AUTOMATIC ICD TO IMPROVE DIAGNOSIS CODING ACCURACY: A LITERATURE STUDY

Tominanto[1*], Linda Widyaningrum[2], Arifah Khoirunnisa'[3], Eva Ayu Anggraini[4]
[1,2,3,4]Faculty of Health Science, Universitas Duta Bangsa Surakarta
*Correspondence Email : tominanto@udb.ac.id

ABSTRACT

ICD coding is usually done by a coder who assigns the ICD code according to the Doctor's clinical diagnosis. However, because coders need to master specific skills, such as knowledge in the field of medicine, coding rules, and medical terminology, manual coding can be costly, time-consuming, and inefficient. Based on this, developing a computationally accurate approach to automatic ICD encoding is imperative. This literature study aims to provide an overview of automatic ICD in terms of the dataset and classification method used. The literature study results show that automatic ICD research to improve the accuracy of diagnosis coding has been done and is still a challenge today. Most of the datasets used are the public MIMIC-III dataset, while automatic ICD classification is the current research trend with deep learning. The deep learning algorithms that are widely used include CNN, RNN, and LSTM. The resulting accuracy based on the dataset and classification method is very diverse. Future research still has many opportunities to contribute and improve the correct classification method to improve automatic ICD accuracy.

KEYWORDS

automatic ICD, diagnosis, dataset, deep learning

INTRODUCTION

The International Classification of Diseases (ICD) is a healthcare classification system administered by the World Health Organization (WHO), which provides a hierarchy of diagnostic codes for diseases, disorders, injuries, signs, symptoms, etc. This system is widely used to report diseases and health conditions, assist in medical cost decisions, and collect morbidity and mortality statistics (Shi, Xie, Hu, Zhang, & Xing, 2017).

The ICD code is unique for each disease used in manual and electronic medical records. Usually, ICD coding is carried out by the Hospital's Coder Division or Medical

Records Unit, which assigns ICD codes to medical records according to the Doctor's clinical diagnosis. However, because coders need to master specific skills, such as knowledge in the field of medicine, coding rules, and medical terminology, manual coding can be costly, time-consuming, and inefficient. Considering this, it is imperative to develop a computationally accurate approach to automatic ICD encoding (Li et al., 2019).

Automatic ICD coding based on medical diagnostic documents remains a challenge. Automatic ICD coding usually has the following problems (Li et al., 2019), (Teng, Ma, Chen, Xiao, & Huang, 2020): 1) The patient's clinical medical record is not always structured, so it is complicated to extract essential and relevant knowledge from various types of medical records effectively, 2) The medical field has many terminologies, which it is difficult for non-professionals to understand the meaning of the terminology so that additional tools are needed to interpret some terms and symptoms and to obtain semantic information from medical records, 3) Doctors usually have their way of describing symptoms so that for the same disease there can be many different ways to describe it, 4) From the aspect of coding rules, there is not necessarily a mapping one by one between the description of the diagnosis and the ICD code, 5) The number of ICD codes for one record varies greatly and can select up to 20 labels per record from an extensive set of labels large (24,478 unique codes for ICD-9 and ICD-9-CM-3). The difficulty and importance of ICD coding make the automated task of the ICD an essential problem.

Several studies on the automatic coding of ICD have been carried out, including Chen et al. (Yun Zhi Chen, Lu, & Li, 2017). Chen et al. In developing the automatic coding ICD, the approach is to improve the Longest Common Subsequence (LCS) method and semantic similarity. The results of his research prove that an F-score of 81% can improve the accuracy of ICD coding. Research on Automatic ICD-9 coding using deep transfer learning (Zeng et al., 2019) shows that deep transfer learning is a crucial element in improving the performance of an automatic ICD-9 coding model. Deep transfer learning has significantly improved the performance of the ICD-9 coding, automatically outperforming hierarchy-based SVM and flat-SVM methods. Another study on automatic ICD was conducted by Teng et al. (Teng et al., 2020), which uses the Medical Topic Mining (MTM) method to predict the ICD code from free-text medical records automatically. The resulting accuracy for disease-specific atrial fibrillation showed F1 scores of 96% and 93.3%, respectively, using the ICD-10 internal data set and the Medical Information Mart for Intensive Care (MIMIC) -III data set.

