REVIEW ARTICLE

Artificial intelligence for prediction of International Classification of Disease codes

Kaitlyn Wallace[1], Jakir Hossain Bhuiyan Masud[2]

[1]University of Chicago, Illinois, USA
[2]Public Health Informatics Foundation, Dhaka, Bangladesh

Correspondence to: Jakir Hossain Bhuiyan Masud, Email: jakir_msd@yahoo.com; jakir@phifbd.org

ABSTRACT

Background: The automatic coding of electronic medical records with ICD (International Classification of Diseases) codes is an area of interest due to its potential in improving efficiency and streamlining processes such as billing and outcome tracking. Artificial intelligence (AI), particularly convolutional neural networks (CNN) have been suggested as a possible mechanism for automatic coding. To this end, a rapid review has been undertaken in order to assess the current use of CNN in predicting ICD codes from electronic medical records.

Methods: After screening PubMed, IEEE Xplore, Scopus, and Google Scholar, 11 studies were analyzed for the use of CNN in predicting ICD codes. We used artificial intelligence and ICD prediction as keywords in the search strategy.

Results: The analysis yielded a recommendation to further explore and research CNN frameworks as a promising lead to automatic ICD coding when paired with word embedding and/or neural transfer learning, while keeping research open to a wide variety of AI techniques.

Conclusion: CNN frameworks are promising for the prediction of ICD codes from clinical notes.

Keywords: artificial intelligence, prediction, ICD, convolutional neural networks

INTRODUCTION

The problem of automatic coding of clinical notes is one that has plagued the fields of healthcare and computer science since the 1990s.[1] Clinical notes are free text descriptions of patient encounters that are an important part of electronic medical records (EMR). Current practice requires manual annotation of clinical notes with ICD codes in order to streamline billing, track outcomes, calculate statistics, and perform public health surveillance.[2] Manual annotation, however, can be time consuming and error-prone, and has inspired efforts to create an automatic coding system. So far, these efforts have been largely unsuccessful due to the large amount of available ICD codes (68,000 in ICD-10 alone), as well as high data heterogeneity.[3] To this end, recent efforts have focused on AI, and more specifically with Convolutional Neural Networks (CNN) frameworks due to their ability to learn without human supervision and computational efficiency.[4]

The ICD code system enables the classification and identification of diseases and health conditions based on agreed-upon criteria. It allows healthcare providers to accurately document and report diagnoses, facilitating appropriate treatment and care planning. The ICD code system is essential for standardized communication, disease classification, epidemiological monitoring, reimbursement processes, research endeavors, and health information management. Its importance lies in facilitating accurate documentation, data analysis, and the provision of quality healthcare services.

While research on ICD prediction using AI has made significant progress, there are still several research gaps that need to be addressed. One of the key challenges is the availability of high-quality labeled datasets for training and evaluating AI models for ICD prediction. AI models used for ICD prediction often operate as

Received: 10 May 2023; Revised version receiving: 22 Jun 2023; Accepted: 22 Jun 2023; Published online: 26 Jun 2023
Supplemental file, and peer review and author response: available at DOI: https://doi.org/10.3329/bsmmuj.v16i2.67235

HIGHLIGHTS

1. A convolutional neural network was proposed to predict ICD codes extensively.
2. This study explores the decision-making method for predicting ICD.
3. AI methods are employed to assess the relationship between the different data sources for predicting ICD Code.

black boxes, making it challenging to understand how they arrive at their predictions. In the medical domain, explainability and interpretability are crucial for gaining trust and acceptance from healthcare professionals. Many AI models for ICD prediction are trained and evaluated on specific datasets or within specific clinical settings. Ensuring the generalizability of these models across different healthcare systems, specialties, and populations is crucial for real-world applicability. Further research is needed to investigate transfer learning techniques and domain adaptation methods to make AI models more robust and adaptable to diverse healthcare contexts. This rapid review analyzes the current literature on the use of CNN frameworks to predict ICD codes from EMRs.

