データセットの非独立性が機械学習モデルの精度に与える影響調査

服部 文志[1,a)]　松下 誠[1,b)]　肥後 芳樹[1,c)]

概要：ソフトウェア開発者のコーディング能力を客観的に判定することを目的として，ソースコードを利用した機械学習・深層学習による判定手法が提案されている．特に深層学習を用いた手法では，ソースコードの構文的・意味的情報を利用することで高い精度で判定を行うことができる．しかし，この深層学習モデルは未知のデータに対する判定精度，つまり汎化性能が低いことが筆者らの先行研究で確認されているが，その原因については明らかになっていない．そこで本研究では，精度低下の原因についての調査を行った．その結果，モデルの性能評価を行う際に利用された，データセットの利用方法に問題があるのではないかということが分かった．データセットの非独立性を極力排除した利用方法について検討を行い，それに従って今回調査を行った結果，先行研究で報告されたモデルの精度は，当初報告された結果と比べて約 30 %低下したことを確認した．

キーワード：コーディング，深層学習，データセット，汎化性能

1. はじめに

ソフトウェア開発企業は，より高品質なソフトウェアを開発するために，より能力の高い開発者を採用する必要がある．実際に，Amazon や Facebook, Google 等の技術系企業では，採用プロセスにおいてコーディングテストを実施することで応募者の技術力を測り，評価基準の 1 つとして利用している [6]．しかし，このような場面でソースコードの評価を人の手で行うと，非常に大きな時間的コストや経済的コストがかかってしまう．

そこで，ソースコードからコーディング能力を機械的に判定する手法が提案されている．槇原 [9] と松井 [11] は，予約語の利用頻度やメトリクスから定量的かつ自動的にコーディング能力を判定する機械学習モデルを作成した．さらに松井 [12] は，ソースコードの構造的情報や意味的情報を含むグラフを構築し，それを学習に利用することでより判定精度の高いモデルを作成した．

しかし，これらのモデルについては汎化性能が低いことが筆者らの先行研究でわかっている [3]．汎化性能とは未知のデータに対する判定能力のことであり，モデルの実用性や信頼性に関わる重要な指標である．松井らの研究では，モデルの学習と評価において同一のデータセットを分割して利用しているため，未知のデータに対するモデルの精度は不明であった．そこで筆者らは新たに 3 つのデータセットを作成し，学習と評価で同一のデータセットを利用した場合と異なるデータセットを利用した場合の精度をそれぞれ測定した．その結果，異なるデータセットを利用した場合に精度が大幅に低下すること，つまり汎化性能が低いことが確認されたが，その原因の解明には至らなかった．

そこで本研究では，従来のモデルで行われた評価をやり直すことによって本来の性能を正しく評価することを目的とした実験を行った．具体的には，従来の評価ではデータセットの利用方法に問題があり，そのため性能が高く評価され過ぎてしまっているのではないかということ，汎化性能が低くなってしまったように見えたのは，高く評価していた理由がなくなってしまったからではないかと考え，データセットを注意深く選ぶようにして評価することとした．

モデルの精度が過剰に高くなってしまう原因として，重複による影響とデータセットの分割方法による影響の 2 つの要因が考えられる．重複による影響は Allamanis[1] により報告されたもので，データセット内に複数のファイルレベルのクローンが存在することによりモデルの精度が最大で 100%上昇してしまう．データセットの分割による影響は医療分野の画像認識モデルで確認されており，訓練セッ

[1] 大阪大学大学院情報科学研究科
[a)] a-hattor@ist.osaka-u.ac.jp
[b)] matusita@ist.osaka-u.ac.jp
[c)] higo@ist.osaka-u.ac.jp

ト・テストセット間で同一の患者のデータを含むように分割した場合，同一の患者のデータを含まない場合に比べて精度が高くなることがわかっている [10]．これらの要因は，どちらもデータセットの非独立性に起因するものである．データセットの非独立性とは，データセット内の各データが互いに独立ではなく相互に何らかの関係を持っていることを意味する．

今回行った実験では，データセットを利用する際に起こりうる，上記にあげた 2 つの非独立性に関する影響を排除するように注意して実施することとして，既存手法を対象に評価を行った．その結果，手法によって構築されたモデルの精度は，当初報告した制度と比較して約 30%減少することが確認できた．