In another study conducted by Cao et al. (Cao et al., 2020) to overcome manual coding is very tedious and error-prone, Cao et al. proposed automation of ICD coding using a Hyperbolic and Co-graph Representation method (HyperCore), using the MIMIC-II and MIMIC-III datasets In, the results are of higher accuracy than the SVM, HA-GRU, and CAML methods. Another study on automatic coding of ICD (Atutxa, de Ilarraza, Gojenola, Oronoz, & Perez-de-Viñaspre, 2019) overcomes the problem of writing a diagnosis, often done in natural or natural language, making automatic coding difficult because it differs from ICD terminology. With deep learning, Atutxa et al. yield an accuracy of 0.838, 0.963, and 0.952 for the multilingual dataset: French, Hungarian and Italian.

Based on the description above, many studies have been carried out to solve the problem of automatic ICD coding. The dataset and methods used are quite diverse and with different accuracy results. This literature study focuses on analyzing the dataset and classification methods used in developing ICD automation to improve the accuracy of diagnosis coding.

RESEARCH METHOD

The approach used in this study is to use a literature study approach. The authors collect data and information related to ICD automation through supporting data from international research journals and international proceedings.

Literature studies have several objectives, namely informing readers of the results of other studies that are closely related to the research being carried out at that time, connecting research with existing literature, and filling gaps in previous studies, literature reviews containing reviews, summaries, and the author's thoughts about several sources of literature on the topics discussed. This literature study aims to determine the dataset and classification method for developing automatic ICDs.

RESULT AND DISCUSSION

A. ICD Dataset

The dataset is a set of data that comes from past information and is managed into information. There are two types of datasets, namely private datasets and public datasets. A private Dataset is a data set that can be taken from an organization we will conduct as an object of research, for example, bank data, hospitals, universities, companies, and so on. A public dataset is a data set we can retrieve from the public repository agreed upon by the researchers. Most of the datasets used in automatic ICD research use public datasets, and some specifically use private datasets. The public dataset used is as follows:

1. MIMIC II

The Medical Information Mart for Intensive Care (MIMIC) II is a public dataset with an extensive database available free of charge consisting of health-related data for more than forty thousand patients living in the critical care unit at Beth Israel Deaconess Medical Center between 2001 and 2012. Cao et al. (Cao et al., 2020) using MIMIC II as the dataset, there were 20533 and 2282 clinical records for training and testing and 5031 unique ICD-9 codes in the dataset. The MIMIC II dataset can be accessed via the https://archive.physionet.org/mimic2/.

2. MIMIC III

MIMIC III is a continuation version of MIMIC-II. MIMIC III is the most popular public dataset in automatic ICD research. MIMIC III, among others, is used as a dataset in studies (Zeng et al., 2019), (Teng et al., 2020), (Cao et al., 2020), (Blanco, Perez, & Casillas, 2020), (Schafer & Friedrich, 2019), (Huang, Osorio, & Sy, 2019), (Yuwen Chen & Ren, 2019), (Sonabend W et al., 2020). The latest version of MIMIC is MIMIC-III v1.4, which consists of 61,532 intensive care units: 53,432 treatments for adults and 8,100 for neonatal patients. The data cover June 2001 - October 2012. Although not identified, the existing database still contains detailed information on the patient's clinical care, so it should be treated with appropriate care. MIMIC III can be accessed via the page https://mimic.physionet.org/gettingstarted/access/. The descriptive data for MIMIC III are shown in Table 1 (Schafer & Friedrich, 2019):

Table 1. MIMIC-III Dataset Descriptive Statistics

	MIMIC-III
Number of records with ICD code	59,652
Number of unique tokens	119,171
Avg. number of tokens / record	1,947
Avg. number of sentences / record	112
Avg. number of labels / record	11.48
Label Density	0.0018
Number of labels in collection	6,918
Number of labels in extended collection	8,790

Table 2 shows the top 10 ICD-9 codes and top 10 ICD-9 categories [10] :

