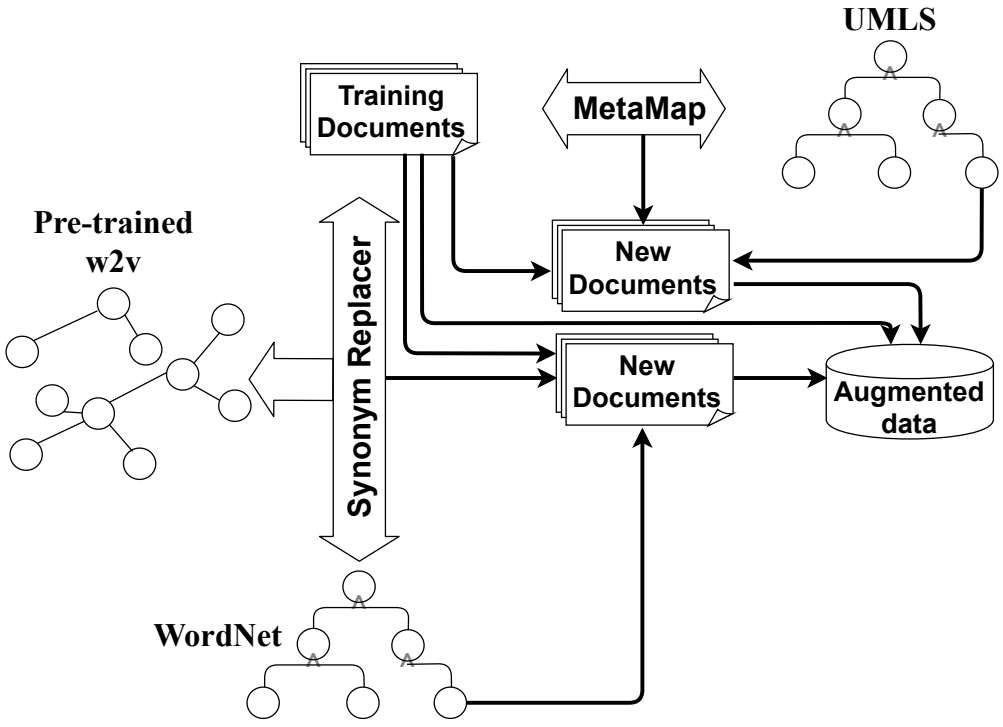


Figure 4: The oversampling on the classes by using $SynName$ and $SciName$ methods for medical document classification in detail

Table 2
Distribution of Intuitive Judgments into Training and Test Sets [27]

Classes	Present		Absent		Questionable		Total	
Diseass	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.

ing and 256 for testing.

Table 2 presents the labels of i2b2 (2008): Present, Absent and Questionable. The data set has sixteen tasks, in which six of the tasks have no Questionable labels and the size of instances for Questionable labels is very small for the other tasks. Hence, we filtered and deleted all instances of this label. Diagnosis, psychology, surgery, pathology, chemistry, genetics, physiology, and metabolism are the PubMed data set labels. The PubMed data set contains 8000 abstract (biomedical literature) documents and each class includes 1000 abstract documents with 70% of those for training and 30% for the testing phase. The data set includes 30178 differ-

ent words. The size of the training documents of the whole data sets will be doubled/tripled by including the newly created notes by using the UMLS and WordNet dictionaries. For example, the number of the train documents in $SynName$ and $SciName$ is increased to (5600×2=) 11200 in PubMed method and (426×2=) 852 in i2b2(2010) by considering the new constructed notes in $SynName$ and $SciName$ methods. The 852 input abstract notes include 14920 different words. The 11200 input abstract notes contain 59151 various words.

5.3. Parameter Settings

Three machine learning approaches are employed to assess the suggested method. Ten percent of the original training set in the all of the tasks are considered as the validation set. The parameter setting in [14] are utilized for the used neural network methods. Early stopping policy is utilized to stop the training phase by including the validation accuracy (three epochs without any improvement on the validation set).

