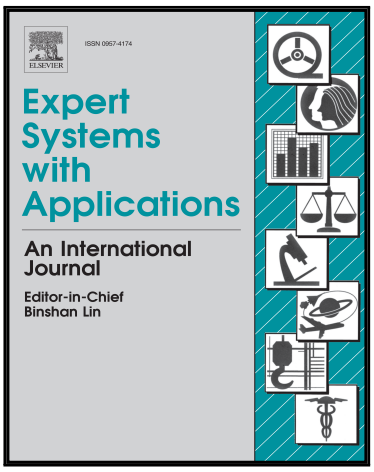

Clinical Text Classification Research Trends: Systematic Literature Review and Open Issues

Ghulam Mujtaba , Liyana Shuib , Norisma Idris , Wai Lam Hoo , Ram Gopal Raj , Kamran Khowaja , Khairunisa Shaikh , Henry Friday Nweke

PII: S0957-4174(18)30611-0
DOI: https://doi.org/10.1016/j.eswa.2018.09.034
Reference: ESWA 12222

To appear in: *Expert Systems With Applications*

Received date: 6 January 2018
Revised date: 14 September 2018
Accepted date: 15 September 2018

Please cite this article as: Ghulam Mujtaba , Liyana Shuib , Norisma Idris , Wai Lam Hoo , Ram Gopal Raj , Kamran Khowaja , Khairunisa Shaikh , Henry Friday Nweke , Clinical Text Classification Research Trends: Systematic Literature Review and Open Issues, *Expert Systems With Applications* (2018), doi: https://doi.org/10.1016/j.eswa.2018.09.034

Highlights

- To review free-text clinical text classification approaches from six aspects.
- In selected studies, mostly content-based and concept-based features were used.
- The datasets used in selected studies were categorized into four distinct types.
- Selected studies used either supervised machine learning or rule-based approaches.
- Ten open research challenges are presented in clinical text classification domain.

Title Page

Clinical Text Classification Research Trends: Systematic Literature Review and Open Issues

Ghulam Mujtaba[1, 5*], Liyana Shuib[1*], Norisma Idris[2], Wai Lam Hoo[1], Ram Gopal Raj[2], Kamran Khowaja[3], Khairunisa Shaikh[4], and Henry Friday Nweke[1,6]

[1]Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia

[2]Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia

[3]Department of Software Engineering, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia

[4]Department of Social and Preventive Medicine, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia

[5]Department of Computer Science, Sukkur IBA University, Sukkur, Pakistan

[6]Department of Computer Science, Ebonyi State University, P.M.B 053, Abakaliki, Ebonyi State, Nigeria

Email addresses: mujtaba@siswa.um.edu.my, liyanashuib@um.edu.my, norisma@um.edu.my, wlhoo@um.edu.my, ramdr@um.edu.my, kamran.khowaja@gmail.com, khairunisashaikh@siswa.um.edu.my, henrynweke@siswa.um.edu.my

Corresponding authors

Author Name: Ghulam Mujtaba
Author Email: mujtaba@siswa.um.edu.my
Cell Number: + (60) 173738760

Author Name: Liyana Shuib
Email: liyanashuib@um.edu.my
Cell Number: + (60) 196649440

Corresponding Authors Affiliation and Postal Address
Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia

Financial Disclosure / Funding Statement

This research was funded by the University Malaya Research Grant–AFR (Frontier Science) Project Number: RG380–17AFR

URL: https://umresearch.um.edu.my/

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Clinical Text Classification Research Trends: Systematic Literature Review and Open Issues

***Abstract*—**The pervasive use of electronic health databases has increased the accessibility of free-text clinical reports for supplementary use. Several text classification approaches, such as supervised machine learning (SML) or rule-based approaches, have been utilized to obtain beneficial information from free-text clinical reports. In recent years, many researchers have worked in the clinical text classification field and published their results in academic journals. However, to the best of our knowledge, no comprehensive systematic literature review (SLR) has recapitulated the existing primary studies on clinical text classification in the last five years. Thus, the current study aims to present SLR of academic articles on clinical text classification published from January 2013 to January 2018. Accordingly, we intend to maximize the procedural decision analysis in six aspects, namely, types of clinical reports, data sets and their characteristics, pre-processing and sampling techniques, feature engineering, machine learning algorithms, and performance metrics. To achieve our objective, 72 primary studies from 8 bibliographic databases were systematically selected and rigorously reviewed from the perspective of the six aspects. This review identified nine types of clinical reports, four types of data sets (i.e., homogeneous–homogenous, homogenous–heterogeneous, heterogeneous–homogenous, and heterogeneous–heterogeneous), two sampling techniques (i.e., over-sampling and under-sampling), and nine pre-processing techniques. Moreover, this review determined bag of words, bag of phrases, and bag of concepts features when represented by either term frequency or term frequency with inverse document frequency, thereby showing improved classification results. SML-based or rule-based approaches were generally employed to classify the clinical reports. To measure the performance of these classification approaches, we used precision, recall, F-measure, accuracy, AUC, and specificity in binary class problems. In multi-class problems, we primarily used micro or macro-averaging precision, recall, or F-measure. Lastly, open research issues and challenges are presented for future scholars who are interested in clinical text classification. This SLR will definitely be a beneficial resource for researchers engaged in clinical text classification.

Keywords: clinical text classification, feature engineering, supervised machine learning, rule-based text classification, performance metrics

1.0 Introduction

The extensive number of electronic health records contain beneficial information in free-text format. Free-text clinical reports have long been recognized to be beneficial for secondary use. Several researchers across the globe have employed text classification to categorize narrative clinical reports into various categories through several machine learning approaches, such as supervised, unsupervised, semi-supervised, ontology-based, rule-based, transfer, reinforcement, and multi-view learning approaches. The *supervised machine learning (SML)* approaches maximize the training data that contain the input and output variables *x* and *y*, respectively. Thereafter, this training data is provided as an input to the learning algorithm to learn the mapping function [$y = f(x)$]. The main goal of supervised learning is to efficiently approximate this mapping function, thereby enabling the accurate prediction of the output variable for new input data (*x*) (Hastie, Tibshirani, & Friedman, 2009). The *unsupervised machine learning (UML)* approaches do not require labeled data and draw the inferences from unlabeled datasets (Ko & Seo, 2000). The crux of the *semi-supervised machine learning (SSML)* approaches is that machine learning algorithms can obtain optimum classification accuracy with only a few labeled instances. These approaches are beneficial where unlabeled data can be obtained easily in huge volumes; nonetheless, the labeling of the collected data is difficult, laborious, and expensive (Settles, 2010; Zhu & Goldberg, 2009). In *ontology-based classification* approaches, domain-related clinical ontologies are prepared with the intervention of domain experts to identify medical-related named entities from clinical reports (Hotho, Maedche, & Staab, 2002). In *rule-based (RB)* approaches, rules are either written manually or generated automatically and verified manually thereafter to save time. The rule-based approach is simple and flexible where rules can be understood and improved over time (Deng, Groll, & Denecke, 2015; MacRae, et al., 2015). *Transfer learning (TL)* is useful when an input data set has inadequately labeled instances to train an accurate model. In such cases, TL translates the capabilities from existing systems to untrained ones. (Pan & Yang, 2010). *Reinforcement learning (RL)* can learn from experience and interaction with the environments to accurately classify the clinical reports by using a system of incentive and punishment. (Kaelbling, Littman, & Moore, 1996). In the *multi-view approaches*, the training

algorithms learn from several views and each view is sufficient for learning the classification rules for classifying the clinical reports (Amini, Usunier, & Goutte, 2009).

In the aforementioned machine learning approaches, the SML approach has been maximized extensively for classifying the clinical reports (Al-garadi, Khan, Varathan, Mujtaba, & Al-Kabsi, 2016; Guzella & Caminhas, 2009; Mirończuk & Protasiewicz, 2018; Nigam, McCallum, Thrun, & Mitchell, 2000). In SML (Hastie, et al., 2009), the set of clinical reports is initially collected from hospitals. Thereafter, these reports are labeled by experts into specific categories (e.g., cancer-positive or cancer-negative). The reports are pre-processed thereafter to remove the unnecessary or noisy information from the reports. After pre-processing, feature engineering is applied to extract the most discriminative features from the clinical reports and form a numeric master feature vector. This master feature vector is provided as an input to learning algorithms (e.g., Naïve Bayes, support vector machines, AdaBoost) to construct and validate the classification model. A machine learning algorithm is either a pre-built or customized code stack in the form of a function or package. This algorithm is used to learn from data to create a model or equation that can eventually be used for classification, segmentation, or prediction. The classification model can be constructed and validated using either random sub-sampling, *k*-fold cross-validation, or leave-one-out techniques (Kohavi, 1995). In random sub-sampling, the shuffled clinical reports instances in the master feature vector will be divided into train and test set (e.g., 70% train set and 30% test set). The *k*-fold splits the shuffled reports into *k* data chunks and perform training on the $(k-1)$ data chunks and test it on the $(k-1)$th chunk. Thereafter, this process is repeated up to *k* times and in each *k*, the subsamples are used only once as the validation set. The advantage of this method is that all clinical reports are used for training and testing, while each report is used for testing only once. The disadvantage of the *k*-fold cross validation is that it is *k* times slower than random subsampling. Leave-one-out is a special case of the *k*-fold. In this technique, we train the model with $(n-1)$ reports and test it on the $(n-1)$th report. We repeat this by leaving out one report each time. This process requires *n* iterations to train and test the classification model. The leave-one-out technique is good for limited data and imbalanced data set and target values. However, this approach is *n* times slower than random subsampling, where *n* shows the number of clinical reports in the data set.

Text classification in the medical domain is more challenging than those in other general domains (e.g., email classification) because of three reasons. First, the narrative clinical report corpus generally contains high levels of noise, sparsity, complex medical vocabularies, medical measures and scores (e.g., blood pressure of 140/65), abbreviations, misspelled words, and poor grammatical sentences (Kaurova, Alexandrov, & Blanco, 2011; H. Nguyen & Patrick, 2016). To address these issues, a detailed pre-analysis of corpora should be performed to reveal the linguistic properties of texts in clinical text classification task. To address such lexical complications, specialized pre-processing techniques are also developed and used to address noise in the corpus (H. Nguyen & Patrick, 2016). Second, the narrative clinical report data sets generally have imbalanced class distribution, where one class, which is of considerable interest (e.g., cancer positive), is insufficiently represented compared with other classes (e.g., cancer negative). Thus, the relatively few positive cases in such data set is likely classified as rare occurrences, ignored, or assumed as outliers and cause more misclassifications compared with the majority class. Hence, such a rare class among the normal instances should be identified. Conversely, any diagnostic errors may bring stress and further complications to the patients. Thus, a classification model should be able to achieve high identification rates in the positive class in the data sets. To address this clinical text classification challenge, several researchers have developed and used data-level (e.g., sampling) and algorithm-level (e.g., cost-sensitive learning) strategies. Lastly, in narrative clinical reports, the experts may use a variety of medical words or phrases interchangeably (e.g., the phrase *heart attack* and *myocardial infarction* can be used interchangeably across various clinical reports). Thus, the task of clinical text classification should address this complex semantic information (e.g., word-level synonymy and polysemy). In general, several researchers have endeavored to overcome the issue of word-level synonymy and polysemy by utilizing specialized medical ontologies such as, systematized nomenclature of medicine-clinical terms (SNOMED-CT) to convert similar terms into unique concept IDs.

In recent years, several articles have been published under the clinical text classification domain to classify free-text clinical reports into specific categories. However, only a few review articles are available in the bibliographic databases that evaluate the existing studies on clinical text classification (Al-garadi, et al., 2016; Cosma, Brown, Archer, Khan, & Pockley, 2017; Holzinger, Schantl,

Schroettner, Seifert, & Verspoor, 2014; Kaurova, et al., 2011; Renganathan, 2017; Spasić, Livsey, Keane, & Nenadić, 2014). These review articles generally focused only on machine learning algorithms and rarely reviewed other aspects of clinical text classification domains, such as data set characteristics, sampling techniques, feature analysis, feature representation, feature reduction, and performance evaluation. Kaurova, et al. (2011) published a short review of articles on clinical text classification from 1995 to 2010. In particular, these authors reviewed various data sets and machine learning algorithms. Al-garadi, et al. (2016) published a systematic literature review (SLR) on articles that used online social network data to track a pandemic. These authors specifically reviewed a variety of machine learning algorithms and feature engineering techniques used in selected primary studies. Although the authors in (Al-garadi, et al., 2016) reviewed text classification techniques, their research is limited to social media sites, such as Twitter and Instagram. Cosma, et al. (2017) published a review of intelligent computation techniques employed to predict prostate cancer. These authors also reviewed metaheuristic optimization methods and machine learning approaches that were employed to detect prostate cancer. Moreover, the aforementioned authors considered cancer data with different modalities, such as text or images, to detect prostate cancer. Renganathan (2017) reviewed text-mining approaches with specific focus on biomedical clustering approaches. This author thoroughly discussed the text clustering and text classification techniques used in clinical report classification.

The preceding review studies have mainly provided an overall review of medical text mining approaches (e.g., clinical information extraction, clinical text classification, and biomedical document clustering) and a brief overview of the clinical text classification literature up to 2013 with specific focus on machine learning algorithms. However, other important aspects of clinical text classification (e.g., dataset characteristics, sampling techniques, feature analysis, feature representation, feature reduction, and performance evaluation) have not been reviewed (or reviewed thoroughly). Therefore, the aim of the current systematic review is to evaluate academic articles on clinical text classification published from January 2013 to January 2018. In particular, this study intends to maximize the procedural decision analysis in six aspects, namely, types of clinical reports, characteristics of the datasets used, pre-processing and sampling techniques, feature engineering, machine learning algorithms, and use of performance measures. To the best of our knowledge, this SLR will be the first that recapitulates the existing contemporary clinical text classification research for future researchers from the aforementioned six aspects. The major contributions of this study are given below:

- This study investigates the different types of clinical reports that have been classified automatically using text classification and natural language processing (NLP) techniques.
- In addition, it reviews the datasets that have been used for clinical report classification. Moreover, it provides a dataset taxonomy based on clinical reports modalities and sources from where the reports were collected.
- Furthermore, it identifies various pre-processing and data sampling techniques that have been used to classify clinical reports and overcome the issue of class imbalance.
- Moreover, it analyzes various feature sets, feature representation schemes, feature reduction techniques, and machine learning algorithms and performance metrics that have been used for automated clinical reports classification.
- Finally, it proposes future research challenges in the clinical text classification domain.

The remainder of this paper is structured as follows. Section 2 presents the research methodology used for selecting the primary studies. Section 3 presents and discusses the review analysis and findings of the selected primary studies. Section 4 presents the discussion and a few observational remarks on the review findings. Section 5 presents the open research issues and challenges. Lastly, Section 6 concludes this review.

2.0 Research Methodology

The research methodology maximizes the SLR guidelines proposed by Kitchenham and Charter (Keele, 2007) for the computer engineering discipline. SLR has four key phases, namely, planning, searching and selection of primary studies, data extraction, and data synthesis. In general, the planning phase identifies the problem statement, review objectives, and review protocols (as discussed in Section 1). The search strategy phase includes the study selection criteria, study selection procedure, formulation of the search keywords and search queries, and quality assessment of the retrieved studies (will be discussed in Section 2.1). The data extraction phase includes the data

extraction strategy from the selected studies (will be discussed in Section 2.2). The final phase, which includes a systematic review, involves data synthesis and critical analysis (will be discussed in detail in Section 3).

2.1 Searching strategy to retrieve primary studies

This review includes the majority of the studies that used clinical reports or medical documents as their source of data for automatic classification of clinical reports by using text classification techniques. Thus, various search keywords are formulated to retrieve the related literature from eight reliable and high-quality academic databases, namely, Web of Science (WoS), Scopus, IEEE Xplore, PubMed, Medline, ScienceDirect, Association for Computing Machinery (ACM), and SpringerLink. Four of the authors (GM, KK, HF, and LS) prepared the list of several relevant keywords to search the relevant literature on "*automated text classification of clinical reports or medical documents*" from the selected databases. Table 1 shows the keywords used to perform queries. Each keyword within the group is paired using the OR operator, whereas the groups are paired using the AND operator (see Table 1) to form a search query. The last row of Table 1 shows how keywords from different groups are concatenated to form a query that was executed in all eight bibliographic databases. Table 1 shows that the query was applied on the article title, article abstract, and article keywords to determine the relevant journal and conference articles from the eight selected bibliographic databases published (in English) from January 2013 to January 2018.

Table 1. Selected keywords in the different groups

Group 1- Keywords related to the text classification domain	Machine learning, text classification, automatic text classification, text analysis, text categorization, text mining, document classification, classification prediction, forecasting, ontology, lexicon, deep learning, natural language processing, NLP
Group 2- Keywords related to medical documents	Medical*documents, clinical reports, plain*text medical reports, free*text medical reports, raw text medical reports, unstructured medical reports, cancer reports, pathology reports, radiology reports, histopathology reports, toxicology reports, autopsy reports, forensic autopsy reports, medical autopsy reports, verbal autopsy reports, post*mortem reports, death reports, death certificates
Group 3- Publication years	January 2013 to January 2018
Group 4- Document types	Journal and Conference Articles
Group 5- Languages	English
Final Search Query	(Group 1) AND (Group 2) AND (Group 3) AND (Group 4) AND (Group 5)

2.2 Search results

The search query, when applied to the selected eight bibliographic databases, retrieved 1729 studies.

Table 2 shows the detailed search results from each database. The search records from each database against the search query were stored in the citation manager software. Subsequently, the duplicate studies across various databases were removed and only distinct copies of each primary study were stored in Endnote. The removal of the duplicate filters excluded 232 studies.

2.3 Screening and selection criteria

After removing the duplicate records, the remaining 1497 studies were screened based on the title, abstract, and keywords of the retrieved articles using the study inclusion and exclusion criteria (see Table 3) by four authors (GM, KK, HF, and LS). For any discrepancies, majority voting was used to include or exclude the article. Moreover, the authors (RG, LS, and NI) took the final decision in case of ties.

Table 2 shows that the article title, abstract, and keywords-based screening process included only 126 articles out of 1497 and excluded the remaining articles. Four primary reasons were use as bases for the exclusion of 1371 articles. First, the aim of the majority of the excluded articles was to extract beneficial information (e.g., concepts, named entity recognition, and relations) from the medical

documents using NLP and text mining techniques. However, their focus was not on the classification of medical documents, such as (Szenasi, Lemnaru, & Barbantan, 2015; Yala, et al., 2017). Second, only a few studies, such as (Abu-El-Haija, et al., 2017), focused on medical document classification, although these studies were unrelated to the automation of the medical document classification. Third, a few of the studies (e.g., (Prabhakar & Rajaguru, 2018)) have used the classification of time series data in the medical field, such as electroencephalography (EEG) signals, and has nothing to deal with text classification. Lastly, a few retrieved studies, such as (Kruthika, Pai, Maheshappa, & Initiative, 2017), focused on medical image classification but not on medical text classification.

Table 2. Search results across nine filters and eight academic databases

Database	Initial Search	After Duplicate Removal	After Abstract Screening	After Full-Text Screening	After Reference Scanning	After Quality Assessment
WoS	278					
Scopus	689					
IEEE Xplore	84					
PubMed	61					
Medline	181	**1497**	**126**	**65**	**72**	**72**
ScienceDirect	29					
ACM	9					
SpringerLink	398					
Total	**1729**					

Table 3. List of the inclusion and exclusion criteria

S. No.	Inclusion Criteria
1	The paper should have automated text classification of clinical reports or medical documents as one of the main topics.
2	Plain-text clinical reports or medical documents dataset should have been used.
3	The paper should describe the use of automated text classification for the classification of clinical reports using machine learning approaches.
4	The paper should be written in English. Nonetheless, papers on the processing of non-English documents were included.
5	Article must be published between 2013 and 2018.
6	Article is either conference proceeding or journal article.
S. No.	**Exclusion Criteria**
1	Studies not primarily aimed to use unstructured clinical documents
2	Studies not primarily aimed to use automated text classification for classification of plain-text clinical reports.

After the abstract-based screening, three of the authors (i.e., GM, KK, and HF) included 126 studies and skimmed through the full text of these studies to determine whether they fit our inclusion criteria (see Table 3). For any discrepancies, the majority voting technique was used to include or exclude the articles for final review. For any ties, the authors (i.e., RG, LS, and NI) made the final decision. Overall, 65 articles were retained after reading the full text of the studies and the remaining articles that did not match our inclusion criteria were excluded. The primary reason for excluding these articles is that they aimed to extract medical concepts, relations, and named entity recognition from medical documents (Jian, et al., 2017; Ju, Duan, & Li, 2016; Lopez-Guede, Moreno-Fernandez-de-Leceta, Martinez-Garcia, & Graña, 2015; Osborne, et al., 2016; Thompson, et al., 2016) using medical ontologies (e.g., SNOMEDT CT, and UMLS) and NER and NLP techniques. Moreover, a few excluded articles have used SML to identify named entities from the medical documents. The authors of these studies have annotated the medical entities with pre-defined categories and constructed a machine learning model thereafter to identify such entities (Diz, Marreiros, & Freitas, 2015; Krawczyk & Woźniak, 2015; Sreejith, Rahul, & Jisha, 2016; Syed & Das, 2015; Teyhouee, McPhee-Knowles, Waldner, & Osgood, 2017; S. Yang, Wei, Guo, & Xu, 2017). Lastly, the

references of the selected 65 articles were scanned to determine any relevant articles that fulfill our inclusion criteria. This reference scanning obtained seven new articles. Thus, 72 articles are included for this review.

