Expert Systems With Applications 213 (2023) 118997

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Review

AI-based ICD coding and classification approaches using discharge summaries: A systematic literature review

Rajvir Kaur *, Jeewani Anupama Ginige, Oliver Obst

School of Computer, Data and Mathematical Sciences, Western Sydney University, Australia

ARTICLE INFO

Keywords:
Computer assisted clinical coding
Clinical classification and coding
Discharge summaries
Natural Language Processing
Machine learning
Deep learning

ABSTRACT

The assignment of codes to free-text clinical narratives have long been recognised to be beneficial for secondary uses such as funding, insurance claim processing and research. The current scenario of assigning clinical codes is a manual process which is very expensive, time-consuming and error prone. In recent years, many researchers have studied the use of Natural Language Processing (NLP), related machine learning and deep learning methods and techniques to resolve the problem of manual coding of clinical narratives and to assist human coders to assign clinical codes more accurately and efficiently. The main objective of this systematic literature review is to provide a comprehensive overview of automated clinical coding systems that utilise appropriate NLP, machine learning and deep learning methods and techniques to assign the International Classification of Diseases (ICD) codes to discharge summaries. We have followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines and conducted a comprehensive search of publications from January, 2010 to December 2021 in four high quality academic databases: PubMed, ScienceDirect, Association for Computing Machinery (ACM) Digital Library, and the Association for Computational Linguistics (ACL) Anthology. We reviewed 6128 publications; 42 met the inclusion criteria. This review identified: 6 datasets having discharge summaries (2 publicly available, 4 acquired from hospitals); 14 NLP techniques along with some other data extraction processes, different feature extraction and embedding techniques. The review also shows that there is a significant increase in the use of deep learning models compared to machine learning. To measure the performance of classification methods, different evaluation metrics are used. Efforts are still required to improve ICD code prediction accuracy, availability of large-scale de-identified clinical corpora with the latest version of the classification system. This can be a platform to guide and share knowledge with the less experienced coders and researchers.

1. Introduction

Documentation related to an episode of care of a patient, commonly referred to as a medical record, contains clinical findings, diagnoses, interventions, laboratory test details and medication details which are invaluable information for clinical decision-making. To carry out meaningful statistical analysis, these medical records are annotated with codes which are called *clinical codes*. The International Classification of Diseases (ICD) provides a taxonomy of classes, each of which identified by a code assigned to an episode of care of a patient, based on which reimbursement is done in some countries (Kaur & Ginige, 2018). Clinical codes are assigned by trained professionals, known as *clinical coders*, who ideally have a sound knowledge of medical terminologies, clinical classification systems, and coding rules and guidelines. The current scenario of assigning clinical codes is a manual process which is very expensive, time-consuming, and error-prone (Xie & Xing, 2018). The wrong assignment of codes leads to issues such as reviewing of the whole process, financial losses, increased labour costs, as well as delays in the reimbursement process. The coded data is not only used by insurance companies for the reimbursement purposes, but also by government agencies and policy makers to analyse healthcare systems, justify investments done in the healthcare industry and plan future investments based on these statistics (Kaur & Ginige, 2018).

With the transition from ICD-9 to ICD-10 in 1992, the number of codes increased from 3882 to approximately 70,000, which further makes manual coding a non-trivial task (Subotin & Davis, 2014). Moreover, the manual assignment of codes is a complex process due to the continuous evolution of rules. As an example, clinical coding competency guidelines in Australia[1] require that an entry-level clinical

* Corresponding author.
E-mail addresses: 18531738@student.westernsydney.edu.au (R. Kaur), j.Ginige@westernsydney.edu.au (J.A. Ginige), O.Obst@westernsydney.edu.au (O. Obst).

[1] https://www2.health.vic.gov.au/health-workforce/health-information-workforce/clinical-coding-workforce

https://doi.org/10.1016/j.eswa.2022.118997
Received 29 March 2022; Received in revised form 6 October 2022; Accepted 7 October 2022
Available online 14 October 2022

coder is able to code a minimum of 5 clinical episodes of care per hour with 85% accuracy. However, in practice, on average, a clinical coder codes 3 to 4 clinical episodes of care, resulting in 15–42 records per day depending on the experience and efficiency of the human coder (Kaur & Ginige, 2018; Santos et al., 2008). The cost incurred in assigning clinical codes and the follow up corrections are estimated to be 25 billion dollars per year in the United States (Farkas & Szarvas, 2008; Xie & Xing, 2018). There are several reasons behind the wrong assignment of codes. First, the assignment of ICD codes to a patient's records could be erroneous due to the subjective nature of human perception (Kaur, 2019). Second, the manual process of assigning codes is a tedious task leading to inability to locate critical and subtle findings due to fatigue. Third, in many cases, physicians often use abbreviations or synonyms in their notes, which causes ambiguity which cannot always be resolved from the context by non-expert readers (Xie & Xing, 2018).

Huang et al. (2019) show that more than 80% of health record data is in an unstructured form which acts as a barrier in an automated clinical decision making process. Unstructured clinical text includes clinical notes, surgical records, discharge summaries, radiology reports, and pathology reports. This unstructured text contains a lot of valuable information but lacks common structural frameworks and may contain errors, such as spelling errors, grammatical errors, and semantic ambiguities, which further increases the complexity of data processing and analysis (Sun et al., 2018). There are various other factors that hinder the automated clinical decision making process:

1. *Idiosyncrasies of medical language*: Free text clinical notes are rife with obscure vocabulary, non-standard syntax, and ambiguous abbreviations. These free text clinical notes are often typed hurriedly, and thus, contain many spelling and grammatical errors (Catling et al., 2018). Moreover, synonyms for clinical concepts are also used interchangeably and negating expressions are placed distantly from the negated concept (Chapman et al., 2001).
2. *Scarcity of electronic health records (EHRs)*: This is a long-standing barrier to automated coding as many hospitals are still using paper records, which limits the availability of training data. There is limited adoption of structured EHRs in developing countries, which leaves clinicians with no choice but to resort to manual consumption of available clinical notes for decision making. Sometimes crucial information about a patient is mostly lost when transcribed into structured EHRs (Gangavarapu et al., 2020).
3. *Label-space problem*: Many disease ontologies contain tens of thousands of labels, and their distribution is highly imbalanced in most datasets, with many absent labels for rare diseases. Many studies have used k-most frequently occurring labels for training and discarded the least occurring labels in order to achieve higher accuracy of their algorithm (Catling et al., 2018). However, this type of negligence would be unacceptable in real healthcare environments where many rare diseases have serious sequelae when neglected.
4. *Requirement of large amount of training data*: It is quite rare to find freely available medical repositories that contain clinical records of a patient. To train a machine learning model, a large amount of training data is required (Kaur & Ginige, 2018).

To reduce coding errors and cost, there is a need for an automated clinical coding system, commonly referred to as computer-assisted clinical coding system that will overcome the manual coding challenges and assist human coders to assign correct clinical codes more quickly and accurately. Several previous studies have addressed the automated ICD coding systems, but are currently not widely used, most likely because the systems are still in development and their performance in a real time scenario is unproven (Stanfill et al., 2010). The automated coding methods make use of Artificial Intelligence (AI) techniques to convert unstructured clinical text to structured text without human interaction (Kaur, 2018). In addition, a wide range of applications in the biomedical domain uses natural language processing (NLP) techniques to manage large volumes of text data by extracting relevant information in a timely manner.

Our contribution

There are only three systematic literature reviews available for clinical coding and classification systems (Burns et al., 2011; Campbell et al., 2001; Stanfill et al., 2010). Two studies (Burns et al., 2011; Campbell et al., 2001) conducted a systematic review to measure the accuracy of routinely collected hospital discharge data in the United Kingdom. The primary objective of these two studies was to identify and investigate the accuracy of hospital episode data and to investigate factors affecting variation in coding. The studies did not focus on automated ICD code assignment using any computer applications. Stanfill et al. (2010) conducted a systematic literature review of studies that use computer application to automatically generate clinical codes or classification from free-text clinical documents. This review includes the studies published prior to March 2009. Mujtaba et al. (2019) focused on clinical text classification studies published from January 2013 to January 2018. In contrast, this paper presents a systematic literature review to investigate all types of clinical reports including pathology reports, radiology reports, autopsy reports, death certificates and other medical reports for text classification, rather than automated ICD coding and classification purposes. This systematic literature review recapitulates the existing studies published from January 2010 to December 2021 on automated ICD code assignment using discharge summaries. Within the broader scope of this review, the work address the following research questions:

1. What are the close ended quality assessment questions that make the study suitable for systematic literature review?
2. What are the different datasets available for ICD code assignment of discharge summaries?
3. Which algorithms are available to assign automated ICD codes?
4. What are the different evaluation metrics used in the studies to evaluate the performance of automated ICD code assignment systems?
5. Which model gives best performance in assigning ICD codes to discharge summaries as it applies to acute admitted care?
6. What are the future directions of research in automated ICD coding?

2. Background

The history of medical coding started as an attempt to avoid the black death known as the bubonic plague, caused by the bacteria *Yersinia pestis*, which arrived in Sicily via ship rats in the year 1347. The outbreaks of plague continued in Europe throughout the next 3 centuries. In the year 1532, the systematic collection of data on causes of death known as *The London Bills of Mortality* began and these data were published weekly. Causes of death found in the Bills included diseases such as jaundice, smallpox, rickets, spotted fever, and plague. In the year 1665, John Graunt, a London merchant, published *Reflections on the Weekly Bills of Mortality* to examine the deaths from plague in the context of all other causes of mortality. In the year 1839, the epidemiologist William Farr prepared a classification system which was primarily based on the anatomical site and consisted of 138 rubrics (Moriyama et al., 2011). After Farr's death in 1883, Jacques Bertillon, a French statistician, prepared a revised list that was adopted by the International Statistical Institute in 1893. The Bertillon Classification was the first standard system implemented internationally to record causes of death known as *the International Classification of Diseases*. Delegates from 26 countries adopted the Bertillon Classification

(ICD-1) in 1900 and subsequent revisions occurred since 1920. After Bertillon's death in 1922, interest grew in using the classification to categorise not only causes of mortality, but also causes of morbidity.

In April 1948, at the Sixth Decennial Revision Conference in Paris, the WHO approved a comprehensive list for classification of causes of illness (morbidity), as well as causes of deaths (mortality). The "Manual of the International Statistical Classification of Diseases, Injuries, and Causes of Death" is generally known as ICD. With the eighth revision, the United States developed its own version, known as ICDA-8 (ICD-Adapted), due to disagreement over the circulatory section. With the WHO's ninth revision (ICD-9) to code and classify mortality data from death certificates, the United States adopted a clinical modification of the international version (ICD-9-CM) as the official system for assigning codes to diagnoses and procedures associated with the hospital utilisation. ICD-9-CM is based on WHO's ICD-9 but provides additional morbidity details and consists of three volumes (Volumes 1 and 2 contain diagnosis codes and volume 3 contains procedure codes). All changes and modifications to ICD-9-CM were done by US governmental agencies: the National Center for Health Statistics (NCHS) and the Centers for Medicare and Medicaid Services (CMS). ICD-10 was adopted by the WHO in 1990, with modifications made by Australia in 1998, and Canada in 2001 (Cartwright, 2013). Although ICD-10 has been used in the United States since 1999 for mortality reporting, ICD-9-CM was used for morbidity until October 2015 due to late adoption of ICD-10 clinical modification. The United States developed a Clinical Modification (ICD-10-CM) for medical diagnoses based on WHO's ICD-10 and CMS developed a new Procedure Coding System (ICD-10-PCS) for inpatient procedures. ICD-10-CM replaces ICD-9-CM, volumes 1 and 2, and ICD-10-PCS replaces ICD-9-CM volume 3.

Many countries extended the ICD-10 classification system to make it suitable for their country specific reporting purposes (Kaur & Ginige, 2018). For example, ICD-10-CM (Clinical Modification) is used in the USA, ICD-10-CA (Canadian Modification) is used in Canada, ICD-10-GM (German Modification) in Germany, and ICD-10-AM (Australian Modification) is used in Australia along with 15 other countries including Ireland, Singapore, and Saudi-Arabia (Cumerlato et al., 2010; Kaur & Ginige, 2018). Twenty-six years after the introduction of ICD-10, in 2018, the next generation of classification ICD-11 was put forward to WHO general assembly for approval and was released in May 2019, but has not been implemented yet in hospital settings (Kaur, 2018; Kaur & Ginige, 2018). ICD-11 increases the complexity by introducing a new code structure, with a new chapter on X-Extension Codes, dimensions of external causes (histopathology, consciousness, temporality, and etiology), and new chapters on sleep–wake disorder, conditions related to sexual health, and traditional medicine conditions (Hargreaves & Njeru, 2014; Reed et al., 2016; World Health Organisation, 2016).

2.1. Comparison of different classification systems

Since the introduction of ICD-10 in 1992, many countries have modified the WHO's ICD-10 classification system to suit their country specific reporting purpose. For example, there are a few major differences between the U.S. and Australian classification systems. Firstly, there are approximately 6606 Australian specific codes that are used in Australia along with 15 other countries that use Australian classification system as their national classification system. For example, in ICD-10, "*contact with venomous spiders*" is coded as $X21$, in ICD-10-CM it is coded as $T63.301A$ whereas in Australia, there is more specificity added in identifying the type of the spider such as funnel web spider, red back spider, or white-tailed and other necrotising spider (Kaur, 2019). Secondly, countries that have developed their own national classification system use different coding practices. For example, in the U.S., Pulmonary oedema is coded as $J81$, whereas in Australia, to assign code for Pulmonary oedema, there is ACS rule 0920 which says, "When *acute pulmonary oedema* is documented without further qualification about the underlying cause, assign *I50.1 Left ventricular failure*".

The health classification system used in Australia includes ICD-10-AM for classifying diseases and other health problems, Australian Classification of Health Interventions (ACHI) for classifying procedures and interventions and Australian Coding Standards (ACS) specifies coding standards. This systematic literature review is not restricted on only the US and Australian health classification system. The above given example is to show the differences in country's specific health classification system. Table 1 shows the comparison of ICD-9-CM, ICD-10-CM and ICD-10-AM diagnosis codes and Table 2 shows the comparison of ICD-9-CM procedure codes, ICD-10-PCS and ACHI codes.

3. Methods

We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) recommendations for reporting in systematic reviews (Moher et al., 2009). After reviewing the PRISMA guidelines, we structured our review into two phase: (see Fig. 1).

1. Search strategy phase
2. Data extraction phase

3.1. Search strategy phase

The search strategy phase includes data sources, formulation of the search keywords and search queries, screening and selection criteria, and quality assessment of the retrieved publications. This phase was designed to identify all potential relevant publications related to the automated clinical coding and classification systems that would leverage on NLP, machine learning and deep learning techniques.

3.1.1. Data sources and search strategies

We conducted a comprehensive search of several databases for publications from January 1, 2010 to December 31, 2021. Since January 2010 onwards, no comprehensive systematic literature review has recapitulated on automated clinical coding using discharge summaries in the last one decade. In this review, the studies were retrieved from four high quality academic databases- PubMed, ScienceDirect, Association for Computing Machinery (ACM) Digital Library, and the Association for Computational Linguistics (ACL) Anthology. We looked for publications which include conference proceedings and journal articles written in English, and excluded those in the form of editorial, letter, note or comment. Our focus is on only acute admitted care, therefore, this systematic literature review only consider studies that uses discharge summaries for automated ICD code assignment. Table 3 shows a list of study selection criteria.

3.1.2. Formulation of keywords and search queries

To perform the search query, each keyword is paired using OR operators, whereas the concepts are paired using AND operator. Table 4 shows the keywords used to perform search query.

3.1.3. Screening and selection criteria

Based on the search query, the publications retrieved from each database were stored in EndNote X9 (Thomson Reuters) reference management software and the *Find Duplicates* function was used to review and delete duplicates. Some manual deletion was also performed. After removing the duplicates, the remaining 2949 publications were screened based on the titles and abstracts to determine if the study is relevant to our review. After the first screening, 230 publications were selected. In the second screening, full text PDF files were obtained using the EndNote *Find Full Text* feature. The full texts that could not be found or obtained because of access restrictions were then requested and attached manually to the list. The third screening was performed to review the full text of the publications. Each paper's *Methodology or Experimental Work* section was reviewed properly to determine if it meets the review criteria. During this screening phase, 182 publications were removed as they did not meet the selection criteria and the remaining 48 studies were eligible for full text assessment.

Table 1
Comparison between ICD-9-CM, ICD-10-CM and ICD-10-AM diagnoses codes.

ICD-9-CM diagnoses codes	ICD-10-CM diagnoses codes	ICD-10-AM diagnoses codes
Used in USA	Currently being used in USA	Currently being used in Australia
Approx. 14,025 codes	Approx. 70,000 diagnoses	Approx. 16,953 disease codes and 2,825 morphology codes
Valid codes have 3 to 5 characters	Valid codes have 3 to 7 characters	Valid codes have 3 to 5 characters
Decimal used after third character	Decimal used after third character	Decimal used after third character
First character is numeric or alpha (E and V only)	First character is always alpha	First character is always alpha
Characters 2 to 5 are numeric	Second character is numeric, 3 to 7 characters are alpha or numeric	Characters 2 to 5 are numeric.

Table 2
Comparison between ICD-9-CM, ICD-10-PCS and ACHI code set.

ICD-9-CM procedure codes	ICD-10-PCS procedure codes	ACHI codes
Approx. 3,824 procedure codes	Approx. 72,000 procedure codes	Approx. 6,248 procedure codes
Valid codes have four digits, all numeric	Valid codes have 7 alphanumeric characters (the letter O and I are not used to avoid confusion with 0 and 1)	Valid codes have 7-digit maximum
Decimal used after second digit (XX.XX format)	No decimals used	Hyphen(-) used after 5 characters (XXXXX-XX format)

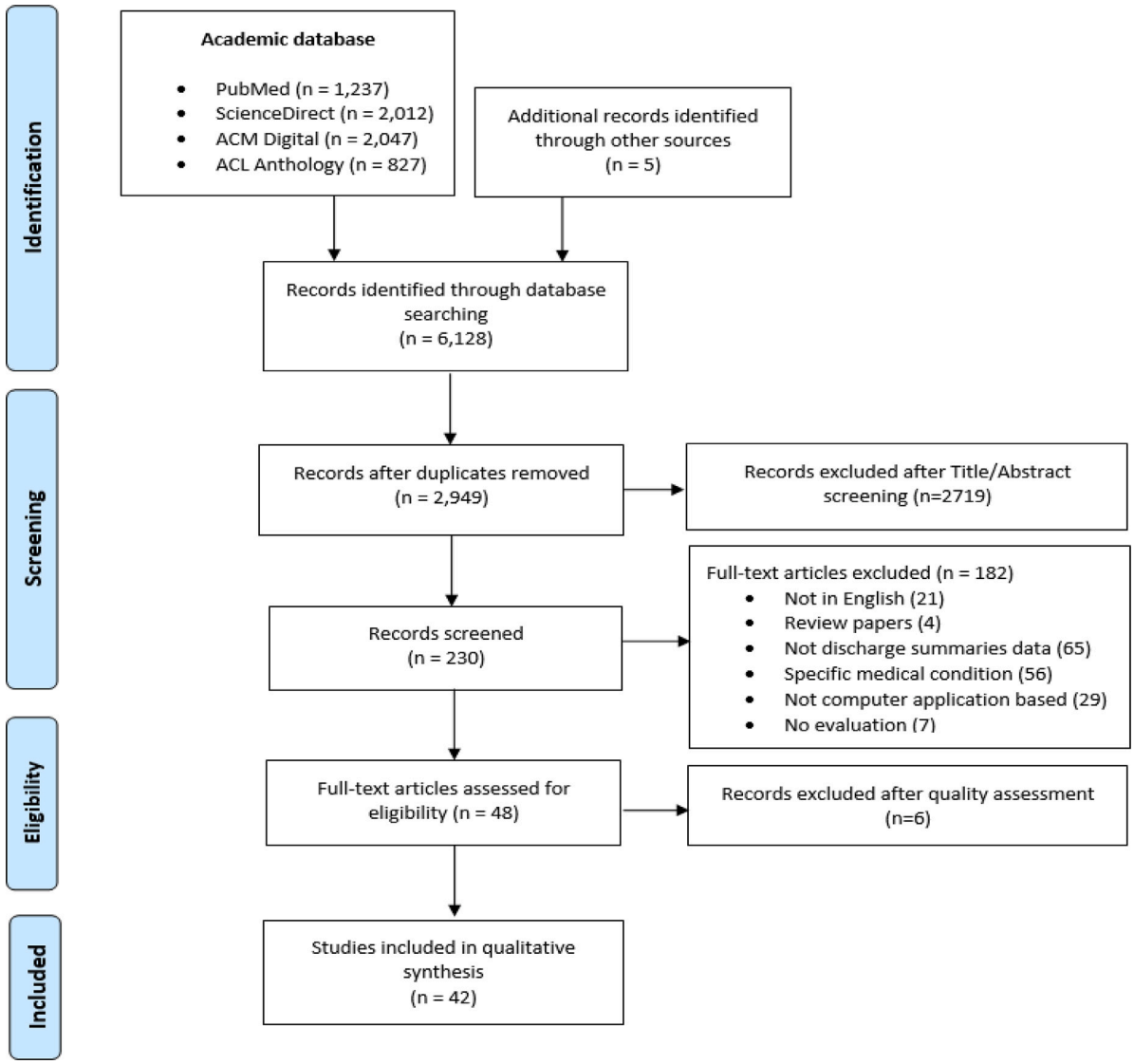

Fig. 1. PRISMA flow diagram.

