AnEMIC: A Framework for Benchmarking ICD Coding Models

Juyong Kim[*1], **Abheesht Sharma**[*2], **Suhas Shanbhogue**[*2], **Pradeep Ravikumar**[1], and **Jeremy C. Weiss**[3]

[1]Machine Learning Department, Carnegie Mellon University
[2]Birla Institute of Technology & Science, Pilani – Goa Campus
[3]National Library of Medicine, National Institutes of Health
{juyongk,pradeepr}@cs.cmu.edu
{f20171014,f20170769}@goa.bits-pilani.ac.in
jeremy.weiss@nih.gov

Abstract

Diagnostic coding, or ICD coding, is the task of assigning diagnosis codes defined by the ICD (International Classification of Diseases) standard to patient visits based on clinical notes. The current process of manual ICD coding is time-consuming and often error-prone, which suggests the need for automatic ICD coding. However, despite the long history of automatic ICD coding, there have been no standardized frameworks for benchmarking ICD coding models.

We open-source an easy-to-use tool named *AnEMIC*, which provides a streamlined pipeline for preprocessing, training, and evaluating for automatic ICD coding. We correct errors in preprocessing by existing works, and provide key models and weights trained on the correctly preprocessed datasets. We also provide an interactive demo performing real-time inference from custom inputs, and visualizations drawn from explainable AI to analyze the models. We hope the framework helps move the research of ICD coding forward and helps professionals explore the potential of ICD coding. The framework and the associated code are available here.

1 Introduction

Diagnostic coding is the task of assigning alphanumeric codes to diagnoses and procedures after a patient visits a healthcare provider. These codes are typically specified by a medical classification standard called the International Classification of Diseases (ICD). Diagnostic coding, or ICD coding, is an integral component of medical billing, and integral to claims paid by health insurance carriers. The diagnostic coding process alone accounts for approximately 21% of medical administrative costs in the US (Tseng et al., 2018). During this process, a professional coder reviews the patient's medical records, including clinical narratives, and manually selects ICD codes. Since the task requires in-depth clinical knowledge and understanding of medical records, and importantly, due to the fact that there are a large number of ICD codes, the task is labor-intensive and error-prone (Manchikanti, 2002).

These difficulties motivate the need for automatic ICD coding systems which perform diagnosis classification given a patient's health record (Kaur et al., 2021; Yan et al., 2022). This has been the subject of considerable research, with some of the early work dating back to the 1990s (Larkey and Croft, 1996), to more recent deep neural NLP approaches. There are a few outstanding and major challenges in the diagnostic coding task. Firstly, the label space, the set of all ICD codes, is large, and the label distribution is highly imbalanced. Secondly, the input text, i.e., the discharge summaries, is noisy and can contain abstruse medical terms, lesser-known abbreviations, misspelt words, etc. Also, they are much longer than what most state-of-the-art models take as input.

Along with those challenges, the absence of a benchmark has impeded the progress of research. Due to privacy restrictions that limit access to even publicly available clinical databases, researchers have to create datasets manually from these, and this results in discrepancies in the actual datasets used in individual papers. For instance, the label set of MIMIC-III top-50 dataset varies among the literature, and some of them are even used incorrectly. Inconsistency in processing the dataset and the inevitable errors introduced as a result of this makes it hard to compare different methods.

In this paper, we introduce a framework for benchmarking automatic ICD coding with the MIMIC clinical database. We name our framework *AnEMIC*, for **An E**rror-reduced **M**IMIC **I**CD **C**oding benchmark. To the best of our knowledge, AnEMIC is the first attempt to collate and benchmark different deep learning approaches for automatic ICD coding with a configurable pipeline.

[*] Equal contribution.

Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 109 - 120

Our contributions can be summarized as follows:

- We provide a pipeline covering the entire process of automatic ICD coding, including preprocessing, training, and evaluation. The whole process is easily configurable with the use of YAML files. We additionally provide key deep learning-based ICD coding models.
- We correct errors in the most widely used datasets and provide benchmark results of the key models on the new datasets.
- We open-source an easy-to-use interactive demo that enables researchers to test their models on custom inputs and visualize input attribution scores for explainability.

The remainder of the paper is organized as follows. In Section 2, we discuss popular automatic ICD coding approaches and datasets. Section 3 details our approaches for preprocessing, training, evaluation, and our demo application. In Section 4, we perform a quantitative and qualitative analysis of AnEMIC. Finally, we conclude with discussion and future work in Section 5.

2 Related Work

2.1 ICD Coding

Over the history of automatic diagnosis coding, approaches have ranged from classical methods such as rule-based approaches (Farkas and Szarvas, 2008), traditional ML models such as SVMs (Perotte et al., 2014), to more recent Deep Learning-based methods. A neural network-based approach was first attempted by Prakash et al. (2017). A prominent deep learning approach is CAML (Mullenbach et al., 2018), which uses a CNN encoder with a unique per-label attention mechanism. Since CAML, there have been many other CNN and RNN-based approaches (Yu et al., 2019; Vu et al., 2020). A few notable CNN based approaches include using dilated convolutional layers (Ji et al., 2020) and multi-filter convolutional layers (Li and Yu, 2020; Luo et al., 2021).

Additionally, researchers have leveraged the hierarchy of ICD codes (Cao et al., 2020; Xie et al., 2019), used external knowledge sources like Wikipedia (Bai and Vucetic, 2019), and knowledge graphs such as UMLS (Yuan et al., 2022) and Freebase (Teng et al., 2020), etc. More recently, there has been an effort to use Transformer-based language models pretrained on clinical datasets, albeit without much success (Pascual et al., 2021; Zhang et al., 2020; Ji et al., 2021). Instead, using a few Transformer encoder layers trained from scratch has proven to be more effective (Biswas et al., 2021).