2.4 Quality Assessment

The quality of the selected 72 studies was assessed against the quality assessment criteria (QAC). QAC was employed to determine if a selected primary study is suitable to address our review objectives. All the authors unanimously made a checklist of close-ended questions to assess the quality of the selected primary studies. Table A1 in Appendix-I shows the list of 10 quality checklist questions. The answer to each question can either be "Yes" or "No," which carry the weights of "1" and "0," respectively. Two groups of authors assessed the selected primary studies (Group 1: GM, LS, and KK; Group 2: NI, HF, and RG). After assessing the quality of primary studies, the results were evaluated and Cohen's κ score was computed for inter-rater agreement. Discrepancies were discussed by the two groups of authors using the Delphi method (Dalkey & Helmer, 1963) until a consensus was reached on the final selection of the primary studies. Lastly, all the authors of the current study set a threshold of 7 to include any study for the review process. This quality assessment process did not exclude any study because all the studies have obtained a score of 7 or above. Table A2 in Appendix-I shows the quality assessment criteria score of the 72 selected studies. Hence, this review involved all the selected 72 studies.

2.4 Data Extraction

Data extracted from the 72 selected primary studies were tabulated and comprised the following six aspects: (1) type of clinical reports, (2) various characteristics of the data sets used in the studies, (3) pre-processing and sampling techniques, (4) feature engineering, (5) machine learning approaches, and (6) performance metrics. Section 3 presents the critical review of these six aspects.

3.0 Review on Clinical Text Classification

This section critically reviews the selected primary studies from six different aspects, namely, types of clinical reports, dataset characteristics, pre-processing and sampling techniques, feature engineering techniques, machine learning approaches, and performance metrics.

3.1 Types of clinical reports used in the selected primary studies

Clinical text classification techniques have been employed in several types of free-text clinical reports, such as pathology reports, radiology reports, autopsy reports, death certificates, and biomedical documents. Overall, nine different types of clinical reports were identified from the literature as shown in Table 4. As shown here, majority of studies employed pathology reports, followed by biomedical documents, radiology reports, and autopsy reports. Most of the pathology reports were used to detect breast cancer or other related cancers via text classification techniques. For instance, Rani, Gladis, and Mammen (2015) used pathology reports to detect cancer stages via text classification techniques. Moreover, Kasthurirathne, et al. (2016) and Kasthurirathne, et al. (2017) investigated the use of non-dictionary-based and dictionary-based text classification approaches to detect cancer from pathology reports. Radiology reports were also used extensively in the field of clinical text classification. G. Zuccon, et al. (2013) used radiology reports to identify limb fractures via text classification techniques. Shin, Chokshi, Lee, and Choi (2017) employed radiology reports relating to brain computed tomography (brain or head CT reports) to identify paediatric traumatic brain injury (TBI). Bates, Fodeh, Brandt, and Womack (2015) used radiology reports to detect the human immunodeficiency virus (HIV) by automated text classification techniques. In addition, researchers have also classified influenza-related clinical reports to detect influenza-like illnesses using supervised machine learning (Pineda, et al., 2015; Ye, Tsui, Wagner, Espino, & Li, 2014). Furthermore, researchers have also used death certificates and autopsy reports to determine cause of death (Butt, Zuccon, Nguyen, Bergheim, & Grayson, 2013; Danso, Atwell, & Johnson, 2014; Mujtaba, Shuib, Raj, Rajandram, & Shaikh, 2018; Yeow, Mahmud, & Raj, 2014). Finally, recent studies collected and combined various clinical reports related to same disease and used those combined reports for developing the classification model (Kavuluru, Rios, & Lu, 2015; Kocbek, et al., 2016; Sarker & Gonzalez, 2015). For instance, Kavuluru, et al. (2015) combined pathology and radiology reports to develop a text classification model to automatically assign ICD-9 codes to electronic medical reports. Kocbek, et al. (2016) combined the pathology reports, radiology reports, and patients' admission-related meta-data to predict the rate of admissions against disease. In

aforementioned studies, authors reported that combining data from various sources or combining features of different reports can produce highly reliable and accurate predictions. Finally, recent studies collected various clinical reports from different sources, combined those reports, and used those combined reports for developing the classification model (Kavuluru, et al., 2015; Kocbek, et al., 2016; Sarker & Gonzalez, 2015). For instance, Kavuluru, et al. (2015) combined pathology and radiology reports to develop a text classification model to automatically assign ICD-9 codes to electronic medical reports. Kocbek, et al. (2016) combined the pathology reports, radiology reports, and patients' admission-related meta-data to predict the rate of admissions against disease. In all these three aforementioned studies, authors reported that combining data from various sources or combining features of different reports can produce highly reliable and accurate predictions.

Table 4. Types of clinical reports

Report Types	Description	Studies	No. of Studies
Influenza Related Reports	This includes emergency Department reports related to Influenza	(MacRae, et al., 2015; Pineda, Tsui, Visweswaran, & Cooper, 2013; Pineda, et al., 2015; Ye, et al., 2014)	**4**
Radiology Reports	This includes the radiology reports related to CT Abdomen, CT Neuro, limb fractures, Cancer, Retrospective Study, Invasive Fungal (IFD) Disease, HIV, Audiologic Data, imaging, and Head CT Reports	(Bates, et al., 2015; Hassanpour & Langlotz, 2016; Mabotuwana, Lee, & Cohen-Solal, 2013; Martinez, et al., 2015; Masino, Grundmeier, Pennington, Germiller, & Crenshaw, 2016; D. H. Nguyen & Patrick, 2014; Shin, et al., 2017; Wagholikar, et al., 2013; K. Yadav, et al., 2016; Y. Zhou, et al., 2014; G. Zuccon, et al., 2013)	**11**
Bio-Medical Documents	This includes Medline abstracts and medical news articles	(Adeva, Atxa, Carrillo, & Zengotitabengoa, 2014; Alghoson, 2014; Farshchi & Yaghoobi, 2013; Fragos & Skourlas, 2016; Jindal & Taneja, 2015; Jo, 2013; Mouriño-García, Pérez-Rodríguez, Anido-Rifón, & Gómez-Carballa, 2016; Parlak & Uysal, 2015, 2016, 2018; Rios & Kavuluru, 2015; H. Y. Zhou, Zhang, Wang, & Zhang, 2015)	**12**
Tweets related to Healthcare	This includes tweets related to Influence like illness (ILI) disease and user comments about hospital services	(Dai & Bikdash, 2015; Greaves, Ramirez-Cano, Millett, Darzi, & Donaldson, 2013; Guido Zuccon, et al., 2015)	**3**
Death Certificates	This includes the death certificate written in English and French and are related to cancer and other diseases	(Butt, et al., 2013; Imane & Mohamed, 2017; Koopman, Karimi, et al., 2015; Koopman, Zuccon, Nguyen, Bergheim, & Grayson, 2015; Wu & Wang, 2017)	**5**
Autopsy Reports	This includes the verbal autopsy reports and forensic autopsy reports	(Danso, Atwell, & Johnson, 2013; Danso, et al., 2014; Kalter, Perin, & Black, 2016; Miasnikof, et al., 2015; Mujtaba, Shuib, Raj, Al-Garadi, et al., 2017; Mujtaba, Shuib, Raj, Rajandram, & Shaikh, 2018; Mujtaba, Shuib, Raj, Rajandram, et al., 2017; Mujtaba, et al., 2016; Yeow, et al., 2014)	**9**
Pathology Reports	This includes pathology reports related to lymphoma, cancer, heart failure patients, stroke and migraine case reports, arthroplasty reports related to hip surgery and reports related to Cervical Spine	(Deng, et al., 2015; Garla, Taylor, & Brandt, 2013; Kasthurirathne, et al., 2016, 2017; Kasthurirathne, Dixon, & Grannis, 2015; Lauren, Qu, Zhang, & Lendasse, 2017; Luo, Sohani, Hochberg, & Szolovits, 2014; Napolitano, Marshall, Hamilton, & Gavin, 2016; Oleynik,	**13**

Report Types	Description	Studies	No. of Studies
		Patrão, & Finger, 2017; Rani, et al., 2015; Saqlain, Hussain, Saqib, & Khan, 2016; Sedghi, Weber, Thomo, Bibok, & Penn, 2016; Yoon, Roberts, & Tourassi, 2017)	
Other Clinical Reports	This includes discharge summaries of patients, nursing care records, patient history reports suffering from diabetes, child abuse consultation reports, and radiotherapy reports	(Afzal, et al., 2013; Amrit, Paauw, Aly, & Lavric, 2017; Barak-Corren, et al., 2017; Buchan, Filannino, & Uzuner, 2017; Clark, Wellner, Davis, Aberdeen, & Hirschman, 2017; Gatta, Vallati, De Bari, & Ozsahin, 2014; Hassanpour, Langlotz, Amrhein, Befera, & Lungren, 2017; Lopprich, et al., 2016; Lucini, et al., 2017; Wang, Coiera, Runciman, & Magrabi, 2017; Wei, Ju, Chun, Hua, & Jin, 2013; L. Zhou, et al., 2015)	**12**
Combination of various Reports	This includes the multi-modality reports where different set of reports belong to same disease were combined for classification task.	(Kavuluru, et al., 2015; Kocbek, et al., 2016; Sarker & Gonzalez, 2015)	**3**

3.2 Review of Dataset and their Characteristics

The free-text clinical report dataset is the essential ingredient of clinical text classification task. Nonetheless, such a dataset is useless on its own until some useful knowledge or patterns are extracted from it. The literature related to clinical text classification shows that authors mostly collected customized datasets of free-text clinical reports from their country. For instance, Ye, et al. (2014) collected the corpus of influenza-related clinical reports to detect influenza. The collected corpus comprised 592 influenza-related reports and 29,092 non-influenza-related reports. For the training set, authors used 468 influenza-related reports and 29,004 non-influenza-related reports to develop a classification model. To test the performance of developed model, authors used the test set that comprised 124 influenza-related reports and 87 non-influenza-related reports. The datasets (related to free-text clinical reports) used in the literature can be categorized into two major categories: homogenous datasets and heterogeneous datasets as shown in Figure 1 (a). The data sources (from where the dataset is collected) can also be categorized into homogenous or heterogeneous sources. This relationship is shown in Figure 1 (b) and described in subsequent paragraphs.

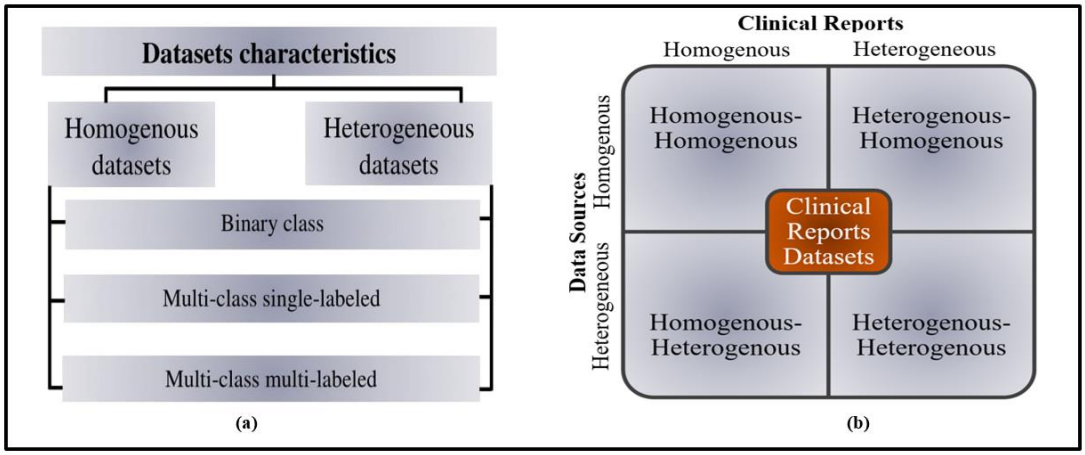

Figure 1. (a) Types of datasets used in selected studies and (b) The dataset and data source matrix

***Homogenous–Homogenous*:** Here, the dataset consists of one type of clinical report (such as pathology report), and the dataset is usually collected from one data source or hospital. In previous studies (MacRae, et al., 2015; Pineda, et al., 2013; Pineda, et al., 2015; Ye, et al., 2014), authors collected influenza-related emergency department reports from one hospital to develop a classification model for detecting influenza-like illnesses. In these studies, authors mentioned that

their results may not be generalized because the constructed classification model was trained on emergency department reports of one hospital.

Homogenous–Heterogeneous: Here, the dataset consists of one type of clinical report (such as pathology reports), and the dataset is usually collected from different data sources or hospitals. For instance, Kasthurirathne, et al. (2016) and Kasthurirathne, et al. (2017) collected 7,000 cancer-related pathology reports from seven different healthcare systems and thirty different hospitals. The experimental findings showed that combining cancer-related pathology reports from various data sources improves generalization, reliability, and classification performance. Wang, et al. (2017) developed a classification model to automate the identification of patients' safety incidents using incident reports. Here, the authors collected 6,000 incident reports from one hospital for training purposes and 5,950 incident reports from another hospital for testing purposes. The experimental findings showed the robustness of using incident reports from different data sources. Hassanpour, et al. (2017) developed a classification model to automatically classify knee magnetic resonance imaging (MRI) reports into positive or negative class. For experiments, 706 reports were collected from Duke and 1748 reports were collected from Stanford healthcare organizations. Authors reported that combining knee MRI reports from two different organizations demonstrates improved classification performance. Barak-Corren, et al. (2017) developed a prediction model to predict the risk of suicidal behaviour of patients. For the experiments, authors collected the narrative clinical notes from a variety of hospitals situated in Boston, USA, to predict the risk of suicidal behaviour of the patients. The homogenous datasets and heterogeneous data sources were collected and used to construct the generalized, accurate, and reliable classification models.

Heterogeneous–Homogenous: Here, the dataset consists of different types of clinical reports (such as pathology and radiology reports), and the dataset is usually collected from one data source or one hospital. Different reports are used for classification because a variety of reports can be prepared by a hospital for reporting the same disease; for instance, cancer cases may be reported into pathology and radiology reports. Thus, combining both of these reports in the auto-prediction model can enhance the prediction accuracy and credibility. For instance, Kavuluru, et al. (2015) combined pathology and radiology reports to develop a text classification model to automatically assign ICD-9 codes to electronic medical reports. The authors collected the pathology and radiology reports from one hospital situated at the United Kingdom. Kocbek, et al. (2016) combined three different clinical reports, namely, pathology, radiology, and patients' admission-related meta-data reports, to predict the rate of admissions against disease. The authors collected these reports from one hospital situated in Australia. The abovementioned studies reported that combining features of different clinical reports produce highly reliable and accurate predictions.

Heterogeneous–Heterogeneous: Here, the dataset consists of different types of clinical reports (such as pathology and radiology reports), and the dataset is usually collected from different data sources or different hospitals. This type of dataset is the most robust dataset for the development of the classification model. Moreover, the results generated from such datasets can be generalized on a wide scale. For instance, Sarker and Gonzalez (2015) collected Twitter tweets and daily strength instances related to adverse drug reaction events. Authors also collected adverse drug events reports from one hospital. Authors combined all these three datasets in the training set and developed a classification model for identifying adverse drug reactions. The experimental results showed that the classification performance significantly benefits from multi corpus training collected from different data sources (such as Twitter, daily strength, and hospital).

Both homogenous and heterogeneous datasets can be further divided into three subtypes, namely, binary class datasets, multi-class single labeled datasets, and multi-class multi-labeled datasets as shown in Figure 1 (a). In *binary class datasets*, reports can be labeled in either of two classes (such as cancer positive or cancer negative). Wagholikar, et al. (2013) and G. Zuccon, et al. (2013) collected radiology reports corpus related to limb fracture. This corpus comprised 99 radiology reports. Each report was labeled with "normal" or "abnormal" class. Of these reports, 90% were used as the training set, and the remaining 10% was used as the test set. In *multi-class single labeled datasets*, clinical reports were composed of more than two categories; however, each report was categorized into one label. For instance, Mujtaba, et al. (2016) developed a classification model to predict CoDs from forensic autopsy reports. Authors collected the dataset from one of the biggest hospitals situated in Kuala Lumpur, Malaysia. The dataset comprised 400 forensic autopsy reports. These 400 reports were labeled into eight different CODs. Authors used tenfold cross validation (Kohavi, 1995;

Refaeilzadeh, Tang, & Liu, 2009) to evaluate the classification model's performance. In studies (Adeva, et al., 2014; Alghoson, 2014; Fragos & Skourlas, 2016; Jindal & Taneja, 2015; Jo, 2013; Mouriño-García, et al., 2016; Parlak & Uysal, 2015, 2016, 2018; Rios & Kavuluru, 2015; H. Y. Zhou, et al., 2015), authors used the subset of the OHSUMED dataset to classify medical abstracts into 23 cardiovascular diseases. This subset of the OHSUMED dataset contains 13,929 Medline abstracts. Nonetheless, each abstract may fall into more than one category, but the authors only considered those Medline abstracts, which fell under one category only. Of these 13,929 Medline abstracts, 6,286 abstracts were used in the training set and the remaining abstracts were used in the test set. In *multi-class multi-labeled datasets*, clinical reports comprised more than two categories, and each report was categorized into more than one class label. For instance, Imane and Mohamed (2017) collected the French Center for Epidemiology and Medical Causes of Death (CépiDC) dataset, which includes death certificates. The dataset contained 65,843 death certificates labeled by 3232 ICD-10 (the international classification of diseases code-version 10) codes. Each certificate may be assigned one or more ICD-10 codes. Of these 65,843 certificates, 52,675 death certificates were used for training purposes, and 13,168 death certificates were used for testing purposes.

Table 5. Related literature based on dataset and data source matrix

Type of Dataset		References
Homogenous - Homogenous	Binary class	(Afzal, et al., 2013; Amrit, et al., 2017; Bates, et al., 2015; Butt, et al., 2013; Dai & Bikdash, 2015; Deng, et al., 2015; Lopprich, et al., 2016; Lucini, et al., 2017; Mabotuwana, et al., 2013; MacRae, et al., 2015; D. H. Nguyen & Patrick, 2014; Pineda, et al., 2013; Pineda, et al., 2015; Saqlain, et al., 2016; Wagholikar, et al., 2013; Wei, et al., 2013; K. Yadav, et al., 2016; Ye, et al., 2014; Guido Zuccon, et al., 2015; G. Zuccon, et al., 2013)
	Multi-class single-label	(Adeva, et al., 2014; Alghoson, 2014; Buchan, et al., 2017; Clark, et al., 2017; Danso, et al., 2013, 2014; Farshchi & Yaghoobi, 2013; Fragos & Skourlas, 2016; Garla, et al., 2013; Gatta, et al., 2014; Greaves, et al., 2013; Hassanpour & Langlotz, 2016; Jindal & Taneja, 2015; Jo, 2013; Kalter, et al., 2016; Kasthurirathne, et al., 2015; Koopman, Karimi, et al., 2015; Koopman, Zuccon, et al., 2015; Lauren, et al., 2017; Luo, et al., 2014; Masino, et al., 2016; Miasnikof, et al., 2015; Mouriño-García, et al., 2016; Napolitano, et al., 2016; Oleynik, et al., 2017; Parlak & Uysal, 2015, 2016, 2018; Rani, et al., 2015; Rios & Kavuluru, 2015; Shin, et al., 2017; Wu & Wang, 2017; Yeow, et al., 2014; Yoon, et al., 2017; H. Y. Zhou, et al., 2015; L. Zhou, et al., 2015; Y. Zhou, et al., 2014)
	Multi-class multi-label	(Imane & Mohamed, 2017)
Homogenous - Heterogeneous	Binary class	(Kasthurirathne, et al., 2016, 2017; Martinez, et al., 2015)
	Multi-class single-label	(Barak-Corren, et al., 2017; Sedghi, et al., 2016)
	Multi-class multi-label	-
Heterogeneous - Homogenous	Binary class	(Hassanpour, et al., 2017)
	Multi-class single-label	(Kavuluru, et al., 2015; Kocbek, et al., 2016; Mujtaba, Shuib, Raj, Al-Garadi, et al., 2017; Mujtaba, Shuib, Raj, Rajandram, & Shaikh, 2018; Mujtaba, Shuib, Raj, Rajandram, et al., 2017; Mujtaba, et al., 2016; Wang, et al., 2017)
	Multi-class multi-label	-
Heterogeneous - Heterogeneous	Binary class	(Sarker & Gonzalez, 2015)
	Multi-class single-label	-
	Multi-class multi-label	-

Table 5 shows the distribution of related literature based on datasets and data source matrix. The majority of studies (57 out of 72) homogenous–homogenous dataset. Of these 57 studies, 36 studies used multi-class single-labeled datasets, 20 used binary class datasets, and only one study used multi-class multi-label datasets. Moreover, five studies used homogenous–heterogeneous dataset. Of these five studies, two used multi-class single-labeled datasets, and three employed binary class datasets. Furthermore, eight studies used heterogeneous–homogenous dataset. Of these eight studies, seven studies used multi-class single-labeled datasets, and one utilized a binary class dataset. Finally, only one study heterogeneous– heterogeneous dataset. Furthermore, several studies have used single-view data to classify clinical reports. For example, pathologists conducting a forensic autopsy examination collect multi-view information (including external examination information, anatomical examination information, eyewitness information, medical history, and summary of the case) on the deceased to determine the cause of death. However, Yeow, et al. (2014) used only single-view data (i.e., summary of the case) of the forensic autopsy reports to predict the cause of death. The developed classification model may be minimally robust and reliable but can be improved by including the features of other views in the decision. To achieve such improvement, Mujtaba, Shuib, Raj, Rajandram, et al. (2017) considered multi-view data (including external examination information, anatomical examination information, eyewitness information, medical history, and summary of the case) to predict the cause of death. Nonetheless, the classification accuracy reported by Mujtaba, Shuib, Raj, Rajandram, et al. (2017) is lower than that of Yeow, et al. (2014). That is, the model of the former is more reliable than that of the latter because of the use of multiple views in the classification decision. In conclusion, one should consider heterogeneous–heterogeneous and multi-view data sets for the development of considerably robust and reliable clinical text classification models.