Table 3
List of the inclusion and exclusion criteria.

No	Inclusion criteria
1.	Study must have been published between January 1, 2010 and December 31, 2021.
2.	Study must have used discharge summaries or other medical reports along with a discharge summary as a dataset and must be written in English.
3.	Study must have used clinical reports or medical documents for automated ICD code assignment.
4.	Study should perform automated ICD code assignment using computational methods such as NLP, machine learning and deep learning techniques.
5.	Study must have evaluated the performance of the proposed system using standard evaluation metrics.
No	**Exclusion criteria**
1.	Study using documents other than discharge summaries such as radiology report, pathology report, death certificate, autopsy, surgical or laboratory report.
2.	Study has not defined any clinical coding or clinical classification system.
3.	Study assigning clinical codes on one specific condition (disease or procedure).
3.	Study focused on medical image classification
4.	Study has not evaluated the performance of automated coding and classification system.
5.	Study focused on medical text snippets.
6.	Study has used time-series data classification in the medical field, such as EEG signals and not associated with text classification

Table 4
Searched keywords using different concepts.

Concepts	Keywords
Concept 1: Keywords related to classification domain	ICD code classification OR ICD code assignment OR clinical text classification OR automatic clinical text classification OR clinical text categorisation OR computer assisted coding OR computer assisted clinical coding
Concept 2: Keywords related to medical documents	medical reports OR clinical reports OR clinical narratives OR electronic health records OR free text clinical reports OR discharge summaries
Search Query	**(Concept 1) AND (Concept 2)**

3.1.4. Quality assessment of the retrieved publications

The quality assessment of the selected publications was one of the essential steps to find out whether or not the study is suitable to address our review objectives. To perform the quality assessment, we formulated a checklist of ten close-ended questions as given below:

Q1: Are research objectives clearly defined?

Q2: Is research methodology well-defined?

Q3: Is the train and test data source clearly defined?

Q4: Are the data pre-processing techniques clearly defined and their selection justified?

Q5: Are the feature extraction techniques or feature engineering clearly described and justified?

Q6: Are the classifiers clearly described?

Q7: Does the study perform the comparison with the existing baseline models?

Q8: Is the performance of the system evaluated and results properly interpreted and discussed?

Q9: Does the study performs an ablation study?

Q10: Does the conclusion reflect the research findings?

The answers to each can be either *Yes* or *No* and each question carries weight of "1". A threshold of 7 was set to include study in the review. During the quality assessment process, 6 studies were excluded as they did not obtain a score of 7. Hence, this review involves 42 selected studies. Table 5 shows the quality assessment criteria score of the selected studies.

3.2. Data extraction phase

The data extraction phase includes review of the selected publications and extraction of the key aspects: Data source, pre-processing techniques, feature extraction and embedding techniques, classification and performance evaluation. Section 4 presents the critical review of these key aspects.

4. Critical review of the selected studies

Various attempts have been made by many researchers to create automated systems for assigning clinical codes to episode of care (Kaur & Ginige, 2018). Research studies have used different methods and techniques to handle and process clinical text. This section critically reviews the selected studies from six different aspects as mentioned above.

4.1. Data source

Clinical documents can be classified into two categories: clinical notes and diagnostic reports. Clinical notes may include a patient's medical history, physical examination history, clinical observations, a summary of diagnostic and therapeutic procedures and treatment plan, whereas diagnostic reports refer to the reports provided by diagnostic services including laboratory reports, radiology reports, and pathology reports. However, in this review, we primarily focus on discharge summaries and other clinical reports used along with discharge summary as the input text data. Fig. 2 shows different types of datasets used in the selected studies. Table 6 shows a list of datasets used by the selected 42 studies. The datasets MIMIC-II and MIMIC-III are publicly available, whereas the remaining four, University of Kentucky (UKY) medical centre, Australian hospital medical records, NYU Langone hospital, and Taiwan hospital discharge notes are private datasets obtained from hospitals.

Apart from that, we found that the majority of the studies ($n = 36$) have focused on the ICD-9-CM classification system, 3 studies (Amoia et al., 2018; Lin et al., 2017; Zhang et al., 2020) predicted ICD-10-CM codes, 2 studies (Kaur & Ginige, 2018, 2019) predicted ICD-10-AM and ACHI codes, and 1 study (Subotin & Davis, 2014) predicted ICD-10-PCS, 1 study (Xu et al., 2019) converted ICD-9-CM codes to ICD-10-CM codes using an online resource[2] before assigning them to discharge summaries. This shows that there is a scarcity of studies relevant to other classification systems including Australian classification systems.

[2] https://www.icd10data.com/

Table 5
Overview of quality scores of studies included in the review — 7 out of 10.

Year	Study	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Score
2013	Perotte et al. (2013)	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	8
2014	Marafino et al. (2014)	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	8
2014	Subotin and Davis (2014)	✓	✓	✗	✓	✓	✓	✗	✓	✗	✓	7
2015	Kavuluru et al. (2015)	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	8
2016	Ayyar and Oliver (2016)	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	8
2017	Prakash et al. (2017)	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	8
2017	Lin et al. (2017)	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	8
2017	Berndorfer and Henriksson (2017)	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	8
2018	Amoia et al. (2018)	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	8
2018	Catling et al. (2018)	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	9
2018	Baumel et al. (2018)	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	9
2018	Mullenbach et al. (2018)	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	9
2018	Samonte et al. (2018)	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	9
2018	Xie and Xing (2018)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
2018	Rios and Kavuluru (2018b)	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	9
2018	Kaur and Ginige (2018)	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	8
2019	Zeng et al. (2019)	✓	✓	✓	✗	✓	✓	✓	✓	✗	✓	8
2019	Kaur and Ginige (2019)	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	8
2019	Xie et al. (2019)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
2019	Falis et al. (2019)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
2019	Huang et al. (2019)	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	9
2019	Li et al. (2019)	✓	✓	✓	✗	✓	✓	✓	✓	✗	✓	8
2019	Rios and Kavuluru (2019)	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	9
2019	Schäfer and Friedrich (2019)	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	9
2019	Xu et al. (2019)	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	8
2019	Du et al. (2019)	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	9
2020	Cao et al. (2020)	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	9
2020	Guo et al. (2020)	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	9
2020	Vu et al. (2020)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
2020	Song et al. (2020)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
2020	Sonabend W et al. (2020)	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	9
2020	Mascio et al. (2020)	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	8
2020	Li and Yu (2020)	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	9
2020	Teng et al. (2020)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
2020	Moons et al. (2020)	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	9
2020	Ji et al. (2020)	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	9
2020	Hsu et al. (2020)	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	9
2020	Zhang et al. (2020)	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	9
2021	Pascual et al. (2021)	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	9
2021	Ji et al. (2021)	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	9
2021	Mayya et al. (2021)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	10
2021	Dong et al. (2021)	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	9

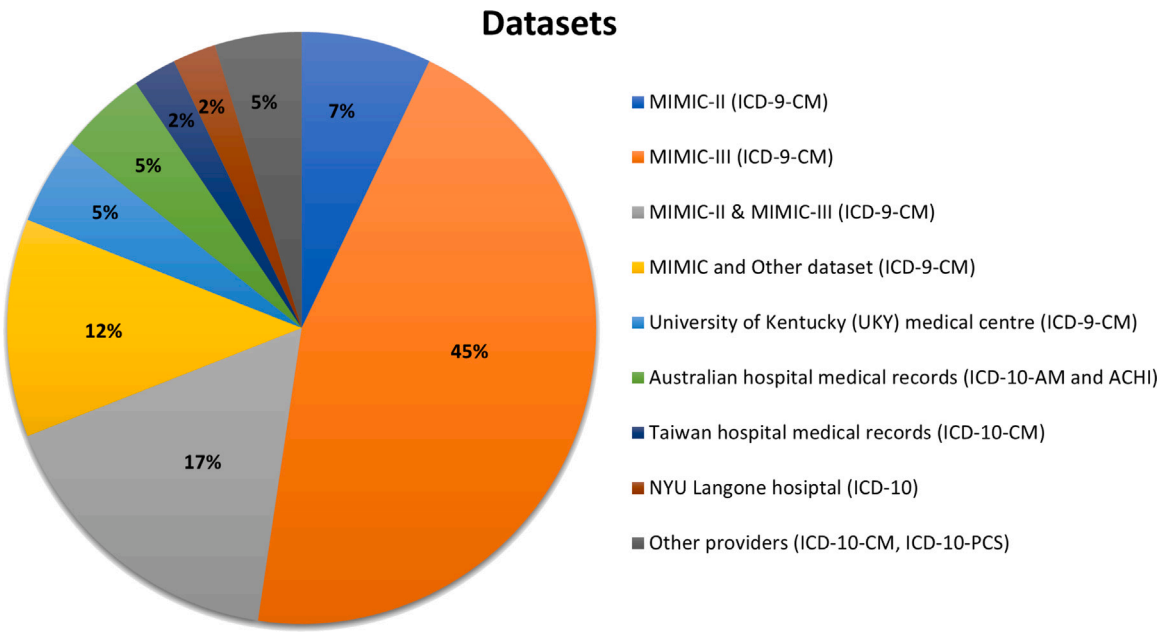

Fig. 2. Types of datasets used by the selected studies for automated ICD code assignment using discharge summaries.

Table 6
Different types of datasets that contain discharge summaries.

Dataset	Coding system	Description	Studies
MIMIC-II	ICD-9-CM	MIMIC-II (Medical Information Mart for Intensive Care II) database contains clinical records from 32,536 subjects collected between 2001 and 2008 from a variety of ICUs (medical, surgical, coronary care, and neonatal) in a single tertiary teaching hospital.	Marafino et al. (2014), Perotte et al. (2013), Du et al. (2019)
MIMIC-III	ICD-9-CM	MIMIC-III database is an extension of MIMIC-II comprising de-identified health-related data associated with over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012.	Ayyar and Oliver (2016), Prakash et al. (2017), Berndorfer and Henriksson (2017), Catling et al. (2018), Samonte et al. (2018), Xie and Xing (2018), Xie et al. (2019), Falis et al. (2019), Huang et al. (2019), Schäfer and Friedrich (2019), Xu et al. (2019), Guo et al. (2020), Song et al. (2020), Teng et al. (2020), Ji et al. (2020), Hsu et al. (2020), Pascual et al. (2021), Ji et al. (2021), Dong et al. (2021)
MIMIC-II and MIMIC-III	ICD-9-CM	Studies have used both MIMIC-II and MIMIC-III dataset.	Baumel et al. (2018), Mullenbach et al. (2018), Rios and Kavuluru (2018b), Li et al. (2019), Cao et al. (2020), Vu et al. (2020), Li and Yu (2020)
MIMIC and other datasets	ICD-9-CM	Studies have used the MIMIC dataset and other sources.	Zeng et al. (2019), Sonabend W et al. (2020), Mascio et al. (2020), Moons et al. (2020), Mayya et al. (2021)
University of Kentucky (UKY) medical centre	ICD-9-CM	The electronic medical records (EMRs) of the UKY medical center in-patient visits with discharge dates in the 2011–2012 two-year period. There are total of 71,461 EMRs having a total of 7,485 unique ICD-9 codes. UKLarge dataset consists of all in-patient visits. A subset of UKLarge, called UKSmall with 1,000 EMRs corresponding to a randomly chosen set of 1,000 in-patient visits from February, 2012.	Kavuluru et al. (2015), Rios and Kavuluru (2019)
Australian hospital medical records	ICD-10-AM and ACHI	A collection of medical records from acute or sub-acute hospitals all over Australia, held by the National Centre for Classification in Health (NCCH).	Kaur (2018), Kaur and Ginige (2019)
Taiwan hospital discharge notes	ICD-10-CM	Discharge notes from The Tri-Service General hospital, Taipei, Taiwan from June 1, 2015 to January 31, 2017	Lin et al. (2017)
NYU Langone Hospital	ICD-10	A total of 7.5 million notes corresponding to visits from about 1 million patients. Over 50% of the notes are progress notes, followed by telephone encounters (10%) and patient instructions (5%).	Zhang et al. (2020)
Other healthcare providers	ICD-10-CM, ICD-10-PCS	Studies have not mentioned data sources	Amoia et al. (2018), Subotin and Davis (2014)

4.2. Pre-processing

Pre-processing removes unwanted or meaningless information from the dataset as clinical narratives may contain high level of noise, sparsity, misspelled words, or grammatical errors. Different pre-processing techniques including tokenization, lowercase conversion, removal of stop words, sentence segmentation, removal of non-alphabetical characters, abbreviation expansion, spelling error detection and correction, negation detection, stemming, lemmatization, parsing, part-of-speech tagging, named entity recognition, and word normalisation were applied in the selected studies as shown in Fig. 3. Table 7 gives the information on different NLP techniques used by the studies. The majority of studies have applied tokenization, followed by removal of stop words, removal of non-alphabetic characters and lowercase conversion. Few studies also used other data processing steps such as regular expression matching (Teng et al., 2020), building a dictionary or vocabulary (Baumel et al., 2018; Lin et al., 2017), removing non-matching terms (Schäfer & Friedrich, 2019), removal of de-identified or confidential information (Huang et al., 2019; Samonte et al., 2018). Studies including (Baumel et al., 2018; Ji et al., 2020; Li & Yu, 2020; Moons et al., 2020; Mullenbach et al., 2018; Prakash et al., 2017; Song et al., 2020; Vu et al., 2020) truncated the documents to a maximum length of 2500 or 4000 tokens in order to reduce the computational cost. Moons et al. (2020), Mullenbach et al. (2018), Song et al. (2020), Teng et al. (2020), Xie et al. (2019) replaced tokens with an 'UNK' token if they appeared in less than three training documents. Some studies (Cao et al., 2020; Li et al., 2019; Subotin & Davis, 2014; Zeng et al., 2019) did not mention any data pre-processing technique.

4.3. Feature extraction

Feature extraction is the process of extracting useful text characteristics. There are two general approaches of feature extraction: expert-driven and fully automated feature extraction (Mujtaba et al., 2019). In expert-driven feature extraction, a group of experts discovers useful and discriminative features from clinical reports, whereas, the fully automated feature extraction approach makes use of algorithms instead of human experts.

Different types of feature extraction techniques/word embeddings used in the selected studies listed in Table 8 includes:

1. Term Frequency-Inverse Document Frequency (TF-IDF) measures the relative importance of a word in a document or corpus.
2. Bag-of-Words (BoW): A vocabulary of unique words is created, where each word represents an independent, and discriminative feature.
3. N-grams are contiguous word sequences up to a length of N. Single words are called 1-grams (or unigrams), sequence of two words are called 2-grams (or bi-grams), sequence of three words are called 3-grams (Tri-grams), and so on.
4. Word2Vec is a method to construct embeddings using two models (Mikolov et al., 2013): Skip-gram and Continuous Bag of

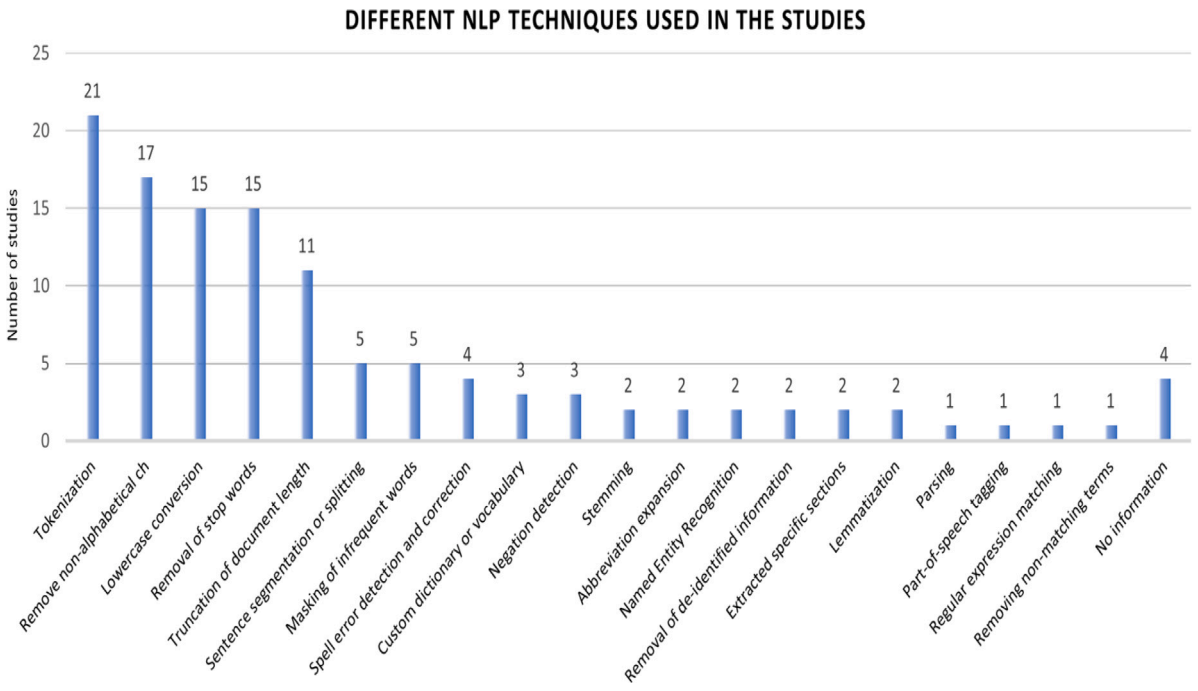

Fig. 3. Different NLP techniques used in the studies to pre-processing the discharge summaries.

Table 7
Different NLP techniques used in the studies for data pre-processing.