METHODS

Four online databases (PubMed, IEEE Xplore, Google Scholar and Scopus) were searched with the keyword "predicting ICD code from clinical notes using CNN." Both authors searched and reviewed the data. The inclusion criteria were applied as follows:

Criteria 1, the focus of the inclusion criteria was on ICD-10 or ICD-9 code prediction using clinical data and/or notes and CNN. It did not include NLP focus, illness progression forecasting, comorbidity focus, or diagnostic prediction without ICD code. Criteria 2, we selected the study based on model development and/or testing. We specifically excluded the papers (systematic reviews, opinions, short communications, case reports, commentaries, research letters, narrative reviews, statement articles, news reports, books, overview pieces, and articles) from our study that were not accessible despite contacting the authors. Criteria 3, English was the reported language. The initial search

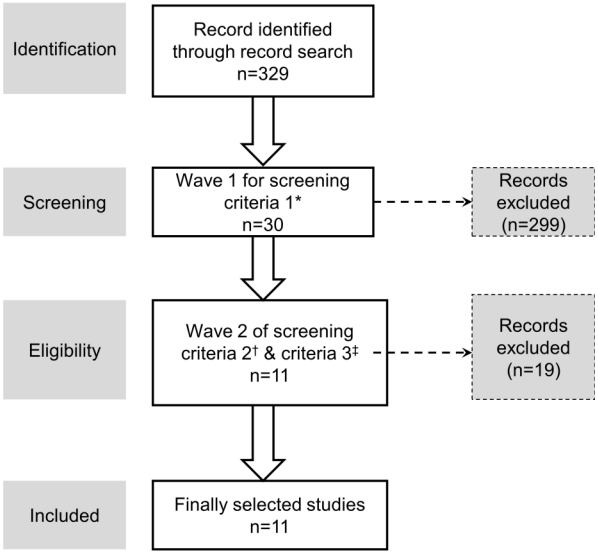

FIGURE 1 Two-wave screening process, yielding 11 studies

yielded a total of 329 records, which were screened in waves **(FIGURE 1)**. The first wave excluded 299 titles on the basis of inclusion criteria 1 (focus), yielding 30 titles. Applying criteria 2 and 3 (study design, and language) excluded a further 19 titles yielding 11 studies which met all three inclusion criteria. Our outcome variable is ICD code. The CNN model performance was measured by accuracy, AUC, precision, recall, and F-measure.

RESULTS

This rapid review yielded 11 studies on the prediction of ICD code from clinical notes using CNN. The characteristics of these papers are described in **TABLE 1**. These studies can be categorized into four general categories: development and testing of specific models, models which combine CNN use with other techniques, and papers which compare CNN-based models with other frameworks.

There are four papers that can be characterized as developing and/or testing specific models. Most notable was the use of MultiResCNN (Multi-Filter Residual Convolutional Neural Network) by Li and Yu.[5] The MultiResCNN model, composed of an input layer, a

TABLE 1 Comparison of model performances

Reference number	Author & Year of publication	Database	Type of Model	Key Findings
5	Fei Li, 2019	Google Scholar	MultiResCNN	Macro-AUC=0.91, Micro-AUC=0.986, Macro-F1=0.085, Micro-F1=0.552
10	Anthony Rios, 2019	PubMed	CNN with supplemental neural transfer learning	Improves F measures by >8% compared to a simple CNN
11	Chin Lin, 2017	PubMed	Word Embedding combined with CNN	Achieves a higher test accuracy (mean AUC=0.9696, mean F-measure=0.9086) than NLP-based approaches (mean AUC range: 0.8183-0.957, mean F-measure range: 0.5050-0.8739)
12	Jinmiao Huang, 2019	Google Scholar	Comparison of CNN, RNN, LSTM, GRU	LSTMs and GRUs achieve better performance metrics than CNNs, but have longer training times
13	Amitabha Karmak-ar , 2018	Google Scholar	CNN, CNN with Attention, LSTM, Hierarchical Models	CNN and CNN with Attention models achieve best F1 score (F1=79.2 and F1=78.2)
6	Min Li, 2018	IEEE Xplore	DeepLabeler (CNN framework with "document to vector" technique)	DeepLabeler outperformed hierarchy-based and flat-SVM by at least 14% MIMIC-II: micro F = 0.335, MIMIC-III: micro F = 0.408
9	Jakir Masud, 2020	Scopus	Word2vector CNN	Best performance: Precision=69%, Recall=89%, F-measure=78%
3	Tal Baumel, 2018	Google Scholar	Compares: SVM (support vector machine), CBOW (continuous bag-of-words), CNN, HA-GRU	HA-GRU performs best in 3 out of 4 performance metrics; CNN performs best in 1 out of 4 performance metrics
1	James Mullenbach, 2018	Google Scholar	CAML and DR-CAML models	On full MIMIC-III set, CAML performs best in 3 out of 4 perfor-mance metrics; DR-CAML performs best in 1 out of 4 performance metrics
7	Christy Li, 2017	Google Scholar	CNN for learning of semantic features from unstructured text input	Tuned CNN outperforms baseline models Accuracy = 96.11, F1 = 80.48% (weighted)
8	Xiaozheng Li, 2019	Google Scholar	CNN framework with word segmentation, word embedding, and model training	One-layer CNN outperforms other models