以降 2 節では，本研究の背景として，コーディング能力判定モデルと汎化性能について説明する．3 節ではデータセットの非独立性がモデルの精度に与え得る影響について説明する．4 節では評価実験の内容と結果について説明し，考察を行う．5 節では競技プログラミングのデータセットを用いた関連研究について紹介する．最後に 6 節ではまとめと今後の課題について述べる．

2. 準備

本節では，今回実験の対象となるコーディング能力判定モデル，評価モデルを対象とした性能指標の 1 つである汎化性能と，これまで行った調査について述べる．

2.1 コーディング能力判定モデル

ソースコードからコーディング能力を判定する手法には，ソースコードのメトリクス等を利用する槇原の手法や [9]，ソースコードのトークン列から Long Short-Term Memory（LSTM）を用いて判定する手法 [5]，ソースコードのグラフ表現を利用する松井の手法などがある [12]．この中でも特に，松井の手法はソースコードの構文的・意味的情報を利用しているため，高い精度で判定を行うことができる．しかしこのモデルは，学習のための訓練セットと評価のためのテストセットで，同一のデータセットを分割したものを利用しており，異なるデータセットのデータに対する判定精度が不明であった．そのため，このモデルを対象とした汎化性能の調査をこれまで行ってきた．以降，本研究ではこのモデルを対象モデルと呼ぶ．

対象モデルは，ソースコードを入力として受け取り，「上級者」，「中級者」，「初級者」のいずれかの判定結果を出力する．このような判定を可能にするために，モデルの学習データとして競技プログラミングのソースコードとレーティングを利用している．競技プログラミングとは，与えられた要件を満たすプログラムを記述する速さや正確さを競う競技である．レーティングとは，競技プログラミングにおける熟練度をあらわす指標である．レーティングが高い上位 20%のソースコードを上級者，レーティングが低い下位 20%のソースコードを初級者，ちょうど中間の 20%のソースコードを中級者として定義する．このモデルの学習の流れを図 1 に示し，以下に各手順の概要を述べる．

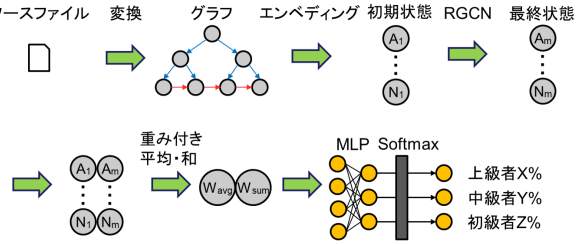

図 1: コーディング能力の判定手順

- **ソースコードをグラフに変換する．**構文木を構築し，各トークンの順序関係や同一変数の利用情報を表すエッジを追加する．
- **エンベディングにより初期状態を取得する．**各ノードを表す文字列に対して Character-level-CNN[19] を適用することでノードの初期状態を得る．
- **グラフの状態を更新し，最終状態を得る．**各ノードが隣接ノードから情報を集約してノードの状態を更新する，という工程を繰り返すことで最終状態が得られる．対象モデルでは，RGCN[14] というグラフネットワークを扱うためのディープラーニングアーキテクチャを採用している．
- **初期状態と最終状態から判定結果を出力する．**各ノードの初期状態と最終状態のベクトルを連結させた後，全ノードの重み付き和と重み付き平均を算出して連結させる．このベクトルを出力数 3 の MLP に入力し，得られる出力をさらに Softmax 層に入力することで，上級者・中級者・初級者である確率の合計値が 100%となるようにする．

2.2 モデルの汎化性能調査

汎化性能とは，機械学習モデルが未知のデータに対して正しく予測できる性能のことである．機械学習における学習とは，訓練セットに対する損失値ができるだけ小さくなるようにモデルのパラメータを設定することである．しかし，訓練セットに対して上手く予測が行えるようなパラメータを学習できたとしても，未知のデータに対して同様に正確に予測を行うことができるとは限らない．例えば，天気を予測する機械学習モデルが過去の天気を 100%当てることができたとしても，未来の天気に対する予測が完璧にできるかどうかはわからない．そのため，汎化性能は機械学習モデルの信頼性や実用性を確認するための重要な指標となっている[*1]．

[*1] https://www.varista.ai/knowledge/generalization-ability/

我々は先行研究において，対象モデルの汎化性能を調査するために，出自の異なる 4 つのデータセットを利用し，相互にモデルの学習と評価を行ってそれぞれの精度を比較した [3]．その結果を図 2 に示す．