Table 2. Admission number for top 10 ICD-9 codes and top 10 ICD-9 categories

ICD-9 Code	Admissions
4019: Hypertension	20,046
4280: Congestive heart failure	12,842
42731: Atrial fibrillation	12,589
41401: Coronary atherosclerosis	12,178
5849: Acute kidney failure	8906
25000: Diabetes Type II	8783
2724: Hyperlipidemia	8503
51881: Acute respiratory failure	7249
5990: Urinary tract infection	6442
53081: Esophageal reflux	6154
ICD-9 Category	**Admissions**
401: Essential hypertension	20,646
427: Cardiac dysrhythmias	16,774
276: Disorders of fluid electrolyte	14,712
272: Disorders of lipoid metabolism	14,212
414: Other chronic ischemic heart disease	14,081
250: Diabetes mellitus	13,818
428: Heart failure	13,330
518: Other diseases of lung	12,997
285: Other and unspecified anemias	12,404
584: Acute kidney failure	11,147

3. MeSH (BioASQ3 dataset)

The automatic ICD study was conducted by Zeng et al. (Zeng et al., 2019) using MeSH data from the Large Scale Biomedical Semantic Indexing Competition (BioASQ3). There are 12,208,342 citations indexed with abstracts and titles stored locally. The main titles (MHs) of MeSH amounted to 27,301, with an average sample of 477.12. Thus, the total number of samples, labels, and the average label sample in the MeSH dataset is much greater than that of the MIMIC-III dataset.

4. CLEF eHealth

CLEF eHealth is a public dataset related to Health data. It can be accessed via the page https://clefehealth.imag.fr/?page_id=215. Automatic ICD research conducted by Atutxa et al. (Atutxa et al., 2019) and Almagro et al. (Almagro, Martínez, Montalvo, & Fresno, 2019) used the CLEF eHealth dataset.

5. Other Dataset

In addition to the four public datasets, some researchers also use others. For example, research (Yu et al., 2019) uses the Xiangya Big Data Platform of Healthcare dataset, (Yuwen Chen & Ren, 2019) use the Chinese TPD dataset, and (Silvestri, Gargiulo, Ciampi, & De Pietro, 2020), (Koopman, Zuccon, Nguyen, Bergheim, & Grayson, 2015), (Ping Chen, Barrera, & Rhodes, 2010) uses private dataset in automatic ICD classification.

B. Classification of Automatic ICD

1. Deep Learning Method

Deep learning is a set of algorithms in machine learning that attempt to learn at multiple levels, according to different levels of abstraction, and usually uses an artificial neural network. The levels in this studied statistical model correspond to different concept levels, where higher-level concepts are defined from lower-level concepts and lower-level concepts can help to define many higher-level concepts (Deng & Yu, 2013).

Most classification methods used in automatic ICD research are deep learning algorithms. Atutxa et al. (Atutxa et al., 2019) used Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) deep learning algorithms by combining them in developing automatic ICD and producing an accuracy of 0.838, 0.963, and 0.952 for a multilingual dataset: French, Hungarian and Italian. Modify the RNN algorithm to the Multilayer Attention Bidirectional Recurrent Neural Network (MA-BiRNN) model on the automatic ICD, and the proposed model achieves 0.639 and 0.766 in F1-score on full-level code and block-level code, respectively (Yu et al., 2019). Figure 1 below is an example of the RNN method on automatic ICD (Yu et al., 2019).