Kaur et al. (2021) and Yan et al. (2022) perform extensive literature reviews of automatic ICD coding approaches. The reader is referred to these surveys for a more detailed description of various architectures and approaches.

2.2 ICD Coding Datasets and Benchmark

Typical ICD coding dataset consists of discharge summaries and the corresponding sets of ICD codes. There are many ICD coding datasets in various languages, but not all are publicly available. The most widely used datasets are from MIMIC-III[1] and MIMIC-II[2] databases. The MIMIC-III clinical database (Johnson et al., 2016) is a collection of medical records from an intensive care unit (ICU) at a hospital between 2001 and 2012. MIMIC-III consists of multiple tables containing diagnosis, procedures, clinical notes, etc., and each patient admission is indicated with an `HADM_ID` identifier. MIMIC-II is a subset of the MIMIC-III dataset and contains medical records between 2001 and 2008 [3].

CAML (Mullenbach et al., 2018) published the preprocessing code of their MIMIC-III full and top-50 datasets, and since then, these have been the most widely used datasets. We correct some errors in preprocessing of CAML and make the process easily configurable. Also, compared to a leaderboard that only manages reported performance, our work provides a framework for benchmarking, i.e., users can run the code to reproduce the results and further perform research on top of it.

3 ICD Coding Benchmark

AnEMIC has been designed so that researchers can easily configure the overall process with config files and therefore, easily start research on ICD coding with minimal code. Also, the architecture has modularity at the center of its design so that researchers can replace one module with another or with their own implementation. Such design enables easy comparison between models and reduces burden while developing new models.

[1] https://physionet.org/content/mimiciii/1.4/

[2] https://archive.physionet.org/mimic2/

[3] There is also the recently released MIMIC-IV database, but clinical notes for this are currently not yet available.

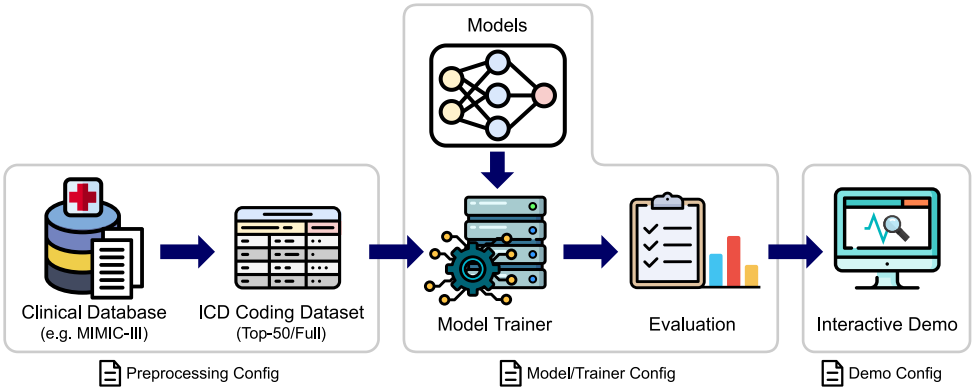

Figure 1: The ICD coding benchmark pipeline of AnEMIC. We provide a pipeline covering the entire process of ICD coding. All steps in the pipeline can be easily configured with YAML files.

Our system also provides an interactive demo for visualizing model predictions with input attribution scores. This demo will help users analyze the performance and interpretability of their models.

In the following subsections, we explain each stage in the pipeline. From now on, we will focus on ICD coding dataset from MIMIC-III since it is the most widely used dataset for this task. Figure 1 illustrates the overall pipeline.

3.1 Data Preprocessing

The first step of the pipeline is to preprocess the available clinical dataset, i.e., the MIMIC-III database. As with other parts of the pipeline, we specify preprocessing-related options in a YAML config file.

Many of the preprocessing steps are inspired by CAML's preprocessing pipeline. However, an important observation to be noted here is that **there are errors in CAML's preprocessing pipeline**. Unfortunately, many subsequent works use CAML's code, and hence, the results obtained by most papers are on the incorrectly preprocessed dataset. This will be discussed later in this subsection and Appendix A.

3.1.1 ICD Code Preprocessing

In the MIMIC-III database, the `DIAGNOSES_ICD` and `PROCEDURES_ICD` tables contain the ICD-9 diagnosis and procedure codes, respectively, of every admission. Since MIMIC-III has ICD-9 codes without the period punctuation (e.g. 4019 instead of 401.9), we reformat those ICD codes to their original format adopting the method of CAML, and use them as labels. ICD-9 codes can have leading and trailing zeros, so care must be taken to retain them when processing. However, in CAML's preprocessing code, some of ICD codes are implicitly treated as integer or floating point numbers[4], resulting in an incorrect set of ICD-9 labels. While correcting this error, we provide an option `incorrect_code_loading` to reproduce the behavior of CAML for researchers who want to make a comparison with previous works.

In addition to the above option, we also provide an option `code_type` to use either diagnosis, procedure, or both types of ICD codes. We set "both" as the default.

3.1.2 Clinical Note Preprocessing

From the `NOTEEVENTS` table of MIMIC-III containing clinical notes in various categories, we select notes belonging to the `Discharge_Summary` category. We provide several options of standard NLP preprocessing for the discharge summary. These can be turned on/off from the config file.

- Convert text to lowercase.
- Remove punctuation marks using `\w+` as the RegEx expression, i.e., retain only alphanumeric characters.
- Either remove numeric characters, or replace all numeric characters with "n".
- Remove stopwords; we use the list of stopwords provided by NLTK, and add common medical terms like "hospital", "admission", "history", etc. to the list.
- Stem or lemmatize the text; we provide popular choices for these such as "WordNet Lemmatizer" and "Porter Stemmer".
- Truncate the text to a maximum length.