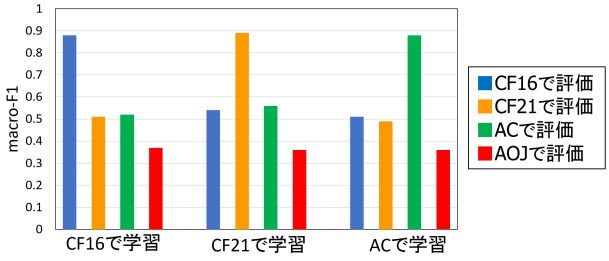

図 2: コーディング能力判定モデルの汎化性能

この結果から，学習と評価で同じデータセットを利用した場合の精度は高いが，異なるデータセットを利用した場合の精度は低くなる，すなわち対象モデルの汎化性能が低いことが確認できた．そこで，このモデルは汎化性能が低い，すなわち，異なるデータセットを利用した際に精度が下がるのではなく，評価を行う際に学習時と同じデータセットを利用していたために，性能が高くなってしまっているのではないか，という仮説を立てた．

3. データセット利用時の影響

本節では，データセットを用いて機械学習モデルを評価する際，その性能が高く評価されてしまう 2 つの原因，データの重複による影響とデータセットの分割方法による影響について述べる．

3.1 重複による影響

Allamanis により，データセット内の重複が機械学習モデルの精度を最大で 100%上昇させることがわかっている [1]．重複とは，ファイルレベルでほぼ酷似したソースコードのことを指す．ソフトウェア開発者は他のソースコードを部分的または全体的にコピーすることが多いため，このような重複が発生してしまう．Allamanis の調査では，github から収集した複数のメジャーなデータセットで重複が存在しており，それにより様々なタスクで精度の上昇が起きていることが報告されている．例えば，Java-Large データセット [2] は全体の 20.2%，Concode[15] データセットは 68.7% のファイルが重複したファイルになっており，Java-Large データセットを利用したメソッド命名タスクでは，重複を除くことで f 値が 12.3%減少することがわかっている．

3.1.1 重複の定義と種類

データセットがモデルの学習を行うための訓練セットと，モデルの精度を測定するためのテストセットに別れている時，以下の 3 種類の重複が発生する．

(1) **訓練重複** 訓練セット内で重複しているファイル
(2) **テスト重複** テストセット内で重複しているファイル
(3) **交差重複** 訓練セットとテストセットの両方に現れるファイル

これを図 3 に示す．

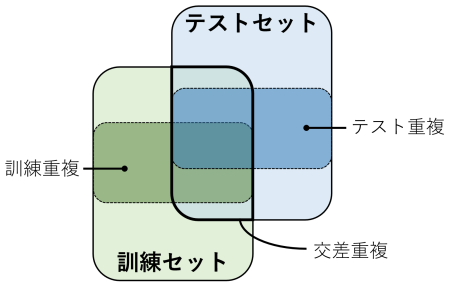

図 3: 重複の種類．[1] の図 1 を日本語に訳して転載．

重複による問題の核は，訓練セットとテストセットに同一のファイルが現れること，つまり交差重複の存在にある．機械学習モデルを学習させる目的は，モデルを用いて新しいコードや見たことのないコードに対して何らかの推測や見識を提供することである．そのためには，機械学習モデルが未知のソースコードによく汎化すること，つまりモデルのユースケースで観察されるようなデータの真の分布を忠実にモデル化することを必要とする．機械学習モデルが真のデータ分布に汎化するためには，その分布から独立に抽出されたデータで学習する必要がある．しかし重複の存在はそれに違反することとなり，モデルが汎化したように見せかけて実際には重複を単に記憶しただけであるため，ユーザが本来観測するユースケース以上の精度となってしまう．

3.2 分割方法による影響

機械学習において，モデルの学習で用いる訓練データとモデルの評価で用いるテストデータは，一般的に 1 つのデータセットをランダムに分割したものが利用される．しかし，データセット中の各データが他のデータと共通の性質を持つ場合，それらが訓練データとテストデータに分割されることで精度の上昇につながってしまう．実際に，医療分野の画像診断モデルにおいて，同一の患者のデータが訓練データとテストデータの両方に含まれる場合と，どちらか一方のみに含まれる場合では，前者のほうが精度が約 40%高くなることが確認されている [10]．このように，訓練データとテストデータの両方に同一のグループのデータが含まれることを避けるようにデータセットを分割する手法は Group K-Fold と呼ばれ，訓練セットとテストセットの独立性を守ることができる．