The used CNN 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 the architecture for each input note is 300 channels × number of terms. The used dropout rate is 50% [14].

HAN [14] is a deep learning architecture proposed for classifying documents. It includes two hierarchies. The first hierarchy checks all of the lines by considering their words and it's input is a word embedding. Next, it utilizes a bidirectional Gated Recurrent Unit (GRU) to employ an attention approach to detect more valuable terms. The output of the first hierarchy is a line embedding which is the input for the second hierarchy to check all of the documents in line level. The dropout value is set to 50% and used for all of the created document embedding. As a last layer, a softmax function is applied to anticipate the label of all of the documents.

The RNN model utilizes a similar attention approach to the used one in a single hierarchy of HAN model. The employed GRU with attention is bidirectional. The hidden cells with size of 200 are used along dropout and softmax. The value of the learning rate for the utilized Adam optimizer is 0.0002. The value of rate for the used dropout is 50%.

6. Results and Discussions

The performance of the approaches is assessed by accuracy and macro F1-measure metrics for the PubMed abstract documents, macro F1-measure metric for the i2b2(2010) and the i2b2(2008) discharge notes.

Three approaches are used for all of the original notes in the used three tasks. The first method ($SynName$) employed a WordNet dictionary to find the whole of the synonyms of the existing terms within a document. Next, a pretrained Glove model (from the GloVe website [1] is used to find the most similar synonyms to substitute the original terms to construct new notes. The employed GloVe model is a vector with 100 dimensions. Ir is trained on Wikipedia data which has six billion tokens (phrases) and 400,000 vocabulary. In $SciName$ (the second approach), UMLS is used to extract scientific names of all of the existing expressions in the notes by considering their concepts to substitute with the original expression in the note. In $synName+SciName$ (the third method), the made documents by the two proposed methods are mixed with the original documents. Whole of the augmentation approaches are used for the training documents only. Next, thirty independent runs are applied to calculate the experimental results for the original test documents.

Tables 3 to 8 compare the statistical results for four methods for the original without augmentation, $SynName$, $SciName$ and $SynName + SciName$ methods. The average, best and accuracies' standard deviation and F1-measure are calculated for each machine learning model and the significance test is done using the experiment results of the 30 runs in comparison with other three methods. The Wilcoxon signed ranks test with significance level of 0.05 is utilized to evaluate if the proposed approaches ($SciName$ and $SynName+SciName$) have made a significant difference in classification performance. In tables 3 to 8, "T" column indicates the significance test of each approach against the previous columns (methods), where "+" presents the proposed approach is significantly better, "=" no significant difference, and "-" significantly less accurate. The best obtained results are bolded in the tables.

Tables 3, 4 and 5 present the obtained results for $SciName$ method in comparison with $SynName$ and original approaches. In table 3, $SciName$ approach has improved the classification performance in ten tasks out of the sixteen tasks by using CNN. The method shows a big improvement in eight tasks (Asthma, Depression, Gallstones, GERD, Gout, Hypercholesterolemia, Hypertension and OA). Table 4 provides the achieved results by RNN. $SciName$ has enhanced F1-measure in eight tasks in which the improvement in six of them (Asthma, CHF, Depression, Gallstones, Gout and Hypertriglyceridemia) is substantial. Comparison of HAN classification performance for the $SciName$ method is shown in table 5. The results demonstrate 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.

The $SciName$ approach is tested on two other data sets (PubMed and i2b2(2010)) too. By analyzing tables 6, 7, and 8, it can be concluded that neural network models are enhanced in F1-measure and accuracy by utilizing the domain-specific ontology to augment the training documents which provides more information in learning process to the used model. The $SciName$ method presents better performance

[1] http://nlp.stanford.edu/data/glove.6B.zip

Table 3
Comparison of CNN classification performance (F1-measure) and standard deviation averages using 30 independent runs for i2b2 2008 data set. 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)	=++

in comparison with $SynName$ in PubMed and i2b2(2010) 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 and 94.66% F1-measure for the i2b2(2010) data set.