Pre-processing techniques	Studies	Count
Tokenization	Perotte et al. (2013), Ayyar and Oliver (2016), Berndorfer and Henriksson (2017), Amoia et al. (2018), Catling et al. (2018), Baumel et al. (2018), Mullenbach et al. (2018), Samonte et al. (2018), Kaur and Ginige (2018), Xie et al. (2019), Falis et al. (2019), Xu et al. (2019), Du et al. (2019), Vu et al. (2020), Mascio et al. (2020), Li and Yu (2020), Teng et al. (2020), Moons et al. (2020), (Ji et al., 2020), Pascual et al. (2021), Ji et al. (2021)	21
Lowercase conversion	Amoia et al. (2018), Mullenbach et al. (2018), Rios and Kavuluru (2018b), Xie et al. (2019), Falis et al. (2019), Schäfer and Friedrich (2019), Vu et al. (2020), Song et al. (2020), Mascio et al. (2020), Li and Yu (2020), Teng et al. (2020), Moons et al. (2020), (Ji et al., 2020), Pascual et al. (2021), Ji et al. (2021)	15
Removal of stop words	Marafino et al. (2014), Kavuluru et al. (2015), Prakash et al. (2017), Berndorfer and Henriksson (2017), Baumel et al. (2018), Mullenbach et al. (2018), Rios and Kavuluru (2018b), Kaur and Ginige (2018), Kaur and Ginige (2019), Huang et al. (2019), Du et al. (2019), Guo et al. (2020), Teng et al. (2020), Hsu et al. (2020), Pascual et al. (2021)	15
Sentence segmentation or splitting	Marafino et al. (2014), Baumel et al. (2018), Kaur and Ginige (2018), Hsu et al. (2020), Zhang et al. (2020)	5
Removal of non-alphabetical characters including punctuation, special characters	Marafino et al. (2014), Baumel et al. (2018), Mullenbach et al. (2018), Amoia et al. (2018), Xie et al. (2019), Falis et al. (2019), Du et al. (2019), Vu et al. (2020), Song et al. (2020), Mascio et al. (2020), Li and Yu (2020), Moons et al. (2020), (Ji et al., 2020), Hsu et al. (2020), Pascual et al. (2021), Mayya et al. (2021), Ji et al. (2021)	17
Abbreviation expansion	Kaur and Ginige (2018), Kaur and Ginige (2019)	2
Spell error detection and correction	Lin et al. (2017), Kaur and Ginige (2018), Kaur and Ginige (2019), Mayya et al. (2021)	4
Negation detection	Marafino et al. (2014), Kaur and Ginige (2018), Guo et al. (2020)	3
Stemming	Kaur and Ginige (2019), Mascio et al. (2020)	2
Lemmatization	Berndorfer and Henriksson (2017), Mascio et al. (2020)	2
Parsing	Sonabend W et al. (2020)	1
Part-of-speech tagging	Berndorfer and Henriksson (2017)	1
Named Entity Recognition	Kavuluru et al. (2015), Sonabend W et al. (2020)	2
Regular expression matching	Teng et al. (2020)	1
Custom dictionary or vocabulary	Lin et al. (2017), Baumel et al. (2018), Mayya et al. (2021)	3
Truncation of document length	Prakash et al. (2017), Mullenbach et al. (2018), Rios and Kavuluru (2019), Vu et al. (2020), Song et al. (2020), Li and Yu (2020), Moons et al. (2020), Ji et al. (2020), Pascual et al. (2021), Ji et al. (2021), Dong et al. (2021)	11
Masking of infrequent words (replace to "Unkown" or "UNK")	Mullenbach et al. (2018), Xie et al. (2019), Song et al. (2020), Teng et al. (2020), Moons et al. (2020)	5
Removal of de-identified or confidential information	Samonte et al. (2018), Huang et al. (2019)	2
Extracted specific sections	Xie and Xing (2018), Mayya et al. (2021)	2
Removing non-matching terms	Schäfer and Friedrich (2019)	1
No information	Subotin and Davis (2014), Zeng et al. (2019), Li et al. (2019), Cao et al. (2020)	4

Words (CBOW). The skip-gram model learns from the existing words in a sentence to predict the next word, whereas the CBOW model uses the neighbouring word to predict the next word (Mujtaba et al., 2019).

5. Doc2Vec and Paragraph2vec (Le & Mikolov, 2014) are variants of word2vec. They focus on predicting words in the document or paragraph. Doc2vec creates a numeric representation of a document irrespective to its length. Paragraph vector learns continuous distributed vector representations for pieces of texts. The text can be of variable-length ranging from sentences to documents.
6. Global Vectors (GloVe) is another commonly used word embedding method proposed by Pennington et al. (2014). GloVe uses global information from the term co-occurrence matrix to learn word embeddings (Khattak et al., 2019).
7. FastText can handle new, out-of-vocabulary terms by extending the skip-gram model with internal sub-word information in the form of character n-grams (Khattak et al., 2019).
8. BERT is a contextualised word representation model based on a multi-layer bi-directional transformer-encoder (Devlin et al., 2019). BERT has two versions: (1) the $\text{BERT}_{\text{BASE}}$ model with 12 layers of transformer blocks, 12 bidirectional self-attention heads, and 110 million parameters, and (2) the $\text{BERT}_{\text{LARGE}}$ model, with 24 layers of transformer blocks, 16 bidirectional self-attention heads, and 340 million parameters. BERT is pre-trained on two unsupervised tasks: masked language model and next sentence prediction. In masked language model, 15% of the tokens are randomly masked and the model is trained to predict the masked tokens. The next sentence prediction task capture more long-term or pragmatic information where the model is given a pair of sentences and is trained to identify when the second one follows the first. Additionally, domain-specific versions of BERT exist:

    - BioBERT (Lee et al., 2019) is a BERT-based model pre-trained on biomedical domain corpora (PubMed abstract and PMC full-text articles) and fine-tuned for biomedical text mining tasks such as named entity recognition, question answering, and relation extraction.
    - ClinicalBERT is trained on clinical text from MIMIC-III database. Alsentzer et al. (2019) presented various versions of BERT including ClinicalBERT, Clinical BioBERT, Discharge Summary BERT, and Discharge Summary BioBERT.

Alternating between synonymous terms is common practice in human language, particularly in clinical texts. For example, "heart attack", "Myocardial Infarction", "MI" denote the same entity of meaning, often referred to as "concept". Concept-based features use semantic lexicons, thesauri or the lexical part of ontologies (e.g. MeSH,[3] SNOMED-CT,[4] or the UMLS metathesaurus[5]) together with semantic taggers like MetaMap[6] or cTAKES.[7]

The majority of analysed studies ($n = 20$) applied Word2Vec embedding, followed by TF-IDF feature representation and BoW. Few studies (Ayyar & Oliver, 2016; Lin et al., 2017; Mascio et al., 2020) used GloVe embeddings, while Samonte et al. (2018) applied topical word embedding. There are a few studies that have not reported the embedding model except dimensions of embedding. Zhang et al. (2020) trained BERT models from scratch on EHR notes. Teng et al. (2020) applied knowledge graph embeddings to extract entities related to ICD-9 from freebase (Bollacker et al., 2008) and their results showed a significant improvement in code prediction.

[3] https://www.nlm.nih.gov/mesh/meshhome.html
[4] https://ontoserver.csiro.au
[5] https://www.nlm.nih.gov/research/umls/index.html
[6] https://metamap.nlm.nih.gov
[7] https://ctakes.apache.org/

4.4. Classification

In this review, the selected 42 studies have assigned ICD codes to discharge summaries either using machine learning or deep learning models.

4.4.1. Traditional machine learning (ML) approach

Machine learning approaches have gained more interest in many clinical studies due to their efficiency and effectiveness (Kaur & Ginige, 2018). Researchers across the globe have employed text classification to categorise clinical narratives into various categories using machine learning approaches including supervised (Hastie et al., 2009), unsupervised (Ko & Seo, 2000), semi-supervised (Zhu & Goldberg, 2009), transfer (Pan & Yang, 2010), and multi-view learning approaches (Amini et al., 2009). The performance of machine learning methods heavily depends on data representation (or features) on which they are applied. Therefore, much of the efforts deploying machine learning algorithms goes into design of pre-processing pipeline and data transformation (Bengio et al., 2013).

Among all the aforementioned ML approaches, the supervised ML approach is used often in clinical text classification research. In the supervised ML approach (Mujtaba et al., 2019), the clinical narratives collected from hospitals are labelled by domain experts into specific categories for further training. For example, the clinical narratives are labelled into two classes, cancer-positive (if the patient is diagnosed with cancer) and cancer-negative (if the patient is not diagnosed with cancer). After labelling, the clinical narratives are pre-processed via NLP techniques so that unnecessary information or noise is removed from them. Then, feature engineering is applied to extract the most discriminative features from the clinical narratives and form a numeric feature vector. This feature vector is then provided as an input to any ML algorithms such as SVM, Decision Trees, and AdaBoost to construct and validate the classification model. The ML algorithms learn from the data and perform classification, segmentation or prediction. The classification model can further be validated through either random sampling, k-fold cross validation, or leave-one-out techniques (Kohavi, 1995). In random sampling, the clinical data are shuffled randomly and divided into training, validation and testing sets. For example, 70% of data is used for training, 20% for testing and the remaining 10% for a validation set. Similarly, in k-fold cross validation, the shuffled data are split into data chunks and training is performed on the $(k-1)$ data chunks and test on the $(k-1)$th chunk. This process is repeated up to k times. The advantage of using k-fold cross validation is that all the clinical narratives are used for training and testing purposes, and each narrative is used for testing only once. However, this method is slower than random sampling. Leave-one-out is a special case of the k-fold, where the model is trained with $(n-1)$ reports and tested on the $(n-1)$th report, leaving one report out each time. However, it is good when the data are limited and imbalanced, but is much slower than random sampling (Mujtaba et al., 2019).

4.4.2. Deep learning

Deep learning is a type of machine learning technique that utilises a multi-layered neural network architecture to learn the hierarchical representation of data (Hasan & Farri, 2019). Deep learning models have demonstrated successful results in many NLP tasks such as language translation (Zhang & Zong, 2015), image captioning (LeCun et al., 2015) and sentiment analysis (Socher et al., 2013). Deep learning works quite well to solve non-linear classification problems and in the recent years, neural network models applied to NLP have achieved promising results over the approaches that use models such as SVM and logistic regression (Goldberg, 2016; Hasan & Farri, 2019). There are various known neural network models that are used for text and document classification such as Convolutional Neural Networks (CNN) (Kim, 2014), Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997), Gated Recurrent Units

Table 8
Feature Extraction techniques and pre-trained embeddings.

Year	Study	TF-IDF	N-grams	BoW	Doc2Vec	Word2Vec	GloVe	W_{embed} dim/Others
2013	Perotte et al. (2013)	✓	–	–	–	–	–	–
2014	Marafino et al. (2014)	✓	✓	–	–	–	–	–
2014	Subotin and Davis (2014)	–	–	✓	–	–	–	–
2015	Kavuluru et al. (2015)	✓	✓	–	–	–	–	–
2016	Ayyar and Oliver (2016)	–	–	–	–	–	✓	–
2017	Prakash et al. (2017)	–	–	✓	–	–	–	–
2017	Lin et al. (2017)	–	–	–	–	–	✓	–
2017	Berndorfer and Henriksson (2017)	✓	–	✓	–	✓(CBOW)	–	–
2018	Amoia et al. (2018)	–	–	✓	–	✓	–	–
2018	Catling et al. (2018)	✓	–	–	–	✓(skip-gram)	–	–
2018	Baumel et al. (2018)	✓	–	✓	–	✓(CBOW)	–	–
2018	Mullenbach et al. (2018)	–	–	✓	–	✓(CBOW)	–	–
2018	Samonte et al. (2018)	–	–	–	–	–	–	✓(Topical W_{embed})
2018	Xie and Xing (2018)	–	–	–	–	–	–	✓(200 dim W_{embed})
2018	Rios and Kavuluru (2018b)	–	–	–	–	–	–	✓(300 dim W_{embed})
2018	Kaur and Ginige (2018)	–	–	✓	–	–	–	–
2019	Zeng et al. (2019)	–	–	–	–	–	–	✓(100 dim W_{embed})
2019	Kaur and Ginige (2019)	✓	–	–	–	–	–	–
2019	Xie et al. (2019)	–	–	–	–	✓(CBOW)	–	–
2019	Falis et al. (2019)	–	–	–	–	✓(CBOW)	–	–
2019	Huang et al. (2019)	✓	–	–	–	✓(CBOW)	–	–
2019	Li et al. (2019)	–	–	–	✓	✓(skip-gram)	–	–
2019	Rios and Kavuluru (2019)	✓	–	–	–	–	–	✓(300 dim W_{embed})
2019	Schäfer and Friedrich (2019)	✓	–	✓	–	–	–	–
2019	Xu et al. (2019)	✓	–	–	–	–	–	✓(256 dim W_{embed})
2019	Du et al. (2019)	–	–	–	–	✓	–	–
2020	Cao et al. (2020)	–	–	–	–	✓	–	–
2020	Guo et al. (2020)	✓	–	–	–	✓(skip-gram)	–	–
2020	Vu et al. (2020)	–	–	–	–	✓(CBOW)	–	–
2020	Song et al. (2020)	–	–	–	–	–	–	✓(200 dim W_{embed})
2020	Sonabend W et al. (2020)	✓	–	–	–	–	–	✓(cui2vec)
2020	Mascio et al. (2020)	–	–	–	–	✓	✓	✓(FastText)
2020	Li and Yu (2020)	–	–	–	–	✓	–	–
2020	Teng et al. (2020)	–	–	–	–	–	–	✓(Knowledge graph)
2020	Moons et al. (2020)	–	–	–	–	–	–	✓(W_{embed})
2020	Ji et al. (2020)	–	–	–	–	✓(100 dim)	–	–
2020	Hsu et al. (2020)	✓	–	–	–	✓	–	–
2020	Zhang et al. (2020)	–	–	–	–	–	–	✓(EHR-BERT)
2021	Pascual et al. (2021)	–	–	–	–	–	–	✓(PubMedBERT)
2021	Ji et al. (2021)	–	–	–	–	✓(CBOW)	–	✓(BERT variants)
2021	Mayya et al. (2021)	–	–	–	–	✓(CBOW)	–	–
2021	Dong et al. (2021)	–	–	–	–	✓(CBOW)	–	✓(BERT)

TF-IDF: Term Frequency with inverse Document Frequency; BoW: Bag-of-words; CBOW: Continuous-bag-of-words; W_{embed}: Word embedding; dim: dimension; BERT: Bidirectional Encoder Representations from Transformers.

(GRUs) (Cho et al., 2014), and Bi-directional Recurrent Neural Networks (BRNN). Though CNN is known for image recognition or visual representation, there are a few studies that have used CNN in sentence classification (Kim, 2014). RNN is known for sentence classification with over sequential input and predict sequential output in NLP and other ML tasks. RNN has a bidirectional structure that incorporates both forward and backward inputs, but suffers from the vanishing gradient problem. To solve this problem, LSTM is used along with RNN (Samonte et al., 2017).

Over the past two decades, researchers explored various machine learning algorithms to assign ICD codes to clinical narratives (Berndorfer & Henriksson, 2017; Kaur & Ginige, 2018, 2019; Kavuluru et al., 2015; Perotte et al., 2013). Despite their research efforts, it is believed that the accuracy of assigning clinical codes can be further improved by deep learning approaches. However, the clinical text poses more challenges than general domain text due to various reasons:

- The free text clinical narratives contain high levels of noise, sparsity, complex medical vocabulary, misspelled words, abbreviations, use of non-standard clinical jargons and grammatical errors (Nguyen & Patrick, 2016).
- Many clinical corpora (or datasets) have imbalanced class distributions. For example, for an average patient, the chances of falling into the class "cancer positive" are less as compared the class "cancer negative". In many cases, a few positive cases in such a dataset are likely classified as a rare occurrence, or ignored because it causes more misclassifications compared to the majority class.
- In many clinical narratives, doctors and physicians prefer to use a variety of medical words or phrases. For example, instead of writing myocardial infarction, the phrase heart attack is used.
- Clinical documents have inconsistent document structure and organisation. Moreover, clinical documents are de-identified and anonymised to ensure patients' data privacy.

Table 10 shows machine learning and deep learning models that were employed for assigning ICD codes to discharge summaries. Notably, in several studies (Amoia et al., 2018; Ayyar & Oliver, 2016; Baumel et al., 2018; Berndorfer & Henriksson, 2017; Catling et al., 2018; Kaur & Ginige, 2018, 2019; Marafino et al., 2014; Moons et al., 2020; Perotte et al., 2013; Subotin & Davis, 2014; Xu et al., 2019) authors did not compare their proposed model with any existing study or algorithm; therefore, the third column value is left empty. A brief overview and comparison of studies is presented in Section 5.

4.5. Evaluation metrics

The performance of clinical text classification models can be measured using standard evaluation metrics which include precision, recall, F-measure (or F-score), accuracy, precision (micro and macro-average), recall (micro and macro-average), F-measure (micro and

Table 9
Confusion metric to compare the predicted value with ground truth value.

		Ground truth	
		Positive	Negative
Predicted	Positive	true positive (TP)	false negative (FN)
	Negative	false posiitve (FP)	true negative (TN)

macro-average), and area under the curve (AUC). These metrics can be computed by using values of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) in the standard confusion matrix as shown in Table 9.

1. *Precision (P)* is the ratio of correct instances retrieved to the total number of retrieved instances and is also known as positive predictive value (PPV).

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

2. *Recall (R)* is the number of correct instances retrieved divided by all correct instances and is also known as sensitivity or true positive rate (TPR).

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

3. *F-score* is weighted average of precision and recall.

$$F = \frac{2\,\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

4. *Accuracy* is the ratio of true instances retrieved to the total number of instances in the dataset.

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{4}$$

5. *Specificity* defines the proportion of negative instances that are correctly predicted as a negative.

$$\text{specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{5}$$

6. *Area Under the Curve (AUC)* measures the ability of a classifier to distinguish between classes. It plots the rate of true positive (TPR) against the rate of false positive (FPR) (Mujtaba et al., 2019).

$$\text{FPR} = 1 - \text{specificity} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{6}$$

The evaluation metrics for binary classification problems include precision, recall (sensitivity), F-score, accuracy, specificity, and AUC. For multi-class problems the performance can be evaluated by measuring micro or macro averaging of precision, recall, F-score, average accuracy, and error rate. Similarly, the performance of multi-label problems can be categorised into three measuring groups: example-based, label-based and ranking-based (Kaur & Ginige, 2019). The detailed description of other classification problems can be found in Gibaja and Ventura (2015), Sokolova and Lapalme (2009).

The instance of multi-label dataset (x_i, Y_i), $i = 1....m$, where $Y_i \subseteq L$ is the set of true labels and $l = (l_j : j = 1....q)$ is the set of all labels, with q the number of distinct labels. The macro-average and micro-average version of B, can be calculated as :

$$B_{\text{macro}} = \frac{1}{q}\sum_{l=1}^{q} B(\text{TP}_l, \text{FP}_l, \text{TN}_l, \text{FN}_l) \tag{7}$$

$$B_{\text{micro}} = B(\sum_{l=1}^{q}\text{TP}_l, \sum_{l=1}^{q}\text{FP}_l, \sum_{l=1}^{q}\text{TN}_l, \sum_{l=1}^{q}\text{FN}_l) \tag{8}$$

Macro-average metric computes the metric independently for each class and then takes the average. It is basically used when all classes need to be treated equally to evaluate the overall performance of the classifier. Micro-average metric aggregates the contribution of individual classes' TP, TN, FP, and FN. It is used to weight each instance or prediction equally.

$$\text{precision}_{\text{macro}} = \frac{\sum_{l=1}^{q}\text{TP}_l/(\text{TP}_l + \text{FP}_l)}{q} \tag{9}$$

$$\text{recall}_{\text{macro}} = \frac{\sum_{l=1}^{q}\text{TP}_l/(\text{TP}_l + \text{FN}_l)}{q} \tag{10}$$

$$F_{1\,\text{macro}} = \frac{2\,\text{precision}_{\text{macro}} \cdot \text{recall}_{\text{macro}}}{\text{precision}_{\text{macro}} + \text{recall}_{\text{macro}}} \tag{11}$$

$$\text{precision}_{\text{micro}} = \frac{\sum_{l=1}^{q}\text{TP}_l}{\sum_{l=1}^{q}(\text{TP}_l + \text{FP}_l)} \tag{12}$$

$$\text{recall}_{\text{micro}} = \frac{\sum_{l=1}^{q}\text{TP}_l}{\sum_{l=1}^{q}(\text{TP}_l + \text{FN}_l)} \tag{13}$$

$$F_{\text{micro}} = \frac{2\,\text{precision}_{\text{micro}} \cdot \text{recall}_{\text{micro}}}{\text{precision}_{\text{micro}} + \text{recall}_{\text{micro}}} \tag{14}$$

Table 11 shows a list of selected studies that used different evaluation metrics. The evaluation metrics employed in the study is presented with tick [✓] mark and the remaining are left empty. To summarise, the majority of studies have evaluated the performance of their model using micro-averaged F1-score, followed by macro-averaged F1-score, and standard F1-score.

5. Discussion

With clinical data mostly being stored in EHRs, applications of machine learning and deep learning models to predict clinical events and outcomes for clinical decision making have sparked widespread interest. Due to high granularity of ICD codes, researchers performed clinical code prediction either based on category-level or full-code prediction. The category-level prediction can also be referred to as group-level or chapter-level prediction in which a set of similar diseases and other underlying health conditions are represented in a unique chapter or a group. This type of problem is a multi-label classification problem where clinical reports can be mapped to more than one group. On the other hand, full code prediction is associated with a five to seven characters (numeric or alphanumeric) code depending upon the coding system being used such as ICD-9-CM, ICD-10-CM or ICD-10-AM. For example, in ICD-9-CM, the first three characters specify the disease category, while the latter two characters provide a more meticulous division of the disease.