3.3 Review of Preprocessing and Sampling Techniques

In clinical text classification, preprocessing involves removing meaningless data from the collected dataset to improve the quality of clinical text classification models. The narrative clinical reports corpus contain high level of noise, sparsity, complex medical vocabularies, medical measures and scores (such as, BP 140/65), abbreviations, misspelled words, and poor grammatical sentences (Kaurova, et al., 2011; H. Nguyen & Patrick, 2016). Therefore, to address the aforementioned issues, the detailed pre-analysis of corpora is needed for revealing linguistic properties of texts in clinical text classification task. In the related literature, preprocessing techniques such as removal of stop words, removal of punctuations or special symbols, removal of empty spaces, case conversion, spell correction, tokenization, stemming, lemmatization, and normalization were applied (as shown in Table 6). Table 6 shows the related literature based on applied preprocessing tasks. Majority of the studies employed basic preprocessing tasks (including stop word removal, removal of punctuation and white spaces, and case conversion) and word tokenization. In addition, these studies reported the effectiveness of these preprocessing techniques on clinical text classification. Nonetheless, few studies (Danso, et al., 2013, 2014; Lauren, et al., 2017; Sarker & Gonzalez, 2015) empirically investigated the presence and absence of stop words and reported that the presence of stop words produces better classification accuracy than their absence. In some studies, researchers demonstrated that applying stemming task with basic preprocessing tasks and word tokenization enhances classification performance (Adeva, et al., 2014; Jo, 2013; Koopman, Karimi, et al., 2015; Koopman, Zuccon, et al., 2015; Sarker & Gonzalez, 2015). Buchan, et al. (2017) and Wang, et al. (2017) applied stemming and lemmatization for clinical text normalization with basic preprocessing tasks and word tokenization. They also reported the effectiveness of using stemming and lemmatization techniques. Nonetheless, Lauren, et al. (2017) applied text classification to classify arthroplasty reports and empirically investigated the effectiveness of stemming and lemmatization to preprocess the arthroplasty reports. Experimental results showed the unsuitability of stemming and lemmatization when applied on psychiatric evaluation reports. Clark, et al. (2017) applied text classification techniques for classifying psychiatric evaluation reports to detect the severity of mental disorders and reported the unsuitability of stemming and lemmatization when applied on psychiatric evaluation reports. Martinez, et al. (2015) and Masino, et al. (2016) applied basic preprocessing techniques with word tokenization to classify radiology reports. In addition to basic preprocessing techniques, researchers also applied few text normalization techniques using regular expressions to convert numbers or dates to common units such as *number* and *date*. The experimental findings showed that such text normalization techniques improve the classification accuracy and overcome the issue of dimensionality. Thus, form the aforementioned literature, it can be concluded that the practitioners should empirically investigate the performance of several preprocessing techniques on collected narrative clinical corpus to evaluate the classification performance.

In general, the narrative clinical reports datasets have imbalanced class distribution where one class which is of more interest (such as, cancer positive) is insufficiently represented compared to other class (for instance, cancer negative). Thus, in such datasets, the relatively low number of positive cases are most likely classified as rare occurrences, ignored, or assumed as outliers and cause more misclassifications compared to the majority class. Hence, there is a critical need to identify such a rare class among the normal instances. Conversely, any diagnostic errors may bring stress and further complications to the patients. Thus, it is crucial that a classification model should be able to achieve higher identification rate on the positive class in the datasets. To address this clinical text classification challenge, several researchers have developed and employed the sampling strategies.

Table 6. Preprocessing tasks used in related literature

Studies	Preprocessing Techniques	Study Count
(Afzal, et al., 2013; Bates, et al., 2015; Dai & Bikdash, 2015; Danso, et al., 2013, 2014; Fragos & Skourlas, 2016; Garla, et al., 2013; Greaves, et al., 2013; Hassanpour & Langlotz, 2016; Jindal & Taneja, 2015; Kalter, et al., 2016; Kasthurirathne, et al., 2015; Kavuluru, et al., 2015; Lopprich, et al., 2016; Napolitano, et al., 2016; D. H. Nguyen & Patrick, 2014; Parlak & Uysal, 2015; Rani, et al., 2015; Saqlain, et al., 2016; Sedghi, et al., 2016; Shin, et al., 2017; Wagholikar, et al., 2013; K. Yadav, et al., 2016; Yoon, et al., 2017; H. Y. Zhou, et al., 2015; Y. Zhou, et al., 2014)	These studies reported that the removal of stop-words, punctuation marks, white spaces, and converting text into lower case improves the classification performance.	26
(Adeva, et al., 2014; Amrit, et al., 2017; Jo, 2013; Kasthurirathne, et al., 2016, 2017; Kocbek, et al., 2016; Koopman, Karimi, et al., 2015; Koopman, Zuccon, et al., 2015; Lucini, et al., 2017; Mujtaba, Shuib, Raj, Rajandram, & Shaikh, 2018; Mujtaba, et al., 2016; Oleynik, et al., 2017; Parlak & Uysal, 2016, 2018; Sarker & Gonzalez, 2015)	These studies reported that the converting the text into lower case, converting the text into word tokens, applying stemming technique, and removing the stop-words, punctuation marks, and white spaces improve the classification performance.	15
(Buchan, et al., 2017; Mujtaba, Shuib, Raj, Al-Garadi, et al., 2017; Mujtaba, Shuib, Raj, Rajandram, et al., 2017; Wang, et al., 2017)	These studies reported that the converting the text into lower case, converting the text into word tokens, applying stemming and lemmatization techniques, spell correction, and removing the stop-words, punctuation marks, and white spaces improve the classification performance.	4
(Imane & Mohamed, 2017; Lauren, et al., 2017; Luo, et al., 2014)	These studies reported that the converting the text into lower case, removing the stop-words, punctuation marks, and white spaces, and converting the text into sentence tokens improve the classification performance.	3
(Martinez, et al., 2015; Masino, et al., 2016)	These studies reported that the converting the text into lower case, converting the text into word tokens, and converting numeric measures and some units into common terms improve the classification performance.	2
(Clark, et al., 2017)	These studies reported that the converting the text into lower case, converting the text into word tokens, applying spell checker, and removing the stop-words, punctuation marks, and white spaces improve the classification performance.	1

Studies	Preprocessing Techniques	Study Count
(Alghoson, 2014; Barak-Corren, et al., 2017; Butt, et al., 2013; Comelli, Agnello, Vitabile, & Ieee, 2015; Danso, et al., 2014; Deng, et al., 2015; Farshchi & Yaghoobi, 2013; Gatta, et al., 2014; Hassanpour, et al., 2017; Miasnikof, et al., 2015; Mouriño-García, et al., 2016; Pineda, et al., 2013; Pineda, et al., 2015; Wei, et al., 2013; Wu & Wang, 2017; Ye, et al., 2014; Yeow, et al., 2014; L. Zhou, et al., 2015; Guido Zuccon, et al., 2015; G. Zuccon, et al., 2013)	These studies have not reported any preprocessing techniques	21

In selected primary studies, various studies have used skewed data for the development of text classification model with relatively low number of positive cases of any given disease against total cases. For instance, Kocbek, et al. (2016) developed a text classification model for classifying breast cancer reports into positive and negative class. In experiments, authors used 177 positive breast cancer reports and 1131 negative breast cancer reports. In such datasets, the class distribution is biased towards the majority class in the sense that the classifier would predict the major class to obtain the overall accuracy. However, in dealing with the highly skewed data, the aim should be accurately predicting the minority class to obtain low false-negative rate and at the same time maintaining the overall accuracy (Witten, Frank, Hall, & Pal, 2016). Thus, to overcome the issue of data skewness, various studies have employed the resampling techniques to distribute the data equally. Among the resampling techniques, researchers have either employed over-sampling technique or under-sampling technique. These techniques are briefly discussed below;

Over-sampling: In this technique, the size of minority class is increased through random replication of positive instance. This technique contributes to balance the class distribution without adding any new information to original dataset (Japkowicz, 2000; Tang & Liu, 2005). For instance, Afzal, et al. (2013) developed an automated case identification system using free-text clinical notes. The dataset used in experiment was highly imbalanced. Authors empirically investigated the use of over-sampling the minority class with cost-sensitive meta classifier and reported the effectiveness of over-sampling towards achieving better classification accuracy. Nonetheless, over-sampling contributes to balance the class-distribution, however, this technique may be more susceptible to model over-fitting because of replicating existing positive cases. To overcome the issue of over-sampling Chawla, Bowyer, Hall, and Kegelmeyer (2002) proposed an alternative of over-sampling technique called SMOTE (Synthetic Minority Over-sampling TEchnique). SMOTE produces new minority class samples by interpolating between preexisting positive instances that lie close together. However, the SMOTE technique was not used in selected primary studies and hence can be a potential future work to test the effectiveness of SMOTE techniques for class distribution.

Under-Sampling: This technique randomly removes the negative cases from the dataset to make dataset balance (Japkowicz, 2000; Tang & Liu, 2005). Amrit, et al. (2017) developed an intelligent system for accurately identifying the child abuse cases. In experiments, authors used plain-text child consultation reports for classifying the child abuse cases. The dataset contained only 5% child abused cases and 95% normal cases. Thus, to address the data skewness, authors employed under-sampling method and remove the cases of majority class. The experimental results showed the improvement in classification accuracy after employing under-sampling technique to overcome the issue of data imbalance. Moreover, Kocbek, et al. (2016) developed a text classification model for classifying breast cancer reports into positive and negative class. In experiments, authors used 177 positive breast cancer reports and 1131 negative breast cancer reports. Hence, to address the data imbalance issue, authors employed under-sampling technique and reported the effectiveness of this technique on classification performance. Nonetheless, in under-sampling techniques the data can be lost, however, it has been proven one of the most successful resampling method (Japkowicz, 2000; Tang & Liu, 2005).

3.4 Feature Engineering

In text classification, one of the key step is feature engineering (Aggarwal & Zhai, 2012a; Domingos, 2012; Heer, Hellerstein, & Kandel, 2015; Jiang, 2012; Tantug, 2010; Witten, et al., 2016; Wolpert & Macready, 1995). Feature engineering is combination of three sub steps namely, feature extraction, feature value representation, and feature selection (as discussed in section 1). Thus, subsequent sub

sections present the review of feature extraction techniques, feature values representation techniques, and feature selection techniques used in selected studies.

3.4.1 Feature Extraction

Feature extraction is the process of extracting useful features from free-text clinical reports. The complete taxonomy of features used in the selected studies is shown in Figure 2. As shown here, the researchers have usually employed and empirically investigated two general approaches of feature extraction, namely, expert-driven (Barak-Corren, et al., 2017; Clark, et al., 2017; Dai & Bikdash, 2015; Sarker & Gonzalez, 2015; Sedghi, et al., 2016) and fully automated feature extraction (Bates, et al., 2015; Comelli, et al., 2015; D. H. Nguyen & Patrick, 2014; Wei, et al., 2013). Thus, this section aims to present details of both of these approaches with their subtypes (as shown in Figure 2). Moreover, it also presents the related literature that compared both of these features to evaluate the classification performance.

3.4.1.1 Fully-automated feature extraction approaches

Here, the features are automatically extracted from given clinical reports by computer programs through the use of various statistical approaches. In these approaches, no human or expert intervention is required. The literature shows that researchers used the automated feature extraction techniques to extract content-based features, concept-based features, structural features, and linguistic features. The content-based features are usually extracted from the content of free-text clinical reports. These features include BoW, *n*-gram, and Word2Vec. In BoW model, the unique words are extracted from all clinical reports available in the dataset irrespective of their categories. All the extracted words are then sorted in ascending order and stored in a list called 'bag of words (BoW)'. In BoW each available word represents an independent, and discriminative feature. An *n*-gram is the contiguous sequence of *n* items (such as words, or characters) from a given sequence of clinical text. They are typically a set of co-occurrence words within a given window. In *n*-gram, *n* may be '1', '2', '3', or any number. When the $n = 1$, it is called unigram, when $n = 2$, it is called bigram, and when $n = 3$, it is called as trigram. Word2Vec exists in two models: skip-gram and continuous bag of words (CBoW) (Goldberg & Levy, 2014). The skip-gram model learns iteratively from existing words available in a sentence to predict the next word. By contrast, the CBoW model uses the neighboring words to predict the current word. In both skip-gram and CBoW, the parameter window size determines the limit on number of words used in the context.

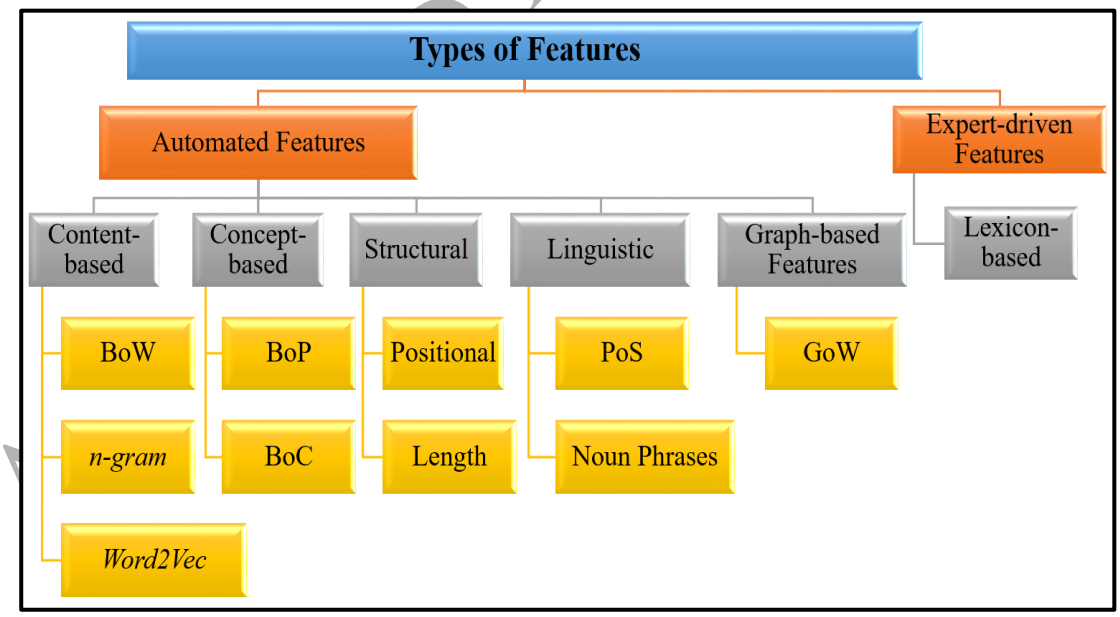

Figure 2. Features used in related literature

****BoW** (Bag of Words), **PoS** (Parts of Speech), **BoP** (Bag of Phrases), **BoC** (Bag of Concepts), **GoW** (Graph of words), **Word2Vec** (Word2Vector)

Medical experts may use different terms to describe same condition in free-text clinical reports. For instance, experts may use the term 'heart attack' or 'Myocardial infarction' interchangeably. Though, both the terms belong to same medical concept however, the content-based feature extraction

techniques cannot identify the relationship between these two terms. Therefore, to overcome this issue, concept-based features are extracted by using the specialized medical ontologies such as, SNOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms) (Donnelly, 2006; Stearns, Price, Spackman, & Wang, 2001). The concept-based features are used to extract medical concepts instead of just terms from clinical reports. In literature two widely used concept-based features were identified namely, Bag of Phrases (BoP), and Bag of Concepts (BoC). In BoP, medical phrases are extracted from the clinical reports through the help of some tool such as, MetaMap (Aronson, 2001). For instance, suppose a sentence 'multiple grazed abrasions over the back of right hand' is available in a clinical report. The MetaMap tool will extract following phrases from the sentence ('multiple', 'grazed', 'abrasion', and 'Back of right hand'). In BoC, specialized medical ontology is used to extract the concept id of similar medical terms and these concept ids are used a s features to differentiate between the classes. For instance, SNOMED-CT is a standardized medical ontology where medical related terms are categorized into medical hierarchical concepts. The root concept is SNOMED CT concept and this concept has many child concepts such as medical conditions, body structures, procedures, etc. In SNOMECTD CT, each medical term has a unique concept id and similar medical terms share the same concept id. Moreover, each medical terms may belong to one or more concepts. For instance, the words 'heart attack' and 'Myocardial infarction' have the same concept id (22298006) and both belong to same parent concept 'Ischemic heart disease' (concept id-414545008). Thus in this particular case, BoC includes the SNOMEDT-CT concept ids as a features.

The Structural features exploit the structure or form of clinical documents for obtaining the discriminative features. These features include length of the clinical reports, number of sentences available in the clinical reports, number of sections available in the clinical reports, and position of the word in a given sentence. Furthermore, the linguistic features are used to determine the correct sense of given words used in the text. These features include parts of speech (POS) features. In POS, each word is represented with its POS tag. Finally, in graph-based features or graph of word (GoW) features, the content- or concept-based features are represented in graphs to capture word order. In GoW, each given clinical report is represented in a graph structure where each vertex of graph contains the distinct feature of the given clinical report and two co-occurring features are connected by an edge. In addition, each edge has the weight that is computed by considering the co-occurrence frequency of neighbouring vertices in the graph. The rationale in using GoW features is to consider the word order in the given clinical reports.

Table 7 shows the type of automated features used in selected primary studies. In most of studies, researchers used BoW features for classifying free-text clinical reports. For instance, in studies (Jindal & Taneja, 2015; Jo, 2013; Parlak & Uysal, 2015, 2016, 2018; H. Y. Zhou, et al., 2015), authors employed BoW features to classify Medline abstracts. Moreover, Garla, et al. (2013) and Kasthurirathne, et al. (2015) extracted and used BoW features for classifying cancer reports. Wu and Wang (2017) and Wang, et al. (2017) extracted BoW features from death certificates to determine ICD-10 codes of reported cause of deaths. Oleynik, et al. (2017) and Kasthurirathne, et al. (2017) classified pathology reports using BoW features. Moreover, Yeow, et al. (2014) used BoW features to determine cause of death from forensic autopsy reports and reported that BoW is useful for predicting CoD. Although the BoW model is simpler and effective in classifying clinical reports, it also suffers from one major limitation. In the BoW model, grammar and word order are disregarded but word frequency is maintained. Therefore, to address the limitations of BoW model, *n*-gram feature extraction technique was proposed (Cavnar & Trenkle, 1994).

Several studies have extracted and used *n*-gram features to classify narrative clinical reports. For instance, Mujtaba, et al. (2016) empirically investigated the performance of *n*-gram features (where $n = 1$ to 3) for the classification of forensic autopsy reports. The experimental findings revealed that unigram outperforms bigram and trigram features, but the performance of bigram was slightly lower than that of unigram. Moreover, Lucini, et al. (2017) investigated the effectiveness of unigram, bigram, and trigram to predict hospital admission using free-text emergency department reports. The experimental results showed that trigram outperforms unigram and bigram. Y. Zhou, et al. (2014) employed *n*-gram features to classify radiology reports; they experimentally investigated *n*-gram from 1 to 8 and reported that *n*-gram (where $n = 4$) obtained the best classification accuracy. Masino, et al. (2016) investigated the character *n*-gram and word *n*-gram features (where $n = 1$ to 3) to classify radiology reports. Their experimental results showed that word bi-gram and word-trigram demonstrate enhanced results. Moreover, Masino, et al. (2016) compared the effectiveness of word-level and character-level unigram, bigram, and trigram features to classify radiology reports. Their

experimental results showed that word-level bigram and trigram outperform the other features. Many studies have demonstrated the effectiveness of the *n*-gram approach, but this technique has three major limitations. First, the *n*-gram approach does not capture word inversion. Second, this approach ignores the word-level synonymy when applied on clinical text reports. Finally, the number of features increases enormously with increasing *n*, thereby resulting in dimensionality. To overcome these issues, researchers employed other kinds of features such as BoP and BoC.