Tables 3, 4 and 5 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. 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 demonstrates the performance of the proposed method 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 the 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 5, 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 in almost all of the tasks.

By checking tables 6 and 7, it can be concluded that neural network models are enhanced in F1-measure and accuracy by utilizing mix of the original documents with the constructed augmented documents from the two proposed augmentation approaches for PubMed data set. The highest accuracy and F1-measure in tables 6 and 7 are achieved by RNN with values 90.80% and 90.91%, respectively. Table 8 presents the statistical results for i2b2(2010) data set. CNN, RNN and HAN present the best F1-measure in the combination method ($SynName + SciName$). The best F1-measure belongs to RNN with 96.43%. In tables 3 to 8, the suggested combined approach shows better performance in comparison with the $SciName$ and the $SynName$ approach [34].

One interesting observation is that the generated documents help improve the performance of the classification although some of them look non-sense to humans. Such results can be due to the fact that the proposed approach works at word level, which benefits from meaningful words even if they form meaningless sentences. The proposed method shows good results on clinical notes (such as i2b2(2008) and i2b2(2010)) and achieves promising performance on related biomedical notes (PubMed set).

7. Further analysis

For further analyzing the introduced $SciName$ and $SynName+SciName$ methods, we compared the obtained experimental results by three different ML models with five different existing methods. The compared methods are divided into two groups. Kappa [27], Solt [26] and Yao [32] used rule-based approaches to classify i2b2 (2008) data set. Meanwhile, Ambert [4] and Garla [15] used automatic feature engineering methods to solve i2b2 (2008) challenge. $SciName$ and $SynName+SciName$ methods utilize an automatic system to enrich the data set.

Table 9 compares the statistical results of the $SynName$, $SciName$ and $SynName + SciName$ with the other five methods. The best results are highlighted in bold. The closest results to the best result 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).

Table 4
Comparison of RNN classification performance (F1-measure) and standard deviation averages using 30 independent runs for i2b2 2008 data set. 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 5
Comparison of HAN classification performance (F1-measure) and standard deviation averages using 30 independent runs for i2b2 2008 data set. 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)**	+++

This performance is impressive considering that our method is fully automatic without using rules or feature engineering.

7.1. Clinical Assessment

In this paper two methods are suggested for augmenting clinical discharge notes. As the second method ($SynName+SciName$) presents better performance, we found those documents which this method predicted correctly but the other

Table 6
Comparison of classification accuracy and standard deviation averages using 30 independent runs for PubMed data set. 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	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	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	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 7
Comparison of classification F1-measure and standard deviation averages using 30 independent runs for PubMed data set. 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	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	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	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 8
Comparison of classification F1-measure and standard deviation averages using 30 independent runs for i2b2 2010 data set. 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	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	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	56.11±18.78 (90.35)	86.90±7.23 (97.48)	+	93.75±2.55 (96.87)	++	**96.27±0.73 (97.51)**	+++

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 in 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 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 synony-

Table 9
Macro-averaged F1 on I2B2 2008 test set. Best scores from our study indicated in bold.

Methods	Kappa [27]	Solt [26]	Yao [32]	Ambert [4]	Superlin [15]	SynName			SciName			SynName+SciName		
Disease						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

mous 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 to 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 may also be used as it may increase the performance as they are more specific compared to UMLS.

7.2. The value of the work

As mentioned early in this paper, in numerous practical researches on modeling, augmenting data is very substantial. We face this situation in practical settings, individual patient's data are often not available in real life to be used as an input to feed data-hungry models (a scarce illness is an example where existing cases are not many). Actually, augmenting and synthesising synthetic data brings data substantial advantages by improving healthcare models study by preserving patient privacy, and it is an encouraging approach for conditions where it is not easy to collect real world data or it is not necessary. In these situations, by considering data augmentation we will be able to generate new synthetic data for training cases. Furthermore, these systems can support doctors in decision making for patients.