Research studies related to automated ICD coding have used different methods and techniques ranging from pattern matching to deep learning approaches to categorise clinical narratives into different categories. In this review, we have considered the studies associated with automated ICD coding of discharge summaries that were published from January 1, 2010 on wards; therefore, this systematic literature review is limited to only classical machine learning and deep learning methods. The studies published prior to the year 2010 are basically focused on pattern matching, rule-based, machine learning or hybrid methods. Based on the close ended quality assessment questions, 42 studies were selected for systematic literature review.

Table 10
List of machine learning classifiers and deep learning models.

Year	Study	ML classifiers & DL models	Compared with other existing studies or algorithms
2013	Perotte et al. (2013)	Flat SVM, Hierarchy-based SVM	–
2014	Marafino et al. (2014)	SVM	–
2014	Subotin and Davis (2014)	Two-level hierarchical classification	–
2015	Kavuluru et al. (2015)	SVM, LR, MNB	BR, copy transformation, ECC
2016	Ayyar and Oliver (2016)	LSTM	–
2017	Prakash et al. (2017)	C-MemNN and A-MemNN	End-to-End Memory Network, KV-MemNNs
2017	Lin et al. (2017)	CNN	SVM, RF, GBM
2017	Berndorfer and Henriksson (2017)	Flat SVM and Hierarchical SVM	–
2018	Amoia et al. (2018)	LR and CNN	–
2018	Catling et al. (2018)	RNN-GRU	–
2018	Baumel et al. (2018)	SVM, CBOW, CNN, HA-GRU	–
2018	Mullenbach et al. (2018)	CAML, DR-CAML	CNN, LR, Bi-GRU, Flat SVM (Perotte et al., 2013), HA-GRU (Baumel et al., 2018), C-MemNN (Prakash et al., 2017; Scheurwegs et al., 2017; Shi et al., 2017)
2018	Samonte et al. (2018)	EnHAN	HAN
2018	Xie and Xing (2018)	Tree-of-sequences LSTM	HierNet (Yan et al., 2015), HybridNet (Hou et al., 2017), BranchNet (Zhu & Bain, 2017), LET (Bengio et al., 2010; Franz et al., 2000; Kavuluru et al., 2013, 2015; Larkey & Croft, 1996; Pestian et al., 2007)
2018	Rios and Kavuluru (2018b)	ZACNN, ZAGCNN	LR (Vani et al., 2017), CNN (Baumel et al., 2018), ACNN (Mullenbach et al., 2018), Match-CNN (Rios & Kavuluru, 2018a)
2018	Kaur and Ginige (2018)	SVM, NB, Decision Tree, kNN, RF, AdaBoost, and MLP	–
2019	Zeng et al. (2019)	Deep transfer learning using multi-scale CNN and Batch normalisation	Flat SVM (Perotte et al., 2013)
2019	Kaur and Ginige (2019)	Binary relevance, Label Power set, ML-kNN	–
2019	Xie et al. (2019)	MSATT-KG	LR, Selected Feature, Bi-GRU, Flat SVM and Hierarchy SVM, Text-CNN, DR-CAML, CAML, LEAM, C-MemNN, Attentive LSTM
2019	Falis et al. (2019)	Multi-view CNN (Ontological attention ensemble mechanism)	CAML (Mullenbach et al., 2018) MVC-LDA and MVC-RLDA (Sadoughi et al., 2018)
2019	Huang et al. (2019)	CNN, LSTM, GRU	Logistic regression, Random Forest, Feed-forward neural network, C-MemNN (Prakash et al., 2017)
2019	Li et al. (2019)	DeepLabeler (CNN and D2V)	Flat SVM and hierarchy-based SVM (Perotte et al., 2013)
2019	Rios and Kavuluru (2019)	CNNs (Transfer Learning)	Logistic regression, LR+L2R+NERC (Kavuluru et al., 2015), averaging ensemble CNNs without transfer learning
2019	Schäfer and Friedrich (2019)	FastText	Binary relevance SVM (Perotte et al., 2013), MT-CNN-net (Du et al., 2019), HA-GRU (Baumel et al., 2018), CAML and DR-CAML (Mullenbach et al., 2018)
2019	Xu et al. (2019)	Text-CNN	–
2019	Du et al. (2019)	ML-Net, ML-CNN, ML-HAN	SVM (Perotte et al., 2013)
2020	Cao et al. (2020)	HyperCore	SVM (Perotte et al., 2013), C-MemNN (Prakash et al., 2017), C-LSTM-ATT (Shi et al., 2017), HA-GRU (Baumel et al., 2018), CAML and DR-CAML (Mullenbach et al., 2018)
2020	Guo et al. (2020)	BiLSTMs	DeepLabeler (Li et al., 2019)
2020	Vu et al. (2020)	LAAT and JointLAAT	LR (Mullenbach et al., 2018), SVM (Perotte et al., 2013), CNN (Mullenbach et al., 2018), Bi-GRU (Mullenbach et al., 2018), C-MemNN (Prakash et al., 2017), C-LSTM-Att (Shi et al., 2017), HA-GRU (Baumel et al., 2018), LEAM (Wang et al., 2018), CAML (Mullenbach et al., 2018), DR-CAML (Mullenbach et al., 2018), MSATT-KG (Xie et al., 2019), MultiResCNN (Li & Yu, 2020)
2020	Song et al. (2020)	ZAGRNN, ZAGRNN with LDAM loss	ZAGCNN (Rios & Kavuluru, 2018b), (Liu et al., 2019), (Xian et al., 2018), (Felix et al., 2018)

(*continued on next page*)

Table 10 (*continued*).

Year	Study	ML classifiers & DL models	Compared with other existing studies or algorithms
2020	Sonabend W et al. (2020)	UNITE	LR, Topic modelling, MLP
2020	Mascio et al. (2020)	ANN, CNN, Bi-LSTM	SVM
2020	Li and Yu (2020)	MultiResCNN	SVM (Flat and Hierarchy) (Perotte et al., 2013), CAML & DR-CAML (Mullenbach et al., 2018), HA-GRU (Baumel et al., 2018), C-MemNN (Prakash et al., 2017), C-LSTM-Att (Shi et al., 2017)
2020	Teng et al. (2020)	G_Coder (Multi-CNN, Graph Presentation, Attention Matching, Adversarial Learning)	C-LSTM-Att (Shi et al., 2017), CAML & DR-CAML (Mullenbach et al., 2018), MultiResCNN (Li & Yu, 2020)
2020	Moons et al. (2020)	CNN, Bi-GRU, DR-CAML, MVC-LDA, MVC-RLDA	–
2020	Ji et al. (2020)	DCAN	CNN (Kim, 2014), C-MemNN (Prakash et al., 2017), Attentive LSTM (Shi et al., 2017), Bi-GRU (Mullenbach et al., 2018), CAML & DR-CAML (Mullenbach et al., 2018), LEAM (Wang et al., 2018), MultiResCNN (Li & Yu, 2020)
2020	Hsu et al. (2020)	CNN, LSTM, GRU, HAN	SVM
2020	Zhang et al. (2020)	AttentionXML (BERT-XML)	LR, Multi-head Attention , BERT, BioBERT, ClinicalBERT
2021	Pascual et al. (2021)	BERT-ICD	C-MemNN (Prakash et al., 2017), LEAM (Wang et al., 2018), CAML (Mullenbach et al., 2018), DR-CAML (Mullenbach et al., 2018), MSATT-KG (Xie et al., 2019), Label Attention (Vu et al., 2020)
2021	Mayya et al. (2021)	Enhanced-CAML (EnCAML)	EnHAN (Samonte et al., 2018), GRU (Huang et al., 2019), Transfer learning (Rios & Kavuluru, 2019), CAML (Mullenbach et al., 2018), DR-CAML (Mullenbach et al., 2018), HA-GRU (Baumel et al., 2018), MultiResCNN (Li & Yu, 2020), BiLSTMs (Guo et al., 2020), Transfer learning (Zeng et al., 2019), DeepLabeler (Li et al., 2019)
2021	Ji et al. (2021)	CNN (retrained)	domain-specific BERT variants (BERT-base, BlueBERT, BioBERT, BioRedditBERT, PubMedBERT, SapBERT, ClinicalBERT), CNN (Kim, 2014), CAML (Mullenbach et al., 2018), MultiResCNN (Li & Yu, 2020), HyperCore (Cao et al., 2020)
2021	Dong et al. (2021)	HLAN	CNN (Gehrmann et al., 2018; Karimi et al., 2017), CAML (Mullenbach et al., 2018), Bi-GRU, HAN (Yang et al., 2016), HA-GRU (Baumel et al., 2018)

SVM: Support Vector Machine; RF: Random Forest; LR: Logistic Regression; NB: Naïve Bayes; MLP: Multi Layer Perceptron; BR: Binary Relevance; ECC: Ensemble of classifier chains; MNB: Multinomial Naïve Bayes; GBM: Gradient Boosting Machine; KV-MemNNs: Key-Value Memory Networks; C-MemNN: Condensed Memory Networks; A-MemNN: Averaged Memory Networks; CNN: Convolutional Neural Networks; RNN: Recurrent Neural Network; GRU: Gated Recurrent Units; HAN: Hierarchical Attention Network; LSTM: Long Short-Term Memory Networks; MSATT-KG: Multi-scale feature attention and structure knowledge graph propagation; HA-GRU: Hierarchical Attention Gated Recurrent Unit; Conv-LSTM: Convolutional LSTM; EnHANs: Enhanced Hierarchical Attention Networks; CAML: Convolutional Attention for Multi-Label classification; DR-CAML: Description Regularized CAML; DCAN: Dilated Convolutional Attention Network; Seg-GRU: Segment-level GRU; MultiResCNN: Multi-Filter Residual Convolutional Neural Network; C-LSTM-Att: Character-aware LSTM-based Attention; TAGs: Term weighting AGgregated using fuzzy Similarity; EnTAGs: Enhanced TAGS; ZAGRNN: Zero-shot Attentive Graph Recurrent Neural Networks; ZAGCNN: Zero-shot Attentive Graph Convolution Neural Networks; UNITE: Unsupervised knowledge integration algorithm; EnCAML: Enhanced Convolutional Attention network; HLAN: Hierarchical Label-wise Attention Network.

Table 11
Studies using different evaluation metrics.

Study	P	R	F1	Acc	AUC	HL	JS	P_{mi}	R_{mi}	$F1_{mi}$	AUC_{mi}	P_{ma}	R_{ma}	$F1_{ma}$	AUC_{ma}	$F1_{diag}$	$F1_{proc}$	P@n	R@n	Sen	Spec	MRR
Perotte et al. (2013)	✓	✓	✓	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Marafino et al. (2014)	✓	✓	✓	✓	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Subotin and Davis (2014)	✓	✓	✓	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	✓
Kavuluru et al. (2015)	–	–	–	–	–	–	–	✓	✓	✓	–	–	–	–	–	–	–	–	–	–	–	–
Ayyar and Oliver (2016)	✓	✓	✓	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Prakash et al. (2017)	–	–	–	–	–	✓	–	–	–	–	–	–	–	–	✓	–	–	✓	–	–	–	–
Lin et al. (2017)	–	–	✓	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Berndorfer and Henriksson (2017)	✓	✓	✓	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Amoia et al. (2018)	–	–	–	–	–	–	–	✓	✓	✓	–	–	–	–	–	–	–	–	–	–	–	–
Catling et al. (2018)	✓	✓	✓	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Baumel et al. (2018)	–	–	–	–	–	–	–	–	–	✓	–	–	–	–	–	–	–	–	–	–	–	–
Mullenbach et al. (2018)	–	–	–	–	–	–	–	–	–	✓	✓	–	–	✓	✓	✓	✓	✓	–	–	–	–
Samonte et al. (2018)	✓	✓	✓	✓	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Xie and Xing (2018)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	✓	✓	–
Rios and Kavuluru (2018b)	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	✓	–	–	–
Kaur and Ginige (2018)	✓	✓	✓	✓	–	✓	✓	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Zeng et al. (2019)	–	–	–	–	–	–	–	✓	✓	✓	–	–	–	–	–	–	–	–	–	–	–	–
Kaur and Ginige (2019)	–	–	–	–	–	✓	✓	✓	✓	✓	–	✓	✓	✓	–	–	–	–	–	–	–	–
Xie et al. (2019)	–	–	–	–	–	–	–	–	–	✓	✓	–	–	✓	✓	✓	✓	✓	–	–	–	–
Falis et al. (2019)	–	–	–	–	–	–	–	✓	✓	✓	–	✓	✓	✓	–	–	–	✓	–	–	–	–
Huang et al. (2019)	✓	✓	✓	✓	–	✓	–	–	–	–	–	–	–	–	✓	–	–	✓	–	–	–	–
Li et al. (2019)	–	–	–	–	–	–	–	✓	✓	✓	–	–	–	–	–	–	–	–	–	–	–	–
Rios and Kavuluru (2019)	–	–	–	–	–	–	–	–	–	✓	–	–	–	✓	–	–	–	–	–	–	–	–
Schäfer and Friedrich (2019)	✓	✓	✓	–	–	–	–	–	–	–	–	–	–	–	–	–	–	✓	–	–	–	–
Xu et al. (2019)	–	–	–	–	–	–	✓	–	–	✓	✓	–	–	✓	✓	–	–	–	–	–	–	–
Du et al. (2019)	✓	✓	✓	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Cao et al. (2020)	–	–	–	–	–	–	–	–	–	✓	✓	–	–	✓	✓	–	–	✓	–	–	–	–
Guo et al. (2020)	–	–	–	–	–	–	–	✓	✓	✓	✓	✓	✓	✓	✓	–	–	–	–	–	–	–
Vu et al. (2020)	–	–	–	–	–	–	–	–	–	✓	✓	–	–	✓	✓	–	–	✓	–	–	–	–
Song et al. (2020)	–	–	–	–	–	–	–	✓	✓	✓	✓	✓	✓	✓	✓	–	–	–	–	–	–	–
Sonabend W et al. (2020)	–	–	✓	–	✓	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Mascio et al. (2020)	–	–	✓avg	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Li and Yu (2020)	–	–	–	–	–	–	–	–	–	✓	✓	–	–	✓	✓	–	–	✓	–	–	–	–
Teng et al. (2020)	–	–	–	–	–	–	–	–	–	✓	✓	–	–	–	–	–	–	✓	–	–	–	–
Moons et al. (2020)	–	–	–	–	–	–	–	–	–	✓	✓	–	–	✓	–	✓micro	✓micro	✓	–	–	–	–
Ji et al. (2020)	–	–	–	–	–	–	–	–	–	✓	✓	–	–	✓	✓	–	–	✓	–	–	–	–
Hsu et al. (2020)	–	–	–	–	–	–	–	–	–	✓	–	–	–	–	–	–	–	–	–	–	–	–
Zhang et al. (2020)	–	–	–	–	–	–	–	–	–	–	✓	–	–	–	✓	–	–	–	–	–	–	–
Pascual et al. (2021)	–	–	–	–	–	–	–	–	–	–	✓	–	–	–	✓	–	–	–	–	–	–	–
Ji et al. (2021)	–	–	–	–	–	–	–	–	–	✓	✓	–	–	✓	✓	–	–	✓	–	–	–	–
Mayya et al. (2021)	–	–	–	–	–	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	–	–	–	–	–	–	–
Dong et al. (2021)	–	–	–	–	–	–	–	✓	✓	✓	✓	✓	✓	✓	✓	–	–	✓	–	–	–	–

P: Precision, R: Recall, F1: F1-score, Acc: Accuracy, AUC: Area under the ROC curve, HL: Hamming Loss, JS: Jaccard Similarity, P_{mi}: Micro-averaged Precision, R_{mi}: Micro-averaged Recall, $F1_{mi}$: Micro-averaged F1-score, AUC_{mi}: Micro-averaged F1 score of Area under the ROC curve (AUC), P_{ma}: Macro-averaged Precision, R_{ma}: Macro-averaged Recall, $F1_{ma}$: Micro-averaged F1-score, AUC_{ma}: Macro-averaged F1 score of Area under the ROC curve (AUC), $F1_{diag}$: Micro-F1 on diagnosis codes, $F1_{proc}$: Micro-F1 on procedure codes, P@n: Precision @ n, R@n: Recall @ n, Sen: Sensitivity, Spec: Specificity, MRR: Mean Reciprocal Rank.

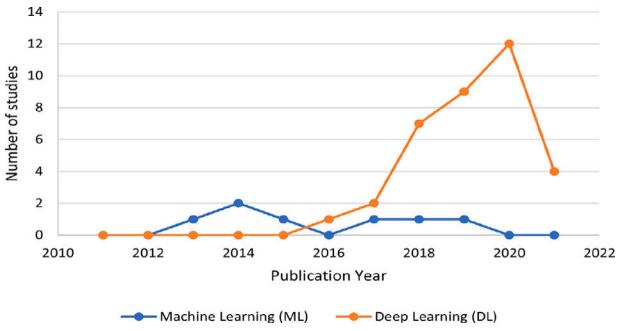

Fig. 4. The distribution of machine learning and deep learning approaches through the period between 2010 and 2020.

Our findings indicate that there is a lack of publicly available data sources. The majority of the studies have used MIMIC-III dataset which is coded using ICD-9 coding standard and is no longer being used for classification purpose since the adoption of ICD's 10th revision. For real-world deployment, clinical codes should be assigned according to current coding standards. In order to use older data, clinical codes should be mapped with the current coding standard. We found that only one study (Xu et al., 2019) has mapped only 32 ICD-9-CM codes to ICD-10-CM. In addition, studies (Falis et al., 2019; Searle et al., 2020) inspected few limitations in the MIMIC-III dataset such as under-coding for specific conditions. We also analysed that due to data imbalance and rare occurrence of diseases in the dataset, studies have reduced the complexity of the codes by predicting the 50 or 100 most common codes.

Our findings also indicate that the majority of studies have pre-processed the discharge summaries using NLP techniques such as tokenization, removal of non-alphabetical characters, stopwords, and lowercase conversion. Out of 42 studies, 11 studies have truncated the long discharge summaries to 2500 or 4000 tokens in order to reduce the computational cost. In terms of feature extraction and pre-trained embeddings, 20 studies have used the word2vec model. Recent studies (Dong et al., 2021; Ji et al., 2021; Pascual et al., 2021; Zhang et al., 2020) show that there is a rising trend towards pre-trained models such as BERT and its variants. Pre-trained models have shown effectiveness in capturing contextual information.

The findings of this review also indicate that deep learning has dominated over the classical machine learning techniques as shown in Fig. 4. The majority of the studies applied deep learning approaches and achieved notable improvements. The main benefit of deep learning is that features are learned automatically from training data rather than being engineered by human experts. However, training deep learning models requires more data, a lot of computational power and have more hyper-parameters than the classical machine learning model.

The "no free lunch" theorem (Wolpert, 1996) states that no single machine learning algorithm performs best in all application areas. If one algorithm performs well with one dataset, it may not do so on another dataset. Though, majority of the studies have used MIMIC-II and MIMIC-III dataset, but the complexity of a problem and experimental settings are different. Thus, statistically comparing the performance of algorithms is infeasible. We analysed that deep learning outperform uniformly compared to classical machine learning. We compared results of the studies where authors compared machine learning and deep learning methods with similar data settings. Based on the results in studies, we find that deep learning algorithms such as HyperCore (Cao et al., 2020), LAAT/JointLAAT (Vu et al., 2020), MSATT-KG (Xie et al., 2019), MultiResCNN (Li & Yu, 2020) consistently outperformed machine learning methods by providing higher scores in terms of standard evaluation metrics. Deep learning algorithms specialised to model sequential data in the form of text such as recurrent neural networks and its advanced variants such as LSTMs and GRUs are more common methods used in the studies. However, studies have also applied CNN by treating text as a 1D sequence. Given LSTMs and GRUs have limited long range dependencies, studies incorporated attention mechanism to further improve the results. Overall the trend is towards applying large-scale pre-trained models such as Transformers and BERT for assigning clinical codes. Table 12 gives a summary of selected studies in terms of their objective, salient finding(s) and limitations found in their proposed methods.

