医療文書における数値表現のトークン化による ICD コード予測と医療タスクへの応用検証

福本拓也[1]　五十嵐正尚[2]　奥村貴史[3]　村松俊平[1]　坂根亜美[1]　堀口裕正[3]　狩野芳伸[1]
[1] 静岡大学 [2] 国立病院機構 [3] 北見工業大学
[1] {tfukumoto,smuramatsu,asakane,kano}@kanolab.net
[2]{igarashi.masanao.th,horiguchi.hiromasa.nz}@mail.hosp.go.jp [3]taka@wide.ad.jp

概要

国立病院機構が構築する大規模匿名化電子カルテデータ NCDA を用いて，自由記述テキスト部から主診断 ICD-10 コードを予測するシステムを構築した．予測には Wikipedia と医療テキストとでそれぞれ事前学習された BERT モデルを fine-tuning して比較した．電子カルテには既存の言語モデルで扱いづらい数値を用いたテキスト表現が多く含まれている．それらの表現を特殊トークンとしてまとめて扱う手法を提案し，予測性能の向上を確認した．この ICD-10 コード予測で学習されたモデルを別の医療タスクに応用し，ICD-10 コード予測による学習が医療言語処理全般に性能を向上させ得るか調査した．

1　はじめに

近年，多くの医療現場で診療記録が電子化されており，情報の集積や管理のし易い環境が整備されている．カルテに記載された病名には ICD-10[1] の分類コードが付与されるが，コーディングと呼ばれる付与作業は診療情報管理士が人手で行っている．

ICD-10 コードとは，「疾病及び関連保健問題の国際統計分類」という WHO によって定められた国際的に用いられる病名コードのことであり（本研究では常に ICD-10 を対象とし，以下 ICD コードと記述する），階層的に分類された病名が記載されている．ICD コードは先頭のアルファベットに何桁かの数字が続き，桁で分類階層を表す記法になっているが，ICD コードは病名と 1 対 1 に対応するわけではない．こうした ICD コードの付与作業は，高度な専門性が必要なうえ曖昧性のある，難易度の高い作業である．基本的にすべてのカルテにコードが付されているため，ICD コードをラベルとみなせば潜在的に膨大な学習データが利用できる．

ICD ではコードが近い疾病は疾病の性質も近くなるように定義されており，少ない桁数に疾病の特徴が圧縮されている．ICD コードの上位桁予測を言語モデルによって解くことができれば，その言語モデルは疾病の特徴を取り込んだモデルになり，医療テキスト処理全般の性能向上が期待できる．

そこで本研究では，まず異なる事前学習済み BERT モデル [2] を fine-tuning して複数の手法で ICD コード分類を行い，ICD コードの予測性能を比較し分析した．電子カルテの自由記載テキストには，日付や時刻，検査数値など数値表現が数多く出現する．しかし，既存の言語処理モデルは数値表現に弱い. 特に医療文書に出てくる検査数値は，医師であれば専門知識によって解釈し適切な診断を下しているが，性別や年齢，身長，体重などの文脈によって数値の示す意味が変わってしまい言語処理タスクとしては難易度が高い．医師がカルテを記載する際には検査結果を具体的な症状などの言葉に変換していることが期待できるため，電子カルテテキストにある数値を特殊トークンに置き換えることで，ICD コード予測の精度が上がるのではないかと考えた．この特殊トークン置換を 3 パターン用意し，ICD コード推測の性能比較に用いた．

さらに，ICD コード予測の fine-tuning によって追加的に学習されたモデルを用いて，別の医療タスクに応用し，ICD コード予測による学習が医療言語モデルに与える影響を調査した．

2　関連研究

これまでに，ICD コードを予測する取り組みはいくつか行われてきている．日本語カルテを対象としたものでは，2016 年に開催された NTCIR-12 の MedNLPDoc Task[3] がある．552 種類の ICD コードを付与した 200 件の模擬カルテを，訓練データとテ

ストデータに分割したうえでテストデータのコードを推測させるものであった．MedNLPDoc では機械学習を用いたチームを抑え，ルールベースを用いたチームが最高性能を達成した [4]．