After note preprocessing, we build the vocabulary and train a Word2Vec model on preprocessed discharge summaries using the Gensim library (Řehůřek and Sojka, 2010). Word2Vec embeddings are used to initialize the embedding layers of models.

[4] Due to not specifying data types when loading tables

3.1.3 Top-k Codes and Data Splitting

Many works report results on two datasets – "MIMIC-III full" and "MIMIC-III top-50". The latter contains the top-50 frequent ICD codes as labels and examples with at least one of these labels.

An important point to note is that MIMIC-III has some duplicate ICD codes, i.e., an ICD code can be repeated multiple times in one admission. These duplicate codes need to be removed when counting the ICD code occurrence. This is another source of error in CAML's code: they do not remove the duplicate codes while counting the ICD codes occurrence, resulting in a change in the top-50 ICD codes. While we correctly select the top-50 ICD codes, we also provide an option `count_duplicate_codes` to reproduce the behavior of CAML.

For data splitting, we use the splits of `HADM_IDs` provided by CAML. They provide separate sets of splits for the full and top-50 datasets, and the split for top-50 dataset has substantially smaller number of examples. To make full use of MIMIC-III, we use the splits of the CAML's full dataset for both versions of our dataset.

As a result of data preprocessing, we have four main variants of the dataset – "MIMIC-III full", "MIMIC-III top-50", "MIMIC-III full (old)", and "MIMIC-III top-50 (old)". Here "(old)" refers to the CAML variants.

3.2 Supported Models

This subsection describes the models we provide in the framework and the criteria for choosing models. To provide researchers with good baselines for ICD coding research, we selected models based on novelty or superior performance. For now, we have chosen a subset of models for which the code is publicly available, but we do plan on implementing other approaches in the near future which have not been open-sourced. The models and the trainer are based on PyTorch.

The models currently supported by the framework are as follows:

- CAML (Mullenbach et al., 2018) is a landmark model in automatic ICD coding which uses a label attention layer. We also implement the vanilla CNN model in the paper and refer to it as CNN.
- MultiResCNN (Li and Yu, 2020) uses multiple CNNs with different filter sizes in parallel.
- DCAN (Ji et al., 2020) uses dilated convolutional layers for ICD coding.

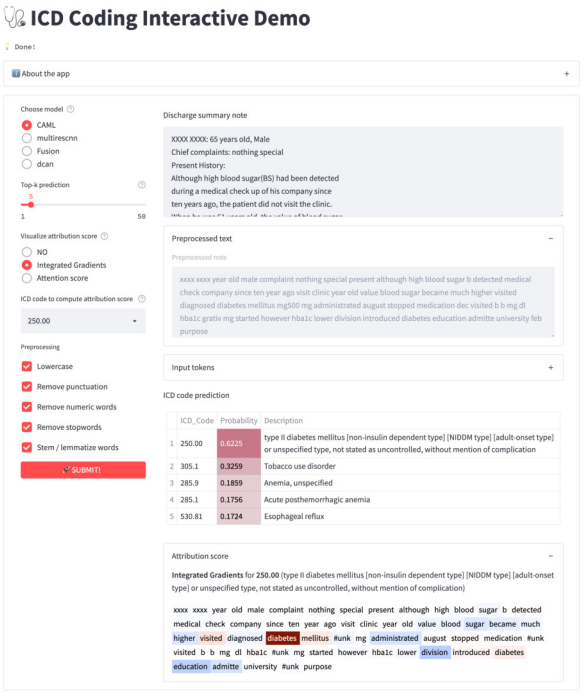

Figure 2: A snapshot of ICD coding interactive demo showing ICD code predictions and the integrated gradient. Input text is extracted from Tsumoto et al. (2019).

- TransICD (Biswas et al., 2021) is the first Transformer-based approach that achieved results comparable to the CNN-based model.
- Fusion (Luo et al., 2021) uses multi-CNN, Transformer encoder, and label attention.

To replicate the author's work in our own system, we re-wired the model from the author's code to make it compatible with our framework. This allows users to also easily tweak the model and its hyperparameters with the config files.

3.3 Training and Evaluation

To train and evaluate the models, we implement a trainer module that manages training and evaluation, with sub-modules for the additional functionalities related to training, such as objective functions, logging, and managing checkpoints. Following the design principle of the framework, the trainer module is also highly configurable so the users can easily customize training and visualize metrics by modifying config files. This also applies to evaluation metrics, and we provide all major evaluation metrics adopted by the automatic ICD coding literature.

3.4 Interactive Demo

In order to enable users to use trained models off-the-shelf, we open source an interactive web ap-

plication based on Streamlit. Using the app, users can feed in a new discharge summary and get the ICD code predictions in real time without writing code to preprocess the input text and to run the models. The app also allows users to change the models and toggle the preprocessing options on the fly so that they can compare models and change preprocessing options.

A major highlight of the app is explainability visualization, i.e., the attribution or importance scores for each word present in the input clinical note. We provide two methods – Integrated Gradients (Sundararajan et al., 2017) and attention scores. Upon choosing the attribution method with an ICD code, the app displays the input tokens with important words highlighted. Note that this interpretability feature is model-agnostic because the explainable AI techniques we use such as integrated gradients are in turn model-agnostic.

A screenshot of the app running on a discharge summary is shown in Figure 2. The bottom of Figure 2 shows the integrated gradient (IG) visualization of ICD code 250.00 "Type II diabetes". We can see that important terms like "diabetes mellitus" exhibit high IG scores[5]. Overall, we expect the interactive demo will be helpful for both researchers who want to validate models, and professionals who want explanations of the model's predictions.