6. Comparison of studies

This section highlights the comparison of 42 selected studies that consider automated ICD coding of discharge summaries. It is difficult to compare different models and find out the best one as each model relies on different parameters and conditions such as type of dataset, data distribution, pre-processing pipeline and evaluation metrics. Therefore, we have compared the studies based on three key aspects: type of dataset, data distribution (train–test split) size and number of clinical codes as given in the Tables 13–17.

Among all the selected studies, the majority of them have experimented with the MIMIC-II and MIMIC-III datasets as they are publicly available. Out of 42 selected studies, 3 studies (Du et al., 2019; Marafino et al., 2014; Perotte et al., 2013) have used the MIMIC-II dataset only, 19 studies (Ayyar & Oliver, 2016; Berndorfer & Henriksson, 2017; Catling et al., 2018; Dong et al., 2021; Falis et al., 2019; Guo et al., 2020; Hsu et al., 2020; Huang et al., 2019; Ji et al., 2020, 2021; Pascual et al., 2021; Prakash et al., 2017; Samonte et al., 2018; Schäfer & Friedrich, 2019; Song et al., 2020; Teng et al., 2020; Xie & Xing, 2018; Xie et al., 2019; Xu et al., 2019) have used the MIMIC-III data only, 7 studies (Baumel et al., 2018; Cao et al., 2020; Li et al., 2019; Li & Yu, 2020; Mullenbach et al., 2018; Rios & Kavuluru, 2018b; Vu et al., 2020) have used both MIMIC-II and MIMIC-III dataset, 5 studies (Mascio et al., 2020; Mayya et al., 2021; Moons et al., 2020; Sonabend W et al., 2020; Zeng et al., 2019) have used MIMIC-III along with other dataset, 6 studies (Amoia et al., 2018; Kaur & Ginige, 2018, 2019; Kavuluru et al., 2015; Lin et al., 2017; Rios & Kavuluru, 2019; Subotin & Davis, 2014; Zhang et al., 2020) used private dataset. Also, a few studies (Cao et al., 2020; Guo et al., 2020; Li & Yu, 2020; Moons et al., 2020; Mullenbach et al., 2018; Prakash et al., 2017; Vu et al., 2020; Xie et al., 2019) have divided MIMIC-III dataset into two common settings, MIMIC-III full and MIMIC-III 50 (or MIMIC-III 100), where MIMIC-III full contains a complete set of ICD codes for 52,722 discharge summaries, MIMIC-III 50 contains the top 50 most frequent codes, and MIMIC-III 100 contains 100 most frequent codes for discharge summaries. However, there might be a slight variation in the number of codes when the MIMIC-III full data are used in the studies, but the process of assigning the codes is same.

Table 12
Summary of studies in terms of salient findings and limitations of their proposed method (s).

Study	Objective	Main finding(s)	Limitation(s)
Perotte et al. (2013)	To explore the traditional machine learning algorithm flat and hierarchical SVM for assigning ICD-9-CM codes to discharge summaries.	The hierarchical SVM outperforms the flat SVM.	Error analysis reveals that gold standard codes are not perfect, and as such the recall and precision are likely underestimated.
Marafino et al. (2014)	To develop a SVM classifier capable of identifying a range of procedures and diagnose in ICU clinical notes for use in risk adjustment.	A classifier using n-gram support vectors improved performance.	Only two diagnoses and two procedures were addressed. Possibility of incorrect measurement of accuracy because of classification error in the assignment of gold standards.
Subotin and Davis (2014)	Design a system for predicting ICD-10-PCS codes from the clinical narratives using several levels of abstraction.	Applied a set of classifiers to identify concepts that appear in EHRs, used General Equivalence Mappings (GEMs) between ICD-9 and ICD-10 codes and estimates probability of ICD-10 codes.	The dataset is not publicly available so it is difficult to make direct comparison with existing studies.
Kavuluru et al. (2015)	To evaluate supervised learning approaches to automatically assign ICD-9-CM diagnosis codes to EMRs.	For shorter narratives, classifier chaining is ideal. For in-patient full EMRs, feature and data selection methods offer high performance for smaller datasets. For large EMR datasets, binary relevance approach with learning-to-rank based code re-ranking offers the best performance.	
Ayyar and Oliver (2016)	To propose a deep learning framework to classify the ICD-9-CM codes to patients' discharge summaries.	Mapped all codes to top level representation (3 digits only), which left 19 top level ICD-9-CM codes. The learning task can be improved by using medical dictionaries to obtain more pertinent word vectors.	Several misspellings in the dataset leads to miss word vector representations. Model need to be able to accurately learn more important words and hold them in memory for longer periods of time.
Prakash et al. (2017)	Introduce condensed memory neural networks (C-MemNNs), an approach to efficiently store condensed representations in memory.	Results improves as the number of hops increases. C-MemNN with five memory hops achieve higher AUC value in comparison to other memory networks.	Predicts only the 50 most-common and the 100 most-common diagnoses codes. The clinical notes and wiki-pages were truncated to 600 words. Training multiple hops with condensed representation gives better results but is computationally expensive.
Lin et al. (2017)	To compare the performance of traditional pipelines (NLP and supervised machine learning models) with word embedding combined with a CNN to identify ICD-10-CM diagnosis codes in discharge notes.	Word embedding combined with a CNN gave higher accuracy than traditional NLP-based approach. Training AUCs of traditional methods were very close to 1 which means that there was no possibility of improvement.	Contain discharge notes from only a single hospital so it cannot confirm how well it would generalise to other data sources. Discharge notes describe only the presence of the disease, but do not include negative statements. All ICD-10-CM codes are truncated to 1-character level only.
Berndorfer and Henriksson (2017)	To explore various text representations and classification models for assigning ICD-9-CM codes to discharge summaries in MIMIC-III	The predictive performance strongly decreases with a lower code frequency. The shallow BoW model performs better on high-frequent codes. For medium and low-frequent codes, the deep W2V representation outperforms BoW. The best performance is obtained by combining models using different representations.	
Amoia et al. (2018)	To build a machine learning based scalable system for predicting ICD-10-CM codes from electronic health records.	Address data imbalance issues by implementing two system architectures using convolutional neural networks and logistic regression models. CNN is more suited for high frequency codes whereas the logistic regression system does better for low frequency codes.	Due to code sparsity issue, the study is restricted to the most frequent codes seen within the first seven months of data.
Catling et al. (2018)	To explore different methods for representing clinical text and labels in hierarchical clinical coding ontologies.	TF-IDF outperformed recurrent neural networks for predicting the disease labels. Learning good representation of the very rare diseases remains a major challenge.	Extract a particular section (i.e. the history of present illness (HoPI)) from each discharge summary, predicts ICD-9-CM codes associated with them and omit lists of other diagnoses. Discharge summaries that do not include HoPI section or contains empty HoPI section were removed and each HoPI document was truncated to 500 tokens.
Baumel et al. (2018)	To investigate four models for assigning multiple ICD-9-CM codes to discharge summaries.	Mapping rare variants using edit distance improved results for CBOW and CNN. HA-GRU gives better results in comparison to other models in the rolled-up ICD-9-CM codes.	
Mullenbach et al. (2018)	To develop convolutional neural network based methods for ICD-9-CM code assignment based on discharge summaries.	CAML model aggregates information across the document using a convolutional neural network, and uses an attention mechanism to select the most relevant segments. The CAML model gives strong results in comparison to other state-of-the-art models.	All discharge summaries are truncated to a maximum length of 2500 tokens because it is difficult to train the model with long sequences, but in the real-world settings it is not an acceptable solution.

(*continued on next page*)

Table 12 (*continued*).

Study	Objective	Main finding(s)	Limitation(s)
Samonte et al. (2018)	To develop a model for automatic multi-class labelling of ICD-9 codes of patient notes	Modified the architecture of hierarchical attention networks (HAN) by adding topical word embedding and a word input. The enhanced hierarchical attention networks (EnHANs) slightly outperforms the baseline model (HAN).	The study considered only top 10 and top 19 ICD-9-CM diagnosis codes.
Xie and Xing (2018)	To design a neural network architecture for automatically perform ICD coding given the diagnosis description	Proposed a tree-of-sequences LSTM architecture to capture the hierarchical relationship among codes and the semantics of each code. Adversarial learning approach improves the matching accuracy by alleviating the discrepancy among the writing styles of diagnosis description. Used an attention mechanism to perform many-to-one and one-to-many mappings between diagnosis descriptions and codes.	Does not perform well on infrequent codes. Less capable to deal with abbreviations.
Rios and Kavuluru (2018b)	To perform a fine-grained evaluation and understand state-of-the-art methods performance on infrequent labels.	The study uses neural architecture that incorporates label descriptors and the hierarchical structure of the label spaces for few and zero-shot prediction.	
Kaur and Ginige (2018)	To perform comparative analysis of pattern matching, rule-based and machine learning approaches to assign ICD-10-AM and ACHI codes to discharge summaries.	Compared the performance of three different approaches: pattern matching, rule-based and machine learning. Decision Tree outperforms all other classifiers on original number of records. The increase in dataset size, the performance of models increased.	Assign clinical codes associated with respiratory and gastrointestinal diseases and interventions only. Increased dataset size by adding 45 discharge summaries by manually mixing and matching certain diagnoses and interventions.
Zeng et al. (2019)	To propose an end-to-end deep transfer learning method to transfer knowledge learned from MeSH source domain into ICD-9-CM codes domain.	Transfer learning improves the performance of ICD-9-CM coding model. Also, multi-scale CNNs is effective in capturing contextual features giving higher results.	
Kaur and Ginige (2019)	To assign ICD-10-AM and ACHI codes to discharge summaries using problem transformation methods and adaptive algorithm.	The results were evaluated using 3 different test ratios. The performance of Binary Relevance and Label Power-set were slightly to moderately improved.	Due to the data scarcity, the dataset was repeated with some minor changes to increase the model performance. More than half of the clinical codes appeared only once in the whole dataset which lowers the learning rate.
Xie et al. (2019)	To predict diagnosis and procedure codes for electronic health record coding with multi-scale feature attention and structured knowledge graph propagation.	The study improved the convolutional attention model by using densely connected CNN and multi-scale feature attention on MIMIC-III dataset. Also, used a graph convolutional neural network to capture the hierarchical relationships among medical codes and the semantics of each code.	
Huang et al. (2019)	To perform an empirical evaluation of deep learning based systems to assign ICD-9-CM codes.	Deep learning algorithms (CNN, LSTM, GRU) outperformed the traditional machine learning algorithms (logistic regression, random forest, feed forward neural networks).	Predicted only top 10 and top 50 common ICD-9-CM diagnosis codes and categories. For top 50 common codes and categories, the model fails to effectively distinguish between 50 different labels due to imbalanced data samples.
Li et al. (2019)	Proposed a deep learning framework called DeepLabeler that combines CNN with document to vector techniques to assign ICD-9-CM codes	DeepLabeler achieved about 15% increase in micro F-measure over flat SVM and hierarchy-based SVM. The model can extract effective representative features and need to train an end-to-end model rather thousands of binary classifiers.	
Rios and Kavuluru (2019)	To study the effect of neural transfer learning for assigning diagnosis codes to electronic medical records (EMRs).	Transfer learning can improve convolution neural networks performance for EMR coding in the presence of data sparsity issues. Improves infrequent label performance by 5%.	Focused on predicting diagnosis codes only rather than considering both diagnosis and procedure codes. Truncated all ICD-9-CM codes to 4-character level (used abc.x instead of abc.xy) and removed all codes than appeared in less than 50 EMRs.
Schäfer and Friedrich (2019)	Introduces support vector machine and FastText with basic UMLS concept mappings into word embedding models to assign ICD code assignment.	FastText achieves high recall results and performance improvement when evaluated with different label count estimation approaches on a hierarchy extended model.	Considered diagnosis codes only.
Xu et al. (2019)	To present a multimodal machine learning model to ICD-10-CM diagnostic codes.	Performed one-to-one mapping of 32 ICD-9-CM codes to ICD-10-CM. Developed an ensemble-based approach that integrates three modality-specific models for improving prediction accuracy.	Focused on predicting 32 ICD-10-CM codes only. Feature dimensions for tabular data are very large and some features are duplicate.

(*continued on next page*)

Table 12 (*continued*).

Study	Objective	Main finding(s)	Limitation(s)
Du et al. (2019)	To propose a deep learning framework for multi-label classification of biomedical texts (ML-Net).	ML-Net combines label prediction and label decision in the same network and determine the output labels based on label confidence scores and document context.	Due to limited computational resources, only 1500 tokens from each clinical notes are taken as input and a thorough hyper-parameter tuning is not performed. The label count prediction network does not work well for labels with a hierarchical structure. The document encoding network can be improved by using an advanced language representation models.
Cao et al. (2020)	To propose a hyperbolic and co-graph representation method (HyperCore) which can jointly exploit code hierarchy and code co-occurrence.	A hyperbolic representation method leverage the code hierarchy in the hyperbolic space. Uses the graph convolutional network (GCN) to capture the co-occurrence correlation. HyperCore outperforms the baselines (Baumel et al., 2018; Mullenbach et al., 2018; Perotte et al., 2013; Prakash et al., 2017; Shi et al., 2017) on MIMIC-III dataset.	Truncated discharge summaries to 2500 tokens to overcome long sequence issues.
Guo et al. (2020)	To propose a disease inference method by extracting symptoms and integrating two symptom representation approaches.	Combines the BiLSTMs model with TF-IDF and Word2Vec for disease inference and extract symptom concepts using MetaMap tool. The combination of two symptom representations improves the performance of disease inference task.	Considers only first three characters (or category-level) of all diagnosis codes that occurs in 50 most-common and 100 most-common diseases.
Vu et al. (2020)	To propose a label attention model which can handle the various lengths as well as the interdependence between text fragments related to ICD codes.	A hierarchical joint learning mechanism (JoinLAAT) improve the performance for infrequent codes, resulting in higher macro-averaged metrics.	
Song et al. (2020)	To propose a latent feature generation framework to improve the prediction on unseen codes without compromising the performance in seen codes.	An adversarial generative model with hierarchical tree structure (AGM-HT) exploits the hierarchical structure of ICD codes to generate semantically meaningful features without any labelled data.	Used only ICD-9-CM diagnosis codes in the study. Documents truncated to 2500 tokens (data pre-processing settings as in Mullenbach et al. (2018))
Sonabend W et al. (2020)	To assign ICD-9-CM codes via unsupervised knowledge integration	Analysed clinical narratives via semantic relevance assessment. Had substantially better performance than logistic regression, topic models and multi-layer perceptron.	Performed automated ICD coding of 6 diagnoses only. The model relies only on concept unique identifier (CUI) and not word embeddings. Also, it assigns only one ICD code at a time rather than assigning multiple codes.
Mascio et al. (2020)	To analyse the impact of numerous word representation methods and classification approaches for four different text classification tasks.	Highlights the efficacy of contextual embeddings compared to traditional methods. Bi-LSTM outperforms contextual embeddings when combined with entity extraction task and specific domain embeddings.	Do not assign any ICD codes, only check the status and temporality of the disease.
Li and Yu (2020)	To propose a multi-filter residual convolution neural network (MultiResCNN) for assigning ICD-9-CM code to discharge summaries.	MultiResCNN combines a multi-filter convolutional layer to capture text patterns with different lengths and a residual convolutional layer to enlarge the receptive field. MultiResCNN improves the convolution attention model proposed by Mullenbach et al. (2018).	Computational cost is high. During preliminary experiments, BERT model did not perform well due to hardware limitations and its fixed-length context.
Teng et al. (2020)	To propose an end-to-end method, graph based coder (G_Coder) for ICD code assignment.	Utilises Multi-CNN to capture local correlation, a knowledge graph combined with attention mechanism to understand the meaning of terminologies and making the coding results interpretable, and adversarial learning to generate adversarial samples. The model outperforms baselines (Li & Yu, 2020; Mullenbach et al., 2018; Shi et al., 2017) for predicting 50 most frequent codes in the dataset.	Only top 50 most common ICD-9-CM codes are considered as the model does not perform well on infrequent codes. For fully automatic coding, infrequent coding has to be considered.
Zhang et al. (2020)	To propose a machine learning model, BERT-XML for large scale automated ICD coding of EHR notes.	BERT-XML combines BERT pretraining with multi-label attention and outperforms other baselines without self-supervised pretraining.	Computational expensive and long training time of 3 weeks. Increase training and inference time when document length increases.
Pascual et al. (2021)	To investigate BERT-based ICD coding model for biomedical language understanding.	Fine-tuning the encoder significantly improves the decoding performance and obtained the best performance after six epochs. Decoder cannot compensate the lack of convergence during the fine-tuning of the encoder.	A substantial gap between the label attention (Vu et al., 2020) and BERT-ICD model. Difficult to fine-tune a BERT encoder on long pieces of text.
Ji et al. (2021)	To conduct a comprehensive quantitative analysis of various contextualised language models' performance used for medical code assignment.	Compared different domain-specific BERT variants. Simple CNN trained from scratch can achieve superior predictive performance.	
Mayya et al. (2021)	To propose the enhanced convolutional attention network for multi-label classification.	Employs multi-channel, variable-sized convolution filters and multiple attention layers. EnCAML model outperforms the state-of-the-art models.	To reduce the computational complexity and cost of training, certain sections were removed from the discharge summaries.
Dong et al. (2021)	To propose a hierarchical label-wise attention network (HLAN) and address the issues of modal explainability and label correlation.	HLAN has more comprehensive explainability and better than CNN-based approaches. The label embedding initialisation effectively boosted the performance by capturing label correlations.	

Table 13
Comparison of studies using MIMIC-II dataset, different algorithms and evaluation metrics.

Train–test split	Study	Algorithms	Performance
D_{Total}: 22,815, D_{Train}: 20,533, D_{Test}: 2,282, Codes: 5031	Perotte et al. (2013)	Flat SVM	P: 86.7%; R: 16%; F-measure: 27.6%
		Hierarchy-based SVM	P: 57.7%, R: 30.0%, F-measure: 39.5%
	Mullenbach et al. (2018)	CAML	AUC_{mi}: 0.966; AUC_{ma}: 0.820; $F1_{ma}$: 0.048; $F1_{mi}$: 0.442; P@8:0.523
		DR-CAML	AUC_{mi}: 0.966; AUC_{ma}: 0.826; $F1_{ma}$: 0.049; $F1_{mi}$: 0.457; P@8: 0.515
	Li et al. (2019)	DeepLabeler (CNN+D2V)	P_{mi}: 0.475; R_{mi}: 0.258; $F\text{-}meas_{mi}$: 0.335
	Cao et al. (2020)	HyperCore	AUC_{mi}: 0.971; AUC_{ma}: 0.885; $F1_{ma}$: 0.070; $F1_{mi}$: 0.477; P@8:0.537
	Vu et al. (2020)	LAAT	AUC_{mi}: 0.973; AUC_{ma}: 0.868; $F1_{ma}$: 0.059; $F1_{mi}$: 0.486; P@5: 0.649; P@8: 0.550; P@15: 0.397
		JointLAAT	AUC_{mi}: 0.972; AUC_{ma}: 0.871; $F1_{ma}$: 0.068; $F1_{mi}$: 0.491; P@5: 0.652; P@8: 0.551; P@15: 0.396
	Li and Yu (2020)	MultiResCNN	AUC_{mi}: 0.968 ± 0.001; AUC_{ma}: 0.850 ± 0.002; $F1_{ma}$: 0.052 ± 0.002; $F1_{mi}$: 0.464 ± 0.002; P@8: 0.544 ± 0.007
D_{Total}: 4191 Neonatal ICU, D_{Total}: 2198 Adult ICU	Marafino et al. (2014)	SVM	P: 0.982; R: 0.952; F1: 0.954; Acc: 0.982 (Ventilation classification task). The study focused on only two diagnoses and two procedures.
D_{Total}: 22,815, D_{Train}: 20,533, D_{Test}: 2,282, Codes: 6,527	Baumel et al. (2018)	SVM	F_{mi}: 28.13%; F_{mi}: 32.50% (Rolled-up codes)
		CBOW	F_{mi}: 30.60%; F_{mi}: 42.06% (Rolled-up codes)
		CNN	F_{mi}: 33.25%; F_{mi}: 46.40% (Rolled-up codes)
		HA-GRU	F_{mi}: 36.60%; F_{mi}: 53.86% (Rolled-up codes)
D_{Train}: 18,822, D_{Test}: 1,711, Codes: 3,118 (Group S: all codes occur > 5 times); 3,459 (Group F: codes between 1 to 5 times); 355 (Group Z: never occur)	Rios and Kavuluru (2018b)	ZACNN	Group S = R@5: 0.135; R@10: 0.247
		ZAGCNN	Group F = R@5: 0.130; R@10: 0.185
			Group Z = R@5: 0.269; R@10: 0.362
			Harmonic Average = R@5: 0.160; R@10: 0.246
			The above given results are of ZAGCNN as it gives better performance than ZACNN.
D_{Total}: 22,815, D_{Train}: 70%, D_{Valid}: 10%, D_{Test}: 20%, Codes: 7024	Du et al. (2019)	ML-Net, ML-CNN,	P: 0.501; R: 0.373; F-score: 0.428
		ML-HAN	ML-CNN-threshold achieved highest F-score among all models

D_{Total}: Total no. of records, D_{Train}: Train data, D_{Valid}: Validation data, D_{Test}: Test data, P_{mi}: Micro-averaged Precision, R_{mi}: Micro-averaged Recall, $F1_{mi}$ or $F\text{-}meas_{mi}$: Micro-averaged F-measure, AUC_{mi}: Micro-averaged F1 score of Area under the ROC curve (AUC), P_{ma}: Macro-averaged Precision, R_{ma}: Macro-averaged Recall, $F1_{ma}$: Micro-averaged F1-score, AUC_{ma}: Macro-averaged F1 score of Area under the ROC curve (AUC), $F1_{diag}$: Micro-F1 on diagnosis codes, $F1_{proc}$: Micro-F1 on procedure codes, P: Precision, R: Recall, F1: F1-score, ACC: Accuracy, P@n: Precision@n; R@n: Recall@n.