8. Conclusions and Future Work

This paper has introduced 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 WordNet dictionary to construct new documents. The introduced approaches are able to increase the accuracy of classification in the neural network models. Experimental results of accuracy and F1-measure show that the proposed approach can improve the efficiency of the CNN, RNN and HAN models by using the suggested ontology-based method ($SciName$) and combination approach of $SynName$ and $SciName$ ($SynName + SciName$) to provide more samples in the training phase.

This work presents promise in employing an ontology-guided data augmentation approach in clinical document classification, however, it is still necessary to do more research to improve the classification accuracy. We will investigate other ways for augmenting data for clinical documents. For example, we will explore to use existing domain-specific dictionaries instead of UMLS to improve the classification precision. Furthermore, based on analyzed documents, we can probably increase the weight of the medication/drug features (as a medically driven - insight) and see if the model performs better.

References

[1] Abdollahi, M., Gao, X., Mei, Y., Ghosh, S., Li, J., 2018. Uncovering discriminative knowledge-guided medical concepts for classifying coronary artery disease notes, in: Australasian Joint Conference on Artificial Intelligence, Springer. pp. 104–110.

[2] Abdollahi, M., Gao, X., Mei, Y., Ghosh, S., Li, J., 2019a. 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), IEEE. pp. 1–8.

[3] Abdollahi, M., Gao, X., Mei, Y., Ghosh, S., Li, J., 2019b. Stratifying risk of coronary artery disease using discriminative knowledge-guided medical concept pairings from clinical notes, in: Pacific Rim International Conference on Artificial Intelligence, Springer. pp. 457–473.

[4] Ambert, K.H., Cohen, A.M., 2009. 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, 590–595.

[5] Aronson, A.R., Lang, F.M., 2010. An overview of metamap: historical perspective and recent advances. Journal of the American Medical Informatics Association 17, 229–236.

[6] Bellazzi, R., Zupan, B., 2008. Predictive data mining in clinical medicine: current issues and guidelines. International journal of medical informatics 77, 81–97.

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

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

[9] Coulombe, C., 2018. Text data augmentation made simple by leveraging nlp cloud apis. arXiv preprint arXiv:1812.04718 .

[10] Dollah, R.B., Aono, M., 2011. Ontology based approach for classifying biomedical text abstracts. International Journal of Data Engineering (IJDE) 2, 1–15.

[11] Eberhart, R.C., Hu, X., 1999. Human tremor analysis using particle swarm optimization, in: Proceedings of the congress on evolutionary computation, IEEE Press Piscataway, NJ. pp. 1927–1930.

[12] Fong, S., Deb, S., Yang, X.S., Li, J., 2014. Feature selection in life science classification: metaheuristic swarm search. IT Professional , 24–29.

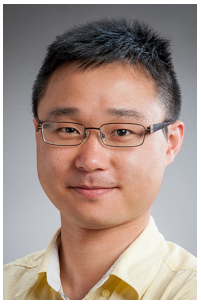
[13] Gaizauskas, R., Barker, E., Paramita, M.L., Aker, A., 2014. Assigning terms to domains by document classification, in: Proceedings of the 4th International Workshop on Computational Terminology (Computerm), pp. 11–21.

[14] Gao, S., Young, M.T., Qiu, J.X., Yoon, H.J., Christian, J.B., Fearn, P.A., Tourassi, G.D., Ramanthan, A., 2017. Hierarchical attention networks for information extraction from cancer pathology reports.