4 Results

In this section, we discuss the quantitative and qualitative results of AnEMIC. On quantitative aspects, we discuss the brief statistics of the datasets and the benchmark results on the our ICD coding datasets. For the qualitative results, we present and analyze some example of interpretability visualization from our demo application.

4.1 Quantitative Results

Dataset Statistics Table 1 shows brief statistics of our ICD coding datasets and the CAML's datasets (old). Our full dataset contains the same number of examples as CAML's full dataset since we used the same data split. However, it has a different set of labels since we corrected the preprocessing of CAML. Our top-50 dataset has the same number of labels as CAML's top-50 dataset, but the label set differs[6]. Also, our top-50 dataset has substantially more examples since the data split of the full dataset is used to make full use of MIMIC-III. It has a slightly less number of examples than the full dataset since examples without any of the top-50 codes are removed.

Dataset	AnEMIC		CAML (old)	
	Full	**Top-50**	**Full**	**Top-50**
# labels	8930	50	8922	50
Mean # labels	15.88	5.73	16.10	5.78
# examples				
- Train set	47723	44728	47723	8066
- Val set	1631	1569	1631	1573
- Test set	3372	3234	3372	1729

Table 1: Statistics of the MIMIC-III full and top-50 datasets. Mean # labels refers to the average number of labels per example.

Benchmark Results To provide the benchmark of our ICD coding datasets, we trained the models introduced in Section 3.2. Hyper-parameters for each model are chosen as reported in the respective paper or code. Note that these hyper-parameters are tuned to CAML datasets, so may not be optimal for our datasets, especially for the top-50 dataset. For DCAN and TransICD model, only the MIMIC-III top-50 experiments was performed, so we use the hyper-parameters for the top-50 dataset in the full dataset experiment. For each model, we ran the experiment three times and computed the mean and variance of the results. Table 2 and 3 shows the benchmark results. Among the models that we implemented, MultiResCNN and Fusion achieved the best test performance on the MIMIC-III full dataset, and DCAN performed best on the MIMIC-III top-50 dataset.

To validate the implementation of key models and the CAML version of dataset, we also ran the same experiments on the CAML version of the datasets. Overall, the results display similar level of performance as reported in the papers. Please see Appendix C for the full results and details of the reproduction experiments.

4.2 Qualitative Analysis

Explainability Visualization Figure 3 shows some examples of explainability visualization from the demo app. For each example, we extract the window around the word with the highest attribution score. In the left figure, for a fixed discharge summary and an ICD code (599.0, *Urinary tract*

[5] Red and blue color in the visualization represent positive and negative scores, respectively.

[6] Please refer to Table 4 in the Appendix to compare.

Model	Macro AUC	Micro AUC	Macro F1	Micro F1	P@8	P@15
CNN	0.835±0.001	0.974±0.000	0.034±0.001	0.420±0.006	0.619±0.002	0.474±0.004
CAML	0.893±0.002	0.985±0.000	0.056±0.006	0.506±0.006	0.704±0.001	0.555±0.001
MultiResCNN	0.912±0.004	0.987±0.000	0.078±0.005	0.555±0.004	0.741±0.002	0.589±0.002
DCAN	0.848±0.009	0.979±0.001	0.066±0.005	0.533±0.006	0.721±0.001	0.573±0.000
TransICD	0.886±0.010	0.983±0.002	0.058±0.001	0.497±0.001	0.666±0.000	0.524±0.001
Fusion	0.910±0.003	0.986±0.000	0.081±0.002	0.560±0.003	0.744±0.002	0.589±0.001

Table 2: Test set results on the MIMIC-III full dataset. The results are shown using the mean±standard deviation format.

Model	Macro AUC	Micro AUC	Macro F1	Micro F1	P@5
CNN	0.913±0.002	0.936±0.002	0.627±0.001	0.693±0.003	0.649±0.001
CAML	0.918±0.000	0.942±0.000	0.614±0.005	0.690±0.001	0.661±0.002
MultiResCNN	0.928±0.001	0.950±0.000	0.652±0.006	0.720±0.002	0.674±0.001
DCAN	0.934±0.001	0.953±0.001	0.651±0.010	0.724±0.005	0.682±0.003
TransICD	0.917±0.002	0.939±0.001	0.602±0.002	0.679±0.001	0.643±0.001
Fusion	0.932±0.001	0.952±0.000	0.664±0.003	0.727±0.003	0.679±0.001

Table 3: Test set results on the MIMIC-III top-50 dataset. The results are shown using the mean±standard deviation format.

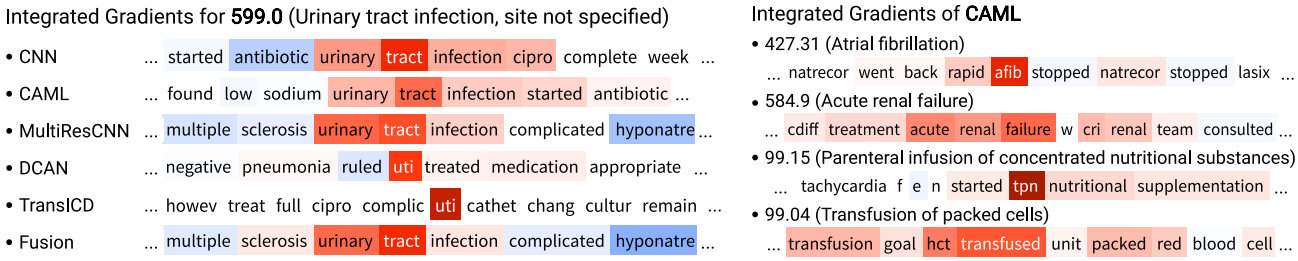

Figure 3: Interpretability visualization examples. **Left**: the integrated gradients of various models on a fixed input and a fixed ICD code (HADM_ID=100020, ICD-9 599.0). **Right**: the integrated gradients of CAML for various ICD codes on a fixed input (HADM_ID=139574).

infection, site not specified), we examine the integrated gradients of various models. From the figure, we can observe that all models correctly attribute their prediction to the words relevant to the diagnosis. In the right figure, for a fixed discharge summary and a model (CAML), we visualize the integrated gradients of some ICD codes that are predicted as positive. As the figure shows, different parts of the input are attributed and they are all semantically relevant to the corresponding ICD code. As both figures illustrate, our interactive demo provides an effective visualization tool for explaining the model's predictions.