Table 13 gives a comparison of studies that used MIMIC-II dataset. Six studies (Cao et al., 2020; Li et al., 2019; Li & Yu, 2020; Mullenbach et al., 2018; Perotte et al., 2013; Vu et al., 2020) have used the same train–test split size and number of codes. Among all these six studies, Vu et al. (2020) give more prominent results using LAAT and JointLAAT model. Baumel et al. (2018) also used the same train–test split size, but with a larger number of codes; they have also used additional data (MIMIC-III) for training. Similarly, Du et al. (2019) have divided MIMIC-II data into training, validation and testing sets and have more than 7000 unique codes. Therefore, we have not compared Baumel et al. (2018) and Du et al. (2019) with the above mentioned six studies. The other two studies, Marafino et al. (2014) and Rios and Kavuluru (2018b), are individual studies and they are focused towards a dedicated task.

Similarly, Table 14 show the comparison of studies using MIMIC-III dataset discharge summaries. Six studies, (Cao et al., 2020; Falis et al., 2019; Li & Yu, 2020; Moons et al., 2020; Vu et al., 2020; Xie et al., 2019), have used same data distribution and unique codes. Among them, Vu et al. (2020) give more prominent results using LAAT and JointLAAT model in all the evaluation metrics except in micro-averaged AUC. Moreover, Mullenbach et al. (2018) have also used the same data split size, but they have higher number of unique codes; therefore, we consider it as a separate study. However, except Moons et al. (2020), the remaining five studies (Cao et al., 2020; Falis et al., 2019; Li & Yu, 2020; Vu et al., 2020; Xie et al., 2019) have compared their models' performance with CAML and DR-CAML method which is developed by Mullenbach et al. (2018). Therefore, we have found that the majority of studies using MIMIC-III data discharge summaries compare their proposed methods with three state-of-the-art models: Flat & Hierarchy-based SVM (Perotte et al., 2013), CAML & DR-CAML (Mullenbach et al., 2018), and HA-GRU (Baumel et al., 2018).

Table 15 shows the comparison of studies that experimented either with top 100 or 50 most frequent codes in the study. For top 50 most frequent ICD codes, LAAT and JointLAAT proposed by Vu et al. (2020) again outperforms among all other models. Similarly, Table 16 shows the list of studies that experimented with less than 50 codes or that have used other datasets along with MIMIC-III. Lastly, Table 17 shows the list of studies that have used private datasets and are considered as individual studies.

Table 14
Comparison of studies using MIMIC-III dataset, different algorithms and evaluation metrics.

Train–test split	Study	Algorithms	Performance
D_{Total}: 59,531, Codes: 1,301	Berndorfer and Henriksson (2017)	Flat SVM	P: 55.10%; R: 33.74%; F1-score: 39.16%
		Hierarchy-based SVM	P: 40.08%; Recall: 41.69%; F1-score: 39.25%
D_{Train}: 37,016, D_{Test}: 1,356, Codes: 4,403 (Group S: all codes occur > 5 times); 4,403 (Group F: codes between 1 to 5 times); 178 (Group Z: never occur)	Rios and Kavuluru (2018b)	ZACNN	Group S = R@5: 0.283; R@10: 0.445
		ZAGCNN	Group F = R@5: 0.166; R@10: 0.216
			Group Z = R@5: 0.428; R@10: 0.495
			Harmonic Average = R@5: 0.252; R@10: 0.337
D_{Total}: 55,172, D_{Train}: 38,588, D_{Valid}: 5,536, D_{Test}: 11,048	Catling et al. (2018)	GRU	P: 0.692; R: 0.705 ; F1: 0.691
			Results for chapter (level 1) label prediction. RNN text representation improved weighted F1 for prediction of 19 disease-category labels to 0.682-0.701 from 0.662-0.682 using tf-idf.
D_{Train}: 36,998, D_{Valid}: 1,632, D_{Test}: 3,372, Codes: 8,921	Mullenbach et al. (2018)	CAML	AUC_{mi}: 0.986; AUC_{ma}: 0.895; $F1_{ma}$: 0.088; $F1_{mi}$: 0.539; $F1_{diag}$: 0.524; $F1_{proc}$: 0.609; P@8: 0.709; P@15: 0.561
		DR-CAML	AUC_{mi}: 0.985; AUC_{ma}: 0.879; $F1_{ma}$: 0.086; $F1_{mi}$: 0.529; $F1_{diag}$: 0.515; $F1_{proc}$: 0.595; P@8: 0.690; P@15: 0.548
D_{Total}: 49,857, Codes: 4,847	Baumel et al. (2018)	SVM	F_{mi}: 22.25%; F_{mi}: 53.02% (Rolled-up codes)
		CBOW	F_{mi}: 30.02%; F_{mi}: 43.30% (Rolled-up codes)
		CNN	F_{mi}: 40.72%; F_{mi}: 52.64% (Rolled-up codes)
		HA-GRU	F_{mi}: 40.52%; F_{mi}: 55.86% (Rolled-up codes)
D_{Total}: 58,976, D_{Train}: 40k, D_{Valid}: 7k, D_{Test}: 12k, Codes: 6,984	Xie and Xing (2018)	Tree of sequences LSTM	Sensitivity: 0.29; Specificity: 0.33
D_{Total}: 59,652, Codes: 6,918, 10-fold cross validation	Schäfer and Friedrich (2019)	FastText	P:0.58; R: 0.668; F1-score: 0.622 ($eFastText\text{-}UMLS_{cardinality}$)
D_{Total}: 52,962, D_{Train}: 47,665, D_{Test}: 5,297, Codes: 6,984	Li et al. (2019)	DeepLabeler (CNN+D2V)	P_{mi}: 0.486; R_{mi}: 0.351; $F\text{-}meas_{mi}$: 0.408
D_{Train}: 47,719, D_{Valid}: 1,631, D_{Test}: 3,372, Codes: 8,929 (Approx.)	Falis et al. (2019)	Ontological Attention	P_{mi}: 0.617; R_{mi}: 0.514; $F1_{mi}$: 0.560; P@8: 0.727; P_{ma}: 0.192; R_{ma}: 0.341; $F1_{ma}$: 0.245; P@8: 0.681
	Xie et al. (2019)	MSATT-KG	AUC_{ma}:91.0%; AUC_{mi}: 99.2%; $F1_{ma}$: 9.0% ; $F1_{mi}$: 55.3%; $F1_{diag}$:54.0%; $F1_{proc}$: 62.3%; P@8: 72.8%; P@15:58.1%
	Cao et al. (2020)	Hypercore	AUC_{mi}: 0.989; AUC_{ma}: 0.930; $F1_{ma}$: 0.090; $F1_{mi}$: 0.551; P@8: 0.722; P@15: 0.579
	Vu et al. (2020)	LAAT	AUC_{ma}: 0.919; AUC_{mi}: 0.988; $F1_{ma}$: 0.099; $F1_{mi}$: 0.575; P@5: 0.813; P@8: 0.738; P@15: 0.591
		JointLAAT	AUC_{ma}: 0.921; AUC_{mi}: 0.988; $F1_{ma}$: 0.107; $F1_{mi}$: 0.575; P@5: 0.806; P@8: 0.735; P@15: 0.590
	Li and Yu (2020)	MultiResCNN	AUC_{ma}: 0.910 ± 0.002, AUC_{mi}: 0.986 ± 0.001, $F1_{ma}$: 0.085 ± 0.007, $F1_{mi}$: 0.552 ± 0.005, P@8: 0.734 ± 0.002, P@15: 0.584 ± 0.001
	Moons et al. (2020)	CNN, GRU, DR-CAML, MVC-LDA, MVC-RLDA	AUC_{mi}: 90.02; $F1_{mi}$: 59.75 (Procedure); $F1_{mi}$: 51.60 (Diagnosis); $F1_{mi}$: 55.03(Both), P@5: 69.77
			$F1_{mi}$: 58.12 (Procedure); AUC_{mi}: 89.93; $F1_{mi}$: 50.70 (Diagnosis); $F1_{mi}$: 51.97 (Both), P@5: 68.53 (MVC-LDA outperforms other models and data includes discharge summary and other notes (radiology, nursing notes etc.)
	Ji et al. (2021)	CNN (retrained)	AUC_{mi}: 0.974; AUC_{ma}: 0.85; $F1_{mi}$: 0.365; $F1_{ma}$:0.059 P@8: 0.49; p@15: 0.394
D_{Train}: 46,157, D_{Valid}: 3,280, D_{Test}: 3,285, Codes: 6,916	Song et al. (2020)	ZAGRNN	P_{mi}: 0.5806; R_{mi}: 0.4494; $F1_{mi}$: 0.5066; AUC_{mi}:0.9667; P_{ma}: 0.3091; R_{ma}: 0.2557; $F1_{ma}$: 0.2799; AUC_{ma}: 0.9403
		ZAGRNN + L_{LDAM}	P_{mi}: 0.5606; R_{mi}: 0.4714; $F1_{mi}$: 0.5122; AUC_{mi}:0.9670; P_{ma}: 0.3172; R_{ma}: 0.2806; $F1_{ma}$: 0.2978; AUC_{ma}: 0.9408
			Th results are for seen codes. For zero-shot codes $F1_{mi}$ improved from 0 to 20.91% and AUC_{mi} improved by 3% and reached 92.18%
D_{Total}: 59,542 D_{Train}: 70%, D_{Valid}: 10%, D_{Test}: 20% (Label-to-Chapter)	Hsu et al. (2020)	CNN, LSTM, GRU, HAN	$F1_{mi}$: 76% (CNN outperforms other models)

D_{Total}: Total no. of records, D_{Train}: Train data, D_{Valid}: Validation data, D_{Test}: Test data, P_{mi}: Micro-averaged Precision, R_{mi}: Micro-averaged Recall, $F1_{mi}$ or $F\text{-}meas_{mi}$: Micro-averaged F-measure, AUC_{mi}: Micro-averaged F1 score of Area under the ROC curve (AUC), P_{ma}: Macro-averaged Precision, R_{ma}: Macro-averaged Recall, $F1_{ma}$: Micro-averaged F1-score, AUC_{ma}: Macro-averaged F1 score of Area under the ROC curve (AUC), $F1_{diag}$: Micro- F1 on diagnosis codes, $F1_{proc}$: Micro-F1 on procedure codes, P: Precision, R: Recall, F1: F1-score, ACC: Accuracy.

Table 15
Comparison of studies using MIMIC-III dataset with top (n = 50, 100) codes, different algorithms and evaluation metrics.

Dataset	Train–test split	Study	Algorithms	Performance
MIMIC-III using top 100 codes	D_{Train}: 80%, D_{Valid}: 10%, D_{Test}: 10%	Prakash et al. (2017)	A-MemNN	AUC_{ma}: 0.720; Average P@5: 0.29; HL:0.11
			C-MemNN	AUC_{ma}: 0.767; Average P@5: 0.32; HL:0.05
	D_{Total}: 46,715	Guo et al. (2020)	BiLSTMs	P_{mi}: 0.496; R_{mi}: 0.564; $F1_{mi}$: 0.528; AUC_{mi}: 0.87; P_{ma}: 0.464; R_{ma}: 0.463; $F1_{ma}$: 0.448; AUC_{ma}: 0.818. (Results are for BiLSTMs + SymVec (TF-IDF+Word2Vec))
	D_{Total}: 59,542, D_{Train}: 70%, D_{Valid}: 10%, D_{Test}: 20%	Hsu et al. (2020)	CNN, LSTM, GRU, HAN	$F1_{mi}$: 51.4% (3-digit); $F1_{mi}$: 50.2% (4-digit) $F1_{mi}$: 55.2% (5-digit) CNN outperforms other models
MIMIC-III using top 50 codes	D_{Train}: 80%, D_{Valid}: 10%, D_{Test}: 10%	Prakash et al. (2017)	A-MemNN	AUC_{ma}: 0.804; Average P@5: 0.40; HL: 0.02
			C-MemNN	AUC_{ma}: 0.833; Average P@5: 0.42; HL: 0.01
		Mullenbach et al. (2018)	CAML, DR-CAML	AUC_{mi}: 0.909; AUC_{ma}: 0.875; $F1_{ma}$: 0.532; $F1_{mi}$: 0.614; P@5:0.609, AUC_{mi}: 0.916; AUC_{ma}: 0.884; $F1_{ma}$: 0.576; $F1_{mi}$: 0.633; P@5: 0.618
		Xie et al. (2019)	MSATT-KG	AUC_{ma}:91.4%; AUC_{mi}: 93.6%; $F1_{ma}$: 63.8% ; $F1_{mi}$: 68.4%; P@5: 64.4%
		Cao et al. (2020)	Hypercore	AUC_{mi}: 0.929; AUC_{ma}: 0.895; $F1_{ma}$: 0.609; $F1_{mi}$: 0.663; P@5: 0.632
	D_{Train}: 8,067, D_{Valid}: 1,574, D_{Test}: 1,730	Vu et al. (2020)	LAAT	AUC_{ma}: 0.925; AUC_{mi}: 0.946; $F1_{ma}$: 0.666; $F1_{mi}$: 0.715; P@5: 0.675; P@8: 0.547; P@15: 0.357
			JointLAAT	AUC_{ma}: 0.925; AUC_{mi}: 0.946; $F1_{ma}$: 0.661; $F1_{mi}$: 0.716; P@5: 0.671; P@8: 0.546; P@15: 0.357
		Moons et al. (2020)	CNN, DR-CAML, GRU, MVC-LDA, MVC-RLDA	$F1_{mi}$: 67.86; $F1_{ma}$: 63.74; AUC_{mi}: 93.47; P@5:63.48 (DR-CAML outperforms other models)
		Li and Yu (2020)	MultiResCNN	AUC_{ma}: 0.899 ± 0.004, AUC_{mi}: 0.928 ± 0.002, $F1_{ma}$: 0.606 ± 0.0011, $F1_{mi}$: 0.670 ± 0.003, P@5: 0.641 ± 0.001
		Pascual et al. (2021)	BERT-ICD	AUC_{ma}: 0.8445; AUC_{mi}: 0.8865
		Ji et al. (2021)	CNN (retrained)	AUC_{ma}:0.908; AUC_{mi}: 0.931; $F1_{ma}$: 0.624; $F1_{mi}$: 0.671; P@5: 0.640
		Dong et al. (2021)	HLAN + label embedding	AUC_{ma}: 88.4 ± 0.5; AUC_{mi}: 91.9 ± 0.3; $F1_{ma}$: 56.8 ± 0.8; $F1_{mi}$: 64.1 ± 1.4; P@5: 62.4 ± 0.6
	D_{Total}: 46,364	Guo et al. (2020)	BiLSTMs	P_{mi}: 0.519; R_{mi}: 0.638; $F1_{mi}$: 0.572; AUC_{mi}: 0.859; P_{ma}: 0.508; R_{ma}: 0.568; $F1_{ma}$: 0.522; AUC_{ma}: 0.823
	D_{Total}: 46,552, D_{Train}: 43,000, D_{Valid}: 1,800, D_{Test}: 1,752	Teng et al. (2020)	G_Coder	$F1_{mi}$: 0.692; AUC_{mi}: 0.933; P@5: 0.653
	D_{Train}: 8,066, D_{Valid}: 1,573, D_{Test}: 1,729	Ji et al. (2020)	DCAN	AUC_{ma}: 90.2 ± 0.6, AUC_{mi}: 93.1 ± 0.1, $F1_{ma}$: 61.5 ± 0.7, $F1_{mi}$: 67.1 ± 0.1, P@8: 64.2 ± 0.2
	D_{Total}: 59,542, D_{Train}: 70%, D_{Valid}: 10%, D_{Test}: 20%	Hsu et al. (2020)	CNN, LSTM, GRU, HAN	$F1_{mi}$: 57.5% (3-digit); $F1_{mi}$:59.5% (4-digit) $F1_{mi}$: 67.4% (5-digit) CNN outperforms other models

D_{Total}: Total no. of records, D_{Train}: Train data, D_{Valid}: Validation data, D_{Test}: Test data, P_{mi}: Micro-averaged Precision, R_{mi}: Micro-averaged Recall, $F1_{mi}$ or F-$meas_{mi}$: Micro-averaged F-measure, AUC_{mi}: Micro-averaged F1 score of Area under the ROC curve (AUC), P_{ma}: Macro-averaged Precision, R_{ma}: Macro-averaged Recall, $F1_{ma}$: Micro-averaged F1-score, AUC_{ma}: Macro-averaged F1 score of Area under the ROC curve (AUC), $F1_{diag}$: Micro-F1 on diagnosis codes, $F1_{proc}$: Micro-F1 on procedure codes, P: Precision, R: Recall, F1: F1-score, ACC: Accuracy.

7. Future research directions

Several research gaps and limitations found in the studies are addressed in this literature review. This section highlights various future research directions where a considerable efforts are required to develop an automated ICD coding system.

1. Data source: One of the main challenges in developing an automated ICD coding system is the lack of publicly available benchmark *Gold Standard* dataset. The *Gold Standard* dataset is created by human experts who have a good knowledge of medical terminologies, clinical classification systems, and coding rules and guidelines. There are a few freely available datasets, such as i2b2,[8] and PhysioNet[9] that includes clinical reports such as discharge summaries, nursing notes, progress notes, radiology reports, and pathology reports. However, the data annotated in PhysioNet (MIMIC-II and MIMIC-III) are based on the ICD-9-CM which is no longer used since the adoption of ICD's 10th revision. Yet, up to date, the majority of the studies are predicting ICD-9-CM codes due to the limited data resources. Out of 42 selected studies, only 1 study (Xu et al., 2019) has mapped 32 ICD-9 codes extracted from MIMIC-III data to ICD-10 codes. Researchers are aiming to improve the performance of their model using ICD-9-CM annotated dataset without considering the change in classification version. As mentioned earlier, the ICD-9-CM codes are quite different from ICD-10-CM codes in the sense that there is a higher number of diagnoses and procedure codes in ICD-10-CM, ICD-10-CM has alphanumeric codes instead of numeric, the order of some chapters is changed, some titles have been renamed, and conditions have been grouped differently. After being aware of all these changes, the researchers are still developing automated ICD coding system based on ICD-9-CM codes. Therefore, this is one of the main problems that needs

[8] https://www.i2b2.org/NLP/DataSets/
[9] https://physionet.org/about/database/

Table 16
Comparison of studies using MIMIC-III and other datasets, algorithms and evaluation metrics.