Pineda, et al. (2015) employed BoP features to extract useful medical phrases using MetaMap tool from influenza-related free-text clinical reports. Kocbek, et al. (2016) used BoP and BoC features to predict the admission against disease via a combination of pathology, radiology, and admission-related patients' data. In studies (Bates, et al., 2015; Comelli, et al., 2015; D. H. Nguyen & Patrick, 2014; Wei, et al., 2013), authors employed BoW and BoC features to classify free-text clinical reports, and they reported that the hybrids of both BoW and BoC features enhances the classification accuracy. In recent studies (Amrit, et al., 2017; Buchan, et al., 2017; Martinez, et al., 2015), researchers used the combination of BoW, n-gram, BoC, structural, and linguistic features to classify clinical reports. Their experimental results showed that structural and linguistic features obtain robust classification accuracy when combined with content-based and concept-based features. Though, the concept-based features are useful in achieving the word-level synonymy and polysemy, in these features, grammar and word order are disregarded. To address this limitation, recently, Yoon, et al. (2017) and Mujtaba, Shuib, Raj, Rajandram, Shaikh, et al. (2018) used GoW features to classify breast cancer reports and forensic autopsy reports respectively. Their experimental results revealed that the proposed GoW features outperformed the traditional features. Nonetheless, GoW is more effective than traditional BoW and *n*-gram but computationally expensive relative to BoW or *n*-gram.

To summarize, it was found that in most of the studies, researchers have extracted content-based (BoW or n-gram) features from clinical reports. Though, the BoW and n-gram feature extraction techniques are simpler and have proven effective in classifying clinical reports. However, these approaches do not consider word-level synonymy and polysemy when applied on narrative clinical reports. To overcome this limitation, the researchers used concept-based (BoC and BoP) features. Nonetheless, BoC and BoP features have proven useful to address the issue of word-level synonymy and polysemy. However, in these features, the grammar and even word order is disregarded but word frequency is kept. Moreover, such features do not capture the word inversion and subset matching. Thus, to overcome the limitations of content-based and concept-based features, researchers employed BoC and GoW features to address the issue of word order, and word-level synonymy and polysemy.

3.4.1.2 Expert-driven Feature Extraction

The groups of experts are responsible for discovering the useful and discriminative features from the clinical reports. Moreover, the experts rank the extracted features on the basis of their discriminative power and store those features in lexicons for classification. This approach requires readily available expert knowledge in the form of decision rules, expert domain knowledge, and human expertise. Table 7 shows various studies that employed expert-driven features for classifying clinical reports. Dai and Bikdash (2015) employed expert-driven approaches to manually extract features (related to medicine and alcohol) from tweets related to influenza. After extracting the features, the authors developed two lexicons: one for storing the features related to medicine and another for storing the features related to alcohol. Finally, the authors used the developed lexicons to classify influenza-related tweets. Sarker and Gonzalez (2015) extracted *n*-gram and BoC features from Twitter tweets, daily strength instances, and free-text clinical reports related to adverse drug reaction. Moreover, the authors developed expert-driven lexicons that contain some useful features to classify adverse drug reaction-related tweets or clinical reports related to adverse drug reaction. The experimental findings showed that a combination of *n*-gram features, BoC features, and manually created expert-driven lexicons leads to enhanced classification accuracy. Deng, et al. (2015) employed expert-driven feature extraction to classify pathology reports and obtained good classification accuracy. Sedghi, et al. (2016) developed expert-driven lexicons with the help of experts for migraine and stroke-related cases. Saqlain, et al. (2016) prepared and used expert-driven lexicons for predicting the heart failure risk of heart patients. In the aforementioned studies, the authors reported that lexicon-based features result in high classification accuracy.

3.4.1.3 Expert-driven versus fully-automated features

To obtain a specific level of classification performance in the clinical text classification domain, several studies empirically investigated the effectiveness of expert-driven and fully automated

approaches. Pineda, et al. (2013) compared the classification performance of expert-driven features and fully automated features to classify influenza-related reports. Their experimental results showed no significant difference between the classification performance obtained through expert-driven and fully automated feature extraction approaches. G. Zuccon, et al. (2013) employed the ATC technique to identify limb fracture radiology reports. The authors developed two different text classification models using expert-driven and fully automated features. Their experimental results showed that the text classification model developed using fully automated features obtained 3% more classification accuracy than that of the expert-driven model. Ye, et al. (2014) employed SML-based ATC techniques to classify influenza-related clinical reports. Features were extracted by domain experts from collected datasets. These extracted features were fed to a classifier to categorize the collected report. To compare the effectiveness of expert-driven features, authors also extracted the BoP features using automated tools, namely, TOPZ and MEDLEE. These tools extracted the medical phrases used in clinical reports. These extracted features were then fed to a classifier to organize the influenza-related reports. The experimental results showed that expert-driven features outperformed TOPZ and MEDLEE features. Koopman, Karimi, et al. (2015) developed an ATC model to classify death certificates through expert-driven and fully automated features. In expert-driven features, experts extracted the useful terms, features, or keywords from death certificates. Conversely, in the automated approach, the BoW and BoC features were extracted from death certificates. The experimental results showed a minute difference in performance between these two approaches; the performance obtained by the expert-driven approach was 1% higher than that of the fully automated approach. Kalter, et al. (2016) developed an ATC model to classify verbal autopsy reports using expert-driven and fully automated features. The experimental results showed that expert-driven features outperform fully automated features. Masino, et al. (2016) developed a text classification model with and without the help of domain expert intervention to classify temporal bone-related radiology reports. Text classifiers with expert intervention obtained the highest classification accuracy than those without expert intervention. Kasthurirathne, et al. (2016) employed expert-driven features and fully automated features to classify cancer-related pathology reports. No significant difference in classification performance was found between the expert-driven and fully automated approaches.

Both expert-driven and fully-automated feature extraction approaches have their own pros and cons. The expert-driven approach is flexible and can easily understand the importance of manually extracted features. Moreover, the misclassification error can be easily fixed when working with expert-driven features. The major drawback of expert-driven approach is that it depends heavily on the deep skills and knowledge of domain experts for robustness and scalability. The expert-driven approach is not purely a scientific activity but more of a balancing act in black art, architecture, design, and development. This approach is time consuming and resource extensive. Finally, this approach is not easily extendable; for any new class or category, experts will be engaged to extend the functionality of the existing model. Nevertheless, this technique is effective in creating baseline results so that further automated methods can be designed and engineered to obtain accuracy similar or better than expert-driven approaches. Conversely, the fully automated feature extraction approaches are less time consuming and do not require any expert intervention to extract useful features from clinical reports. Nonetheless, the major limitation of these approaches is their requirement for a large number of labeled clinical reports for extracting useful features that correlate well with the class. Moreover, in medical domains, one cannot rely only on fully automated techniques, so a robust comparison of fully automated and expert-driven approaches is needed to evaluate their performance differences.

Table 7. Features, feature representation and feature selection schemes used in selected studies

Study	Features	Feature Representation	Feature Selection
(Afzal, et al., 2013)	BoW	BR	Chi-Square
(Garla, et al., 2013)	Bow	BR	--
(Jo, 2013)	BoW	TF	--
(Pineda, et al., 2013)	*n*-gram	BR	--
(Wei, et al., 2013)	BoW and BoC	TF	IG
(Butt, et al., 2013)	*n*-gram and BoC	BR and TF	--
(G. Zuccon, et al., 2013)	*n*-gram and BoC	TF	--
(Danso, et al., 2013)	*n*-gram and LGF	N-TFiDF	--
(Farshchi & Yaghoobi, 2013)	BoW, STF, LGF	TF and TFiDF	--
(Greaves, et al., 2013)	EDF	TF	IG

Study	Features	Feature Representation	Feature Selection
(Wagholikar, et al., 2013)	EDF	BR	--
(Danso, et al., 2014)	*n*-gram	BR, TF, TFiDF and N-TFiDF	LSFS
(Gatta, et al., 2014)	BoW	TFiDF	--
(Yeow, et al., 2014)	BoW	TF	--
(Y. Zhou, et al., 2014)	*n*-gram	BR	--
(Luo, et al., 2014)	GoW	TF	--
(D. H. Nguyen & Patrick, 2014)	BoW and BoC	BR	--
(Alghoson, 2014)	EDF	BR	--
(Ye, et al., 2014)	*n*-gram and EDF	BR	--
(Adeva, et al., 2014)	*n*-gram	TFiDF	Chi-Square
(Jindal & Taneja, 2015)	BoW	TF	--
(Kasthurirathne, et al., 2015)	BoW	TF	--
(Parlak & Uysal, 2015)	BoW	TF	--
(H. Y. Zhou, et al., 2015)	BoW	TF	--
(Rani, et al., 2015)	*n*-gram	BR	--
(Pineda, et al., 2015)	*n*-gram	BR	--
(Bates, et al., 2015)	BoW and BoC	BR	MI
(Comelli, et al., 2015)	BoW and BoC	BR	--
(Kavuluru, et al., 2015)	*n*-gram and BoC	BR	BNSS
(Koopman, Karimi, et al., 2015)	*n*-gram and BoC	BR	--
(Koopman, Zuccon, et al., 2015)	*n*-gram and BoC	BR	IG
(Martinez, et al., 2015)	BoW, BoS, BoP, BoC, and STF	BR	PC
(Rios & Kavuluru, 2015)	W2V	TF	--
(Guido Zuccon, et al., 2015)	*n*-gram and STF	BR	--
(Dai & Bikdash, 2015)	EDF	BR	ED
(Deng, et al., 2015)	EDF	BR	--
(MacRae, et al., 2015)	EDF	BR	--
(Sarker & Gonzalez, 2015)	EDF	TFiDF	--
(Miasnikof, et al., 2015)	BoW and EDF	BR	--
(L. Zhou, et al., 2015)	*n*-gram	TF	--
(Kasthurirathne, et al., 2016)	BoW	TF	IG
(Lopprich, et al., 2016)	BoW	TF	--
(Parlak & Uysal, 2016)	BoW	TF	GI and DFS
(Mujtaba, et al., 2016)	*n*-gram	BR, TF, TFiDF and N-TFiDF	--
(Hassanpour & Langlotz, 2016)	*n*-gram	TFiDF	--
(Masino, et al., 2016)	*n*-gram	BR and TF	--
(Napolitano, et al., 2016)	BoS and BoP	BR and TF	--
(Mouriño-García, et al., 2016)	BoC	TF	--
(K. Yadav, et al., 2016)	*n*-gram and BoC	TF	--
(Kocbek, et al., 2016)	BoP and BoC	TF	IG
(Fragos & Skourlas, 2016)	BoW and BoP	TFiDF	--
(Kalter, et al., 2016)	EDF	BR	--
(Saqlain, et al., 2016)	EDF	TF	--
(Sedghi, et al., 2016)	EDF	BR	--
(Imane & Mohamed, 2017)	BoW	TFiDF	--
(Kasthurirathne, et al., 2017)	BoW	TF	IG
(Oleynik, et al., 2017)	BoW	TFiDF	--
(Wang, et al., 2017)	BoW	BR, TF and TFiDF	--
(Wu & Wang, 2017)	BoW	TFiDF	--
(Hassanpour, et al., 2017)	*n*-gram	TFiDF	--
(Yoon, et al., 2017)	GoW	TF	--
(Lauren, et al., 2017)	W2V	TF	--
(Shin, et al., 2017)	W2V	BR, TF, TFiDF and N-TFiDF	--
(Amrit, et al., 2017)	BoW and STF	BR, TF and TFiDF	--
(Buchan, et al., 2017)	*n*-gram, BoC and LGF	N-TFiDF	--
(Barak-Corren, et al., 2017)	EDF	BR	--
(Clark, et al., 2017)	EDF	BR and TF	MI
(Lucini, et al., 2017)	*n*-gram	BR, TF and TFiDF	Chi-Square
(Mujtaba, Shuib, Raj, Rajandram, & Shaikh, 2018)	*n*-gram	BR, TF, TFiDF and N-TFiDF	IG, Chi-Square and PC

Study	Features	Feature Representation	Feature Selection
(Mujtaba, Shuib, Raj, Rajandram, et al., 2017)	EDF	TF	IG, Chi-Square and PC
(Mujtaba, Shuib, Raj, Al-Garadi, et al., 2017)	*n*-gram and BoC	TF	--
(Parlak & Uysal, 2018)	BoW	TF and TFiDF	DFS

****BoW** (Bag of Words), **NGF** (*n*-gram Features), **BoS** (Bag of Sentences), **BoP** (Bag of Phrases), **BoC** (Bag of Concepts), **GoW** (Graph of words), **W2V** (Word2Vector), **STF** (Structural Features), **LGF** (Linguistics Features), **EDF** (Expert-driven Features)

** **BR** (Binary Representation), **TF** (Term Frequency), **TFiDF** (Term Frequency with inverse Document Frequency), **N-TFiDF** (Normalized TFiDF)

****IG** (Information Gain), **Chi** (Chi-Square), **PC** (Pearson Correlation), **GI** (Gini-Index), **LSFS** (Local Semi-Supervised Feature Selection), **ED** (Expert-driven), **MI** (Mutual Information), **MDA** (Multiple Discriminant Analysis), **BNSS** (Bi-Normal Separation Score), **PCA** (Principal Component Analysis), **DFS** (Distinguishing Feature Selector)

To summarize, in the field of clinical text classification, researchers have comprehensively investigated the performance of classification models using expert-driven features and fully automated features. Moreover, varying results were obtained. Few studies showed that expert-driven features outperform fully automated features, and several studies reported that automated features outperform expert-driven features. Some studies reported no significant difference between the classification performance obtained through expert-driven and fully automated features. Therefore, one should empirically investigate the performance of both approaches on free-text clinical reports to evaluate the superior one.

3.4.2 Review of Feature Representation Techniques

Feature representation (also called term-weighting) is an important step after extracting the features from clinical reports. This step is responsible for assigning the numeric value to each extracted feature for linear algebraic methods to learn the classification rules (Debole & Sebastiani, 2004). In selected studies, researchers employed either binary representation (BR), term frequency (TF), term frequency with inverse document frequency (TFiDF), or normalized TFiDF (N-TFiDF) for feature value representation. In BR, the feature contains the value of either '0' or '1', where '1' denotes the presence of a feature in the clinical report and '0' denotes its absence. In TF, the feature value is computed by its frequency of occurrence in a clinical report. However, if the reports belonging to different classes contain the same feature, then that feature may not be a discriminative feature. To address this, TFiDF feature representation scheme was introduced where the feature f is a discriminative feature if it frequently occurs in clinical reports belonging to one class and less frequently available in reports belonging to another class. Finally, in N-TFiDF term frequency and document frequency are combined with a normalized factor such as, the length of the clinical reports to ensure features found in long and short clinical reports are equally important.

Table 7 shows the study-wise frequency distribution of each feature representation technique. As can be seen here, in most of the studies, researchers have used BR or TF technique. Moreover, in (Butt, et al., 2013; Masino, et al., 2016; Napolitano, et al., 2016) authors compared the performance of BR and TF to classify clinical reports and reported that BR outperformed TF. Clark, et al. (2017) compared the performance of BR and TF to classify psychiatric evaluation reports. The experimental results showed that TF outperformed BR. Kavuluru, et al. (2015) compared the performance of BR and TFiDF to classify pathology and radiology reports. The experimental results showed that there was no significant difference between BR and TFiDF results. In (Parlak & Uysal, 2016, 2018) authors compared the performance of TF and TFiDF to classify Medline abstracts. The experimental results showed that TF outperformed TFiDF. Farshchi and Yaghoobi (2013) extracted the BoW features from dataset of medical news articles and represented the extracted features using TF and TFiDF feature representation techniques. The experimental results showed that there was no significant difference between the results obtained through TF and TFiDF. Amrit, et al. (2017) compared BR, TF, and TFiDF to classify child abuse consultation reports and reported that TF outperformed. Lucini, et al. (2017) compared BR, TF, and TFiDF to classify emergency department reports and reported that TFiDF outperformed. Wang, et al. (2017) compared BR, TF, and TFiDF to classify incident reports and reported that BR outperformed. Danso, et al. (2014) compared BR, TF, TFiDF, and N-TFiDF to determine cause of death from verbal autopsy reports. Their experimental results showed that N-TFiDF outperformed. Moreover, Mujtaba, Shuib, Raj, Rajandram, and Shaikh (2018) compared BR, TF, TFiDF, and N-TFiDF to categorize forensic autopsy reports. Their experimental results revealed that TF and TFiDF outperformed.

It can be noted from above discussion that the choice of feature representation schemes affects the classification results. This is because all four feature representation schemes have a different design philosophy. Thus, it is always better to empirically investigate the use of all four types of feature representation schemes on clinical datasets to see which one obtain the better classification accuracy. To summarize, mostly the researchers have either employed BR or TF approach to represent the features. Though, the BR approach is easy to compute and it constructs a basic binary numeric vector to differentiate between the two documents. Nonetheless, it can only be suitable for the datasets with controlled terminologies with slight conceptual differences. Conversely, the TF, TFiDF, N-TFiDF approaches may be suitable the datasets with uncontrolled vocabulary. Moreover, these approaches can easily compute the similarity between two documents. However, there are two major limitations of these approaches. First, these approaches cannot capture the position in the text. Finally, these approaches cannot capture the semantics, and co-occurrences in different clinical reports.

3.4.3 Feature Selection Techniques

Feature selection techniques discover the most relevant subset of features following certain selection criteria (Guyon & Elisseeff, 2003). Thus, feature selection is widely used for efficient clinical text classification. Nonetheless, in the related literature on clinic text classification, very few studies have employed feature selection to examine the effect of various subsets of features on classification accuracy. Most features for construction of text classification models have been used. In the related literature, the following feature selection techniques were employed.

Information Gain (IG): It identifies the significance of a given feature *f* in a feature vector, and the expected reduction in entropy caused by segregating the data sample according to *f* (Y. Yang & Pedersen, 1997).

Chi-square (chi): The Chi-square test (χ^2) is the statistical test that measures the relevance of feature *f* with class *c* (Y. Yang & Pedersen, 1997).

Pearson Correlation (PC): It is a commonly used method for reducing feature dimensionality and evaluating the discrimination power of a feature in classification methods. It is also a straightforward method for choosing significant features. Pearson correlation measures the relevance of a feature by computing the Pearson correlation between it and a class. Pearson correlation coefficient measures the linear correlation between two attributes (Benesty, Chen, Huang, & Cohen, 2009).

Local Semi-Supervised Feature Selection (LSFS): This technique defines a margin for each data sample in the dataset and selects the most result-oriented features by increasing the margins with reference to a feature weight vector (Xu, King, Lyu, & Jin, 2010).

Expert-driven (ED): In ED, experts manually rank the discriminative features from the set of given features.

Mutual Information (MI): It computes the amount of information of a feature *f* contributes in accurate classification decision (Cover & Thomas, 2012).

Gini-Index (GI): It is a non-purity split method. It is widely used in decision trees. It calculates the heterogeneity from the sum of squared probabilities of each class from one (Loh, 2011).

Distinguishing Feature Selector (DFS): It aims is to select the most informative features while removing irrelevant ones with reference to some pre-determined conditions (Uysal & Gunal, 2012).

Principal Component Analysis (PCA): It is a statistical method that utilizes orthogonal transformation to transform a set of observations of correlated features into principal components. In general, the count of principal components is less than or equal to original number of observations (Wold, Esbensen, & Geladi, 1987).

Multiple Discriminant Analysis (MDA): It is a statistical technique that is used to minimize the differences between variables so as to categorize them into a broad groups (Hair, Black, Babin, Anderson, & Tatham, 1998).

Bi-Normal Separation Score (BNSS): It is defined as $F^{-1}(tp) - F^{-1}(fp)$ where F^{-1} is the inverse cumulative probability of standard normal distribution (Forman, 2003).