5 Conclusions and Future Work

In this work, we present AnEMIC, a comprehensive framework for automatic diagnostic coding. It serves as a standardized benchmark for ICD coding on MIMIC-III by correcting errors in existing datasets and providing popular deep learning-based models. Our framework has a modularized and easy-to-use config-based design, and researchers can easily experiment by writing config files or adding custom submodules. We also provide an interactive app for performing real-time inference and visualization for model explainability.

AnEMIC is under active development and welcomes contributions from the community. Upcoming updates to our pipelines include adding more recent approaches and models, especially those that incorporate additional sources of external knowledge, as well as supporting other datasets like the MIMIC-II dataset.

References

Tian Bai and Slobodan Vucetic. 2019. Improving medical code prediction from clinical text via incorporating online knowledge sources. In *The World Wide Web Conference*, WWW '19, page 72–82, New York, NY, USA. Association for Computing Machinery.

Biplob Biswas, Thai-Hoang Pham, and Ping Zhang. 2021. Transicd: Transformer based code-wise attention model for explainable icd coding.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. 2020. HyperCore: Hyperbolic and co-graph representation for automatic ICD coding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3105–3114, Online. Association for Computational Linguistics.

Richárd Farkas and György Szarvas. 2008. Automatic construction of rule-based icd-9-cm coding systems. In *BMC bioinformatics*, volume 9, pages 1–9. Springer.

Shaoxiong Ji, Erik Cambria, and Pekka Marttinen. 2020. Dilated convolutional attention network for medical code assignment from clinical text. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 73–78, Online. Association for Computational Linguistics.

Shaoxiong Ji, Matti Hölttä, and Pekka Marttinen. 2021. Does the magic of bert apply to medical code assignment? a quantitative study. *Comput. Biol. Med.*, 139(C).

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Rajvir Kaur, Jeewani Anupama Ginige, and Oliver Obst. 2021. A systematic literature review of automated ICD coding and classification systems using discharge summaries. *CoRR*, abs/2107.10652.

Leah S Larkey and W Bruce Croft. 1996. Combining classifiers in text categorization. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 289–297.

Fei Li and Hong Yu. 2020. Icd coding from clinical text using multi-filter residual convolutional neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8180–8187.

Junyu Luo, Cao Xiao, Lucas Glass, Jimeng Sun, and Fenglong Ma. 2021. Fusion: Towards automated ICD coding via feature compression. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2096–2101, Online. Association for Computational Linguistics.

Laxmaiah Manchikanti. 2002. Implications of fraud and abuse in interventional pain management. *Pain Physician*, 5(3):320.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 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)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

Damian Pascual, Sandro Luck, and Roger Wattenhofer. 2021. Towards BERT-based automatic ICD coding: Limitations and opportunities. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 54–63, Online. Association for Computational Linguistics.

Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. 2014. Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association*, 21(2):231–237.

Aaditya Prakash, Siyuan Zhao, Sadid A Hasan, Vivek Datla, Kathy Lee, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2017. Condensed memory networks for clinical diagnostic inferencing. In *Thirty-first AAAI conference on artificial intelligence*.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Fei Teng, Wei Yang, Li Chen, LuFei Huang, and Qiang Xu. 2020. Explainable prediction of medical codes with knowledge graphs. *Frontiers in Bioengineering and Biotechnology*, 8:867.

Phillip Tseng, Robert S Kaplan, Barak D Richman, Mahek A Shah, and Kevin A Schulman. 2018. Administrative costs associated with physician billing and insurance-related activities at an academic health care system. *Jama*, 319(7):691–697.

Shusaku Tsumoto, Tomohiro Kimura, Haruko Iwata, and Shoji Hirano. 2019. Estimation of disease code from electronic patient records. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 2698–2707. IEEE.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. A label attention model for icd coding from clinical text. In *Proceedings of the Twenty-Ninth*

International Joint Conference on Artificial Intelligence, IJCAI-20, pages 3335–3341. International Joint Conferences on Artificial Intelligence Organization. Main track.

Xiancheng Xie, Yun Xiong, Philip S. Yu, and Yangyong Zhu. 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*, CIKM '19, page 649–658, New York, NY, USA. Association for Computing Machinery.

Chenwei Yan, Xiangling Fu, Xien Liu, Yuanqiu Zhang, Yue Gao, Ji Wu, and Qiang Li. 2022. A survey of automated international classification of diseases coding: development, challenges, and applications. *Intelligent Medicine*.

Ying Yu, Min Li, Liangliang Liu, Zhihui Fei, Fang-Xiang Wu, and Jianxin Wang. 2019. Automatic icd code assignment of chinese clinical notes based on multilayer attention birnn. *Journal of Biomedical Informatics*, 91:103114.

Zheng Yuan, Chuanqi Tan, and Songfang Huang. 2022. Code synonyms do matter: Multiple synonyms matching network for automatic ICD coding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 808–814, Dublin, Ireland. Association for Computational Linguistics.