Dataset	Train–test split	Study	Algorithms	Performance
MIMIC-III Less than 50 codes	D_{Train}: 39,541, D_{Valid}: 13,181, Codes: Top 19	Ayyar and Oliver (2016)	LSTM	P: 0.799; R: 0.685; F1-score: 0.708
	D_{Total}: 49,857, Codes: 19	Samonte et al. (2018)	EnHANs	P: 0.910; R: 0.540; F1-score: 0.500; Acc: 0.900 The results described in the analysis section are not matching with the result table provided in the study
	D_{Train}: 31,155, D_{Valid}: 4,484, D_{Test}: 9,020, Codes: 32 (mapped from ICD-9 to ICD-10)	Xu et al. (2019)	Ensemble-based (Text-TF-IDF-CNN)	$F1_{mi}$: 0.7633; $F1_{ma}$: 0.6867; AUC_{mi}: 0.9541; AUC_{ma}: 0.9337; JSC: 0.1806 (Text data) and 0.3105 (tabular data) The combined model of Text-TF-IDF-CNN, Label Smoothing (LS), diagnosis-based ranking (DR) and Tabular data (TD) achieves highest score among all the model tested in the study.
	D_{Total}: 52,726, Top-10-code: 40,562, Top-50-code: 49,354, Top-10-cat: 44,419, Top-50-cat: 51,034	Huang et al. (2019)	CNNs, LSTM RNNs, GRU RNNs	Top-10-code: P: 0.7502; R: 0.6519; F1: 0.6957; ACC: 0.8967 (GRU RNNs) Top-10-cat: P: 0.7580; R: 0.6941; F1: 0.7233; ACC: 0.8588 (GRU RNNs) AUC_{ma}: 0.8599; P@5: 0.8109; HL: 0.0645 (GRUs)
	Shielding data D_{Train}: 4,574; D_{Test}: 322; D_{Valid}: 153, Codes: 20	Dong et al. (2021)	HLAN + label embedding	AUC_{ma}: 93.5 ± 2.5; AUC_{mi}: 96.9 ± 0.7; $F1_{ma}$: 56.3 ± 2.4; $F1_{mi}$: 74.6 ± 1.6; P@1: 81.2 ± 1.2
MIMIC-III used with other datasets	D_{Total}: 58,929, Codes: 6,984, D_{Train}: 47,665, D_{Test}: 5,297 D_{Total}: 12,208,342 (BioASQ3), MeSH labels: 27,301	Zeng et al. (2019)	Deep transfer learning	P_{mi}: 0.483; R_{mi}: 0.371; F_{mi}: 0.42
	D_{Total}: 52,691 (MIMIC-III), D_{Total}: 193,677 (PHS)	Sonabend W et al. (2020)	UNITE	AUC: 0.91 (PHS) and 0.92 (MIMIC) over six diseases, comparable to LR and MLP
	D_{Total}: 53,423 (MIMIC-III), 10-Fold cross validation	Mascio et al. (2020)	ANN, CNN, RNN BiLSTM	$F1\text{-score}_{averaged}$: 94.5% (Status); $F1\text{-score}_{averaged}$: 97.9% (Temporality); $F1\text{-score}_{averaged}$: 98.7% (ShARe Negation); $F1\text{-score}_{averaged}$: 97.3% (ShARe Uncertainity) Custom Bi-LSTM model outperformed all other models using MIMIC dataset and Word2Vec embedding

D_{Total}: Total no. of records, D_{Train}: Train data, D_{Valid}: Validation data, D_{Test}: Test data, P_{mi}: Micro-averaged Precision, R_{mi}: Micro-averaged Recall, $F1_{mi}$ or $F\text{-meas}_{mi}$: Micro-averaged F-measure, AUC_{mi}: Micro-averaged F1 score of Area under the ROC curve (AUC), P_{ma}: Macro-averaged Precision, R_{ma}: Macro-averaged Recall, $F1_{ma}$: Micro-averaged F1-score, AUC_{ma}: Macro-averaged F1 score of Area under the ROC curve (AUC), $F1_{diag}$: Micro- F1 on diagnosis codes, $F1_{proc}$: Micro-F1 on procedure codes, P: Precision, R: Recall, F1: F1-score, ACC: Accuracy.

to be resolved by creating a benchmark dataset annotated with the latest version of the classification system that can be used for future research purposes.

2. Crowd-sourcing platform: Another possible research area is in developing a crowd-sourcing clinical coding and classification platform where the experts can guide and share their views, ideas and knowledge with the less experienced coders and researchers. A study by Searle et al. (2020) has found that frequently assigned codes in MIMIC-III data display signs of undercoding up to 35%. No other study has attempted to validate the MIMIC-III data due to time consuming factor and costly nature of the endeavour. For example, if two clinical coders, worked 38 h a week re-coding all 52,726 admission notes at a rate of 5 min and $3 per document, that would amount to approximately $316,000 (US) and approximately 115 weeks to create a gold standard dataset (Searle et al., 2020). Therefore, if a crowdsourced knowledge-based platform was created, then the problem of overcoding, undercoding, miscoding and lack of data sources using the latest coding version could be resolved. Also, this would help in developing an automated ICD coding system in real-time hospital settings.
3. National adoption and coding rules: Many countries follow their country specific classification system such as Australia, Canada and the US. There are a few codes that are used in a specific country only. For example, the Australian ICD includes more specificity regarding spider injuries than in Canadian and the US ICD classification system. Therefore, research studies conducted in different countries may have another new challenge if the data are coded in different classification version. Apart from that, various studies have focused on automated ICD code prediction despite considering the use of coding rules.
4. Reducing the complex problem: In this review, we have also found that a few studies have either truncated the codes to n-digits (3 or 4) only, predicted only top 50 or 100 most frequent codes (Guo et al., 2020; Prakash et al., 2017) or removed the rare occurring codes from the data in order to reduce complexity of the problem. One of the reasons behind this is lower number of reports or the rare codes in the reports. Thus, researchers should consider using multi-modal data or reports in order to reduce the complexity.
5. Transfer learning approach for automated ICD coding: In many machine learning methods, the training and testing sets are drawn from the same underlying dataset with the same distribution. However, in the real-world applications, it is difficult to collect sufficient training data to train a model. In such cases, transfer learning can learn from one related task and apply that knowledge to a target task. This approach has been proven very effectively and applied widely in biomedical research. A few studies, (Rios & Kavuluru, 2019; Zeng et al., 2019), have applied transfer learning for automated ICD-9 coding and improved classification performance. Thus, researchers may investigate various transfer learning approaches for automated ICD coding task.
6. Active learning and reinforcement learning approaches for clinical classification problems: One of the major challenges in using machine learning or deep learning approaches is to train the model with the lower number of reports, imbalanced or rare classes. Active learning has helped to discover rare class

Table 17
Comparison of studies using different datasets, algorithms and evaluation metrics.

Dataset	Train–test split	Study	Algorithms	Performance
Australian hospital medical records	D_{Total}: 190 and 235, Train–Test: 80-20%	Kaur and Ginige (2018)	SVM, Naive Bayes, Decision Tree, kNN, RF, AdaBoost, MLP	P: 0.9206; R: 0.8505; F-score: 0.8730; Acc: 0.7920; HL: 0.0877; JS: 0.7453 (For data190 using Decision Tree) P: 0.9239; R: 0.9201; F-score: 0.9141; Acc: 0.8611; HL: 0.0945; JS: 0.8294 (For data235 using AdaBoost)
	D_{Total}: 190 and 380, Codes: 420	Kaur and Ginige (2019)	Binary Relevance, Label Power-set, ML-kNN	P_{ma}: 0.3723; R_{ma}: 0.3739; $F1_{ma}$: 0.3717; P_{mi}: 0.9547; R_{mi}: 0.9476; $F1_{mi}$: 0.9511; HL: 0.0016; JS: 0.9121 Repetitive task gives better result than other two tasks using Binary Relevance classifier
10 Healthcare providers and covers 17 months of data University of Kentucky (UKY) medical center	Train/dev: 16 months data, Test: 1 month data, Codes: 3000	Amoia et al. (2018)	LR+CNN	P_{mi}: 0.681; R_{mi}: 0.616; $F1_{mi}$: 0.646
	D_{Total}: 71,463 EMRs, Codes: 7,485	Rios and Kavuluru (2019)	Transfer Learning (CNNs)	F_{mi}: 0.567; F_{ma}: 0.286
	D_{Total}: 827, D_{Train}: 727, D_{Test}: 100, Codes: 56 (UKSmall) D_{Total}: 71,463, D_{Train}: 66,463, D_{Valid}: 2000, D_{Test}: 3000, Codes: 1231 (UKLarge)	Kavuluru et al. (2015)	Binary relevance, copy transformation, ECC	CMC data: P_{mi}: 0.88, R_{mi}: 0.82, F_{mi}: 0.85 (ECC + SVM), UKSmall: F_{mi}: 0.44, F_{ma}: 0.32 (BNS+OTS+MLPTO), UKLarge: F_{mi}: 0.479, F_{ma}: 0.211 (LR+L2R+NERC)
CodiEsp (English version)	D_{Total}:1000, D_{Train}: 500, D_{Valid}: 250, D_{Test}: 250, Codes: 1,767	Moons et al. (2020)	CNN, DR-CAML, GRU, MVC-LDA, MVC-RLDA	$F1_{mi}$: 12.52 (CNN); $F1_{ma}$: 11.03 (GRU); AUC_{mi}: 50.54 (GRU); P@5: 7.96 (CNN) Most categories led to lower performance of all models due to insufficient amount of training data.
The Tri-Service General Hospital, Taipei, Taiwan	D_{Total}: 103,390, 5-fold cross validation	Lin et al. (2017)	CNN	F-measure: 0.9086; AUC: 0.9696
The NYU Langone Hospital EHR	D_{Total}: 7.5 million notes, D_{Train}: 70%, D_{Valid}: 10%, D_{Test}: 20%, Codes: 2,292	Zhang et al. (2020)	BERT-XML	AUC_{mi}: 0.970; AUC_{ma}: 0.927 For EHR BERT Big model with maximum input length of 1024.
Private data (Individual clinical records)	D_{Total}: 28,536, Thousand EHRs for development testing and evaluation, each, and rest for training along with 175,798 outpatient surgery EHRs with ICD-9 procedures codes; Codes: 5,650 (Unique PCS codes)	Subotin and Davis (2014)	Two-level hierarchical classification	MRR: 0.572 (All data, all features and l_2 regularisation)

D_{Total}: Total no. of records, D_{Train}: Train data, D_{Valid}: Validation data, D_{Test}: Test data, P_{mi}: Micro-averaged Precision, R_{mi}: Micro-averaged Recall, $F1_{mi}$ or $F\text{-meas}_{mi}$: Micro-averaged F-measure, AUC_{mi}: Micro-averaged F1 score of Area under the ROC curve (AUC), P_{ma}: Macro-averaged Precision, R_{ma}: Macro-averaged Recall, $F1_{ma}$: Micro-averaged F1-score, AUC_{ma}: Macro-averaged F1 score of Area under the ROC curve (AUC), $F1_{diag}$: Micro- F1 on diagnosis codes, $F1_{proc}$: Micro-F1 on procedure codes, P:Precision, R: Recall, F1: F1-score, ACC: Accuracy, HL: Hamming Loss, AUROC: Area under the ROC curve, JS: Jaccard Similarity, ECC: Ensemble of classifier chains, MRR: Mean Reciprocal Rank

(Hospedales et al., 2013), imbalanced class and other biomedical classification problems (Flores et al., 2021). Similarly, reinforcement learning has also proven to be suitable for imbalanced data classification problem as it can pay more attention to minority classes by giving higher rewards to them (Lin et al., 2020). Active learning and reinforcement learning can also be beneficial where unlabelled data can be obtained easily as the annotation of clinical reports is difficult, laborious and expensive (Mujtaba et al., 2019). Therefore, the researchers could also investigate various active learning and reinforcement learning algorithms for clinical report classification problems and improve the efficiency and classification model performance.

8. Conclusions

This study provides a comprehensive overview of automated ICD coding system based on discharge summaries. In this systematic literature review, we have addressed six research questions. A total of 42 studies have been selected from 4 different academic databases. The selected studies have been reviewed based on five key aspects: dataset, preprocessing techniques, feature extraction techniques, classification methods and evaluation metrics. This review basically focuses on publications that have used discharge summaries or other medical reports along with discharge summaries as the dataset. The majority of studies have used publicly available datasets MIMIC-II and MIMIC-III that are coded using ICD-9-CM codes. Various preprocessing techniques have been applied to remove unwanted or meaningless information from the discharge summaries which has helped to obtain improved classification results. For feature representation Word2Vec embedding and TF-IDF have been determined to be beneficial. Also, we found that the trend is towards applying large-scale pre-trained models such as Transformers and BERT. In addition, the majority of studies have compared the performance of their proposed deep learning model with the state-of-the-art method which resulted in improved classification results. The comparison of the selected studies is conducted based on three key aspects: dataset, train–test split size, and number of clinical codes. To evaluate the performance of a model, the majority of studies have used micro- and macro-averaging of precision, recall, and F1-score. Lastly, we have addressed various future research directions where a considerable effort are required in order to develop an automated ICD coding system.

CRediT authorship contribution statement

Rajvir Kaur: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **Jeewani Anupama Ginige:** Resources, Writing – review & editing, Supervision. **Oliver Obst:** Resources, Validation, Writing – review & editing, Investigation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

This work is supported by Western Sydney University, Australia Post Graduate Research Scholarship.

References

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. In *Proceedings of the 2nd clinical natural language processing workshop* (pp. 72–78). Minneapolis, Minnesota, USA: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/W19-1909, https://aclanthology.org/W19-1909.

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, vol. 22* (pp. 28–36). Curran Associates, Inc., https://www.scopus.com/inward/record.uri?eid=2-s2.0-80052204125&partnerID=40&md5=80adc5df88d419284cc58c7f379c7d91.

Amoia, M., Diehl, F., Gimenez, J., Pinto, J., Schumann, R., Stemmer, F., Vozila, P., & Zhang, Y. (2018). Scalable wide and deep learning for computer assisted coding. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 3 (industry papers)* (pp. 1–7). New Orleans - Louisiana: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N18-3001, https://www.aclweb.org/anthology/N18-3001.

Ayyar, S., & Oliver, I. (2016). Tagging patient notes with ICD-9 codes. In *29th conference on neural information processing systems*.

Baumel, T., Nassour-Kassis, J., Cohen, R., Elhadad, M., & Elhadad, N. (2018). Multi-label classification of patient notes: Case study on ICD code assignment. In *Workshops at the thirty-second AAAI conference on artificial antelligence*. https://www.aaai.org/ocs/index.php/WS/AAAIW18/paper/view/16881/15610.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828. http://dx.doi.org/10.1109/TPAMI.2013.50.

Bengio, S., Weston, J., & Grangier, D. (2010). Label embedding trees for large multi-class tasks. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, & A. Culotta (Eds.), *Advances in neural information processing systems, vol. 23* (pp. 163–171). Curran Associates, Inc., https://proceedings.neurips.cc/paper/2010/file/06138bc5af6023646ede0e1f7c1eac75-Paper.pdf.

Berndorfer, S., & Henriksson, A. (2017). Automated diagnosis coding with combined text representations. *Studies in Health Technology and Informatics*, *235*, 201–205. http://dx.doi.org/10.3233/978-1-61499-753-5-201.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on management of data* (pp. 1247—1250). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/1376616.1376746.

Burns, E., Rigby, E., Mamidanna, R., Bottle, A., Aylin, P., Ziprin, P., & Faiz, O. (2011). Systematic review of discharge coding accuracy. *Journal of Public Health*, *34*(1), 138–148. http://dx.doi.org/10.1093/pubmed/fdr054.

Campbell, S. E., Campbell, M. K., Grimshaw, J. M., & Walker, A. E. (2001). A systematic review of discharge coding accuracy. *Journal of Public Health*, *23*(3), 205–211. http://dx.doi.org/10.1093/pubmed/23.3.205.

Cao, P., Chen, Y., Liu, K., Zhao, J., Liu, S., & Chong, W. (2020). Hypercore: Hyperbolic and co-graph representation for automatic ICD coding. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 3105–3114). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.282, https://www.aclweb.org/anthology/2020.acl-main.282.

Cartwright, D. J. (2013). ICD-9-CM to ICD-10-CM codes: What? why? how? *Advances in Wound Care*, *2*, 588–592. http://dx.doi.org/10.1089/wound.2013.0478.

Catling, F., Spithourakis, G. P., & Riedel, S. (2018). Towards automated clinical coding. *International Journal of Medical Informatics*, *120*, 50–61. http://dx.doi.org/10.1016/j.ijmedinf.2018.09.021, http://www.sciencedirect.com/science/article/pii/S1386505618304039.

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., & Buchanan, B. G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, *34*(5), 301–310. http://dx.doi.org/10.1006/jbin.2001.1029, http://www.sciencedirect.com/science/article/pii/S1532046401910299.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1724–1734). Doha, Qatar: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/D14-1179, https://aclanthology.org/D14-1179.

Cumerlato, M., Best, L., & Saad, B. (2010). *Fundamentals of morbidity coding using ICD-10-AM, ACHI and ACS* (7th ed.). National Centre for Classification in Health.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *1*, In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423.

Dong, H., Suárez-Paniagua, V., Whiteley, W., & Wu, H. (2021). Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. *Journal of Biomedical Informatics*, *116*, Article 103728. http://dx.doi.org/10.1016/j.jbi.2021.103728, https://www.sciencedirect.com/science/article/pii/S1532046421000575.

Du, J., Chen, Q., Peng, Y., Xiang, Y., Tao, C., & Lu, Z. (2019). ML-Net: Multi-label classification of biomedical texts with deep neural networks. *Journal of the American Medical Informatics Association*, *26*(11), 1279—1285. http://dx.doi.org/10.1093/jamia/ocz085, http://dx.doi.org/10.1093/jamia/ocz085.

Falis, M., Pajak, M., Lisowska, A., Schrempf, P., Deckers, L., Mikhael, S., Tsaftaris, S., & O'Neil, A. (2019). Ontological attention ensembles for capturing semantic concepts in ICD code prediction from clinical text. In *Proceedings of the tenth international workshop on health text mining and information analysis* (pp. 168–177). Hong Kong: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D19-6220, https://www.aclweb.org/anthology/D19-6220.

Farkas, R., & Szarvas, G. (2008). Automatic construction of rule-based ICD-9-CM coding systems. *BMC Bioinformatics*, *9*(3), S10. http://dx.doi.org/10.1186/1471-2105-9-S3-S10.

Felix, R., B G, V. K., Reid, I., & Carneiro, G. (2018). Multi-modal cycle-consistent generalized zero-shot learning. In *15th European conference on computer vision* (pp. 21–37). http://dx.doi.org/10.1007/978-3-030-01231-1_2.

Flores, C. A., Figueroa, R. L., & Pezoa, J. E. (2021). Active learning for biomedical text classification based on automatically generated regular expressions. *IEEE Access*, *9*, 38767–38777. http://dx.doi.org/10.1109/ACCESS.2021.3064000.

Franz, P., Zaiss, A., Schulz, S., Hahn, U., & Klar, R. (2000). Automated coding of diagnoses–three methods compared. In *AMIA 2000, American medical informatics association annual symposium* (pp. 250–254).

Gangavarapu, T., Jayasimha, A., Krishnan, G. S., & S., S. K. (2020). Predicting ICD-9 code groups with fuzzy similarity based supervised multi-label classification of unstructured clinical nursing notes. *Knowledge-Based Systems*, *190*, Article 105321. http://dx.doi.org/10.1016/j.knosys.2019.105321, http://www.sciencedirect.com/science/article/pii/S0950705119305982.

Gehrmann, S., Dernoncourt, F., Li, Y., Carlson, E. T., Wu, J. T., Welt, J., Foote, J., Jr., Moseley, E. T., Grant, D. W., Tyler, P. D., & Celi, L. A. (2018). Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PLoS One*, *13*(2), 1–19. http://dx.doi.org/10.1371/journal.pone.0192360.

Gibaja, E., & Ventura, S. (2015). A tutorial on multilabel learning. *ACM Computing Surveys*, *47*(3), http://dx.doi.org/10.1145/2716262, https://doi-org.ezproxy.uws.edu.au/10.1145/2716262.

Goldberg, Y. (2016). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, *57*(1), 345–420, http://dl.acm.org/citation.cfm?id=3176748.3176757.