英語カルテでの取り組みでは，Bagheri ら [5] が HA-GRU[6] を用いた ICD コードの 1 桁予測と 3 桁予測に取り組んでおり，1 桁予測について Accuracy で 72.5，F1 スコアで 43.5，3 桁予測については Accuracy で 23.7 の最高性能を達成，F1 スコアで 19.8 を報告している．特定の種類の疾病に限定した ICD コード予測として，Sammani ら [7] が BiGRU[8] を用いて，出現頻度上位 10 種類の 3 桁分類に限定し予測を行った．その結果，F1 スコアで 76-99 を達成し，上位 10 種類のさらに小分類にあたる 4 桁分類 16 種でも，F1 スコア 87-98 を達成した．

数値表現のトークン化に関して，Loukas ら [9] は，上場企業の定期報告書に対する固有表現抽出タスクにおいて，数値表現を単一のトークンに置き換えることで BERT モデルの性能向上を達成している．数値を <NUM> トークンに置き換える手法と，1,234.5 のような数値表現を X,XXX.X に置き換える手法の 2 種類を試しており，後者がより高い性能を示した．

3 提案手法

本研究では，いくつかの異なる事前学習を行った BERT モデルを用いて，電子カルテテキストから ICD コードを予測する多クラス分類を行う．具体的には，ICD コードの先頭のアルファベットを予測する 1 桁分類と，先頭のアルファベットとそれに続く数字 2 桁を予測する 3 桁分類の二種類を行う．ICD-10 の分類詳細は付録の表 4 を参照されたい．電子カルテでは主診断と副診断それぞれに ICD コードが割り振られているが，今回は主診断のみを分類する．システムの概要を図 2 に示す．

3.1 数値表現のトークン化

電子カルテのテキストには数値に関する表現が大きく 3 つある．1 つ目はカルテを記載するときや過去の診察について記載するときに現れる日付表現，2 つ目は検査時刻や経過時間などを表す時間表現，3 つ目は検査数値である．前処理として，正規表現を用いて日付表現を <DATE>, 時間表現を <TIME> トークンに変換し, その後残りの数値表現を全て <NUM> トークンに変換する．123.45 のような数値は整数部と小数部をまとめて 1 トークンとする．変換されたテキストの例を図 1 に示す．

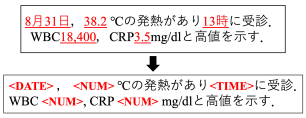

図 1 数値表現トークン化の例

3.2 BERT モデルによる分類

事前学習済み BERT モデルを fine-tuning することによって多クラス分類を行う．東北大学が公開している日本語 Wikipedia で事前学習された bertbase-japanese-whole-word-masking[1)]（以下 cl-tohoku）と，東京大学が公開している 1.2 億行の日本語大規模カルテ記録で事前学習された UTH-BERT[2)]（以下 UTH）とを用い比較する．cl-tohoku のトークナイザは MeCab-Unidic+WordPiece，UTH は MeCab(ipadic-NEologd+万病辞書) + WordPiece を用いる．fine-tuning には，これら事前学習済みモデルの最終層に全結合層を加え，クラス数分の出力ノードを用意して学習した．

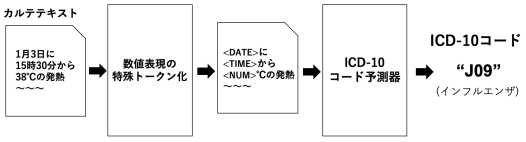

図 2 システム概要図

3.3 データセット

国立病院機構の電子カルテ集積基盤 NCDA[10] の 2022 年 10 月末時点での匿名化データのうち，21 病院，58,397 人の患者を対象とする．そのうち，ICD コードが主診断として記録されていて疑い診断ではない患者を抽出したところ，主診断の総数は 105,817 個となった．「紹介状参照」など他の文書を参照する記述 1 行のみのデータは除去した．患者に主診断の ICD コードが複数紐づくことがある．対象患者に紐づけられたすべての主診断 ICD コードについて，1 桁分類では最初のアルファベットが，3 桁分類では 3 桁すべてが一致する場合に限定することで，推測対象の ICD コードが一意に定まる患者のみとした．1 文書の各行を <NL> トークンで連結して 1 文とし，患者 1 人分をまとめて 1 つの文書として扱った．

1） https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking
2） https://ai-health.m.u-tokyo.ac.jp/home/research/uth-bert