Zachariah Zhang, Jingshu Liu, and Narges Razavian. 2020. BERT-XML: Large scale automated ICD coding using BERT pretraining. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 24–34, Online. Association for Computational Linguistics.

A Notes on ICD Code Preprocessing

In CAML's preprocessing pipeline, there are two errors. Firstly, when they load the `DIAGNOSES_ICD` and `PROCEDURES_ICD` tables into Pandas dataframes, the ICD codes are loaded without specifying a data type, `dtype` in the `pd.read_csv()` method, resulting in the loss of some of leading zeros (e.g. 0040 $\rightarrow$ 40). This affects more than 190 codes out of 8930 in MIMIC-III. Also, when they store the converted ICD codes (with period) into a file and re-read it, data type is not specified, resulting in that some of the codes are converted as floating number and lose leading and trailing zeros. This also affects many ICD codes. For example, a major top-50 ICD code, 93.90 is not selected.

Secondly, MIMIC-III has duplicate ICD codes in the `DIAGNOSES_ICD` and `PROCEDURES_ICD` table, i.e., an ICD code can be repeated in one admission[7]. While preprocessing, CAML's code does not remove such duplicate codes, and as a result of this, some ICD codes were selected as top-50 incorrectly.

As a result, CAML's MIMIC-III full dataset has 8922 labels, while our correctly fixed dataset has 8930 labels. Moreover, our MIMIC-III top-50 dataset has ICD codes 93.90, V45.82, and CAML's dataset has 33.24, 45.13 instead.

Table 4 lists the ICD codes in CAML's, our, and TransICD's MIMIC-III top-50 datasets. TransICD (Biswas et al., 2021) corrected the first mentioned error, i.e., loading ICD codes incorrectly, but counts duplicate ICD codes when choosing top-50 codes, resulting in another incorrect set of top-50 codes.

B Sample Configuration File

Figure 4 shows the YAML config files for preprocessing our MIMIC-III full dataset, to show the configurable pipeline of AnEMIC. Users can create their own ICD coding datasets with, for example, different top-k or word stemmer, by customizing options in the config file. Also, for more customized behavior, users can implement submodules of the pipeline – for example, tokenizer and embedding trainer, and register in the `ConfigMapper` to be used in the config file.

C Reproduction Results on the CAML's Dataset

In this section, we describe the reproduction experiments and explain the results. To ensure that our framework correctly re-implemented the old, CAML version of the datasets and the key models, we trained the models on the old datasets and compared the results with the ones reported in the papers. As in the benchmark experiments, for each configuration, we ran experiments three times and computed the mean and the standard deviation. To make a fair comparison between the models, we created three sets of the old datasets and used each of them for each run of model training. Effectively, the runs will have different weight initialization, including the embedding matrix.

The results are shown in Table 5 and 6. Overall, our reproduction shows similar performance as reported in the papers and preserves the relative order

[7] For example, ICD code 33.24 appears 11 times in the admission with `HADM_ID=193989`.

No.	CAML		TransICD		AnEMIC	
1	401.9	20053	401.9	20053	401.9	20046
2	38.93	14444	38.93	14444	38.93	12866
3	428.0	12842	428.0	12842	428.0	12842
4	427.31	12594	427.31	12594	427.31	12589
5	414.01	12179	414.01	12179	414.01	12178
6	96.04	9932	96.04	9932	96.04	9493
7	96.6	9161	96.6	9161	96.6	9102
8	584.9	8907	584.9	8907	584.9	8906
9	250.00	8784	250.00	8784	250.00	8783
10	96.71	8619	96.71	8619	272.4	8503
11	272.4	8504	272.4	8504	96.71	8426
12	518.81	7249	518.81	7249	518.81	7249
13	99.04	7147	99.04	7147	99.04	7102
14	39.61	6809	39.61	6809	39.61	6781
15	599.0	6442	599.0	6442	599.0	6442
16	530.81	6156	530.81	6156	530.81	6154
17	96.72	5926	96.72	5926	96.72	5815
18	272.0	5766	272.0	5766	272.0	5766
19	285.9	5296	285.9	5296	285.9	5295
20	88.56	5240	88.56	5240	88.56	5045
21	244.9	4788	244.9	4788	244.9	4785
22	486	4733	486	4733	486	4732
23	38.91	4575	38.91	4575	285.1	4499
24	285.1	4499	285.1	4499	38.91	4449
25	36.15	4390	36.15	4390	36.15	4387
26	276.2	4358	276.2	4358	276.2	4358
27	496	4296	496	4296	496	4296
28	99.15	4172	99.15	4172	99.15	4162
29	995.92	3792	995.92	3792	995.92	3792
30	V58.61	3698	V58.61	3698	V58.61	3697
31	507.0	3592	507.0	3592	507.0	3592
32	038.9	3580	038.9	3580	038.9	3580
33	88.72	3500	88.72	3500	585.9	3367
34	585.9	3367	585.9	3367	403.90	3350
35	403.90	3350	403.90	3350	311	3347
36	311	3347	311	3347	88.72	3305
37	305.1	3272	305.1	3272	305.1	3272
38	37.22	3248	37.22	3248	412	3203
39	412	3203	412	3203	37.22	3147
40	**33.24**	**3188**	**33.24**	**3188**	39.95	3133
41	39.95	3178	39.95	3178	287.5	3002
42	287.5	3002	287.5	3002	410.71	3001
43	410.71	3001	410.71	3001	276.1	2985
44	276.1	2985	276.1	2985	V45.81	2943
45	V45.81	2943	V45.81	2943	424.0	2876
46	424.0	2878	424.0	2878	V15.82	2741
47	**45.13**	**2849**	**45.13**	**2849**	511.9	2693
48	V15.82	2741	V15.82	2741	**93.90**	**2656**
49	511.9	2693	511.9	2693	**V45.82**	**2651**
50	**37.23**	**2659**	**93.90**	**2663**	**37.23**	**2619**
51	**V45.82**	**2651**	**37.23**	**2659**	**33.24**	**2607**
52	403.91	2566	**V45.82**	**2651**	403.91	2566
53	V29.0	2529	403.91	2566	**45.13**	**2552**
54	424.1	2517	V29.0	2529	V29.0	2529
55	785.52	2501	424.1	2517	424.1	2517
56	V58.67	2497	785.52	2501	785.52	2501
57	427.89	2396	V58.67	2497	V58.67	2497
58	327.23	2328	427.89	2396	427.89	2396
59	997.1	2313	327.23	2328	327.23	2328
60	99.55	2304	997.1	2313	997.1	2313
61	**93.9**	**2233**	99.55	2304	99.55	2275