Table 7 shows the distribution of related studies based on feature selection techniques used. As shown here, a few studies have employed feature selection techniques to discover discriminative feature subsets. Among those studies, IG, chi square, and Pearson correlation (PC) were mainly used. For instance, Mujtaba, Shuib, Raj, Rajandram, and Shaikh (2018) compared three feature selection techniques, namely, information gain (IG), chi square, and PC for discovering discriminative feature subsets for the classification of forensic autopsy reports. Their experimental findings showed that IG

and chi-square outperform PC. Kasthurirathne, et al. (2016) compared the performance of manually ranked features by experts with that of IG (fully automated feature selection scheme) to classify cancer reports. N significant difference was found between the expert-driven feature ranking and IG techniques. Amrit, et al. (2017) used GI and chi-square feature selection schemes to classify child abuse consultation reports; their findings demonstrated that chi square outperforms GI. Parlak and Uysal (2016) and Parlak and Uysal (2018) compared GI and DFS feature selection schemes to classify Medline abstracts. They revealed that DFS outperforms the GI technique. Buchan, et al. (2017) applied PCA and MI feature selection schemes to classify diabetic patient history reports; PCA was found to outperform MI. It can be inferred from Table 7 that IG, and Chi-square are most widely employed feature selection schemes for clinical text classification. Moreover, both techniques obtain better classification results. This is because IG, and chi-square favor the most frequently used clinical terms in the available dataset. Moreover, both of these techniques also use category information to discover useful feature subset. In addition, these techniques also take into account the information of clinical term absences to determine the category probability (Y. Yang & Pedersen, 1997). The possible reason for poor performance of MI or GI is that these techniques are bias towards low frequent term features (Y. Yang & Pedersen, 1997). In few studies, the PC performed worse than Chi-square. The possible reason may be because it is more appropriate for dichotomous data (Nicolosi, 2008; Sebastiani, 2002). Moreover, in few studies PC showed the lowest results, this may be because PC selects the features that are most indicative of membership only, whereas Chi-square and IG consider the terms features most indicative of membership and non-membership and that may be useful for classification performance (Forman, 2003; Y. Yang & Pedersen, 1997).

3.5 Review of Machine Learning Approaches

In selected primary studies, the clinical reports were classified using either RB learning approach or SML approach (as shown in Figure 3). In SML, both generative and discriminative algorithms were employed (Aggarwal & Zhai, 2012b; Alabbas, Al-Khateeb, & Mansour, 2016; Witten, et al., 2016). The generative algorithms learn the joint probability distribution $p(x, y)$. These algorithms do not focus on differences between the classes. Conversely, these algorithms try to build a model that is representative of particular class. Naïve Bayes (NB) is a good example of generative supervised machine learning algorithm (Aggarwal & Zhai, 2012b; Sebastiani, 2002; Witten, et al., 2016). On the other hand, the discriminative algorithms learn the conditional probability distribution $p(y|x)$. These algorithms learn the hard and soft boundary between class. The discriminative algorithms highlight the differences between two classes. The examples of discriminative algorithms include, support vector machine, linear regression, decision trees, and neural networks (Aggarwal & Zhai, 2012b; Sebastiani, 2002; Witten, et al., 2016).

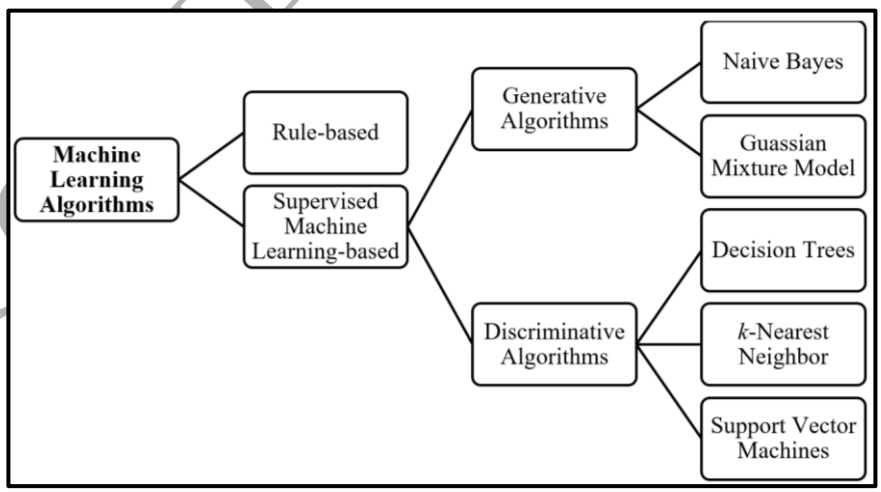

Figure 3. Machine learning algorithms used in related literature

Table 8 shows the machine learning algorithms that were employed in the experiments along with the preeminent algorithm that obtained the highest classification results. Notably, in several studies, authors did not compare various machine learning algorithms but only employed one algorithm. In such studies, the third column value is empty. Moreover, different studies have used different clinical reports dataset and thus, it is not viable to compare the performance of classification models across

the selected studies. For instance, Butt, et al. (2013) employed support vector machine (SVM), naïve bayes (NB), decision tree (DT), and AdaBoost text classifiers to classify cancer-related death certificates. Their findings showed that SVM outperforms NB, DT, and AdaBoost by obtaining 98% F-measure. Farshchi and Yaghoobi (2013) classified medical news articles utilizing artificial neural network (ANN) via back propagation, k-nearest neighbour (*k*NN), NB, and SVM. ANN using back propagation obtained the highest overall accuracy of 86%. Pineda, et al. (2013) employed efficient Bayesian multivariate classification (EBMC) to classify influenza-related clinical reports. Moreover, to demonstrate the effectiveness of EMBC, authors compared its performance with those of NB, baysian network (BN), random forest (RF), SVM, liner regression (LR), and ANN. EMBC was revealed to outperform all other classifiers by obtaining 99% area under the curve (AUC). Afzal, et al. (2013) modified the C4.5 decision tree algorithm to develop a new classifier called My C to classify hepatobiliary disease and renal failure reports. This classifier builds a decision tree by recursively splitting samples using the chi-square test results. To show the effectiveness of My C, its performance was compared with those of C4.5 decision tree, SVM, and Ripper. My C obtained the highest sensitivity of 94%. Moreover, authors experimentally proved that the proposed My C classifier is computationally faster than existing decision tree algorithms. Kasthurirathne, et al. (2015) investigated the performances of LR, NB, *k*NN, RF, and DT text classifiers to classify cancer reports. The LR, *k*NN, RF, and DT classifiers obtained the highest accuracy with no significant differences among them. Pineda, et al. (2015) compared the performance of seven different text classifiers, namely, NB, BN, EBMC, RF, SVM, LR, and ANN, to classify influenza-related clinical reports. They reported that the NB, LR, SVM, and ANN classifiers almost obtained similar results. In studies (Danso, et al., 2013, 2014), authors compared three different text classifiers (SVM, NB, and RF) for the classification of verbal autopsy reports; SVM obtained the highest accuracy of 83%. Mujtaba, Shuib, Raj, Rajandram, and Shaikh (2018) compared the performance of six different classifiers, namely, NB, SVM, *k*NN, DT, RF, and ensemble voting classifier to classify forensic autopsy reports; SVM and RF obtained the highest accuracy of 78%. Kasthurirathne, et al. (2016) and Kasthurirathne, et al. (2017) investigated the performance of NB, LR, DT, RF, and *k*NN text classifiers to classify cancer-related pathology reports. They reported that DT and RF outperform the other techniques by obtaining 90% F-measure. Kasthurirathne, et al. (2017) developed and compared the classification models with and without domain-related ontologies to classify cancer-related pathology reports. No significant difference in classification performance was observed when models were developed with or without domain-related ontologies.

Table 8. Table showing the study-wise preeminent classifier and other compared classifiers

Study	**Preeminent Classifier**	**Compared with**
(Pineda, et al., 2013)	EBMC	NB, BN, RF, SVM, LR, and ANN
(Comelli, et al., 2015)	*k*NN	--
(Jo, 2013)	Proposed Table-based	*k*NN, NB, ANN, and SVM
(Greaves, et al., 2013)	NB	DT, Bagging, and SVM
(Farshchi & Yaghoobi, 2013)	ANN	*k*NN, NB, and SVM
(Butt, et al., 2013)	SVM	NB, DT, and AdaBoost
(Afzal, et al., 2013)	My C	DT, SVM, and Ripper
(Danso, et al., 2013)	SVM	ZeroR
(G. Zuccon, et al., 2013)	RB	NB, and SVM
(Garla, et al., 2013)	SVM	RB
(Wagholikar, et al., 2013)	NB	RB
(Wei, et al., 2013)	SVM	PLS-DA
(Ye, et al., 2014)	BN tuned with expert	BN tuned with TOPZ, and with MEDLEE
(Gatta, et al., 2014)	ESA	Rocchio, and NB
(Yeow, et al., 2014)	CBR	--
(Danso, et al., 2014)	SVM	NB, and RF
(D. H. Nguyen & Patrick, 2014)	SVM	--
(Luo, et al., 2014)	SVM	--

Study	Preeminent Classifier	Compared with
(Alghoson, 2014)	RB	--
(Adeva, et al., 2014)	SVM	KNN, NB, and Rocchio
(Sarker & Gonzalez, 2015)	SVM	NB, and MEM
(Pineda, et al., 2015)	NB, LR, SVM, and ANN	NB, BN, EBMC, RF, SVM, LR, and ANN
(Parlak & Uysal, 2015)	BN	DT, and RF
(Koopman, Zuccon, et al., 2015)	SVM	--
(Koopman, Karimi, et al., 2015)	RB	SVM
(Kasthurirathne, et al., 2015)	LR, *k*NN, RF, and DT	LR, NB, *k*NN, RF, and DT
(Jindal & Taneja, 2015)	L-*k*NN	*k*NN
(Guido Zuccon, et al., 2015)	SVM	NB, LR, J48, RF, and LMT
(Y. Zhou, et al., 2014)	DLM and NB	DLM and NB
(L. Zhou, et al., 2015)	DT	SVM, *k*NN, and RIPPER
(H. Y. Zhou, et al., 2015)	*k*NN	--
(Rios & Kavuluru, 2015)	CNN	NB, SVM, LR, AdaBoost, and Voted Classifier
(Rani, et al., 2015)	RF	J48, NB, and LAD Tree
(Miasnikof, et al., 2015)	NB	OTM, and Inter-VA4
(Martinez, et al., 2015)	SVM	BN, NB, and RF
(MacRae, et al., 2015)	RB	Human Expert Classification
(Kavuluru, et al., 2015)	LR	NB, SVM, and LR
(Dai & Bikdash, 2015)	NB	--
(Bates, et al., 2015)	SVM	--
(Deng, et al., 2015)	RB	--
(Napolitano, et al., 2016)	*k*NN	PAUM, and NB
(Kasthurirathne, et al., 2016)	RF, and DT	LR, NB, and KNN
(Mujtaba, et al., 2016)	SVM, and RF	NB, KNN, DT, and ensemble-Voted classifier
(Mouriño-García, et al., 2016)	BN	--
(Kocbek, et al., 2016)	SVM	SVM, and NB
(Parlak & Uysal, 2016)	BN	DT
(K. Yadav, et al., 2016)	DT	--
(Sedghi, et al., 2016)	PART	NB, and SVM
(Saqlain, et al., 2016)	NB	LR, SVM, ANN, RF, and DT
(Masino, et al., 2016)	SVM and LR	DT, RF, and NB
(Lopprich, et al., 2016)	SVM	MEM
(Kalter, et al., 2016)	RB	Compare with fully automated
(Hassanpour & Langlotz, 2016)	SVM	--
(Fragos & Skourlas, 2016)	Extended lf-igf-*k*NN	*k*NN, and lf-igf *k*NN
(Yoon, et al., 2017)	RF	NB, and LR
(Oleynik, et al., 2017)	SVM	--
(Kasthurirathne, et al., 2017)	RF, and DT	LR, NB, and *k*NN
(Mujtaba, Shuib, Raj, Rajandram, & Shaikh, 2018)	SVM, and RF	NB, KNN, DT, and ensemble-Voted classifier
(Buchan, et al., 2017)	NB	MaxEnt, and SVM
(Amrit, et al., 2017)	NB	RF, and SVM
(Lucini, et al., 2017)	SVM	DT, RF, Random Trees, AdaBoost, LR, and NB

Study	Preeminent Classifier	Compared with
(Wu & Wang, 2017)	CNN	LR, NB, and SVM
(Wang, et al., 2017)	SVM	LR
(Shin, et al., 2017)	CNN	LR, RF, and SVM
(Mujtaba, Shuib, Raj, Rajandram, et al., 2017)	RF, and SVM	NB, SVM, KNN, and DT
(Mujtaba, Shuib, Raj, Al-Garadi, et al., 2017)	SVM	NB, and RF
(Lauren, et al., 2017)	ELM	--
(Imane & Mohamed, 2017)	DT	SVM, and AdaBoost
(Hassanpour, et al., 2017)	SVM	--
(Clark, et al., 2017)	ANN	--
(Barak-Corren, et al., 2017)	NB	--
(Parlak & Uysal, 2018)	BN	DT

NB (Naive Bayes), **BN** (Bayesian Network), **SVM** (Support Vector Machine), **RF** (Random Forest), **DT** (Decision Tree), ***k*NN** (k-Nearest Neighbor), **ANN** (Artificial Neural Network), **CNN** (Convolutional Neural Network), **LR** (Linear Regression), **EBMC** (Efficient Bayesian Multivariate Classification), **ESA** (Entropy Scoring Algorithm), **CBR** (Case-based Reasoning), **RB** (Rule-based), and **AUC** (Area Under the Curve)

As shown in Table 8, the majority of the studies used SML-based algorithms to classify free-text clinical reports. However, these algorithms are characterized by two major limitations. First is the knowledge bottleneck, in which a decent SML algorithm requires a large number of labeled clinical reports for constructing an accurate classification model (Hastie, et al., 2009). Hence, many believe that the quality of SML-based algorithms heavily depends on data rather than algorithms. Another major limitation of SML-based algorithms is difficulty to fix reported quality bugs (Hastie, et al., 2009). The developed model is usually a black box, and no direct expert intervention is available to fix the problem unless the constructed model is retrained with new features. However, in such models, there is no guarantee that the reported issue will be fixed well with retraining because the learning process needs to balance all the features in the newly constructed model. Thus, to overcome the aforementioned limitations of the SML-based algorithms, researchers developed clinical text classification models using rule-based (RB) algorithms.

Alghoson (2014) developed a RB classifier to classify Medline abstracts and obtained 60% precision. Moreover, Deng, et al. (2015) developed a RB classifier to classify pathology reports and obtained 91% F-measure. Koopman, Karimi, et al. (2015) developed a RB classifier to classify death certificates and compared its performance with that of SVM. Their experimental results showed a minor difference in performance between these two classifiers. The RB classifier obtained 95% F-measure, and SVM obtained the 94% F-measure. Kalter, et al. (2016) developed a RB classifier to classify verbal autopsy reports. In the developed classifier, the authors used the rules defined by domain experts to determine the CoD. Moreover, the results of the developed classifier were compared with two fully automated systems, namely, Tariff and Inter-VA4. The developed RB classifier was revealed to outperform the automated systems by obtaining 80% overall accuracy. The abovementioned studies showed good classification accuracy using RB classifiers, but this approach has its own advantages and disadvantages. The RB approach is flexible, with rules that are easy to understand. Misclassification results are easier to fix in the RB approach than with other approaches. However, the major limitation of this approach is that it depends more on the deep skills and knowledge of domain experts and rule designers for robustness and scalability. Moreover, this approach is not purely a scientific activity but more of a balancing act in architecture, design, and development.

The frequency count of preeminent machine learning algorithms in the selected studies is shown in Figure 4. In most of the related studies, a customized dataset was used. Thus, comparing the performance values across different related studies is inadvisable. Nevertheless, when performance was analyzed, most of the studies found that the SVM algorithm outperforms many other algorithms, followed by NB, RF, DT, RB, BN, and *k*NN. The least used classifiers were LR and ANN. SVM with appropriate kernel function (such as poly kernel and RBF kernel) can learn good classification rules on linear and non-linear data. Moreover, SVM exhibits enhanced performance with high-dimensional data. The limitations of SVM include memory requirement, complexity, and interpretability

(Cristianini & Shawe-Taylor, 2000). In many comparative studies, *k*NN showed the lowest classification performance. The *k*NN algorithm computes the similarity between a new clinical report and a training set of clinical reports. The *k*-most similar cases are retrieved in descending order. The new clinical report is assigned with a class label that belongs to majority of the retrieved *k* reports (Fukunaga, 2013). The modest classification performance of *k*NN may be due to the linear scaling of features, which possibly inaccurately computed the *k*NN distance measures. Moreover, this assumption of linear scaling becomes misleading when the master feature vector contains non-discriminative features (Fukunaga, 2013; Hastie, et al., 2009). Note that only three studies have used the convolutional neural network (CNN) algorithm to classify clinical reports. CNN is a deep learning-based algorithm and it is capable of learning complex features from the clinical data set compared with traditional generative and discriminative machine learning algorithms. Although CNN showed considerable classification accuracy, its major limitation is the amount of data that it needs to learn the complex features from a clinical report dataset. Therefore, CNN will show limited classification accuracy with only a few instances in the training set. The reason is that CNN has to learn several feature weights to determine the most discriminative and result-oriented features for classifying the clinical reports. Thus, enormous training data are required to achieve this objective. In the clinical text classification domain, the clinical reports related to any particular disease may be generally unavailable in large volumes. In such cases, pre-trained deep learning models or TL can be maximized to classify clinical reports compared with CNN (Do & Ng, 2006).

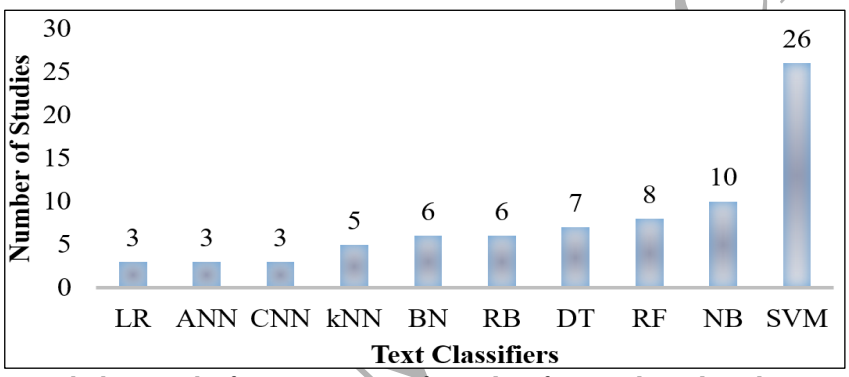

Figure 4. Graph showing the frequency count of text classifiers used in selected primary studies

3.6 Review of Performance Metrics

The performance of constructed clinical text classification model can be measured through the use of various performance metrics. These performance metrics include, accuracy, precision, recall, F-measure, specificity, sensitivity, micro and macro averaging of accuracy, precision, recall, and F-measure, and area under the curve. The values of these metrics can be computed by using the values of true positive (TP), false positive (FP), false negative (FN), and true negative (TN) from confusion matrix. In selected primary studies, different types of performance metrics were employed to evaluate the classification performance. The commonly used performance metrics for binary class problems were precision, recall, F-measure, accuracy, area under curve, sensitivity, and specificity. However, in multi-class problems the commonly used performance metrics were micro or macro-averaging of precision, recall, and F-measure are used. These performance metrics are briefly described in subsequent paragraphs. However, the detailed discussion on these performance metrics can be found in (Sokolova & Lapalme, 2009).

- ***Precision:*** It is the ratio of correctly predicted positive clinical reports to the total positively predicted clinical reports. It is also known as positive predictive value (PPV). It is formally defined as following:

$$precision = \frac{TP}{TP + FP}$$

- ***Recall:*** It is the ratio of correctly predicted positive clinical reports to the all clinical reports in actual positive class. It is also known as true positive rate (TPR) or sensitivity. It is formally defined as following:

$$recall = \frac{TP}{TP + FN}$$

- ***F-measure:*** It is the weighted average of precision and recall. It is formally defined as following:

$$F-measure=\frac{2\times(precision\times recall)}{(precision+recall)}$$

- ***Accuracy:*** It is the most widely used performance metric. It is the ratio of correctly predicted clinical reports to the total clinical reports. It is formally defined as following:

$$accuracy=\frac{(TP+TN)}{(TP+FP+FN+TN)}$$

- ***Receiver Operating Characteristic Curve (ROC curve):*** It plots the rate of true positive (TPR) against the rate of false positive (FPR). The TPR and FPR are defined as following

$$TPR=\frac{TP}{TP+FN} \text{ and } FPR=\frac{FP}{FP+TN}$$

Figure 5 shows a typical ROC graph where two ROC curves are shown for Algorithm 1, and 2. The purpose of this graph is to have a model be at the upper left corner and getting no false positives (a perfect classifier (obtaining no false positives). The area under the ROC curve is called receiver operating characteristic area under curve (AUCROC) is just the area under the ROC curve. The higher it is, the better the model is. AUROC is used to compute the goodness of clinical report classification model by plotting a particular curve and computing area under that curve. The value of '1' for AUC shows the classifier performance is good. Conversely, when AUC value is 0.5 or lower than that shows the poor performance of clinical report classifier (Fawcett, 2006; Hand & Till, 2001; Provost, Fawcett, & Kohavi, 1998). It is formally defined as following:

$$AUC=\frac{S_0-n_0(n_0+1)/2}{n_0n_1}$$

Here, n_0 and n_1 denote the count of positive and negative clinical reports respectively, and $S_0=\sum r_i$ where r_i is the rank of i_{th} positive sample in ranked list.