3.3 データセットの非独立性

ここまで説明してきた 2 つの要因は，どちらもデータ

セットの非独立性に起因するものである．データセットの非独立性は筆者らが新たに定義した用語であり，「データセットが非独立である」とは，「データセット内の各データが独立ではなく，他のデータと共通の性質や特性を持っている」ことを意味する．このような性質は，機械学習における重要な仮定に違反している．それは，「各データはすべて同じ確率分布から得られ，互いに独立である (i.i.d)」というものである [1]．これは不合理な仮定ではなく，機械学習の研究と実践において広く使用されている．

対象モデルの学習や評価で利用されているデータセットは競技プログラミングから収集したものであり，非独立な性質を持っている．それは，同一のユーザが提出したソースコードが多く存在していることと，同一の問題に提出されたソースコードが多く存在していることである．同一のユーザが提出したソースコードは，構文的特徴や識別子名などの特徴が似たものになっていると考えられる．同一の問題に提出されたソースコードは，使用されるアルゴリズムやデータ構造などの特徴が似たものになっていると考えられる．さらに，データセットには同一のユーザが同一の問題に提出したソースコードも複数存在している．これは，提出したソースコードが間違っていた場合に，ソースコードを修正して再提出をすることが可能になっているからである．このようなソースコードは修正量が少なく，酷似した状態にある．そこで本研究では，同一のユーザが同一の問題に提出したソースコードが複数存在した場合，それらを互いに「重複したソースコード」とみなす．

4. 実験

この節では，前節で説明したデータセットの非独立性による精度への影響を，対象モデルを用いて実証的に調査する．そのために，データセット内の重複の有無や訓練セット・テストセットの分割方法を変えて精度を測定・比較する．

4.1 使用するデータセット

本研究では，CF16，CF21，AC の 3 つのデータセットを使用する．各データセットの概要を表 1 に示す．

表 1: 使用する各データセットの概要

データセット名	CF16	CF21	AC
収集元ドメイン	Codeforces	Codeforces	AtCoder
ファイル数	1,644,636	1,752,427	1,459,964
収集期間	2016/5/19〜2016/11/15	2021/12/1〜2021/12/31	2021/1/1〜2021/6/1

CF16 と CF21 は，ロシア最大級の競技プログラミングサイトである Codeforces[*2]から収集したデータセットであるが，それぞれ収集期間が異なっている．一方，AC は日本最大級の競技プログラミングサイトである AtCoder[*3]から収集したデータセットである．CF16 は堤によって作成されたデータセットであり [17]，松井のモデルの学習や評価で使用されたものである．CF21 と AC は筆者らが汎化性能調査のために先行研究にて作成したデータセットである．

表 2: CF16 の重複統計

重複度	1	2	3	4	5 以上
グループ数	193,667	69,905	34,222	18,996	31,130
ファイル数	193,667	139,810	102,666	75,984	231,092
割合	26%	19%	14%	10%	31%

表 3: CF21 の重複統計

重複度	1	2	3	4	5 以上
グループ数	244,051	80,458	38,224	20,642	32,431
ファイル数	244,051	160,916	114,672	82,568	232,895
割合	29%	19%	14%	10%	28%

表 4: AC の重複統計

重複度	1	2	3	4	5 以上
グループ数	344,856	51,694	17,188	7,487	8,880
ファイル数	344,856	103,388	51,564	29,488	59,837
割合	58%	18%	9%	5%	10%

これらのデータセットの重複度合いを表 2〜4 に示す．重複度とは，同一のユーザが同一の問題に提出したファイル数を表す．各データセットの総ファイル数は表 1 より少なくなっているが，これは対象モデルの学習と評価で利用する上級者・中級者・初級者のファイルのみを対象としているためである．