multi-filter convolutional layer, a residual convolutional layer, an attention layer, and an output layer outperformed all other models tested (SVM (Support Vector Machines), HA-GRU (Hierarchical Attention Gated Recurrent Unit), CAML (Convolutional Attention network for Multi-Label Classification), DR-CAML (Description Regularized CAML)) in prediction of ICD-9 codes from the MIMIC-III dataset in all performance metrics, with an macro-AUC of 0.85, a micro-AUC of 0.968, a macro-F1 of 0.052, and a micro-F1 of 0.464.[5]

The MultiResCNN model outperformed the next models in this category– CAML (Convolutional Attention Network for Multi-Label Classification) and DR-CAML (Description Regularized CAML) proposed by Mullenbach et al. in 2018.[1] Mullenbach et. al developed CAML and DR-CAML from a traditional CNN augmented with an attention mechanism and embedding of label descriptions. In this case, the CAML model (without augmentation with embedding of label descriptions) outperformed the DR-CAML model in 3 out of 4 performance metrics on the full MIMIC-III dataset, with DR-CAML achieving only a slightly higher macro-AUC. Both models, however, outperformed the other models tested (including a traditional CNN and the Bi-GRU (Bi-Gated Recurrent Unit).[1]

Though not directly compared to the MultiResCNN, CAML, and DR-CAML models, Li and Fei et al.'s model DeepLabeler warrants discussion as well; by using a CNN combined with a "Document to Vector" technique, DeepLabeler achieves a micro F-measure of 0.335 on the MIMIC-II dataset and a micro F-measure of 0.408 on the MIMIC-III dataset, outperforming both hierarchy-based SVM and flat-SVM models by at least 14%.[6] Li and Fei et al. also concluded that the most effective component of their model was the convolutional neural network, supporting the hypothesis that CNN frameworks may be particularly useful for ICD coding of clinical notes.

Secondarily, five papers were included which tested CNN frameworks in conjunction with other AI techniques. Li and Konomis et al., for example, used a

CNN framework optimized to learn semantic features from unstructured textual input.[7] This model achieved a higher F1 score on all top 10 most common disease classes in the MIMIC dataset in comparison with SVM, MLP (multilayer perceptron), LR (logistic regression), and RF (random forest) models, and achieved the best results on all other performance metrics in aggregate except one (the LR-based model achieved a higher FNR). This study is limited by the fact that the framework was only used to predict the 10 most common ICD codes in the MIMIC-III rather than all possible codes.

Similarly, Li and Wang et al. used a CNN framework augmented with word segmentation, word embedding, and model training. Their fine-tuned single-layer CNN outperformed all other tested models (including LSTMs, GRUs, and more), but is limited by its application in Chinese EMRs (non-English language).[8]

The final three papers in this category describe the potential of the use of CNN for ICD code prediction. Masud and Lin describe enhanced performance of physicians using AI-assisted ICD-10 code prediction through the implementation of word embedding word 2 vector CNN.[9] The application is limited by the use of medication list rather than unstructured clinical notes. The other two papers (Rios and Kavaluru and Lin et al.) describe the potential for combining CNN with other AI methods such as word embedding and neural transfer learning, both of which enhance the performance of a CNN in prediction ICD-10 codes.[10, 11] Combining word embedding with a CNN provides a higher test accuracy (mean AUC=0.9696, mean F-measure=0.9086) than the competing NLP (natural language processing) approach (mean AUC range: 0.8183-0.9571; mean F-measure range: 0.5050-0.8739), while supplementing CNN with neural language transfer (in this case, supplemented with EMR data from PubMed indexed biomedical research abstracts) increases macro and micro F-scores by >8% compared to other transfer learning methods.