Journal of the American Medical Informatics Association 25, 321–330.
[15] Garla, V.N., Brandt, C., 2012. Ontology-guided feature engineering for clinical text classification. Journal of biomedical informatics 45, 992–998.
[16] Jungiewicz, M., Smywiński-Pohl, A., 2019. Towards textual data augmentation for neural networks: synonyms and maximum loss. Computer Science 20.
[17] Kobayashi, S., 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. arXiv preprint arXiv:1805.06201 .
[18] Le Guennec, A., Malinowski, S., Tavenard, R., 2016. Data augmentation for time series classification using convolutional neural networks.
[19] Malhotra, P., TV, V., Vig, L., Agarwal, P., Shroff, G., 2017. Timenet: Pre-trained deep recurrent neural network for time series classification. arXiv preprint arXiv:1706.08838 .
[20] Quijas, J.K., 2017. Analysing the effects of data augmentation and free parameters for text classification with recurrent convolutional neural networks. The University of Texas at El Paso.
[21] Rosario, R.R., 2017. A Data Augmentation Approach to Short Text Classification. Ph.D. thesis. UCLA.
[22] Salamon, J., Bello, J.P., 2017. Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Processing Letters 24, 279–283.
[23] Sánchez, D., Batet, M., Viejo, A., 2014. Utility-preserving privacy protection of textual healthcare documents. Journal of biomedical informatics 52, 189–198.
[24] Shah, F.P., Patel, V., 2016. A review on feature selection and feature extraction for text classification, in: Wireless Communications, Signal Processing and Networking (WiSPNET), International Conference on, IEEE. pp. 2264–2268.
[25] Shivade, C., Malewadkar, P., Fosler-Lussier, E., Lai, A.M., 2015. Comparison of umls terminologies to identify risk of heart disease using clinical notes. Journal of biomedical informatics 58, S103–S110.
[26] Solt, I., Tikk, D., Gál, V., Kardkovács, Z.T., 2009. Semantic classification of diseases in discharge summaries using a context-aware rule-based classifier. Journal of the American Medical Informatics Association 16, 580–584.
[27] Uzuner, Ö., 2009. Recognizing obesity and comorbidities in sparse data. Journal of the American Medical Informatics Association 16, 561–570.
[28] Wagholikar, K.B., Sundararajan, V., Deshpande, A.W., 2012. Modeling paradigms for medical diagnostic decision support: a survey and future directions. Journal of medical systems 36, 3029–3049.
[29] Wong, S.C., Gatt, A., Stamatescu, V., McDonnell, M.D., 2016. Understanding data augmentation for classification: when to warp?, in: 2016 international conference on digital image computing: techniques and applications (DICTA), IEEE. pp. 1–6.
[30] Wu, D., Warwick, K., Ma, Z., Gasson, M.N., Burgess, J.G., Pan, S., Aziz, T.Z., 2010. 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, 109–116.
[31] Yadav, M., Malhotra, P., Vig, L., Sriram, K., Shroff, G., 2016. Ode-augmented training improves anomaly detection in sensor data from machines. arXiv preprint arXiv:1605.01534 .
[32] Yao, L., Mao, C., Luo, Y., 2019. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. BMC medical informatics and decision making 19, 71.
[33] Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J.F., Hua, L., 2012. Data mining in healthcare and biomedicine: a survey of the literature. Journal of medical systems 36, 2431–2448.
[34] Zhang, X., Zhao, J., LeCun, Y., 2015. Character-level convolutional networks for text classification, in: Advances in neural information processing systems, pp. 649–657.

Mahdi Abdollahi received a B.Sc. degree in software engineering and a M.Sc. degree in Computer Science from Iran, in 2008 and 2013, respectively. He had worked as a lecturer at Shahid Madani University of Azarbaijan for three years. He is currently a PhD researcher in the School of Engineering and Computer Science, Victoria University of Wellington, Wellington, New Zealand. His research focuses on artificial intelligence such as text mining, text classification, evolutionary computation, parallel computing, different global optimization techniques and applications. He has been researching such topics as feature selection, multi-objective optimization, global optimization, computational intelligence, swarm intelligence, analyzing data, natural language processing, and machine learning. He is a regular reviewer of Computers Mathematics with Applications, the Journal of Supercomputing and Information Sciences journals.