これらの表から，どのデータセットにおいても重複が多数存在していることがわかる．各データセットに注目すると，AC は重複度 1 のファイルが 58%を占めており，重複しているソースコードはデータセット全体の半分以下である．一方，CF16 と CF21 は重複度が 2 以上のファイルの割合がともに 70%を超えており，非常に多くの重複ファイルが存在している．さらに，CF16 と CF21 は重複度が 5 以上の割合がともに 30%前後あり，非常に深刻な重複の問題を抱えている．また，CF16 と CF21 の重複割合の分布は非常に似ており，同じドメインから収集したデータセットは収集時期が異なっていても重複の度合いが似ると考えられる．

4.2 評価指標

本研究では，評価指標として macro-F1 を使用する．macro-F1 は各クラスにおける F 値の平均をとったもので

[*2] https://codeforces.com/

[*3] https://atcoder.jp/

あり，本研究では以下の式で求めることができる．

$$macro\text{-}F1 = \frac{F1_{high} + F1_{mid} + F1_{low}}{3}$$

ここで，式中の $F1_{high}$ は上級者の F 値，$F1_{mid}$ は中級者の F 値，$F1_{low}$ は初級者の F 値を表している．

4.3 重複と分割方法の組み合わせ

ここでは，重複や分割方法に関する用語について述べる．まず，重複に関するテストの用語を定める．これらの用語は，Allamanis の用語定義を参考にして定義したものである．

- **完全非重複テスト** 訓練セットとテストセットの両方から重複を取り除いて行うテスト．ここで，「重複を取り除く」とは，同一のユーザが同一の問題に提出したソースコードのうち，最新のもののみを残して残りをすべて除去することである．
- **訓練重複テスト** 訓練セットに重複があり，テストセットからは重複を取り除いて行うテスト．
- **交差重複テスト** 訓練セットの重複と交差重複を含み，テスト内の重複は含まないテストセットを使用して行うテスト．
- **完全重複テスト** すべての重複を含む状態で行うテスト．つまり，データセットに全く加工をしていない元のままで学習やテストが行われる．重複を意識せずに行われるテストは完全重複テストとなる．

次にデータセットの分割方法に関する用語を紹介する．

- **ランダム分割** データセット中のデータをランダムに訓練・検証・テストセットに分割する方法．ここで，分割比率はいずれの手法においても訓練: 検証: テスト =8:1:1 になるように分割する．
- **ユーザ分割** 同一のユーザが提出したソースコードが，訓練・検証・テストセットのいずれかの集合にのみ含まれるように分割する手法．つまり，同一のユーザのソースコードが訓練セットとテストセットの両方に現れることはない．
- **問題分割** 同一の問題に提出されたソースコードが，訓練・検証・テストセットのいずれかの集合にのみ含まれるように分割する手法．

これらの重複と分割方法の組み合わせを図 4 に示す．

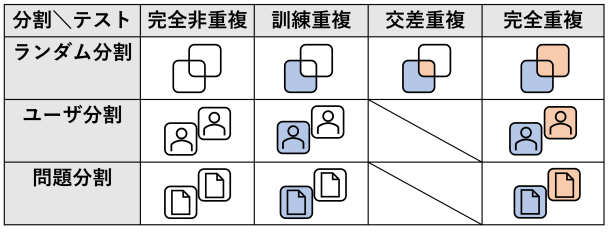

図 4: 重複と分割方法の組み合わせ

図中の各記号は訓練セットとテストセットのベン図として解釈できる．左下の集合が訓練セット，右上の集合がテストセットを表しており，集合の網掛けは重複があることを，空白は重複が排除されていることを意味している．また，各集合内の人型記号や紙型記号はそれぞれ，ユーザ分割と問題分割を意味している．

本研究では，この図で示した各パターンでの対象モデルの精度を測定し比較することで，データセットの非独立性に起因する各要因の影響を調査する．

4.4 実験結果

まず初めに，CF16 のみに注目して各テストパターンにおけるモデルの精度の測定結果を表 5 に示す．

表 5: CF16 における各条件下での精度．表中の値は macro-F1.