Finally, three papers can be described as "comparison papers," in which multiple models are tested against one another. For example, Huang et al. tested a variety of models, including RNN-based frameworks, CNN-based frameworks, LSTMs and GRUs, finding that CNN -based frameworks were significantly outperformed by LSTMs and GRUs, though Huang et al. did note that CNN-based frameworks had a significantly shorter training time.[12]

On the other hand, Karmakar found both simple CNNs and CNNs with attention to be more promising than and to outperform LSTMs and hierarchical models in F1 scores when classifying MIMIC clinical notes into Level 1 ICD-9 codes.[13] Similarly, Baumel et al. evaluated the performance of multiple types of AI frameworks, finding a simple CNN to be the second highest performing model behind HA-GRU (Hierarchical Attention-bidirectional Gated Recurrent Unit) with micro-F values of 33.25% and 40.72% for ICD-9 codes in MIMIC (Medical Information Mart for Intensive Care)-II and MIMIC-III, respectively, and 46.40% and 52.64% for rolled-up ICD-9 codes for MIMIC-II and MIMIC-III, respectively, in contrast to HA-GRU's micro -F value of 36.60% and 40.52% for ICD codes in MIMIC -II and MIMIC-III, and 53.86% and 55.86% for rolled-up ICD-9 codes for MIMIC-II and MIMIC-III, respectively.[3]

DISCUSSION

The major finding of this rapid review is the current limitations on AI use for prediction of ICD codes. A focus on CNN-based frameworks returned a relatively small number of studies (eleven studies were included from the final screening). These studies returned some disagreement on the effectiveness of the use of CNN frameworks for the prediction of ICD codes from clinical notes. The majority of the studies indicate that CNN frameworks are promising for this purpose (Li and Yu, Li and Fei et al., Li and Konomis et al., Rios and Kavaluru, Karmarkar, in particular).[5, 6, 7, 10, 13] However, several studies also contracted this claim; in particular, Huang et al. found CNNs to be significantly outperformed by LSTM- and GRU-based frameworks, and Baumel et al.'s HA-GRU model achieved better performance metrics than a simple CNN framework.[12, 3] These conflicting results suggest that while future work should likely focus on CNN-based models, other

frameworks (in particular, NLP-based frameworks), should not be ruled out entirely.

The second major finding of this rapid review is the importance of augmentation. In all successful models, the CNN-based framework was supported by other techniques such as word embedding, neural transfer learning, attention mechanisms, "document to vector" techniques, word segmentation, and more, all of which improved the performance of a simple CNN for the prediction of ICD codes from clinical notes. These studies indicate that CNN frameworks (particularly those supplemented with word embedding and/or neural transfer learning) have the potential to surpass NLP-based frameworks in performance.

Therefore, the strongest recommendation that can be made on the data available is continuing research into CNN frameworks supplemented and/or supported by other techniques such as word embedding and neural transfer learning, while comparing and/or remaining cognizant of other approaches to the ICD coding problem.

This analysis is limited by the small number of studies that were able to be included by discussion of all three categories. This limits the recommendations that can be made on the available data. Secondarily, for the purpose of time constraints, systematic reviews, statement articles, case-studies, etc. were excluded from study; this may have caused some studies to have been excluded from the current rapid review. Finally, this analysis is limited by the basic AI knowledge of the reviewer; though the data extraction was generally quickly performed, some descriptions of the type of AI models may have been oversimplified or underexplained in the interest of accuracy.

Conclusion

Although CNN frameworks are promising for the prediction of ICD codes from clinical notes, the current research is limited. Though the studies included are promising, there are conflicting results on whether or not CNN frameworks are the only and/or best framework for this task. While these studies also offer leads in the areas of augmentation techniques such as word embedding and neural transfer learning, further research is necessary to expand and solidify these leads, as well as providing stronger evidence for the high performances of CNN frameworks in predicting ICD codes from clinical notes.