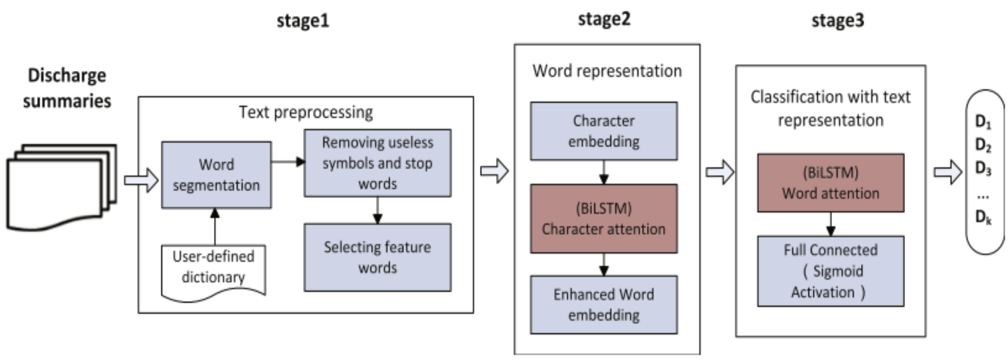

Figure 1. The architecture of hierarchial attention RNN

The deep learning RNN algorithm was also used by (Blanco et al., 2020), yielding an accuracy of 39.55% and 47.28% F-Score for the Osakidetza and MIMIC-III datasets, respectively. Meanwhile, (Huang et al., 2019) used RNN and CNN in classifying automatic ICD. The best models could predict the top 10 ICD-9 codes with 0.6957 F 1 and 0.8967 accuracy and estimate the top 10 ICD-9 categories with 0.7233 F 1 and 0.8588 accuracy. The Long Short-Term Memory (LSTM) deep learning algorithm, a modification of the RNN on automatic ICD, produces an accuracy of 0.61 on the MIMIC III dataset and 0.97 on the Chinese TPD dataset. Another deep learning method is Medical Topic Mining (MTM), with an accuracy of 0.96 with the ICD-10 internal dataset and 0.93 with the MIMIC-III dataset (Teng et al., 2020).

2. Other methods

Methods other than deep learning used in automatic ICD research include the Cross-lingual Language Model (XLM) algorithm with an accuracy of 0.76 (Silvestri et al., 2020), the Longest Common Subsequence (LCS) algorithm and semantic similarity with an accuracy of 0.81 (Yun Zhi Chen et al., 2017), Hyperbolic and Co-graph Representation method (HyperCore) with accuracy results on the MIMIC-II dataset 0.47 and MIMIC-III 0.51 (Cao et al., 2020). Research (Almagro et al., 2019) using a cross-lingual approach based on Machine Translation methods to code death certificates with ICD-10 using supervised learning resulted in an accuracy of 0.90. The Algorithm Support Vector Machine (SVM) in the study (Koopman et al., 2015) influenced the automatic ICD-10 classification of cancers from free-text death certificates with an accuracy of 0.94. Meanwhile, (Schafer & Friedrich, 2019) combining the Support Vector Machines (SVM), FastText, and Unified Medical Language System (UMLS) algorithm with the MIMIC-III dataset can produce an accuracy of 0.62. The semantics analysis method with an accuracy of 0.60 in the study (Ping Chen et al., 2010) included dependency parsing of clinical records obtained from testing and training data sets and calculating semantic matching scores.

CONCLUSION

Automatic ICD research to improve the accuracy of diagnosis coding has been carried out and is still a challenge today. Most of the datasets used are the public MIMIC-III dataset, while the automatic ICD classification is the current research trend with deep learning. The deep learning algorithms that are widely used include CNN, RNN, and LSTM. The resulting accuracy based on the dataset and classification method is very diverse. There have been many attempts to improve accuracy by modifying and improving the classification method used.

Future work still has many opportunities to contribute and improve the correct classification method to improve automatic ICD accuracy. Another opportunity that can be done is to improve the preprocessing stage and feature of the extraction dataset and the formation of new datasets to improve classification accuracy in determining automatic diagnosis codes.

REFERENCES

Almagro, Mario, Martínez, Raquel, Montalvo, Soto, & Fresno, Víctor. (2019). A cross-lingual approach to automatic ICD-10 coding of death certificates by exploring machine translation. *Journal of Biomedical Informatics*, *94*(May), 103207. https://doi.org/10.1016/j.jbi.2019.103207

Atutxa, Aitziber, de Ilarraza, Arantza Díaz, Gojenola, Koldo, Oronoz, Maite, & Perez-de-Viñaspre, Olatz. (2019). Interpretable deep learning to map diagnostic texts to ICD-10 codes. *International Journal of Medical Informatics*, *129*(May), 49–59. https://doi.org/10.1016/j.ijmedinf.2019.05.015

Blanco, Alberto, Perez, Alicia, & Casillas, Arantza. (2020). Extreme multi-label ICD classification: sensitivity to hospital service and time. *IEEE Access*, *8*, 1–1. https://doi.org/10.1109/access.2020.3029429

Cao, Pengfei, Chen, Yubo, Liu, Kang, Zhao, Jun, Liu, Shengping, & Chong, Weifeng. (2020). HyperCore: Hyperbolic and Co-graph Representation for Automatic ICD Coding. *Proceedings Ofthe 58th Annual Meeting Ofthe Association for Computational Linguistics*, 3105–3114. https://doi.org/10.18653/v1/2020.acl-main.282