- ***Specificity:*** It measures the proportion of negative clinical reports that are correctly predicted a negative. It is formally defined as following:

$$specificity=\frac{TN}{TN+FP}$$

- ***Micro- and Macro-average of Precision, Recall and F-measure:*** In micro averaging of precision, recall, and F-measure, individual TP, FP, and FN of the system for different sets are summed up and then apply them to get the statistics. Conversely, in macro averaging of precision, recall, and F-measure, simply the average of precision, recall or F-measure of the system on different sets is taken. The formal definitions of micro and macro averaging of precision, recall, and F-measure can be found in (Sokolova & Lapalme, 2009).

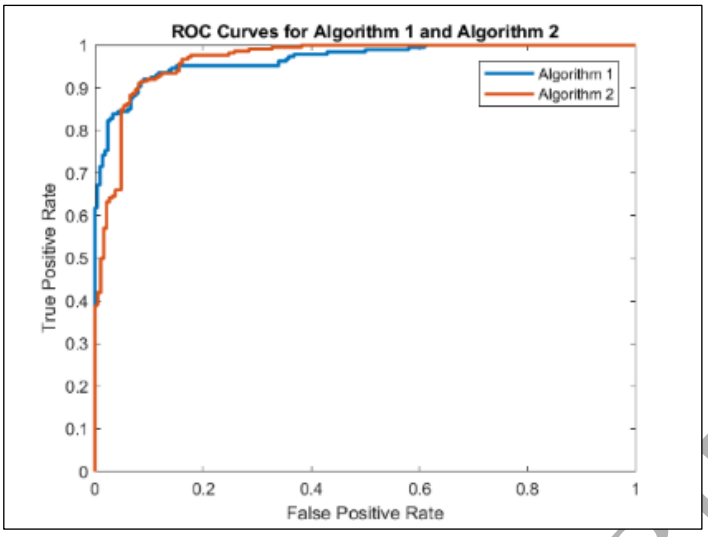

Figure 5. Example of ROC curve

Table 9 shows the frequency count of performance measures used in each study. Majority of the studies either employed precision, recall, and F-measure or used F-measure and accuracy for binary class problems. For multi-class problems, the studies either used micro-averaging or macro-averaging. In general, macro-averaging is used to determine overall performance of a system across sets of data. Conversely, micro-averaging is effective when the datasets vary in size (Sokolova & Lapalme, 2009). Though, the analysis showed that the most commonly employed performance metrics were precision, recall, and F-measure, but these metrics alone may not be sufficient to evaluate classifier performance correctly. For instance, the dataset was imbalanced in various studies. In such cases, the AUC should be the correct performance metric for evaluating classification performance correctly because AUC is suitable in computing the classification performance pertaining to individual class (Provost & Fawcett, 1997; Provost, et al., 1998). For instance, Sarker and Gonzalez (2015) collected three different datasets (i.e., Twitter tweets, daily strength instances, and clinical reports) to develop a classification model for predicting adverse drug events. Moreover, the authors used accuracy and F-measure metrics to evaluate classification performance. The Twitter dataset comprised 11.4% tweets that mention ADR and 88.6% tweets that do not mention ADR. Moreover, daily strength dataset contained 23.7% instances that mention ADR and 76.3% instances that do not mention ADR. Finally, the clinical reports were composed of 29.0% ADR mentions and 71.0% that do not mention ADR. In the above example, all three datasets were imbalanced in nature. In such cases, accuracy or F-measure metrics may be biased toward the majority class. Thus, AUC is a correct choice in such cases to determine the performance of classifiers accurately. Ye, et al. (2014) collected the corpus of influenza-related clinical reports to develop a classifier for classifying influenza-related clinical reports. The collected corpus was imbalanced in nature and comprised 592 influenza-related reports and 29,092 non-influenza-related reports. Thus, the authors employed AUC to address the class imbalance problem and evaluate the performance of classifiers accurately.

Table 9. Frequency count of performance metrics used in each selected primary study

Study	Metrics	Count
(Alghoson, 2014; Barak-Corren, et al., 2017; Butt, et al., 2013; Imane & Mohamed, 2017; Jindal & Taneja, 2015; Kocbek, et al., 2016; Koopman, Karimi, et al., 2015; Lauren, et al., 2017; Lucini, et al., 2017; Luo, et al., 2014; Martinez, et al., 2015; Mouriño-García, et al., 2016; Napolitano, et al., 2016; Wang, et al., 2017; L. Zhou, et al., 2015)	Recall, Precision, and F-Measure	**14**
(Bates, et al., 2015; Farshchi & Yaghoobi, 2013; Fragos & Skourlas, 2016; Greaves, et al., 2013; Jo, 2013; Lopprich, et al., 2016; Parlak & Uysal, 2016, 2018; Sarker & Gonzalez, 2015; Wagholikar, et al., 2013; Wei, et al., 2013; Guido Zuccon, et al., 2015; G. Zuccon, et al., 2013)	Accuracy and F-Measure	**13**
(Comelli, et al., 2015; Dai & Bikdash, 2015; Gatta, et al., 2014;	Accuracy	**9**

Study	Metrics	Count
Kalter, et al., 2016; Rani, et al., 2015; Shin, et al., 2017; Wu & Wang, 2017; Yeow, et al., 2014; Y. Zhou, et al., 2014)		
(Danso, et al., 2013, 2014; Koopman, Zuccon, et al., 2015; Mujtaba, Shuib, Raj, Al-Garadi, et al., 2017; Mujtaba, Shuib, Raj, Rajandram, & Shaikh, 2018; Mujtaba, Shuib, Raj, Rajandram, et al., 2017; Mujtaba, et al., 2016)	Macro Averaging of Accuracy, Recall, Precision, and F-Measure	**7**
(Amrit, et al., 2017; Hassanpour, et al., 2017; Kasthurirathne, et al., 2015; Parlak & Uysal, 2015; Saqlain, et al., 2016; Ye, et al., 2014)	Accuracy, Recall, Precision, and F-Measure	**6**
(Garla, et al., 2013; Hassanpour & Langlotz, 2016; Kasthurirathne, et al., 2016, 2017; Masino, et al., 2016; K. Yadav, et al., 2016)	F-Measure, Sensitivity, and Specificity	**6**
(Deng, et al., 2015; D. H. Nguyen & Patrick, 2014; Oleynik, et al., 2017; Rios & Kavuluru, 2015; H. Y. Zhou, et al., 2015)	F-Measure	**5**
(Afzal, et al., 2013; Clark, et al., 2017; MacRae, et al., 2015; Miasnikof, et al., 2015; Sedghi, et al., 2016)	Sensitivity and Specificity	**5**
(Adeva, et al., 2014; Buchan, et al., 2017; Kavuluru, et al., 2015; Yoon, et al., 2017)	Micro Averaging of Accuracy, Recall, Precision, and F-Measure	**4**
(Lucini, et al., 2017; Pineda, et al., 2013)	Accuracy, F-Measure, and AUC	**2**
(Pineda, et al., 2015)	AUC	**1**

Several studies employed simple precision, recall, and F-measure for multi-class classification (Farshchi & Yaghoobi, 2013; Gatta, et al., 2014; Jo, 2013). However, the suitable performance metrics for multi-class classification problems are micro- and macro-averaging precision, recall, and F-measure (Sokolova & Lapalme, 2009). Mujtaba, et al. (2016) developed a clinical text classification model to determine the CoD from a forensic autopsy dataset that comprised eight different CoDs. To evaluate the classification performance, authors employed macro-averaging precision, recall, and F-measure. Danso, et al. (2014) developed a clinical text classification model for determining the CoD from a verbal autopsy dataset that comprised 16 different CoDs. To evaluate the classification performance, authors employed macro-averaging precision, recall, and F-measure. Yoon, et al. (2017) developed a clinical text classification model to determine the cancer stage from pathology reports. The dataset comprised pathology reports related to cancer. These reports were related to four different stages of cancer: Grades I, II, III, and IV. Thus, to evaluate the performance of classifiers, authors used micro-averaging precision, recall, and F-measure.

4.0 Discussion

This study comprehensively reviewed academic articles on clinical text classification published from January 2013 to January 2018. In particular, the current research maximized the procedural decision analysis in six aspects, namely, types of clinical reports, characteristics of data sets used, pre-processing and sampling techniques, feature engineering, machine learning approaches, and use of performance measures.

The findings of this review indicate that clinical text classification has been employed in various application domains, including breast cancer, lung cancer, influenza-like illness, child abuse, autopsy, and adverse drug reaction events. In each domain, various types of clinical reports were used to develop a clinical text classification model, including radiology reports, pathology reports, cancer reports, biomedical documents, MRI reports, autopsy reports, influenza-related reports, death

certificates, and healthcare-related social networking site instances (e.g., Twitter and DailyStrength). In many cases, these reports are collected from healthcare organizations, such as hospitals. The data set size and quality are positively correlated with the classifier performance. Several studies have collected the homogenous–homogenous data set to develop a classification model. However, note that several hospitals may have different medical documentation systems and patterns or styles, thereby possibly producing hurdles in generalizing constructed classifier to multiple hospitals. Moreover, one disease may be reported in a variety of reports. For example, cancer patients' findings can be reported in pathology and radiology reports. Hence, the practitioners in clinical text classification are suggested to use the heterogeneous–heterogeneous reports to develop a classification model. The major limitation in the clinical text classification research is the collection of a balanced data set with a sufficient sample size for a training set to enable the text classifiers to effectively learn from training sets and determine the category of the test set. Many studies have been observed to be using imbalanced data sets, in which the samples of predicting classes vary in size (Afzal, et al., 2013; Amrit, et al., 2017; Kocbek, et al., 2016). In these cases, many studies (e.g., (Sarker & Gonzalez, 2015)) have also been observed to lack the appropriate emphasis on employing the suitable validation approach to evaluate the classification performance. Therefore, the suggestion is that in the case of imbalanced class distribution data sets, the appropriate sampling techniques (e.g., over-sampling, under-sampling, or SMOTING (Chawla, et al., 2002; Japkowicz, 2000; Tang & Liu, 2005)) should be utilized to balance the class distribution in the data sets. For imbalanced datasets, the appropriate performance metrics (e.g., AUC) should be used to accurately evaluate the classification performance (Provost & Fawcett, 1997; Provost, et al., 1998). Several studies have been observed to employ inappropriate performance metrics for the multi-class classification problem. Thus, the recommendation for multi-class classification problems is that appropriate measures, such as micro and macro-averaging precision, recall, F-measure, and overall accuracy, are the correct choice to accurately measure the classification performance (Sokolova & Lapalme, 2009).

This study identified several features that were used in the selected studies for classifying clinical reports. The most commonly used features were BoW, *n*-gram, BoP, and BoC. In several studies, features were extracted by human experts. For clinical text classification, one should not depend only on content-based features, such as BoW, BoP, and *n*-gram. These features may produce classification results with limited accuracy because these features exhibit two major limitations. First, grammar and even word order are disregarded in these features, although word frequency is retained (Cavnar & Trenkle, 1994; Papadakis, Giannakopoulos, & Paliouras, 2016; Sebastiani, 2002; C. S. Yadav, Sharan, & Joshi, 2014). In addition, these features do not capture the word inversion and subset matching (Cavnar & Trenkle, 1994; Malliaros & Skianis, 2015; Papadakis, et al., 2016; Witten, et al., 2016). Second, various experts may use different vocabulary terms to report any event or information in the clinical report. Thus, these features do not consider word-level synonymy and polysemy when applied in clinical text reports (C. S. Yadav, et al., 2014). To address the first limitation, the addition of BoC features may enhance classification accuracy (Mouriño-García, et al., 2016; Mujtaba, Shuib, Raj, Al-Garadi, et al., 2017). Various studies have used the SNOMED CT ontology (Stearns, et al., 2001) to extract the medical concepts and related synonyms. To address the second limitation, recent studies have employed graph of word (GoW) features (Chakravarthy, Venkatachalam, & Telang, 2010; Malliaros & Skianis, 2015; Mujtaba, Shuib, Raj, Rajandram, Shaikh, et al., 2018). These studies have shown that GoW coupled with BoC and BoP produced improved classification accuracy. Therefore, the investigation of the use of various features of clinical reports is suggested to produce improved classification accuracy. Moreover, the fusion of BoP, BoC, and GoW may produce the better classification results compared with BoW and *n*-gram (Bates, et al., 2015; Comelli, et al., 2015; Luo, et al., 2014; Mujtaba, Shuib, Raj, Al-Garadi, et al., 2017; Mujtaba, Shuib, Raj, Rajandram, Shaikh, et al., 2018; D. H. Nguyen & Patrick, 2014; Wei, et al., 2013). However, the fusion of BoP, BoC, and GoW will produce enormous features, thereby resulting in the curse of dimensionality. Thus, machine learning algorithms need an extensive computational power in learning classification rules with such an enormous number of features. Thus, either the information gain or chi-square feature selection scheme should be maximized to overcome the dimensionality issue (Forman, 2003; Uysal & Gunal, 2012; Witten, et al., 2016). Note that all the available features may not contribute positively toward classification accuracy. Such non-discriminative features may also cause model overfitting (Forman, 2003). Therefore, the use of the appropriate feature selection scheme is suggested to determine a suitable subset of features that can improve classification performance, reduce noise, and reduce computation time in developing the classification model (Hall & Smith, 1998). To determine the optimum number of features in a feature subset, researchers should hypothesize that different feature subsets will obtain different classification performance results.

Thus, evaluating this proposition entails researchers to use progressive sampling (Beleites, Neugebauer, Bocklitz, Krafft, & Popp, 2013) to discover the optimal learning curve (Figueroa, Zeng-Treitler, Kandula, & Ngo, 2012). This learning curve depicts the classification performance that pertains to different feature subset sizes. To discover the optimal feature subset, a subset of five features may be initially selected to evaluate the classification performance. The features may be increased up to the point at which no further improvement in the classification performance is obtained. In addition, the classification performance using "all" features should be investigated.

The "no free lunch" theorem (Wolpert & Macready, 1995) indicates that no single machine learning algorithm performs best in all application areas. Hence, a variety of machine learning algorithms should be employed to evaluate which algorithm outperforms on the collected data set. Each selected primary study has used its own customized data set and different experimental settings. Thus, statistically comparing the performance values across the studies is infeasible. Nonetheless, when the outcome of the different studies was analyzed, the results showed that SVM showed better results in the SML-based algorithms followed by NB, RF, and DT (Hassanpour & Langlotz, 2016; Hassanpour, et al., 2017; Lucini, et al., 2017). Several studies have also reported that the RB machine learning algorithms obtained better classification results (Kalter, et al., 2016; MacRae, et al., 2015). Researchers in a few studies have developed and evaluated the performance of the RB and SML-based algorithms to classify clinical reports. The experimental results of these studies showed fluctuating results, whereby a few studies showed that the RB algorithms outperformed the SML-based algorithms. By contrast, a few studies have reported that the SML-based algorithms outperformed the RB algorithms. Therefore, empirically investigating the performance of the RB and SML-based classifiers is suggested to evaluate which one performs better on a collected data set.

Out of the 72 selected primary studies, 3 studies (Rios & Kavuluru, 2015; Shin, et al., 2017; Wu & Wang, 2017) have employed deep learning (specifically CNN) for the classification of clinical reports and compared the outcome of deep learning with that of shallow learning (e.g., LR, SVM, and RF). The experimental results reported that the deep learning classification method outperformed the shallow learning classification methods. Deep learning enables computational methods with a few processing layers to learn data representation with different levels of abstraction (LeCun, Bengio, & Hinton, 2015; Zhang, Pueyo, Wendt, Najork, & Broder, 2017). The main benefit of deep learning is that the features are not engineered by human experts. Conversely, these features are learned automatically from training data through general-purpose learning processes. Deep learning algorithms may prove beneficial in free-text clinical document classification with high-dimensional data, in which human-engineered features may not imitate learning vectors from the training data.

This review observed that only 5 studies (Afzal, et al., 2013; Butt, et al., 2013; Luo, et al., 2014; K. Yadav, et al., 2016; L. Zhou, et al., 2015) out of the 72 selected ones have provided error analysis in detail for misclassification. Six common misclassification reasons were discussed in these studies. First, general problems were observed with report ambiguity. In a few cases, intra-class or inter-class reports were ambiguous in nature. Thus, the classifier encountered difficulty in differentiating between the two classes. Second, spelling issues were noted in the clinical reports. Third, improper labeling or labeling errors were observed in the data sets in a few cases. Fourth, negation was excessive in a few reports, thereby presenting difficulty for the classifiers to identify the negation context in the report. Fifth, the authors in several studies have developed their own lexicons for each type of clinical reports that facilitate classification. In these lexicons, the authors indexed all the relevant terms that can classify reports in a specific category. However, several synonyms in a few studies were missed in the constructed lexicons, thereby causing the misclassification of the clinical reports. Lastly, only a few studies have used medical ontologies for extracting the relevant medical concepts from the extracted clinical terms from the reports. These studies reported that a few concepts may not be selected by commonly used ontologies, such as SNOMED CT and MetaMap, thereby causing classification errors.

5.0 Future Research Directions

Several research gaps were identified from our review. This section highlights various future research directions, in which considerable effort is required to improve the performance of clinical text classification systems. These research directions are presented as follows.

(1) Quality of the data sets: The majority of the studies have used the homogenous–homogenous clinical reports. However, several hospitals may have different medical documentation systems and

patterns or styles, thereby producing hurdles in generalizing constructed classifiers to multiple hospitals. Thus, collecting clinical reports from more than one organizations to develop more generic classification models is consistently preferable. Moreover, one disease may be reported in a variety of reports. For example, cancer patients' findings can be reported in pathology and radiology reports. Thus, researchers should consider multi-modal reports in the future to develop accurate and generic classification models. Moreover, class distribution in many data sets was observed to be imbalanced and numerous studies have employed sampling techniques to avoid majority class bias. In the future, researchers may employ and investigate various sampling techniques and their effect on classification performance.

(2) Big data in clinical text classification: To overcome the challenges of generalizing constructed classifiers, researchers should collect clinical reports from multiple institutions. Moreover, multi-modal reports (e.g., pathology and radiology reports for cancer) should be collected and used in the classification process. Thus, such multi-modal data require big data tools and techniques to overcome the heterogeneity issue. However, combining different data sets collected from multiple institutions is a major challenge primarily because of patient privacy and security concerns (Coates, Souhami, & El Naqa, 2016). Thus, state-of-the-art big data tools require further research effort for appropriate data-sharing protocols (Coates, et al., 2016). For example, IBM has developed the "SystemT" big data tool (Chiticariu, et al., 2010) for text mining to improve the accuracy, productivity, expressivity, and customizability of text mining applications. This tool provides a declarative rule and optimization engines to generate high-performance algebraic regular expressions with execution plans. However, such big data tools entail further research and development effort to focus on the appropriate data-sharing protocols to overcome the heterogeneity issue in cases of multi-modal narrative clinical reports data sets.

(3) Publicly available datasets: Although this review showed that only a few studies have not reported in detail the text classification methods that have been used, such studies also reported excellent results. These results may be the result of publication bias, in which experiments that obtain low results may not be disclosed. To address this issue, a few standard data sets for benchmarking are required and will be proven to add value for the proposed benchmarking methods. To the best of our knowledge, only a few datasets are freely available, such as OHSUMED (http://davis.wpi.edu/xmdv/datasets/ohsumed.html), i2b2 (https:// www.i2b2.org/index.html), and PhysioNet (http://physionet.org/mimic2/mimic2_clinical_overview.shtml#clinical-database-categories). However, various other domains, such as autopsy reports, radiology reports, or heterogeneous data sets, remain desirable for further research. Thus, future researchers may focus on creating publicly available corpora in the clinical text classification domain.

(4) Quality of features and dynamic updating the feature set: Several studies have reported that the fusion of content-based features (e.g., BoW, *n*-gram, and BoP) and conceptual features (e.g., BoC) produced the improved classification results. Moreover, the review of selected studies proved that expert-driven features and lexicon-based features prepared by human experts yielded improved classification performance. Thus, for any new clinical data set, a variety of features, feature representation techniques, and feature selection techniques should be investigated empirically to evaluate the sets of features, feature representation techniques, and feature selection techniques outperform. Moreover, future researchers may contribute to designing methods that enable the incremental addition or removal of features without rebuilding the entire model to keep up with the new trends in clinical report classification.

(5) Deep learning in clinical text classification: Figure 4 shows that only 3 out of the 72 studies have used deep learning. Thus, future researchers can use and investigate deep learning algorithms (e.g., CNN, recurrent neural network, and recursive neural network) for the classification of clinical reports. Deep learning enables computational methods with a few processing layers to learn data representation with different levels of abstraction (LeCun, et al., 2015; Zhang, et al., 2017). The main benefit of deep learning is that the features are not engineered by human experts. Conversely, these features are learned automatically from training data through general-purpose learning processes. Deep learning algorithms may prove beneficial in free-text clinical document classification with high-dimensional data, in which human-engineered features may not imitate learning vectors from the training data.