Acknowledgments

We thank the reviewers for their comments.

Author Contributions

Conception and design: JHBM, KW. Acquisition, analysis and interpretation of data: KW, JHBM. Manuscript drafting and revising it critically: JHBM, KW. Approval of the final version of the manuscript: JHBM, KW. Guarantor accuracy and integrity of the work: JHBM, KW.

Funding

This study did not receive any funding.

Conflict of Interest

The authors declare no conflict of interest.

Ethical Approval

No ethical approval was taken for conducting this study.

ORCID iD

Jakir Hossain Bhuiyan Masud http://orcid.org/0000-0002-4542-3862

REFERENCES

1. Mullenbach J, Wiegreffe S, Duke J, Sun J, Eisenstein J. Explainable prediction of medical codes from clinical text. arXiv preprint arXiv:1802.05695. 2018 Feb 15. DOI: https://doi.org/10.48550/arXiv.1802.05695
2. (From the AAP Division of Health Care Finance. ICD-10-CM used to document diagnoses but also affects payment. https://www.aappublications.org/news/2019/02/06/coding020619. Published July 29, 2020. Accessed August 4, 2022.
3. Baumel T, Nassour-Kassis J, Cohen R, Elhadad M, Elhadad N. Multi-label classification of patient notes a case study on ICD code assignment. arXiv preprint arXiv:1709.09587. 2017 Sep 27. DOI: https://doi.org/10.48550/arXiv.1709.09587
4. Dertat A. Applied deep learning—Part 4: Convolutional neural networks. 2017. URL: https://towardsdatascience. com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2. 2017.. Accessed August 4, 2020.
5. Li F, Yu H. ICD coding from clinical text using multi-filter residual convolutional neural network. Inproceedings of the AAAI conference on artificial intelligence 2020 Apr 3 (Vol. 34, No. 05, pp. 8180-8187). DOI: https://doi.org/10.1609/aaai.v34i05.6331.
6. Li M, Fei Z, Zeng M, Wu FX, Li Y, Pan Y, Wang J. Automated ICD-9 coding via a deep learning approach. IEEE/ACM transactions on computational biology and bioinformatics. 2018 Mar 20;16(4):1193-202. DOI: https://doi.org/10.1109/TCBB.2018.2817488.
7. Li C, Konomis D, Neubig G, Xie P, Cheng C, Xing E. Convolutional neural networks for medical diagnosis from

admission notes. arXiv preprint arXiv:1712.02768. 2017 Dec 6. DOI: https://doi.org/10.48550/arXiv.1712.02768

8. Li X, Wang H, He H, Du J, Chen J, Wu J. Intelligent diagnosis with Chinese electronic medical records based on convolutional neural networks. BMC bioinformatics. 2019 Dec;20:1-2. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2617-8

9. Masud JH, Lin MC. Predicting Diagnosis Code from Medication List of an Electronic Medical Record Using Convolutional Neural Network. InMIE 2020 Jun 16 (pp. 1355-1356). DOI: https://doi.org/10.3233/SHTI200439

10. Rios A, Kavuluru R. Neural transfer learning for assigning diagnosis codes to EMRs. Artificial intelligence in medicine. 2019 May 1;96:116-22. DOI: https://doi.org/10.1016/j.artmed.2019.04.002.

10. Lin C, Hsu CJ, Lou YS, Yeh SJ, Lee CC, Su SL, Chen HC. Artificial intelligence learning semantics via external resources for classifying diagnosis codes in discharge notes. Journal of medical Internet research. 2017 Nov 6;19(11):e380. DOI: https://doi.org/10.2196/jmir.8344

12. Huang J, Osorio C, Sy LW. An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes. Computer methods and programs in biomedicine. 2019 Aug 1;177:141-53. DOI: https://doi.org/10.1016/j.cmpb.2019.05.024

13. Karmakar A. Classifying medical notes into standard disease codes using machine learning. arXiv preprint arXiv:1802.00382. 2018 Feb 1. DOI: https://doi.org/https://doi.org/10.48550/arXiv.1802.00382.