4 実験と結果

4.1 ICD-10 コードの 1 桁分類

対象データ ICD コードの先頭アルファベット 1 桁を分類する．先頭アルファベットが傷病及び死亡の外因に属する患者数は少なかった（V,W,Y が 0 人，X が 1 人）ためこれら 4 クラスを除外し，ICD コードの種類は 22 種類，患者数は 29,566 人，約 368 万行のテキストとなった．最もサンプル数が多いクラスは J で 2,927 人，最も少ないクラスは O で 270 人であった．1 桁分類の分布を付録の図 3 に示す．

実験設定と評価結果 学習時の Epoch 数は 10，中間層は 768 次元，optimizer には Adam を使用し，学習率は 1e-5, 最大系列長は 200 トークン，損失関数には交差エントロピーを用いた．性能の評価には Micro-Precision, Micro-Recall，Micro-F1,Macro-F1 を使用した．訓練データとテストデータを 8:2 とする 5 分割交差検証を行い，それぞれの評価値の平均を最終的な評価とした．実験では，数値表現の置き換えを以下の 3 パターン比較した．

- なし 数値表現置き換えなし
- **日時** <DATE>,<TIME> のみ置き換え
- **日時数** <DATE>,<TIME>,<NUM> の置き換え

実験の評価結果を表 1 に示す．± の後に続く値は，クラス毎の最大最少スコアのレンジを示す．1 桁分類では，数値表現をそのまま用いた **UTH-なし**が最も高い性能であった．

考察 **cl-tohoku** では <DATE>,<TIME>,<NUM> を使うことによる影響がほとんどなかったことから，一般的なテキストで事前学習した言語モデルは，推測に数値表現を利用していないと考えられる．**UTH** に関しては <DATE>,<TIME> による影響はほとんどなかったものの，<NUM> を用いると明確に性能が低下した．その原因として，電子カルテで事前学習した **UTH** が検査数値として出現した数値表現の意味をある程度学習できていたところ，特殊トークンに置き換えたせいでその情報が利用できなくなった可能性がある．別の原因として，薬品名と共起する成分容量の数字（たとえば「薬剤名 20mg」）などが消えてしまい，薬品名の認識に悪影響を及ぼした可能性がある．この場合は，薬品名に隣接する数値表現を NUM トークンに置き換えないようにすれば，悪影響を抑えることができるかもしれない．1 桁分類は疾患の部位や系統に関する分類であるため，数値表現が手掛かりになりづらく，全体的に置き換えによる差分が小さくなったと考えらえる．

表 1 ICD コード 1 桁分類の評価結果

		Micro-Pre.	Micro-Rec	Micro-F1	Macro-F1
cl-tohoku	なし	70.9±3.5	59.4±7.4	64.6±5.6	59.6±11.8
UTH	なし	**71.0**±3.4	**63.1**±2.2	**66.8**±2.1	**63.5**±1.9
cl-tohoku	**日時**	70.8±2.5	59.1±8.9	64.4±5.1	59.5±10.2
UTH	**日時**	**71.0**±2.1	63.0±1.8	**66.8**±1.4	**63.5**±2.1
cl-tohoku	**日時数**	71.0±1.5	59.2±10.3	64.6±6.1	59.7±11.6
UTH	**日時数**	70.7±0.8	62.3±2.3	66.2±1.5	63.0±2.2

4.2 ICD-10 コードの 3 桁分類

対象データ ICD-10 コードの先頭から 3 文字にあたる，アルファベット 1 桁+数字 2 桁からなる 3 桁分類を対象とする．学習データが十分に見込めるよう，サンプル数が 100 以上のクラスを用い，コードの種類は 67 種類，患者数は 16,731 人, 約 197 万行のテキストとなった．クラス当たりサンプル数は T78 が最大の 2,254 人，M17 と L72 で共に最小の 101 人であった．付録の図 4 に分布を示す．

実験設定と評価結果 1 桁分類と同様の条件で，**cl-tohoku** と **UTH** を用いて実験を行った．評価結果を表 2 に示す．Micro-Presicion と Micro-Recall については，DATE,TIME,NUM トークンを全て用いた **cl-tohoku** の性能が最も高かった．また，Micro-Recall,Micro-F1,Macro-F1 については DATE,TIME トークンを用いた **UTH** の性能が最も高くなった．両モデルに共通して，日付表現と時間表現を特殊トークンに置き換えたときに Micro-Precision 以外の指標が改善した．