分割＼重複	完全非重複	訓練重複	交差重複	完全重複
ランダム分割	0.75	0.76	0.87	0.88
ユーザ分割	0.59	0.59	-	0.61
問題分割	0.73	0.73	-	0.74

完全非重複テスト（）の結果と，訓練重複テスト（）の結果から，訓練セットにおける重複の有無はモデルの精度にはあまり影響を与えていないことがわかる．また，交差重複テスト（）と完全重複テスト（）の結果から，テストセットにおける重複の有無もモデルの精度にあまり影響を与えていないことがわかる．一方，訓練重複テスト（）と交差重複テスト（）の結果から，交差重複の存在によってモデルの精度が高くなっていることがわかる．また，分割方法に着目すると，ランダム分割はユーザ分割や問題分割よりも精度が高くなっていることがわかる．

4.4.1 交差重複の影響

各データセットにおける交差重複によるモデルの精度の変化率を表 6 に示す．

表 6: 交差重複による精度の変化率．表中の値は macro-F1.

データセット			Δ (,)
CF16	0.87	0.76	**-12.6%**
CF21	0.88	0.82	**-6.8%**
AC	0.87	0.82	**-5.7%**

この結果から，いずれのデータセットにおいても重複によってモデルの精度が高くなっており，重複を排除することでその精度が低下することが確認できた．変化の一番小さかった AC においても精度が 5%以上低くなっており，重複による影響が顕著であることがわかる．特に CF16 は精度が一番低くなっており，交差重複を排除することによってモデルの精度が 10%以上も低くなっていることがわかる．CF21 は重複統計において CF16 と同程度の重複割合であったが，精度が低くなった割合は CF16 の半分程度で

あり，重複率と重複の排除による精度の低下は必ずしも比例するとは限らないことがわかる．

4.4.2 ユーザ分割の影響

各データセットにおけるユーザ分割によるモデルの精度の変化率を表 7 に示す．

表 7: ユーザ分割による精度の変化率．表中の値は macro-F1.

データセット			Δ (,)
CF16	0.76	0.59	**-22.4%**
CF21	0.82	0.66	**-19.5%**
AC	0.82	0.62	**-24.4%**

いずれのデータセットにおいても，ユーザ分割をすることでランダム分割より精度が約 20%前後減少している．つまり，同一のユーザが提出したソースコードが訓練セットとテストセットの両方に存在していたことで，モデルの精度が高くなっていたことがわかる．このことから，対象モデルは未知のユーザが書いたソースコードに対しては，コーディング能力を正確に判定しにくいといえる．

4.4.3 問題分割の影響

各データセットにおける問題分割によるモデルの精度の変化率を表 7 に示す．

表 8: 問題分割による精度の変化率．表中の値は macro-F1.

データセット			Δ (,)
CF16	0.76	0.73	**-3.9%**
CF21	0.82	0.82	**0.0%**
AC	0.82	0.82	**0.0%**

CF16 では問題分割によって精度の低下が起きているが，CF21 と AC では問題分割による精度の低下は起きていない．この表の値は macro-F1 の小数第 3 位を四捨五入した値になっており，実際にはわずかに精度の変化はあったものの，ほぼ誤差とみなすことができる程度のものであった．この結果から CF21 と AC で学習したモデルは，未知の問題に対して提出されたソースコードに対しても正確にコーディング能力を判定できるといえる．

4.5 結果の考察

本研究の結果から，以下の 3 つのことがわかった．

(1) 競技プログラミングのデータセットにおいても，交差重複の存在によってモデルの精度が大幅に上がること
(2) 同一のユーザが提出したソースコードが訓練・テストセットの両方に存在することで，モデルの精度が大幅に上がること
(3) 同一の問題に提出されたソースコードが訓練・テストセットの両方に存在することによる影響は，データセットによって異なること

これらの結果から，対象モデルの性能はデータセットの非独立性によって高くなっていることがわかった．特に重複とユーザ分割による影響が大きいことが確認できた．重複とユーザ分割の両方を考慮した精度の変化率を表 9 に示す．

表 9: 重複とユーザ分割の両方による精度の変化率．表中の値は macro-F1.