Dr Xiaoying Gao received her BE and ME from China in 1990 and 1992, and PhD from the University of Melbourne, Australia in 2000. She is currently an Associate Professor in the School of Engineering and Computer Science, Victoria University of Wellington, Wellington, New Zealand. Her research interests include web intelligence, text mining, text classification, clustering, and evolutionary computation. She has over 80 fully refereed international journals and conferences, including the top conferences and journals such as KDD and KAIS. She is a key member of the Web Intelligence Consortium (WIC), and she serves as an associate editor for Web Intelligence: An International Journal, and a vice-chair (2008) and workshop co-chair (2020, 2021) for the IEEE/WIC/ACM Joint Conferences Series on Web Intelligence and Intelligent Agent Technology.



Dr. Yi Mei (M'09-SM'18) received the BSc and PhD degrees from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Senior Lecturer at the School of Engineering and Computer Science, Victoria University of Wellington, Wellington, New Zealand. His research interests include evolutionary scheduling and combinatorial optimisation, machine learning, genetic programming, and hyper-heuristics. He has over 100 fully referred publications, including the top journals in EC and Operations Research such as IEEE TEVC, IEEE TCYB, Evolutionary Computation Journal, European Journal of Operational Research, ACM Transactions on Mathematical Software. He serves as a Vice-Chair of the IEEE CIS Emergent Technologies Technical Committee, and a member of Intelligent Systems Applications Technical Committee. He is an Editorial Board Member/Associate Editor of three International Journals, and a guest editor of a special issue of the Genetic Programming Evolvable Machine journal. He serves as a reviewer of over 30 international journals.

Dr Shameek Ghosh is the Co-founder and Chief Technology Officer of Medius Health in Sydney, Australia. An alumnus of the University of Technology Sydney (UTS), he completed his PhD in Data Mining and Machine Learning in 2018. His research interests include knowledge graphs, data mining, machine learning, and their applications in health, insurance, and finance. He has 30 referred publications, including premier journals and conferences like IEEE JBHI, Elsevier JBI, ACM CIKM, and AMIA.

Dr. Jinyan Li is a Professor of Data Science and Program Leader of Bioinformatics at the Data Science Institute, Faculty of Engineering IT, University of Technology Sydney, Australia. He has been actively working on data mining and bioinformatics for 20 years. He has published 240 papers, including 130 papers in prestigious journals of data mining, machine learning, and computational biology. He is widely known for his pioneering research on the theories and algorithms of Emerging Patterns (EPs). One of these papers has received 1300 Google Scholar citations. Jinyan has a Bachelor degree of Science (Applied Mathematics) from National University of Defense Technology (China), a Master degree of Engineering (Computer Engineering) from Hebei University of Technology (China), and a PhD degree (Computer Science) from the University of Melbourne (Australia). More details of his research can be found at http://www.uts.edu.au/staff/jinyan.li

Dr Michael Narag has a background in medicine having completed his medical degree at the University of Santo Tomas in the Philippines. Prior to coming to Australia in 2009, he was an emergency medicine physician. Michael has been in the digital health industry for the past 8 years, working in both private and government organisations. His work is mainly on the development of digital technology in the areas of medical diagnosis, preventive medicine, chronic disease and epidemiology. Currently, he is managing a project in a federal government agency ensuring digital health capability in the healthcare sector. Michael also completed a double master's degree in public health and health management (MPH, MHM) at the University of New South Wales. He is also a member of the Society of Artificial Intelligence in Medicine, Surgery and Healthcare (AMSAH), International Epidemiological Association (IEA), Australasian Epidemiological Association (AEA), and Philippine Medical Association (PMA).

Conflict of Interest and Authorship Conformation Form

Please check the following as appropriate:

- o All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.
- o This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.
- o The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript
- o The following authors have affiliations with organizations with direct or indirect financial interest in the subject matter discussed in the manuscript:

Author's name	Affiliation
Mahdi Abdollahi	Victoria University of Wellington, Wellington, New Zealand
Xiaoying Gao	Victoria University of Wellington, Wellington, New Zealand
Yi Mei	Victoria University of Wellington, Wellington, New Zealand
Shameek Ghosh	Medius Health, Sydney, Australia
Jinyan Li	University of Technology Sydney, Sydney, Australia
Michael Narag	Medius Health, Sydney, Australia