表 2 ICD コード 3 桁分類の評価結果

		Micro-Pre.	Micro-Rec	Micro-F1	Macro-F1
cl-tohoku	なし	84.8±4.6	49.6±21.5	62.5±15.1	37.3±25.3
UTH	なし	84.8±2.7	50.3±4.0	63.1±4.0	36.4±9.2
cl-tohoku	**日時**	84.8±5.0	51.1±24.1	63.5±16.9	39.5±43.9
UTH	**日時**	85.2±2.1	**52.1**±3.8	**64.7**±2.9	**39.9**±6.3
cl-tohoku	**日時数**	**85.6**±2.5	**52.1**±19.6	63.9±24.3	39.3±25.9
UTH	**日時数**	84.9±4.9	51.1±5.44	63.8±2.8	37.4±4.6

考察 **cl-tohoku** では，NUM トークン置き換えによって性能がさらに向上した．これは，一般テキストで事前学習した BERT モデルが扱えなかった数値表現を置き換えることによる効果と考えられる．**UTH** では NUM トークン置き換えで性能が下がってしまった．これは前述の 1 桁分類と同様の理由だと考える．

また，**cl-tohoku** はクラスごとの性能ばらつきが大きい．クラスによっては，一般テキストに表れづら

い表現があるためではないかと考えられる．医療系の語彙セットを持たない **cl-tohoku-日時数**が，**UTH-日時**の性能かなり近いことから，ICD コード予測は具体的な数値表現がなくても適切な語彙セットを持っていれば対応できる可能性がある．そのため，日本語 Wikipedia と電子カルテの両方で事前学習を行ったモデルを作ることで，より良い ICD コード予測を行えると期待できる．

4.3 ICD コード追加学習の応用検証

前節の ICD コード予測を追加的な学習と考え，この学習により疾病の類似度に関する知識を獲得できたと期待する．その応用検証として，8 つの病気，症状に罹患しているかを Twitter 投稿から推測するマルチラベル分類を行う，NTCIR-13 の MedWeb[11] タスクを実行し評価する．

4.4 データセット

MedWeb タスクは，ツイートに 8 つの病気または症状（インフルエンザ，下痢／腹痛，花粉症，咳／喉の痛み，頭痛，熱，鼻水／鼻づまり，風邪）の罹患の有無を割り当てたマルチラベル分類タスクである．データセットは日本語，英語，中国語の 3 言語が用意されており，各言語につき学習データ 1,920 ツイート，テストデータ 640 ツイートで構成される．本研究では日本語データのみを用いる．

4.5 実験

前述の 3 桁分類で fine-tuning したモデルの最終層を除去し，新たに最終層を追加して MedWeb データで fine-tuning する．ベースラインは ICD コード予測を行っていないオリジナルの **cl-tohoku** と **UTH** とした．ICD コード予測の学習を行ったモデルとして，特殊トークン置き換えをしていない **cl-tohoku-なし**および **UTH-なし**と，3 桁分類の性能が高かった **cl-tohoku-日時数**（DATE,TIME,NUM トークンを置き換え）および **UTH-日時**（DATE,TIME を置き換え）の計 6 つのモデルを用いて比較実験する．

4.6 結果

MedWeb 実験の評価結果を表 3 に示す．

考察　**cl-tohoku** では，ICD コード予測で学習したモデルのほうが，学習してないモデルよりも性能が高くなった．これは ICD コード予測によって医学知識をある程度学習できたことが要因と考えられる．

表 3　マルチラベル分類の結果

		Micro-Pre.	Micro-Rec.	Micro-F1	Macro-F1
cl-tohoku	ICD 未学習	87.2±6.3	90.3±13.7	88.7±5.3	85.1±8.3
UTH	ICD 未学習	87.8± 3.0	85.0± 4.1	86.4±1.8	85.1±1.7
cl-tohoku	ICD	**88.2**± 3.0	**93.3**±2.7	**90.7**±2.1	**89.3** ±3.3
UTH	ICD	87.5± 7.4	85.0±11.6	86.2±4.9	85.3±5.5
cl-tohoku	ICD+**日時数**	87.8±2.9	**93.3**±6.0	90.5±1.4	89.0±2.5
UTH	ICD+**日時**	**88.2**±3.3	84.4±5.9	86.2±3.3	85.1±3.7