Table 4: Top-61 frequency ICD codes from differently processed datasets. The frequency of each code to select the top-50 labels is shown next to each code. Note the frequencies of ICD codes are affected by preprocessing method and error. The top-50 ICD codes that are not contained in all three top-50 sets are marked in bold.

```yaml
paths:
  mimic_dir: &mimic_dir datasets/mimic3/csv
  static_dir: &static_dir datasets/mimic3/static
  dataset_dir: &dataset_dir datasets/mimic3_full
  word2vec_dir: &word2vec_dir datasets/mimic3_full/word2vec

preprocessing:
  name: mimic_iii_preprocessing_pipeline
  params:
    paths:
      mimic_dir: *mimic_dir
      static_dir: *static_dir
      save_dir: *dataset_dir
      diagnosis_code_csv_name: DIAGNOSES_ICD.csv.gz
      procedure_code_csv_name: PROCEDURES_ICD.csv.gz
      noteevents_csv_name: NOTEEVENTS.csv.gz
      train_json_name: train.json       # will be saved
      val_json_name: val.json           # will be saved
      test_json_name: test.json         # will be saved
      label_json_name: labels.json      # will be computed and saved
      label_freq_json_name: null
    dataset_metadata:
      column_names:
        subject_id: SUBJECT_ID
        hadm_id: HADM_ID
        chartdate: CHARTDATE
        charttime: CHARTTIME
        storetime: STORETIME
        category: CATEGORY
        description: DESCRIPTION
        cgid: CGID
        iserror: ISERROR
        text: TEXT
        icd9_code: ICD9_CODE
        labels: LABELS
    dataset_splitting_method:
      name: caml_official_split
      params:
        hadm_dir: *static_dir
        train_hadm_ids_name: train_full_split.json
        val_hadm_ids_name: val_full_split.json
        test_hadm_ids_name: test_full_split.json
    clinical_note_preprocessing:
      to_lower:
        perform: true
      remove_punctuation:
        perform: true
      remove_numeric:
        perform: true
        replace_numerics_with_letter: null
      remove_stopwords:
        perform: true
        params:
          stopwords_file_path: null
          remove_common_medical_terms: true
      stem_or_lemmatize:
        perform: true
        params:
          stemmer_name: nltk.WordNetLemmatizer
      truncate:
        perform: false
    incorrect_code_loading: false
    count_duplicate_codes: false
    code_preprocessing:
      top_k: 0                          # enter 0 for all codes
      code_type: both
      add_period_in_correct_pos:
        perform: true
    train_embed_with_all_split: false
    tokenizer:
      name: spacetokenizer
      params: null
    embedding:
      name: word2vec
      params:
        embedding_dir: *word2vec_dir
        pad_token: "<pad>"
        unk_token: "<unk>"
        word2vec_params:
          vector_size: 100
          min_count: 3
          epochs: 5
```

Figure 4: The YAML config file for preprocessing the MIMIC-III full dataset.

of performance among the models, illustrating that our code can be used in the research of automatic ICD coding.

Despite the effort of re-implementing the existing datasets and key models, there is a minor difference from the CAML's preprocessing, specifically in training vocabulary and embeddings, that may affect the results. In our preprocessing, the vocabulary and embeddings are trained together from Gensim's word2vec training, which means

Model		Macro AUC	Micro AUC	Macro F1	Micro F1	P@8	P@15
CNN	Repr	0.833±0.003	0.974±0.000	0.027±0.005	0.419±0.006	0.612±0.004	0.467±0.001
	Orig	0.806	0.969	0.042	0.419	0.581	0.443
CAML	Repr	0.880±0.003	0.983±0.000	0.057±0.000	0.502±0.002	0.698±0.002	0.548±0.001
	Orig	0.895	0.986	0.088	0.539	0.709	0.561
MultiResCNN	Repr	0.905±0.003	0.986±0.000	0.076±0.002	0.551±0.005	0.738±0.003	0.586±0.003
	Orig	0.910±0.002	0.986±0.001	0.085±0.007	0.552±0.005	0.734±0.002	0.584±0.001
DCAN	Repr	0.837±0.005	0.977±0.001	0.063±0.002	0.527±0.002	0.721±0.001	0.572±0.001
	Orig	Not available					
TransICD	Repr	0.882±0.010	0.982±0.001	0.059±0.008	0.495±0.005	0.663±0.007	0.521±0.006
	Orig	Not available					
Fusion	Repr	0.910±0.003	0.986±0.000	0.076±0.007	0.555±0.008	0.744±0.003	0.588±0.003
	Orig	0.915	0.987	0.083	0.554	0.736	N/A

Table 5: Reproduced test set results on the MIMIC-III full (old) dataset. For each model, the upper row (Repr) shows the reproduction results in mean±standard deviation, and the lower row (Orig) shows the results in the original papers.

