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