UTH では，ICD コード予測で学習してもしなくてもほとんど変化が無かった．ICD コード予測で学習したモデルの疾病の類似度に関する知識は増えていない可能性が高い．

ICD コード予測をしたモデルを用いて MedWeb で学習しなおした場合，**cl-tohoku** でも **UTH** でも数値表現の特殊トークン置き換えを取り入れると性能が下がってしまった．これは，MedWeb タスクのデータに数値に関する情報が少なかったことから数値トークンが活用されずかえって悪影響を及ぼしたのではないかと考えられる．

5 結論

数値表現を 3 種類の特殊トークンに置き換えることによる ICD-10 コード予測の性能向上を試みた．時間と日付に関する数値表現の置き換えは ICD コードの上位 3 桁分類において，Wikipedia で事前学習した言語モデルと医療テキストで事前学習した言語モデルのどちらでも性能向上に寄与した．全数値表現を置き換えると，Wikipedia で事前学習した言語モデルでは性能向上が見られたが，医療テキストで事前学習したモデルでは性能が低下した．

また，ICD-10 コード予測によって fine-tuning したモデルを他の医療タスクに適用する実験を行った．医療テキストで事前学習した言語モデルでは性能に変化が見られなかったが，Wikipedia で事前学習した言語モデルでは事前に ICD-10 コード予測による追加的な学習を約 197 万行の医療テキストで行うことで，約 1.2 億行の医療テキストで事前学習した言語モデルを超える精度を出すことができた．

今後の課題として，数値表現の置き換え条件を詳細にすることで ICD-10 コード予測精度の向上を目指す．さらに，ICD-10 コード予測をシングルラベルからマルチラベル分類に変えることによる後段タスクへの影響や，日本語 Wikipedia と電子カルテテキストの両方で事前学習したモデルの性能調査などを行いたい．

謝辞

本研究は厚生労働科学研究費補助金 21HA2015, JSPS 科研費 JP22H00804, JP21K18115, JST AIP 加速課題 JPMJCR22U4, およびセコム科学技術財団特定領域研究助成の支援をうけた.

参考文献

[1] World Health Organization. ICD-10 : international statistical classification of diseases and related health problems : tenth revision, 2nd ed. 2004. https://apps.who.int/iris/handle/10665/42980.

[2] Devlin Jacob, Chang Ming-Wei, Lee Kenton, and Toutanova Kristina. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, 2019.

[3] Eiji Aramaki, Mizuki Morita, Yoshinobu Kano, and Tomoko Ohkuma. Overview of the NTCIR-12 MedNLP-Doc Task. In **Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-12)**, pp. 71–75, 2016.

[4] Masahito Sakishita and Yoshinobu Kano. Inference of ICD Codes by Rule-Based Method from Medical Record in NTCIR-12 MedNLPDoc. In **Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-12)**, pp. 80–84, 2016.

[5] Ayoub Bagheri., Arjan Sammani., Peter G. M. Van Der Heijden., Folkert W. Asselbergs., and Daniel L. Oberski. Automatic ICD-10 Classification of Diseases from Dutch Discharge Letters. In **conjunction with the 13th International Joint Conference on Biomedical Engineering Systems and Technologies-BIOSTEC 2020**, pp. 281–289, 2020.

[6] Du Yong, Wang Wei, and Wang Liang. Hierarchical recurrent neural network for skeleton based action recognition. In **Proceedings of the IEEE conference on computer vision and pattern recognition**, pp. 1110–1118, 2015.

[7] Arjan Sammani, Ayoub Bagheri, Peter G. M. van der Heijden, Anneline S. J. M. te Riele, Annette F. Baas, C. A. J. Oosters, Daniel Oberski, and Folkert W. Asselbergs. Automatic multilabel detection of ICD10 codes in Dutch cardiology discharge letters using neural networks. **npj Digital Medicine**, Vol. 5, No. 37, 2021.

[8] Deng Yaping, Lu Wang, Jia Hao, Tong Xiangqian, and Feng Li. A Sequence-to-Sequence Deep Learning Architecture Based on Bidirectional GRU for Type Recognition and Time Location of Combined Power Quality Disturbance. **IEEE Transactions on Industrial Informatics**, Vol. 15, No. 8, pp. 4481–4493, 2019.