Model		Macro AUC	Micro AUC	Macro F1	Micro F1	P@5
CNN	Repr	0.892±0.003	0.920±0.003	0.583±0.006	0.652±0.008	0.627±0.007
	Orig	0.876	0.907	0.576	0.625	0.620
CAML	Repr	0.865±0.017	0.899±0.008	0.495±0.035	0.593±0.020	0.597±0.016
	Orig	0.875	0.909	0.532	0.614	0.609
MultiResCNN	Repr	0.898±0.006	0.928±0.003	0.590±0.012	0.666±0.013	0.638±0.005
	Orig	0.899±0.004	0.928±0.002	0.606±0.011	0.670±0.003	0.641±0.001
DCAN	Repr	0.915±0.002	0.938±0.001	0.614±0.001	0.690±0.002	0.653±0.004
	Orig	0.902±0.006	0.931±0.001	0.615±0.007	0.671±0.001	0.642±0.002
TransICD	Repr	0.895±0.003	0.924±0.002	0.541±0.010	0.637±0.003	0.617±0.005
	Orig	0.894±0.001	0.923±0.001	0.562±0.004	0.644±0.003	0.617±0.003
Fusion	Repr	0.904±0.002	0.930±0.001	0.606±0.009	0.677±0.003	0.640±0.001
	Orig	0.909	0.933	0.619	0.674	0.647

Table 6: Reproduced test set results on the MIMIC-III top-50 (old) dataset. For each model, the upper row (Repr) shows the reproduction results in mean±standard deviation, and the lower row (Orig) shows the results in the original papers.

that rare words in the corpus are replaced with the UNK token before training word2vec. In CAML's preprocessing, the embeddings are trained without replacing UNK tokens, and later, the embeddings of the frequent words are extracted. Also, in our code, only the train corpus is used to train the embedding, while the CAML's code uses the whole corpus. Furthermore, when choosing words for the vocabulary, CAML's code counts the number of documents, i.e., discharge summary note, that each word appears in, while our code uses the total occurrences of each word. Here, both codes use only the train corpus.

D More Attribution Scores of MIMIC-III

Table 7∼10 show more examples of interpretability visualization. When the model predicted an ICD code correctly, then the relevant part of the input text is attributed. The cases when a model does not predicted are the second and third row of Table 8.

Intergrated Gradients for **428.0** (Congestive heart failure unspecified), HADM_ID=158682	
CNN	hypoventilation syndrome chronic diastolic heart failure hypothyroidism irritable bowel syndrome
CAML	hypoventilation syndrome chronic diastolic heart failure hypothyroidism irritable bowel syndrome
MultiResCNN	hypoventilation syndrome chronic diastolic heart failure hypothyroidism irritable bowel syndrome
DCAN	hypoventilation syndrome chronic diastolic heart failure hypothyroidism irritable bowel syndrome
TransICD	obes hypoventil syndrom chronic diastol heart failur hypothyroid irrit bowel syndrom vitamin
Fusion	hypoventilation syndrome chronic diastolic heart failure hypothyroidism irritable bowel syndrome

Table 7: Integrated gradients of various models on a fixed input and a fixed ICD code

Intergrated Gradients for **285.9** (Anemia, unspecified), HADM_ID=100408	
CNN	p mv repair htn lipid chronic anemia persistent afib chf arthritis
CAML	physician name9 pre information name9 pre name3 lf r division cardiothoracic
MultiResCNN	cardiothoracic allergy recorded known allergy drug attending name3 lf asymptomatic
DCAN	p mv repair htn lipid chronic anemia persistent afib chf arthritis
TransICD	p mv repair htn lipid chronic anemia persist afib chf arthriti tonsillectomi
Fusion	p mv repair htn lipid chronic anemia persistent afib chf arthritis

Table 8: Integrated gradients of various models on a fixed input and a fixed ICD code

Integrated Gradients of **Fusion**, HADM_ID=148372
96.04 (Insertion of endotracheal tube)
unsuccessfully repeat abg paco2 ph intubated arterial line finally placed successfully
38.91 (Arterial catheterization)
repeat abg paco2 ph intubated arterial line finally placed successfully extubated
427.31 (Atrial fibrillation)
perioperative pe anticoagulation atrial fibrillation anticoagulation hypertension diabetes type
250.00 (Diabetes mellitus without mention of complication, type ii or unspecified type)
fibrillation anticoagulation hypertension diabetes type ii obstructive sleep apnea hypercholesterolemia
401.9 (Unspecified essential hypertension)
anticoagulation atrial fibrillation anticoagulation hypertension diabetes type ii obstructive sleep

Table 9: Integrated gradients of Fusion for various ICD codes on a fixed input

Integrated Gradients of **MultiResCNN**, HADM_ID=135796
414.01 (Coronary atherosclerosis of native coronary artery)
niacin attending name3 lf cad vessel cad aortic stenosis toxic multinodular goiter
427.31 (Atrial fibrillation)
operative dysphagia post operative atrial fibrillation h 1st degree av block p peg
96.6 (Enteral infusion of concentrated nutritional substances)
ir post pyloric tube placed feeding eventually peg placed picc placed pt screened
38.93 (Venous catheterization, not elsewhere classified)
placed feeding eventually peg placed picc placed pt screened rehab c rehad
584.9 (Acute renal failure, unspecified)
protection mri brain performed cu arf elevation creatinine pt subsequently reintubated

Table 10: Integrated gradients of Fusion for various ICD codes on a fixed input

Acknowledgements

We acknowledge the support of Center for Machine Learning and Health at Carnegie Mellon University. This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health.