Guo, D., Duan, G., Yu, Y., Li, Y., Wu, F.-X., & Li, M. (2020). A disease inference method based on symptom extraction and bidirectional long short term memory networks. *Methods*, *173*, 75–82. http://dx.doi.org/10.1016/j.ymeth.2019.07.009, http://www.sciencedirect.com/science/article/pii/S1046202319301033.

Hargreaves, J., & Njeru, J. (2014). ICD-11: A dynamic classification for the information age. *HIM-Interchange*.

Hasan, S. A., & Farri, O. (2019). Clinical natural language processing with deep learning. In S. Consoli, D. Reforgiato Recupero, & M. Petković (Eds.), *Data science for healthcare: methodologies and applications* (pp. 147–171). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-05249-2_5.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Overview of supervised learning. In *The elements of statistical learning: data mining, inference, and prediction* (pp. 9–41). New York, NY: Springer New York, http://dx.doi.org/10.1007/978-0-387-84858-7_2.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Hospedales, T. M., Gong, S., & Xiang, T. (2013). Finding rare classes: Active learning with generative and discriminative models. *IEEE Transactions on Knowledge and Data Engineering*, *25*(2), 374–386. http://dx.doi.org/10.1109/TKDE.2011.231.

Hou, S., Feng, Y., & Wang, Z. (2017). Vegfru: A domain-specific dataset for fine-grained visual categorization. In *2017 IEEE international conference on computer vision* (pp. 541–549). http://dx.doi.org/10.1109/ICCV.2017.66.

Hsu, C.-C., Chang, P.-C., & Chang, A. (2020). Multi-label classification of ICD coding using deep learning. In *2020 international symposium on community-centric systems* (pp. 1–6). http://dx.doi.org/10.1109/CcS49175.2020.9231498.

Huang, J., Osorio, C., & Sy, L. (2019). An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes. *Computer Methods and Programs in Biomedicine*, *177*, http://dx.doi.org/10.1016/j.cmpb.2019.05.024.

Ji, S., Cambria, E., & Marttinen, P. (2020). Dilated convolutional attention network for medical code assignment from clinical text. In *Proceedings of the 3rd clinical natural language processing workshop* (pp. 73–78). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.clinicalnlp-1.8, https://www.aclweb.org/anthology/2020.clinicalnlp-1.8.

Ji, S., Hölttä, M., & Marttinen, P. (2021). Does the magic of BERT apply to medical code assignment? A quantitative study. *Computers in Biology and Medicine*, *139*(C), http://dx.doi.org/10.1016/j.compbiomed.2021.104998.

Karimi, S., Dai, X., Hassanzadeh, H., & Nguyen, A. (2017). Automatic diagnosis coding of radiology reports: A comparison of deep learning and conventional classification methods. In *BioNLP 2017* (pp. 328–332). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/W17-2342, http://aclweb.org/anthology/W17-2342.

Kaur, R. (2018). *A comparative analysis of selected set of natural language processing (NLP) and machine learning (ML) algorithms for clinical coding using clinical classification standards*. (Master of Research thesis), [Penrith, N.S.W.] : Western Sydney University, http://hdl.handle.net/1959.7/uws:49614, https://trove.nla.gov.au/work/235151724.

Kaur, R. (2019). Distributed knowledge based clinical auto-coding system. In *Proceedings of the 57th annual meeting of the association for computational linguistics: student research workshop* (pp. 1–9). Florence, Italy: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P19-2001, https://www.aclweb.org/anthology/P19-2001.

Kaur, R., & Ginige, J. A. (2018). Comparative analysis of algorithmic approaches for auto-coding with ICD-10-AM and ACHI.. *Studies in Health Technology and Informatics*, *252*, 73–79.

Kaur, R., & Ginige, J. A. (2019). Analysing effectiveness of multi-label classification in clinical coding. In *ACSW 2019, Proceedings of the Australasian computer science week multiconference*. New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3290688.3290728.

Kavuluru, R., Han, S., & Harris, D. (2013). Unsupervised extraction of diagnosis codes from EMRs using knowledge-based and extractive text summarization techniques. In O. R. Zaïane, & S. Zilles (Eds.), *Advances in artificial intelligence* (pp. 77–88). Berlin, Heidelberg: Springer Berlin Heidelberg.

Kavuluru, R., Rios, A., & Lu, Y. (2015). An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artificial Intelligence in Medicine*, *65*(2), 155–166. http://dx.doi.org/10.1016/j.artmed.2015.04.007, http://www.sciencedirect.com/science/article/pii/S0933365715000482, Intelligent healthcare informatics in big data era.

Khattak, F. K., Jeblee, S., Pou-Prom, C., Abdalla, M., Meaney, C., & Rudzicz, F. (2019). A survey of word embeddings for clinical text. *Journal of Biomedical Informatics: X*, *4*, Article 100057. http://dx.doi.org/10.1016/j.yjbinx.2019.100057, https://www.sciencedirect.com/science/article/pii/S2590177X19300563.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1746–1751). Doha, Qatar: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/D14-1181, https://www.aclweb.org/anthology/D14-1181.

Ko, Y., & Seo, J. (2000). Automatic text categorization by unsupervised learning. In *Proceedings of the 18th conference on computational linguistics - vol. 1* (pp. 453–459). Stroudsburg, PA, USA: Association for Computational Linguistics, http://dx.doi.org/10.3115/990820.990886.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th international joint conference on artificial intelligence – vol. 2* (pp. 1137–1143). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., http://dl.acm.org/citation.cfm?id=1643031.1643047.

Larkey, L. S., & Croft, W. B. (1996). Combining classifiers in text categorization. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 289–297). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/243199.243276.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In E. P. Xing, & T. Jebara (Eds.), *Proceedings of machine learning research*: *vol. 32*, *Proceedings of the 31st international conference on machine learning, no. 2* (pp. 1188–1196). Bejing, China: PMLR, https://proceedings.mlr.press/v32/le14.html.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, http://dx.doi.org/10.1093/bioinformatics/btz682.

Li, M., Fei, Z., Zeng, M., Wu, F., Li, Y., Pan, Y., & Wang, J. (2019). Automated ICD-9 coding via a deep learning approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *PP*, 1. http://dx.doi.org/10.1109/TCBB.2018.2817488.

Li, F., & Yu, H. (2020). ICD coding from clinical text using multi-filter residual convolutional neural network. (pp. 8180–8187). http://dx.doi.org/10.1609/aaai.v34i05.6331, https://ojs.aaai.org/index.php/AAAI/article/view/6331,

Lin, E., Chen, Q., & Qi, X. (2020). Deep reinforcement learning for imbalanced classification. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, 1–15.

Lin, C., Hsu, C.-J., Lou, Y.-S., Yeh, S.-J., Lee, C.-C., Su, S.-L., & Chen, H.-C. (2017). Artificial intelligence learning semantics via external resources for classifying diagnosis codes in discharge notes. *Journal of Medical Internet Research*, *19*(11), Article e380. http://dx.doi.org/10.2196/jmir.8344, http://www.jmir.org/2017/11/e380/.

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., & Yu, S. (2019). *Large-scale long-tailed recognition in an open world* (pp. 2532–2541). http://dx.doi.org/10.1109/CVPR.2019.00264.

Marafino, B. J., Davies, J. M., Bardach, N. S., Dean, M. L., & Dudley, R. A. (2014). N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit. *Journal of the American Medical Informatics Association*, *21*(5), 871–875. http://dx.doi.org/10.1136/amiajnl-2014-002694.

Mascio, A., Kraljevic, Z., Bean, D., Dobson, R., Stewart, R., Bendayan, R., & Roberts, A. (2020). Comparative analysis of text classification approaches in electronic health records. In *Proceedings of the 19th SIGBioMed workshop on biomedical language processing* (pp. 86–94). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.bionlp-1.9, https://www.aclweb.org/anthology/2020.bionlp-1.9.

Mayya, V., S., S. K., Krishnan, G. S., & Gangavarapu, T. (2021). Multi-channel, convolutional attention based neural model for automated diagnostic coding of unstructured patient discharge summaries. *Future Generation Computer Systems*, *118*, 374–391. http://dx.doi.org/10.1016/j.future.2021.01.013, https://www.sciencedirect.com/science/article/pii/S0167739X21000236.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems, vol. 26*. Curran Associates, Inc., https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, T. P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*(7), 1–6. http://dx.doi.org/10.1371/journal.pmed.1000097.

Moons, E., Khanna, A., Akkasi, A., & Moens, M.-F. (2020). A comparison of deep learning methods for ICD coding of clinical records. *Applied Sciences*, *10*(15), http://dx.doi.org/10.3390/app10155262, https://www.mdpi.com/2076-3417/10/15/5262.

Moriyama, I. M., Loy, R. M., Robb-Smith, A. H. T., Rosenberg, H. M., & Hoyert, D. L. (2011). *History of the statistical classification of diseases and causes of death*: *DHHS publication No. (PHS) 2011-1125*, National Center for Health Statistics (U.S.), https://stacks.cdc.gov/view/cdc/5928.

Mujtaba, G., Shuib, L., Idris, N., Hoo, W. L., Raj, R. G., Khowaja, K., Shaikh, K., & Nweke, H. F. (2019). Clinical text classification research trends: Systematic literature review and open issues. *Expert Systems with Applications*, *116*, 494–520, http://www.sciencedirect.com/science/article/pii/S0957417418306110.

Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., & Eisenstein, J. (2018). Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 1101–1111). New Orleans, Louisiana: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N18-1100, https://www.aclweb.org/anthology/N18-1100.

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). New York, NY, USA: ACM, http://dx.doi.org/10.1145/2939672.2939720, http://doi.acm.org/10.1145/2939672.2939720.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. http://dx.doi.org/10.1109/TKDE.2009.191.

Pascual, D., Luck, S., & Wattenhofer, R. (2021). Towards BERT-based automatic ICD coding: Limitations and opportunities. In *Proceedings of the 20th workshop on biomedical language processing* (pp. 54–63). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.bionlp-1.6, https://aclanthology.org/2021.bionlp-1.6.

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/D14-1162, https://aclanthology.org/D14-1162.

Perotte, A., Pivovarov, R., Natarajan, K., Weiskopf, N., Wood, F., & Elhadad, N. (2013). Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association*, *21*(2), 231–237. http://dx.doi.org/10.1136/amiajnl-2013-002159.

Pestian, J. P., Brew, C., Matykiewicz, P., Hovermale, D., Johnson, N., Cohen, K. B., & Duch, W. (2007). A shared task involving multi-label classification of clinical free text. In *Biological, translational, and clinical language processing* (pp. 97–104). Prague, Czech Republic: Association for Computational Linguistics, https://www.aclweb.org/anthology/W07-1013.

Prakash, A., Zhao, S., Hasan, S. A., Datla, V., Lee, K., Qadir, A., Liu, J., & Farri, O. (2017). Condensed memory networks for clinical diagnostic inferencing. In *Proceedings of the thirty-first AAAI conference on artificial intelligence* (pp. 3274–3280). AAAI Press.

Reed, G. M., Drescher, J., Krueger, R. B., Atalla, E., Cochran, S. D., First, M. B., Cohen-Kettenis, P. T., Arango-de Montis, I., Parish, S. J., Cottler, S., Briken, P., & Saxena, S. (2016). Disorders related to sexuality and gender identity in the ICD-11: revising the ICD-10 classification based on current scientific evidence,

best clinical practices, and human rights considerations. *World Psychiatry*, *15*(3), 205–221. http://dx.doi.org/10.1002/wps.20354, https://onlinelibrary.wiley.com/doi/abs/10.1002/wps.20354.

Rios, A., & Kavuluru, R. (2018a). EMR coding with semi-parametric multi-head matching networks. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 2081–2091). New Orleans, Louisiana: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N18-1189, https://www.aclweb.org/anthology/N18-1189.

Rios, A., & Kavuluru, R. (2018b). Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 3132–3142). Brussels, Belgium: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D18-1352, https://www.aclweb.org/anthology/D18-1352.

Rios, A., & Kavuluru, R. (2019). Neural transfer learning for assigning diagnosis codes to EMRs. *Artificial Intelligence in Medicine*, *96*, 116–122. http://dx.doi.org/10.1016/j.artmed.2019.04.002, http://www.sciencedirect.com/science/article/pii/S0933365718304378.

Sadoughi, N., Finley, G., Fone, J., Murali, V., Korenevski, M., Baryshnikov, S., Axtmann, N., Miller, M., & Suendermann-Oeft, D. (2018). Medical code prediction with multi-view convolution and description-regularized label-dependent attention. arXiv preprint arXiv:1811.01468.

Samonte, M. J. C., Gerardo, B. D., Fajardo, A. C., & Medina, R. P. (2018). ICD-9 tagging of clinical notes using topical word embedding. In *Proceedings of the 2018 international conference on internet and e-business* (pp. 118–123). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3230348.3230357.

Samonte, M. J. C., Gerardo, B. D., & Medina, R. P. (2017). Towards enhanced hierarchical attention networks in ICD-9 tagging of clinical notes. In *Proceedings of the 3rd international conference on communication and information processing* (pp. 146–150). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3162957.3163030.

Santos, S., Murphy, G., Baxter, K., & Robinson, K. M. (2008). Organisational factors affecting the quality of hospital clinical coding. *Health Information Management Journal*, *37*(1), 25–37.

Schäfer, H., & Friedrich, C. M. (2019). UMLS mapping and word embeddings for ICD code assignment using the MIMIC-III intensive care database. In *2019 41st annual international conference of the IEEE engineering in medicine and biology society* (pp. 6089–6092).

Scheurwegs, E., Cule, B., Luyckx, K., Luyten, L., & Daelemans, W. (2017). Selecting relevant features from the electronic health record for clinical code prediction. *Journal of Biomedical Informatics*, *74*, 92–103. http://dx.doi.org/10.1016/j.jbi.2017.09.004, https://www.sciencedirect.com/science/article/pii/S1532046417302010.

Searle, T., Ibrahim, Z., & Dobson, R. (2020). Experimental evaluation and development of a silver-standard for the MIMIC-III clinical coding dataset. In *Proceedings of the 19th SIGBioMed workshop on biomedical language processing* (pp. 76–85). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.bionlp-1.8, https://www.aclweb.org/anthology/2020.bionlp-1.8.

Shi, H., Xie, P., Hu, Z., Zhang, M., & Xing, E. P. (2017). Towards automated ICD coding using deep learning. CoRR abs/1711.04075, arXiv:1711.04075.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1631–1642). Association for Computational Linguistics, http://aclweb.org/anthology/D13-1170.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. http://dx.doi.org/10.1016/j.ipm.2009.03.002, http://www.sciencedirect.com/science/article/pii/S0306457309000259.

Sonabend W, A., Cai, W., Ahuja, Y., Ananthakrishnan, A., Xia, Z., Yu, S., & Hong, C. (2020). Automated ICD coding via unsupervised knowledge integration (UNITE). *International Journal of Medical Informatics*, *139*, Article 104135. http://dx.doi.org/10.1016/j.ijmedinf.2020.104135, http://www.sciencedirect.com/science/article/pii/S1386505619313024.

Song, C., Zhang, S., Sadoughi, N., Xie, P., & Xing, E. (2020). Generalized zero-shot text classification for ICD coding. In C. Bessiere (Ed.), *Proceedings of the twenty-ninth international joint conference on artificial intelligence* (pp. 4018–4024). International Joint Conferences on Artificial Intelligence Organization, http://dx.doi.org/10.24963/ijcai.2020/556, Main track.

Stanfill, M. H., Williams, M., Fenton, S. H., Jenders, R. A., & Hersh, W. R. (2010). A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association*, *17*(6), 646–651. http://dx.doi.org/10.1136/jamia.2009.001024.

Subotin, M., & Davis, A. (2014). A system for predicting ICD-10-PCS codes from electronic health records. In *Proceedings of BioNLP 2014* (pp. 59–67). Baltimore, Maryland: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/W14-3409, https://www.aclweb.org/anthology/W14-3409.

Sun, W., Cai, Z., Li, Y., Liu, F., Fang, S., & Wang, G. (2018). Data processing and text mining technologies on electronic medical records: A review. *Journal of Healthcare Engineering*, *2018*, https://www.ncbi.nlm.nih.gov/pubmed/29849998.

Teng, F., Yang, W., Chen, L., Huang, L., & Xu, Q. (2020). Explainable prediction of medical codes with knowledge graphs. *Frontiers in Bioengineering and Biotechnology*, *8*, 867. http://dx.doi.org/10.3389/fbioe.2020.00867, https://www.frontiersin.org/article/10.3389/fbioe.2020.00867.

Vani, A., Jernite, Y., & Sontag, D. (2017). Grounded recurrent neural networks. arXiv:1705.08557.

Vu, T., Nguyen, D. Q., & Nguyen, A. (2020). A label attention model for ICD coding from clinical text. In C. Bessiere (Ed.), *Proceedings of the twenty-ninth international joint conference on artificial intelligence* (pp. 3335–3341). International Joint Conferences on Artificial Intelligence Organization, http://dx.doi.org/10.24963/ijcai.2020/461, Main track.

Wang, G., Li, C., Wang, W., Zhang, Y., Shen, D., Zhang, X., Henao, R., & Carin, L. (2018). Joint embedding of words and labels for text classification. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 2321–2331). Melbourne, Australia: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P18-1216, https://www.aclweb.org/anthology/P18-1216.

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, *8*(7), 1341–1390. http://dx.doi.org/10.1162/neco.1996.8.7.1341.

World Health Organisation (2016). *ICD-11 Revision Conference*: *Report*, Tokyo, Japan: WHO, Published at: https://cdn.who.int/media/docs/default-source/classification/icd/icd11/icd11-revision-conference-report-oct2016.pdf.

Xian, Y., Lorenz, T., Schiele, B., & Akata, Z. (2018). *Feature generating networks for zero-shot learning* (pp. 5542–5551). http://dx.doi.org/10.1109/CVPR.2018.00581.

Xie, P., & Xing, E. (2018). A neural architecture for automated ICD coding. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 1066–1076). Melbourne, Australia: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P18-1098, https://www.aclweb.org/anthology/P18-1098.

Xie, X., Xiong, Y., Yu, P. S., & Zhu, Y. (2019). EHR coding with multi-scale feature attention and structured knowledge graph propagation. In *Proceedings of the 28th ACM international conference on information and knowledge management* (pp. 649–658). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3357384.3357897.

Xu, K., Lam, M., Pang, J., Gao, X., Band, C., Mathur, P., Papay, F., Khanna, A. K., Cywinski, J. B., Maheshwari, K., Xie, P., & Xing, E. P. (2019). Multimodal machine learning for automated ICD coding. In F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, & J. Wiens (Eds.), *Proceedings of machine learning research*: *vol. 106*, *Proceedings of the 4th machine learning for healthcare conference* (pp. 197–215). Ann Arbor, Michigan: PMLR, http://proceedings.mlr.press/v106/xu19a.html.

Yan, Z., Zhang, H., Piramuthu, R., Jagadeesh, V., DeCoste, D., Di, W., & Yu, Y. (2015). HD-CNN: Hierarchical deep convolutional neural networks for large scale visual recognition. In *2015 IEEE international conference on computer vision* (pp. 2740–2748). http://dx.doi.org/10.1109/ICCV.2015.314.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 1480–1489). San Diego, California: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N16-1174, https://aclanthology.org/N16-1174.

Zeng, M., Li, M., Fei, Z., Yu, Y., Pan, Y., & Wang, J. (2019). Automatic ICD-9 coding via deep transfer learning. *Neurocomputing*, *324*, 43–50. http://dx.doi.org/10.1016/j.neucom.2018.04.081, http://www.sciencedirect.com/science/article/pii/S0925231218306246, Deep Learning for Biological/Clinical Data.

Zhang, Z., Liu, J., & Razavian, N. (2020). BERT-XML: Large scale automated ICD coding using BERT pretraining. In *Proceedings of the 3rd clinical natural language processing workshop* (pp. 24–34). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.clinicalnlp-1.3, https://aclanthology.org/2020.clinicalnlp-1.3.

Zhang, J., & Zong, C. (2015). Deep neural networks in machine translation: An overview. *IEEE Intelligent Systems*, *30*(5), 16–25. http://dx.doi.org/10.1109/MIS.2015.69.

Zhu, X., & Bain, M. (2017). B-CNN: Branch convolutional neural network for hierarchical classification. arXiv:1709.09890.

Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *3*(1), 1–130.