[9] Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. FiNER: Financial Numeric Entity Recognition for XBRL Tagging. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 4419–4431, 2022.

[10] Natsuko Kanazawa1, Takuaki Tani, Shinobu Imai, Hiromasa Horiguchi, Kiyohide Fushimi, and Norihiko Inoue. Existing Data Sources for Clinical Epidemiology: Database of the National Hospital Organization in Japan. In **Clinical Epidemiology**, Vol. 2022:14, pp. 689–698, 2022.

[11] Shoko Wakamiya, Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, and Eiji Aramaki. Overview of the NTCIR-13 MedWeb Task. In **Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-13)**, pp. 40–49, 2017.

A　付録 (Appendix)

表 4　ICD-10 コードの章レベルの分類体系

ICD-10	グループ 1	分類見出し
A00-B99	全身症	感染症及び寄生虫症
C00-D48	全身症	新生物
D50-D89	全身症	血液および造血器の疾患並びに免疫機構の障害
E00-E90	全身症	内分泌，栄養及び
F00-F99	解剖学的系統別の疾患	精神及び行動の障害
G00-G99	解剖学的系統別の疾患	神経系の疾患
H00-H59	解剖学的系統別の疾患	眼及び付属器の疾患
H60-H95	解剖学的系統別の疾患	耳及び乳様突起の疾患
I00-I99	解剖学的系統別の疾患	循環器系の疾患
J00-J99	解剖学的系統別の疾患	呼吸器系の疾患
K00-K93	解剖学的系統別の疾患	消化器系の疾患
L00-L99	解剖学的系統別の疾患	皮膚及び皮下組織の疾患
M00-M99	解剖学的系統別の疾患	筋骨格系及び結合組織の疾患
N00-N99	解剖学的系統別の疾患	腎尿路生殖器系の疾患
O00-O99	分娩・奇形・新生児疾患	妊娠，分娩及び産褥
P00-p96	分娩・奇形・新生児疾患	周産期に発生した病態
Q00-Q99	分娩・奇形・新生児疾患	先天奇形，変形及び染色体異常
R00-R99		症状，徴候及び異常臨床初見・異常検査初見で他に分類されないもの
S00-T98		損傷，中毒及びその他の外因の影響
V01-Y98		傷病及び死亡の外因
Z00-Z99		健康状態に影響を及ぼす要因及び保険サービスの利用
U00-U99		特殊用目的コード

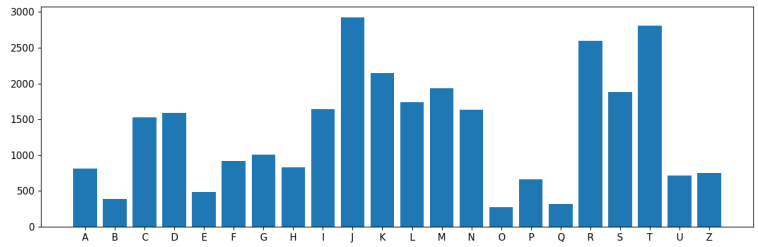

図 3　1 桁分類に用いるデータの分布

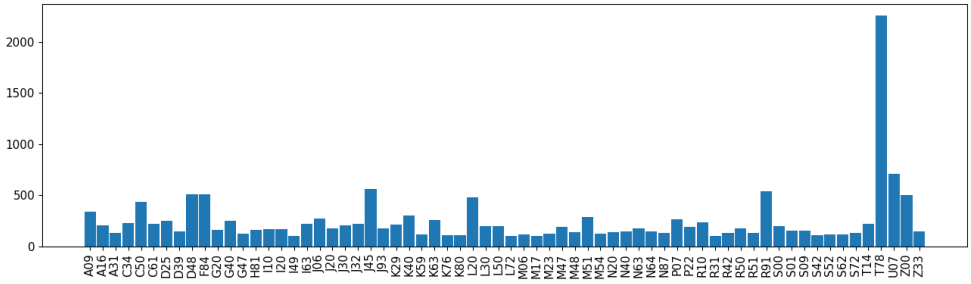

図 4　3 桁分類に用いるデータの分布