(6) Unsupervised clustering approaches in clinical text classification: The majority of the selected primary studies used the SML approaches to construct a clinical text classification model. Although these approaches produced better results, one major limitation of these approaches is the labeling of clinical reports to construct a training set. This labeling requires expert intervention to label each report or instance in a particular category. For example, Sarker et al. (Sarker & Gonzalez, 2015) collected Twitter tweets and DailyStrength instances to develop a classification model for predicting adverse drug events. The Twitter data set comprised 10822 tweets. Of these Tweets, 9583 were manually labeled by experts into two classes, namely, ADR and no ADR, to create a trainset. Moreover, the authors collected 10617 instances from DailyStrength. Of these instances, 8104 were manually labeled by experts into two classes, namely, ADR and no ADR, to create a trainset. Thus, a considerable effort was exerted in the preparation of the trainset. Therefore, to avoid such labeling efforts, future research can maximize the UML-based approaches (e.g., clustering) to classify clinical reports.

(7) Active learning approaches in clinical text classification: Out of the 72 selected studies, only 1 (D. H. Nguyen & Patrick, 2014) has used active learning to classify cancer-related radiology reports. Active learning is an SSML approach and the main concept behind this approach is that text classifiers can obtain optimum classification accuracy with only a few labeled instances. Active learning is beneficial where unlabeled data can be obtained easily in huge volumes. Nonetheless, the labeling of collected data is difficult, laborious, and expensive. Thus, future researchers may investigate various active learning algorithms to classify clinical reports.

(8) Adapting graph-based approaches: Recent studies have used graph-based approaches or GoW features to overcome the limitations of the content-based features (e.g., BoW, BoP, and *n*-gram). Yoon, et al. (2017) used the GoW features to classify breast cancer-related pathology reports. Accordingly, the authors converted each trainset report into a graph. For classification, the authors converted the unlabeled reports into graphs and applied graph similarity metric (such as edge matching) to determine the category of the unlabeled reports. The experimental findings reported that GoW outperformed the simple BoW. Thus, future research may focus on GoW features. In GoW, many researchers have maximized the Bow and BoC features. Researchers may also focus on other types of similarity metrics (e.g., edge, weight, and vertices metrics) and uniqueness metrics (e.g., edges unmatched and vertices unmatched) to classify clinical reports.

(9) Use of ontology: Existing studies have either used the SML-based or rule-based approaches. Thus, future researchers can emphasize classifying clinical reports using ontology. Moreover, an adaptive ontology can be planned and created from the classification results that can be developed and customized based on the end user's reports.

(10) Language-based barriers: The majority of the selected primary studies used clinical reports written in English. Thus, the constructed classifiers can only classify English clinical reports. Only 1 study out of the selected 72 (Wei, et al., 2013) has developed a classifier that can categorize clinical reports written in Chinese using the SVM classifier. Therefore, future studies should identify and propose the features that assist in classifying clinical reports written in non-English languages.

(11) Reinforcement learning approach for classifying narrative clinical reports: One of the major challenges of using the SML algorithms for classifying clinical reports is how the system can learn from interaction with the environment. Evidently, difficulty is encountered in obtaining training data that are sufficiently representative of the minority class (e.g., cancer positive). The model should learn from experience to accurately predict a particular class from clinical reports (Huys, Maia, & Frank, 2016). Moreover, the automatic classification of clinical reports entails enormous challenges because this process involves various heterogeneous data that are closely linked. Therefore, the effective prediction of clinical report category from clinical report data set requires the accurate and systematic identification of multiple interacting levels that will enable mapping those levels together. Thus, the development of the RL algorithms for classifying clinical reports may realistically maximize the efficiency and improve the classification model performance.

(12) Transfer learning approach for classifying narrative clinical reports: The TL approach uses previously learned knowledge to improve the performance of the related or different domains. In this case, TL translates the capabilities from existing systems to untrained systems. In clinical text classification, the development of the TL approach is vital because it may facilitate the reduction of

training time, produce robust systems, and source the target environment for classifying the clinical reports. However, developing effective TL approaches is extremely challenging in clinical text classification. Typical TL approaches that enable sources to new target transfer or features and classification model transfer are highly desirable.

6.0 Conclusion

This comprehensive study presented a critical analysis of the clinical text classification domain by combining major research endeavors to assist researchers in this domain, thereby acquiring an improved awareness of the existing related solutions. Articles on clinical text classification published in 2013–2018 were comprehensively reviewed. A total of 72 primary studies were rigorously selected from 8 different academic databases. The selected primary studies were reviewed from six aspects: types of clinical reports used, data set characteristics, pre-processing and sampling techniques, feature engineering, text classification techniques, and performance metrics. In clinical text classification, various types of free-text clinical reports were used, of which the most extensively employed clinical reports were pathology reports, radiology reports, and Medline biomedical documents. In the majority of the selected primary studies, the authors used their own data sets, which primarily comprised only one type of reports, and were collected from only one organization. In the collected data sets, data were typically imbalanced, although only a few studies have used the correct performance metrics for measuring classification performance. Various pre-processing techniques were applied to remove noisy or irrelevant terms from the data sets. However, several text normalization techniques may prove beneficial to obtain improved classification results. In features, a combination of BoW, BoP, and BoC showed improved results. The representation of these features in graph structure further enhances the performance but also increases computational time. Apart from these features, expert-driven features also performed well. In the majority of the studies, the BR, TF, and TFiDF feature representation techniques were determined to be beneficial. To remove the redundant or non-discriminative features, various studies used different types of feature selection schemes. In many studies, chi-square and information gain showed better results. For text classification, the majority of the studies have used either the SML approaches or rule-based approaches. Among the SML algorithms, SVM obtained better results followed by RF, DT, and *k*NN. The rule-based classifiers also showed promising results. A few studies used rule-based and SML-based classifiers and obtained fluctuating results. In recent publications, researchers have used NLP and deep learning approaches to classify free-text clinical reports. To reduce the time of preparing the training set, recent studies have used semi-supervised active learning approaches for free-text clinical report classification. In performance metrics, the majority of the studies used precision, recall, F-measure, and accuracy to measure the classification performance in the binary-class classification problem. In the multi-class classification problem, the authors used micro- and macro-averaging of precision, recall, and F-measure. We believe that this comprehensive review will provide a profound understanding of the clinical text classification domain and afford valuable insights to researchers in this field.

References

Abu-El-Haija, M., Kumar, S., Szabo, F., Werlin, S., Conwell, D., Banks, P., Morinville, V. D., & Comm, N. P. (2017). Classification of Acute Pancreatitis in the Pediatric Population: Clinical Report From the NASPGHAN Pancreas Committee. *Journal of Pediatric Gastroenterology and Nutrition, 64*, 984-990.

Adeva, J. G., Atxa, J. P., Carrillo, M. U., & Zengotitabengoa, E. A. (2014). Automatic text classification to support systematic reviews in medicine. *Expert Systems with Applications, 41*, 1498-1508.

Afzal, Z., Schuemie, M. J., van Blijderveen, J. C., Sen, E. F., Sturkenboom, M., & Kors, J. A. (2013). Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records. *Bmc Medical Informatics and Decision Making, 13*, 11.

Aggarwal, C. C., & Zhai, C. (2012a). *Mining text data*: Springer Science & Business Media.

Aggarwal, C. C., & Zhai, C. (2012b). A survey of text classification algorithms. *Mining text data*, 163-222.

Al-garadi, M. A., Khan, M. S., Varathan, K. D., Mujtaba, G., & Al-Kabsi, A. M. (2016). Using online social networks to track a pandemic: A systematic review. *Journal of Biomedical Informatics, 62*, 1-11.

Alabbas, W., Al-Khateeb, H. M., & Mansour, A. (2016). Arabic text classification methods: Systematic literature review of primary studies. In *Information Science and Technology (CiSt), 2016 4th IEEE International Colloquium on* (pp. 361-367): IEEE.

Alghoson, A. M. (2014). Medical Document Classification Based on MeSH. In R. H. Sprague (Ed.), *2014 47th Hawaii International Conference on System Sciences* (pp. 2571-2575). New York: Ieee.

Amini, M., Usunier, N., & Goutte, C. (2009). Learning from multiple partially observed views-an application to multilingual text categorization. In *Advances in neural information processing systems* (pp. 28-36).

Amrit, C., Paauw, T., Aly, R., & Lavric, M. (2017). Identifying child abuse through text mining and machine learning. *Expert Systems with Applications, 88*, 402-418.

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium* (pp. 17): American Medical Informatics Association.

Barak-Corren, Y., Castro, V. M., Javitt, S., Hoffnagle, A. G., Dai, Y., Perlis, R. H., Nock, M. K., Smoller, J. W., & Reis, B. (2017). Predicting Suicidal Behavior From Longitudinal Electronic Health Records. *American Journal of Psychiatry, 174*, 154-162.

Bates, J., Fodeh, S. J., Brandt, C. A., & Womack, J. A. (2015). Classification of radiology reports for falls in an HIV study cohort. *Journal of the American Medical Informatics Association, 23*, e113-e117.

Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C., & Popp, J. (2013). Sample size planning for classification models. *Analytica chimica acta, 760*, 25-33.

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing* (pp. 1-4): Springer.

Buchan, K., Filannino, M., & Uzuner, O. (2017). Automatic prediction of coronary artery disease from clinical narratives. *Journal of Biomedical Informatics, 72*, 23-32.

Butt, L., Zuccon, G., Nguyen, A., Bergheim, A., & Grayson, N. (2013). Classification of cancer-related death certificates using machine learning. *Australasian Medical Journal, 6*, 292-300.

Cavnar, W. B., & Trenkle, J. M. (1994). N-gram-based text categorization. *Ann Arbor MI, 48113*, 161-175.

Chakravarthy, S., Venkatachalam, A., & Telang, A. (2010). A graph-based approach for multi-folder email classification. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on* (pp. 78-87): IEEE.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research, 16*, 321-357.

Chiticariu, L., Krishnamurthy, R., Li, Y., Raghavan, S., Reiss, F. R., & Vaithyanathan, S. (2010). SystemT: an algebraic approach to declarative information extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 128-137): Association for Computational Linguistics.

Clark, C., Wellner, B., Davis, R., Aberdeen, J., & Hirschman, L. (2017). Automatic classification of RDoC positive valence severity with a neural network. *Journal of Biomedical Informatics*.

Coates, J., Souhami, L., & El Naqa, I. (2016). Big data analytics for prostate radiotherapy. *Frontiers in oncology, 6*.

Comelli, A., Agnello, L., Vitabile, S., & Ieee. (2015). *An Ontology-Based Retrieval System for Mammographic Reports*. New York: Ieee.

Cosma, G., Brown, D., Archer, M., Khan, M., & Pockley, A. G. (2017). A survey on computational intelligence approaches for predictive modeling in prostate cancer. *Expert Systems with Applications, 70*, 1-19.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*: John Wiley & Sons.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*: Cambridge university press.

Dai, X., & Bikdash, M. (2015). Hybrid classification for tweets related to infection with influenza. In *Conference Proceedings - IEEE SOUTHEASTCON* (June ed., Vol. 2015-June).

Dalkey, N., & Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. *Management science, 9*, 458-467.

Danso, S., Atwell, E., & Johnson, O. (2013). Linguistic and statistically derived features for cause of death prediction from verbal autopsy text. In *Language processing and knowledge in the web* (pp. 47-60): Springer.

Danso, S., Atwell, E., & Johnson, O. (2014). A comparative study of machine learning methods for verbal autopsy text classification. *arXiv preprint arXiv:1402.4380*.

Debole, F., & Sebastiani, F. (2004). Supervised term weighting for automated text categorization. In *Text mining and its applications* (pp. 81-97): Springer.

Deng, Y., Groll, M. J., & Denecke, K. (2015). Rule-based Cervical Spine Defect Classification Using Medical Narratives. In *Studies in health technology and informatics* (Vol. 216, pp. 1038).

Diz, J., Marreiros, G., & Freitas, A. (2015). Using data mining techniques to support breast cancer diagnosis. In *New Contributions in Information Systems and Technologies* (pp. 689-700): Springer.

Do, C. B., & Ng, A. Y. (2006). Transfer learning for text classification. In *Advances in Neural Information Processing Systems* (pp. 299-306).

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*, 78-87.

Donnelly, K. (2006). SNOMED-CT: The advanced terminology and coding system for eHealth. *Stud Health Technol Inform, 121*, 279.

Farshchi, S. M. R., & Yaghoobi, M. (2013). Categorization of Medical Documents Using Hybrid Competitive Neural Network with String Vector, a Novel Approach. In Z. Y. Du (Ed.), *Intelligence Computation and Evolutionary Computation* (Vol. 180, pp. 1045-1054). Berlin: Springer-Verlag Berlin.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters, 27*, 861-874.

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., & Ngo, L. H. (2012). Predicting sample size required for classification performance. *Bmc Medical Informatics and Decision Making, 12*, 8.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research, 3*, 1289-1305.

Fragos, K., & Skourlas, C. (2016). Smoothing Class Frequencies for KNN Medical Article Classification. In *Proceedings of the 20th Pan-Hellenic Conference on Informatics* (pp. 79): ACM.

Fukunaga, K. (2013). *Introduction to statistical pattern recognition*: Academic press.

Garla, V., Taylor, C., & Brandt, C. (2013). Semi-supervised clinical text classification with Laplacian SVMs: an application to cancer case management. *Journal of Biomedical Informatics, 46*, 869-875.

Gatta, R., Vallati, M., De Bari, B., & Ozsahin, M. (2014). The impact of different training sets on medical documents classification. In (Vol. 1213, pp. 1-5).

Goldberg, Y., & Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., & Donaldson, L. (2013). Use of Sentiment Analysis for Capturing Patient Experience From Free-Text Comments Posted Online. *Journal of Medical Internet Research, 15*, 9.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research, 3*, 1157-1182.

Guzella, T. S., & Caminhas, W. M. (2009). A review of machine learning approaches to spam filtering. *Expert Systems with Applications, 36*, 10206-10222.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (1998). *Multivariate data analysis* (Vol. 5): Prentice hall Upper Saddle River, NJ.

Hall, M. A., & Smith, L. A. (1998). Practical feature subset selection for machine learning.

Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine learning, 45*, 171-186.

Hassanpour, S., & Langlotz, C. P. (2016). Predicting High Imaging Utilization Based on Initial Radiology Reports: A Feasibility Study of Machine Learning. *Academic Radiology, 23*, 84-89.

Hassanpour, S., Langlotz, C. P., Amrhein, T. J., Befera, N. T., & Lungren, M. P. (2017). Performance of a Machine Learning Classifier of Knee MRI Reports in Two Large Academic Radiology Practices: A Tool to Estimate Diagnostic Yield. *AJR Am J Roentgenol, 208*, 750-753.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Overview of supervised learning. In *The elements of statistical learning* (pp. 9-41): Springer.

Heer, J., Hellerstein, J. M., & Kandel, S. (2015). Predictive Interaction for Data Transformation. In *CIDR*.

Holzinger, A., Schantl, J., Schroettner, M., Seifert, C., & Verspoor, K. (2014). Biomedical text mining: state-of-the-art, open problems and future challenges. In *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics* (pp. 271-300): Springer.

Hotho, A., Maedche, A., & Staab, S. (2002). Ontology-based text document clustering. *KI, 16*, 48-54.

Huys, Q. J., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature neuroscience, 19*, 404.

Imane, A., & Mohamed, B. A. (2017). Multi-label Categorization of French Death Certificates using NLP and Machine Learning. In *Proceedings of the 2nd international Conference on Big Data, Cloud and Applications* (pp. 29): ACM.

Japkowicz, N. (2000). The class imbalance problem: Significance and strategies. In *Proc. of the Int'l Conf. on Artificial Intelligence*.

Jian, Z., Guo, X., Liu, S., Ma, H., Zhang, S., Zhang, R., & Lei, J. (2017). A cascaded approach for Chinese clinical text de-identification with less annotation effort. *Journal of Biomedical Informatics, 73*, 76-83.

Jiang, J. (2012). Information extraction from text. In *Mining text data* (pp. 11-41): Springer.

Jindal, R., & Taneja, S. (2015). A Lexical Approach for Text Categorization of Medical Documents. In P. Samuel (Ed.), *Proceedings of the International Conference on Information and Communication Technologies, Icict 2014* (Vol. 46, pp. 314-320). Amsterdam: Elsevier Science Bv.

Jo, T. (2013). Application of Table based Similarity to Classification of Bio-Medical Documents. *2013 Ieee International Conference on Granular Computing (Grc)*, 162-166.

Ju, M., Duan, H., & Li, H. (2016). Lexical characteristics analysis of Chinese clinical documents. In *Proceedings - 2015 7th International Conference on Information Technology in Medicine and Education, ITME 2015* (pp. 121-125).

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of artificial intelligence research, 4*, 237-285.

Kalter, H. D., Perin, J., & Black, R. E. (2016). Validating hierarchical verbal autopsy expert algorithms in a large data set with known causes of death. *J Glob Health, 6*, 010601.

Kasthurirathne, S. N., Dixon, B. E., Gichoya, J., Xu, H. P., Xia, Y. N., Mamlin, B., & Grannis, S. J. (2016). Toward better public health reporting using existing off the shelf approaches: A comparison of alternative cancer detection approaches using plaintext medical data and non-dictionary based feature selection. *Journal of Biomedical Informatics, 60*, 145-152.

Kasthurirathne, S. N., Dixon, B. E., Gichoya, J., Xu, H. P., Xia, Y. N., Mamlin, B., & Grannis, S. J. (2017). Toward better public health reporting using existing off the shelf approaches: The value of medical dictionaries in automated cancer detection using plaintext medical data. *Journal of Biomedical Informatics, 69*, 160-176.

Kasthurirathne, S. N., Dixon, B. E., & Grannis, S. J. (2015). Evaluating Methods for Identifying Cancer in Free-Text Pathology Reports Using Various Machine Learning and Data Preprocessing Approaches. In (Vol. 216, pp. 1070).

Kaurova, O., Alexandrov, M., & Blanco, X. (2011). Classification of free text clinical narratives (short review). *Business and Engineering Applications of Intelligent and Information Systems*, 124.

Kavuluru, R., Rios, A., & Lu, Y. (2015). An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artif Intell Med, 65*, 155-166.

Keele, S. (2007). Guidelines for performing systematic literature reviews in software engineering. In *Technical report, Ver. 2.3 EBSE Technical Report. EBSE*: sn.

Ko, Y., & Seo, J. (2000). Automatic text categorization by unsupervised learning. In *Proceedings of the 18th conference on Computational linguistics-Volume 1* (pp. 453-459): Association for Computational Linguistics.

Kocbek, S., Cavedon, L., Martinez, D., Bain, C., Mac Manus, C., Haffari, G., Zukerman, I., & Verspoor, K. (2016). Text mining electronic hospital records to automatically classify admissions against disease: Measuring the impact of linking data sources. *Journal of Biomedical Informatics, 64*, 158-167.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, pp. 1137-1145): Montreal, Canada.

Koopman, B., Karimi, S., Nguyen, A., McGuire, R., Muscatello, D., Kemp, M., Truran, D., Zhang, M., & Thackway, S. (2015). Automatic classification of diseases from free-text death certificates for real-time surveillance. *Bmc Medical Informatics and Decision Making, 15*, 10.

Koopman, B., Zuccon, G., Nguyen, A., Bergheim, A., & Grayson, N. (2015). Automatic ICD-10 classification of cancers from free-text death certificates. *International journal of medical informatics, 84*, 956-965.

Krawczyk, B., & Woźniak, M. (2015). Hypertension type classification using hierarchical ensemble of one-class classifiers for imbalanced data. In *ICT innovations 2014* (pp. 341-349): Springer.

Kruthika, K., Pai, A., Maheshappa, H., & Initiative, A. s. D. N. (2017). Classification of Alzheimer and MCI Phenotypes on MRI Data Using SVM. In *International Symposium on Signal Processing and Intelligent Recognition Systems* (pp. 263-275): Springer.

Lauren, P., Qu, G., Zhang, F., & Lendasse, A. (2017). Discriminant document embeddings with an extreme learning machine for classifying clinical narratives. *Neurocomputing*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*, 436-444.

Loh, W. Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1*, 14-23.

Lopez-Guede, J. M., Moreno-Fernandez-de-Leceta, A., Martinez-Garcia, A., & Graña, M. (2015). Lynx: Automatic elderly behavior prediction in home telecare. *Biomed Res Int, 2015*.

Lopprich, M., Krauss, F., Ganzinger, M., Senghas, K., Riezler, S., & Knaup, P. (2016). Automated Classification of Selected Data Elements from Free-text Diagnostic Reports for Clinical Research. *Methods of Information in Medicine, 55*, 373-380.

Lucini, F. R., Fogliatto, F. S., da Silveira, G. J., Neyeloff, J. L., Anzanello, M. J., Kuchenbecker, R. d. S., & Schaan, B. D. (2017). Text mining approach to predict hospital admissions using early medical records from the emergency department. *International journal of medical informatics, 100*, 1-8.