Chen, Ping, Barrera, Araly, & Rhodes, Chris. (2010). Semantic analysis of free text and its application on automatically assigning ICD-9-CM codes to patient records. *Proceedings of the 9th IEEE International Conference on Cognitive Informatics, ICCI 2010*, 68–74. https://doi.org/10.1109/COGINF.2010.5599783

Chen, Yun Zhi, Lu, Hui Juan, & Li, Lan Juan. (2017). Automatic ICD-10 coding algorithm using an improved longest common subsequence based on semantic similarity. *PLoS ONE*, *12*(3), 1–17. https://doi.org/10.1371/journal.pone.0173410

Chen, Yuwen, & Ren, Jiangtao. (2019). Automatic ICD code assignment utilizing textual descriptions and hierarchical structure of ICD code. *Proceedings - 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019*, 348–353. https://doi.org/10.1109/BIBM47256.2019.8983078

Deng, Li, & Yu, Dong. (2013). Deep learning: Methods and applications. *Foundations and Trends in Signal Processing*, *7*(3–4), 197–387. https://doi.org/10.1561/2000000039

Huang, Jinmiao, Osorio, Cesar, & Sy, Luke Wicent. (2019). An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes. *Computer Methods and Programs in Biomedicine*, *177*, 141–153. https://doi.org/10.1016/j.cmpb.2019.05.024

Koopman, Bevan, Zuccon, Guido, Nguyen, Anthony, Bergheim, Anton, & Grayson, Narelle. (2015). Automatic ICD-10 classification of cancers from free-text death certificates. *International Journal of Medical Informatics*, *84*(11), 956–965. https://doi.org/10.1016/j.ijmedinf.2015.08.004

Li, Min, Fei, Zhihui, Zeng, Min, Wu, Fang Xiang, Li, Yaohang, Pan, Yi, & Wang, Jianxin. (2019). Automated ICD-9 coding via a deep learning approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *16*(4), 1193–1202. https://doi.org/10.1109/TCBB.2018.2817488

Schafer, Henning, & Friedrich, Christoph M. (2019). UMLS mapping and Word embeddings for ICD code assignment using the MIMIC-III intensive care database. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 6089–6092. https://doi.org/10.1109/EMBC.2019.8856442

Shi, Haoran, Xie, Pengtao, Hu, Zhiting, Zhang, Ming, & Xing, Eric P. (2017). Towards automated ICD coding using deep learning. *ArXiv*, 1–11.

Silvestri, Stefano, Gargiulo, Francesco, Ciampi, Mario, & De Pietro, Giuseppe. (2020). Exploit Multilingual Language Model at Scale for ICD-10 Clinical Text Classification. *Proceedings - IEEE Symposium on Computers and Communications*, *2020-July*. https://doi.org/10.1109/ISCC50000.2020.9219640

Sonabend W, Aaron, Cai, Winston, Ahuja, Yuri, Ananthakrishnan, Ashwin, Xia, Zongqi, Yu, Sheng, & Hong, Chuan. (2020). Automated ICD coding via unsupervised knowledge integration (UNITE). *International Journal of Medical Informatics*, *139*(February), 104135. https://doi.org/10.1016/j.ijmedinf.2020.104135

Teng, Fei, Ma, Zheng, Chen, Jie, Xiao, Ming, & Huang, Lufei. (2020). Automatic Medical Code Assignment via Deep Learning Approach for Intelligent Healthcare. *IEEE Journal of Biomedical and Health Informatics*, *24*(9), 2506–2515. https://doi.org/10.1109/JBHI.2020.2996937

Yu, Ying, Li, Min, Liu, Liangliang, Fei, Zhihui, Wu, Fang Xiang, & Wang, Jianxin. (2019). Automatic ICD code assignment of Chinese clinical notes based on multilayer attention BiRNN. *Journal of Biomedical Informatics*, *91*(February), 103114. https://doi.org/10.1016/j.jbi.2019.103114

Zeng, Min, Li, Min, Fei, Zhihui, Yu, Ying, Pan, Yi, & Wang, Jianxin. (2019). Automatic ICD-9 coding via deep transfer learning. *Neurocomputing*, *324*, 43–50. https://doi.org/10.1016/j.neucom.2018.04.081