Luo, Y., Sohani, A. R., Hochberg, E. P., & Szolovits, P. (2014). Automatic lymphoma classification with sentence subgraph mining from pathology reports. *Journal of the American Medical Informatics Association, 21*, 824-832.

Mabotuwana, T., Lee, M. C., & Cohen-Solal, E. V. (2013). An ontology-based similarity measure for biomedical data - Application to radiology reports. *Journal of Biomedical Informatics, 46*, 857-868.

MacRae, J., Love, T., Baker, M. G., Dowell, A., Carnachan, M., Stubbe, M., & McBain, L. (2015). Identifying influenza-like illness presentation from unstructured general practice clinical narrative using a text classifier rule-based expert system versus a clinical expert. *Bmc Medical Informatics and Decision Making, 15*, 78.

Malliaros, F. D., & Skianis, K. (2015). Graph-based term weighting for text categorization. In *Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on* (pp. 1473-1479): IEEE.

Martinez, D., Ananda-Rajah, M. R., Suominen, H., Slavin, M. A., Thursky, K. A., & Cavedon, L. (2015). Automatic detection of patients with invasive fungal disease from free-text computed tomography (CT) scans. *Journal of Biomedical Informatics, 53*, 251-260.

Masino, A. J., Grundmeier, R. W., Pennington, J. W., Germiller, J. A., & Crenshaw, E. B. (2016). Temporal bone radiology report classification using open source machine learning and natural langue processing libraries. *Bmc Medical Informatics and Decision Making, 16*, 65.

Miasnikof, P., Giannakeas, V., Gomes, M., Aleksandrowicz, L., Shestopaloff, A. Y., Alam, D., Tollman, S., Samarikhalaj, A., & Jha, P. (2015). Naive Bayes classifiers for verbal autopsies: comparison to physician-based classification for 21,000 child and adult deaths. *BMC Medicine, 13*, 9.

Mirończuk, M. M., & Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications*.

Mouriño-García, M., Pérez-Rodríguez, R., Anido-Rifón, L., & Gómez-Carballa, M. (2016). Bag-of-Concepts Document Representation for Bayesian Text Classification. In *Computer and Information Technology (CIT), 2016 IEEE International Conference on* (pp. 281-288): IEEE.

Mujtaba, G., Shuib, L., Raj, R. G., Al-Garadi, M. A., Rajandram, R., & Shaikh, K. (2017). Hierarchical text classification of autopsy reports to determine MoD and CoD through term-based and concepts-based features. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 10357 LNAI, pp. 209-222).

Mujtaba, G., Shuib, L., Raj, R. G., Rajandram, R., & Shaikh, K. (2018). Prediction of cause of death from forensic autopsy reports using text classification techniques: A comparative study. *Journal of Forensic and Legal Medicine, 57*, 41-50.

Mujtaba, G., Shuib, L., Raj, R. G., Rajandram, R., Shaikh, K., & Al-Garadi, M. A. (2017). Automatic ICD-10 multi-class classification of cause of death from plaintext autopsy reports through expert-driven feature selection. *PLoS ONE, 12*, 27.

Mujtaba, G., Shuib, L., Raj, R. G., Rajandram, R., Shaikh, K., & Al-Garadi, M. A. (2018). Classification of forensic autopsy reports through conceptual graph-based document representation model. *Journal of Biomedical Informatics, 82*, 88-105.

Mujtaba, G., Shuib, L., Raj, R. G., Rajandram, R., Shaikh, K., & Ieee. (2016). Automatic Text Classification of ICD-10 Related CoD from Complex and Free Text Forensic Autopsy Reports. *2016 15th Ieee International Conference on Machine Learning and Applications (Icmla 2016)*, 1055-1058.

Napolitano, G., Marshall, A., Hamilton, P., & Gavin, A. T. (2016). Machine learning classification of surgical pathology reports and chunk recognition for information extraction noise reduction. *Artificial Intelligence in Medicine, 70*, 77-83.

Nguyen, D. H., & Patrick, J. D. (2014). Supervised machine learning and active learning in classification of radiology reports *J Am Med Inform Assoc, 21*, 893-901.

Nguyen, H., & Patrick, J. (2016). Text Mining in Clinical Domain: Dealing with Noise. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 549-558): ACM.

Nicolosi, N. (2008). Feature selection methods for text classification. In: November.

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning, 39*, 103-134.

Oleynik, M., Patrão, D. F. C., & Finger, M. (2017). Automated Classification of Semi-Structured Pathology Reports into ICD-O Using SVM in Portuguese. In *Studies in health technology and informatics* (Vol. 235, pp. 256-260).

Osborne, J. D., Wyatt, M., Westfall, A. O., Willig, J., Bethard, S., & Gordon, G. (2016). Efficient identification of nationally mandated reportable cancer cases using natural language processing and machine learning. *Journal of the American Medical Informatics Association, 23*, 1077-1084.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*, 1345-1359.

Papadakis, G., Giannakopoulos, G., & Paliouras, G. (2016). Graph vs. bag representation models for the topic classification of web documents. *World Wide Web, 19*, 887-920.

Parlak, B., & Uysal, A. K. (2015). Classification of Medical Documents According to Diseases. In *2015 23rd Signal Processing and Communications Applications Conference* (pp. 1635-1638). New York: Ieee.

Parlak, B., & Uysal, A. K. (2016). The impact of feature selection on medical document classification. In *Iberian Conference on Information Systems and Technologies, CISTI* (Vol. 2016-July).

Parlak, B., & Uysal, A. K. (2018). On feature weighting and selection for medical document classification. In *Studies in Computational Intelligence* (Vol. 718, pp. 269-282).

Pineda, A. L., Tsui, F.-C., Visweswaran, S., & Cooper, G. F. (2013). Detection of patients with influenza syndrome using machine-learning models learned from emergency department reports. *Online journal of public health informatics, 5*.

Pineda, A. L., Ye, Y., Visweswaran, S., Cooper, G. F., Wagner, M. M., & Tsui, F. (2015). Comparison of machine learning classifiers for influenza detection from emergency department free-text reports. *Journal of Biomedical Informatics, 58*, 60-69.

Prabhakar, S. K., & Rajaguru, H. (2018). Comparison of Fuzzy Output Optimization with Expectation Maximization Algorithm and Its Modification for Epilepsy Classification. In *Proceedings of International Conference on Cognition and Recognition* (pp. 263-272): Springer.

Provost, F. J., & Fawcett, T. (1997). Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. In *KDD* (Vol. 97, pp. 43-48).

Provost, F. J., Fawcett, T., & Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. In *ICML* (Vol. 98, pp. 445-453).

Rani, G. J. J., Gladis, D., & Mammen, J. (2015). Classification and Prediction of Breast Cancer Data derived Using Natural Language Processing. *Proceeding of the Third International Symposium on Women in Computing and Informatics (Wci-2015)*, 250-255.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. In *Encyclopedia of database systems* (pp. 532-538): Springer.

Renganathan, V. (2017). Text mining in biomedical domain with emphasis on document clustering. *Healthc Inform Res, 23*, 141-146.

Rios, A., & Kavuluru, R. (2015). Convolutional neural networks for biomedical text classification: Application in indexing biomedical articles. In *BCB 2015 - 6th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics* (pp. 258-267).

Saqlain, M., Hussain, W., Saqib, N. A., & Khan, M. A. (2016). Identification of Heart Failure by Using Unstructured Data of Cardiac Patients. In *Proceedings of the International Conference on Parallel Processing Workshops* (Vol. 2016-September, pp. 426-431).

Sarker, A., & Gonzalez, G. (2015). Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics, 53*, 196-207.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR), 34*, 1-47.

Sedghi, E., Weber, J. H., Thomo, A., Bibok, M., & Penn, A. M. (2016). A new approach to distinguish migraine from stroke by mining structured and unstructured clinical data-sources. *Network Modeling Analysis in Health Informatics and Bioinformatics, 5*, 30.

Settles, B. (2010). Active learning literature survey. *University of Wisconsin, Madison, 52*, 11.

Shin, B., Chokshi, F. H., Lee, T., & Choi, J. D. (2017). Classification of radiology reports using neural attention models. In *Neural Networks (IJCNN), 2017 International Joint Conference on* (pp. 4363-4370): IEEE.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*, 427-437.

Spasić, I., Livsey, J., Keane, J. A., & Nenadić, G. (2014). Text mining of cancer-related information: review of current status and future directions. *International journal of medical informatics, 83*, 605-623.

Sreejith, S., Rahul, S., & Jisha, R. (2016). A real time patient monitoring system for heart disease prediction using random forest algorithm. In *Advances in Signal Processing and Intelligent Recognition Systems* (pp. 485-500): Springer.

Stearns, M. Q., Price, C., Spackman, K. A., & Wang, A. Y. (2001). SNOMED clinical terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium* (pp. 662): American Medical Informatics Association.

Syed, H., & Das, A. K. (2015). Vector Space Models for Encoding and Retrieving Longitudinal Medical Record Data. In *VLDB Workshop on Big Graphs Online Querying* (pp. 3-15): Springer.

Szenasi, G., Lemnaru, C., & Barbantan, I. (2015). Concept Extraction from Medical Documents A Contextual Approach. In R. Potolea (Ed.), *2015 Ieee 11th International Conference on Intelligent Computer Communication and Processing* (pp. 13-17). New York: Ieee.

Tang, L., & Liu, H. (2005). Bias analysis in text classification for highly skewed data. In *Data Mining, Fifth IEEE International Conference on* (pp. 4 pp.): IEEE.

Tantug, A. C. (2010). Document categorization with modified statistical language models for agglutinative languages. *International Journal of Computational Intelligence Systems, 3*, 632-645.

Teyhouee, A., McPhee-Knowles, S., Waldner, C., & Osgood, N. (2017). Prospective Detection of Foodborne Illness Outbreaks Using Machine Learning Approaches. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation* (pp. 302-308): Springer.

Thompson, P., Batista-Navarro, R. T., Kontonatsios, G., Carter, J., Toon, E., McNaught, J., Timmermann, C., Worboys, M., & Ananiadou, S. (2016). Text Mining the History of Medicine. *PLoS ONE, 11*, 33.

Uysal, A. K., & Gunal, S. (2012). A novel probabilistic feature selection method for text classification. *Knowledge-Based Systems, 36*, 226-235.

Wagholikar, A., Zuccon, G., Nguyen, A., Chu, K., Martin, S., Lai, K., & Greenslade, J. (2013). Automated classification of limb fractures from free-text radiology reports using a clinician-informed gazetteer methodology. *Australas Med J, 6*, 301-307.

Wang, Y., Coiera, E., Runciman, W., & Magrabi, F. (2017). Using multiclass classification to automate the identification of patient safety incident reports by type and severity. *Bmc Medical Informatics and Decision Making, 17*, 84.

Wei, Z., Ju, Z. X., Chun, X., Hua, J., & Jin, P. (2013). An automatic electronic nursing records analysis system based on the text classification and machine learning. In *Proceedings - 2013 5th International Conference on Intelligent Human-Machine Systems and Cybernetics, IHMSC 2013* (Vol. 2, pp. 494-498).

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems, 2*, 37-52.

Wolpert, D. H., & Macready, W. G. (1995). No free lunch theorems for search. In: Technical Report SFI-TR-95-02-010, Santa Fe Institute.

Wu, H., & Wang, M. D. (2017). Infer Cause of Death for Population Health Using Convolutional Neural Network. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (pp. 526-535): ACM.

Xu, King, I., Lyu, M. R.-T., & Jin, R. (2010). Discriminative semi-supervised feature selection via manifold regularization. *IEEE Transactions on Neural networks, 21*, 1033-1047.

Yadav, C. S., Sharan, A., & Joshi, M. L. (2014). Semantic graph based approach for text mining. In *Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on* (pp. 596-601): IEEE.

Yadav, K., Sarioglu, E., Choi, H. A., Cartwright, W. B. t., Hinds, P. S., & Chamberlain, J. M. (2016). Automated Outcome Classification of Computed Tomography Imaging Reports for Pediatric Traumatic Brain Injury. *Acad Emerg Med, 23*, 171-178.

Yala, A., Barzilay, R., Salama, L., Griffin, M., Sollender, G., Bardia, A., Lehman, C., Buckley, J. M., Coopey, S. B., Polubriaginof, F., Garber, J. E., Smith, B. L., Gadd, M. A., Specht, M. C., Gudewicz, T. M., Guidi, A. J., Taghian, A., & Hughes, K. S. (2017). Using machine learning to parse breast pathology reports. *Breast Cancer Research and Treatment, 161*, 203-211.

Yang, S., Wei, R., Guo, J., & Xu, L. (2017). Semantic Inference on Clinical Documents: Combining Machine Learning Algorithms with an Inference Engine for Effective Clinical Diagnosis and Treatment. *IEEE Access, 5*, 3529-3546.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Icml* (Vol. 97, pp. 412-420).

Ye, Y., Tsui, F., Wagner, M., Espino, J. U., & Li, Q. (2014). Influenza detection from emergency department reports using natural language processing and Bayesian network classifiers. *Journal of the American Medical Informatics Association, 21*, 815-823.

Yeow, W. L., Mahmud, R., & Raj, R. G. (2014). An application of case-based reasoning with machine learning for forensic autopsy. *Expert Systems with Applications, 41*, 3497-3505.

Yoon, H. J., Roberts, L., & Tourassi, G. (2017). Automated Histologic Grading from Free-Text Pathology Reports using Graph-of-Words Features and Machine Learning. *2017 Ieee Embs International Conference on Biomedical & Health Informatics (Bhi)*, 369-372.

Zhang, A., Pueyo, L. G., Wendt, J. B., Najork, M., & Broder, A. (2017). Email Category Prediction.

Zhou, H. Y., Zhang, Q. R., Wang, H. X., & Zhang, D. (2015). Feature selection in medical text classification based on differential evolution algorithm. In *Electronics, Information Technology and Intellectualization - International Conference on Electronics, InformationTechnology and Intellectualization, EITI 2014* (pp. 79-82).

Zhou, L., Baughman, A. W., Lei, V. J., Lai, K. H., Navathe, A. S., Chang, F., Sordo, M., Topaz, M., Zhong, F., Murrali, M., Navathe, S., & Rocha, R. A. (2015). Identifying Patients with Depression Using Free-text Clinical Documents. *Stud Health Technol Inform, 216*, 629-633.

Zhou, Y., Amundson, P. K., Yu, F., Kessler, M. M., Benzinger, T. L., & Wippold, F. J. (2014). Automated classification of radiology reports to facilitate retrospective study in radiology. *J Digit Imaging, 27*, 730-736.

Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning, 3*, 1-130.

Zuccon, G., Khanna, S., Nguyen, A., Boyle, J., Hamlet, M., & Cameron, M. (2015). Automatic detection of tweets reporting cases of influenza like illnesses in Australia. *Health information science and systems, 3*, S4.

Zuccon, G., Wagholikar, A. S., Nguyen, A. N., Butt, L., Chu, K., Martin, S., & Greenslade, J. (2013). Automatic Classification of Free-Text Radiology Reports to Identify Limb Fractures using Machine Learning and the SNOMED CT Ontology. *AMIA Jt Summits Transl Sci Proc, 2013*, 300-304.

Appendix-I

Table A1. List of Quality Checklist Questions

S. No.	Quality Checklist Question
QAC 1	Are research objectives are clearly stated?
QAC 2	Is methodology well-defined?
QAC 3	Is the number of training and testing data identified?
QAC 4	Are the pre-processing techniques used in the study clearly described and their selection

	justified?
QAC 5	Are the features used for clinical reports classification described clearly?
QAC 6	Is the process of feature engineering clearly stated to transform clinical reports into numeric vector?
QAC 7	Are the classifiers used in study clearly described?
QAC 8	Does the study perform the comparison of proposed approach with existing baseline approaches?
QAC 9	Were the performance measures fully defined?
QAC 10	Are results properly interpreted and discussed and does the conclusion reflect the research findings?

Table A2. Quality Assessment Criteria of 72 Studies

Study	QAC1	QAC2	QAC3	QAC4	QAC5	QAC6	QAC7	QAC8	QAC9	QAC10	Overall Score
(Afzal, et al., 2013)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Butt, Zuccon, Nguyen, Bergheim, & Grayson, 2013)	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	9
(Danso, et al., 2013)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Farshchi & Yaghoobi, 2013)	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	9
(Garla, Taylor, & Brandt, 2013)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Greaves, Ramirez-Cano, Millett, Darzi, & Donaldson, 2013)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Jo, 2013)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Mabotuwana, Lee, & Cohen-Solal, 2013)	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	9
(Arturo Lopez Pineda, et al., 2013)	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	9
(Wagholikar, et al., 2013)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Wei, Ju, Chun, Hua, & Jin, 2013)	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	9
(G. Zuccon, et al., 2013)	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	9
(Adeva, Atxa, Carrillo, & Zengotitabengoa, 2014)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Alghoson, 2014)	✓	✓	✓	✗	✗	✓	✓	✗	✓	✓	7
(Danso, et al., 2014)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Gatta, Vallati, De Bari, & Ozsahin, 2014)	✓	✓	✓	✗	✓	✓	✓	✗	✓	✓	8
(Luo, Sohani, Hochberg, & Szolovits, 2014)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Nguyen & Patrick, 2014)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Ye, Tsui, Wagner, Espino, & Li, 2014)	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	9
(Yeow, Mahmud, & Raj, 2014)	✓	✓	✗	✗	✓	✓	✓	✗	✓	✓	7
(Y. Zhou, et al., 2014)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Bates, Fodeh,	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10

Study	QAC1	QAC2	QAC3	QAC4	QAC5	QAC6	QAC7	QAC8	QAC9	QAC10	Overall Score
Brandt, & Womack, 2015)											
(X. Dai & Bikdash, 2015a)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Deng, et al., 2015)	✓	✓	✗	✓	✓	✓	✓	✗	✓	✓	7
(Jindal & Taneja, 2015)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Kasthurirathne, Dixon, & Grannis, 2015)	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	9
(Kavuluru, Rios, & Lu, 2015)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Koopman, Karimi, et al., 2015)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Koopman, Zuccon, Nguyen, Bergheim, & Grayson, 2015)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(MacRae, et al., 2015)	✓	✓	✓	✗	✗	✗	✓	✓	✓	✓	7
(Martinez, et al., 2015)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Miasnikof, et al., 2015)	✓	✓	✓	✗	✗	✗	✓	✓	✓	✓	7
(Parlak, Uysal, & Ieee, 2015)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(A. L. Pineda, et al., 2015)	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	9
(Rani, Gladis, & Mammen, 2015)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Rios & Kavuluru, 2015)	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	9
(Sarker & Gonzalez, 2015)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(H. Y. Zhou, Zhang, Wang, & Zhang, 2015)	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	9
(L. Zhou, et al., 2015b)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Guido Zuccon, et al., 2015)	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	9
(Fragos & Skourlas, 2016)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Saeed Hassanpour & Curtis P Langlotz, 2016)	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	9
(Kalter, Perin, & Black, 2016)	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	9
(S. N. Kasthurirathne, et al., 2016)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Kocbek, et al., 2016)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Lopprich, et al., 2016)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Masino, Grundmeier, Pennington, Germiller, & Crenshaw, 2016)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Mouriño-García,	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	9

Study	QAC1	QAC2	QAC3	QAC4	QAC5	QAC6	QAC7	QAC8	QAC9	QAC10	Overall Score
Pérez-Rodríguez, Anido-Rifón, & Gómez-Carballa, 2016)											
(G. Mujtaba, et al., 2016)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Napolitano, Marshall, Hamilton, & Gavin, 2016)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Parlak & Uysal, 2016a)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Saqlain, Hussain, Saqib, & Khan, 2016)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Sedghi, Weber, Thomo, Bibok, & Penn, 2016)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(K. Yadav, et al., 2016)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	9
(Amrit, Paauw, Aly, & Lavric, 2017)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Barak-Corren, et al., 2017)	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	9
(Buchan, Filannino, & Uzuner, 2017)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Clark, Wellner, Davis, Aberdeen, & Hirschman, 2017)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Hassanpour, Langlotz, Amrhein, Befera, & Lungren, 2017)	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	9
(Imane & Mohamed, 2017)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(S. N. Kasthurirathne, et al., 2017)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Lauren, et al., 2017)	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	9
(Lucini, et al., 2017)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(G. Mujtaba, L. Shuib, R. G. Raj, M. A. Al-Garadi, et al., 2017b)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Ghulam Mujtaba, et al., 2017)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(G. Mujtaba, L. Shuib, R. G. Raj, R. Rajandram, et al., 2017a)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Oleynik, Patrão, & Finger, 2017)	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	9
(Shin, Chokshi, Lee, & Choi, 2017)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Wang, et al., 2017)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Wu & Wang,	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	9

Study	QAC1	QAC2	QAC3	QAC4	QAC5	QAC6	QAC7	QAC8	QAC9	QAC10	Overall Score
2017)											
(Yoon, Roberts, Tourassi, et al., 2017)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
(Parlak & Uysal, 2018)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10