データセット			Δ (,)
CF16	0.87	0.59	**-32.2%**
CF21	0.88	0.66	**-25.0%**
AC	0.87	0.62	**-28.7%**

この表から，重複を除いてデータセットを適切に分割することで，モデルの精度が約 30%低下することが確認できた．逆に言えば，先行研究で報告されていた対象モデルの精度は，本来の性能よりも 50%近く高くなっていたことになる．このようなモデルの精度の変化は，他のモデルやデータセットでも発生している可能性がある．

(1) の結果については，競技プログラミングのデータセットを使用する多くのタスクに適用されると考えられる．どのようなモデルであっても，訓練セットと重複した（同一のユーザが同一の問題に提出した）ソースコードが実際のユースケースにおいて入力されることはほとんどないと考えられる．交差重複の存在は，モデルの汎化性能ではなくモデルの記憶力を評価することにつながってしまう．そのため，モデルの真の性能を確かめるうえで交差重複を除くことは重要である．

(2) の結果については，コーディング能力判定という今回の調査で対象としたタスクに固有のものであると考えられる．本研究では，コーディング能力の基準として競技プログラミングのレーティングを使用しており，レーティングはソースコードに依存するものではなく，そのソースコードを書いたユーザに依存するものである．そのため，同じユーザが書いたソースコードを訓練セット内で学習しているか否かはモデルの精度に大きく影響を与えると考えられる．他のタスクにおいては，ソースコードを書くユーザに依存するような指標を予測したい場合には，ユーザ分割は非常に有効であると考えられる．

(3) の結果については，対象とするタスクやデータセットによって変化すると考えられる．コーディング能力の判定においては，同一のアルゴリズムやデータ構造を学習しているかどうかはモデルの精度にあまり影響を与えないことがわかった．しかし，カテゴリ分類のようにソースコードの機能を学習させるようなタスクにおいては，同じ機能をもつソースコードを学習しているかどうかという点は非常に重要である．そのため，問題分割によって同一の問題に提出されたソースコードが訓練・テストセットの両方に

出現しないようにすることで，そのモデルの性能を正確に判断することができる．

4.6 妥当性の脅威

本研究では，モデルの学習や評価において交差検証を行っていない．交差検証とは，データセット中のデータを k 個に分割してそのうちの 1 つをテストセットに，残りの k-1 個を訓練セットとしてモデルの学習と評価を行う．そして，これを k 個のデータすべてが 1 回ずつテストセットとなるように k 回学習と評価を行い，それらの精度の平均を取る手法である．本研究では，重複の有無や分割方法による複数の組合せパターンでの学習を 3 つのデータセットそれぞれで行っているため，モデルの学習に非常に時間がかかってしまい，1 つの組み合わせパターンに対して複数回の学習が必要となる交差検証を行うことができなかった．そのため，1 つの組み合わせパターンに対しては，訓練・テストセットは 1 通りの分割でしか精度を測定できていない．訓練・テストセットを同じ条件で異なる分割にした場合，本研究で紹介した結果とは精度が異なる可能性がありえる．

5. 関連研究

競技プログラミングから収集したデータセットは，本研究で使用したもの以外にも多くあり，様々なタスクで利用されている．その例を以下に示す．

- **GoogleCodeJam データセット [20]** GoogleCodeJam[*4]は，Google により毎年開催される世界最大規模の競技プログラミングのコンテストである．このデータセットは，Zhao らによって作成されたもので，12 の問題から 1,669 のソースコードを集めている．Zhao らはこのデータセットを利用して，ソースコードの類似性判定を行う深層学習モデルを作成した．
- **POJ-104[8]** POJ-104 は教育的なオンラインジャッジシステムから収集したデータセットで，104 個の問題に対してそれぞれ 500 個の解答ソースコードを含んでおり，計 52,000 個のソースコードから構成されている．このデータセットは，コード分類 [16] やコードの類似性判定 [18] の研究に利用されている．
- **APPS[4]** APPS データセットは，Codeforces や Kattis[*5]など，様々な競技プログラミングサイトから収集した 10,000 の問題と 232,421 個のソースコードで構成されている．このデータセットは，GPT 等の大規模言語モデルにおけるソースコード生成の正確性を評価ことを目的として，Hendrycks らによって作成された．そのため，解答を判定するためのテストケースも含まれており，生成されたプログラムの正確性を厳密に評価することができる．
- **CodeNet[13]** CodeNet は IBM によって作られたデータセットであり，日本の競技プログラミングサイトである AtCoder と Aizu Online Judge からソースコードを収集している．このデータセットは 4,053 個の問題から 13,916,868 個のソースコードを集めており，非常に大規模なデータセットになっている．CodeNet は Li らによって作成されたコード生成モデルである AlphaCode[7] の学習データの一部として利用されている．

これらのデータセットは，いずれも同一のユーザによって提出されたソースコードや，同一の問題に提出されたソースコードを含んでいる．つまり非独立な性質を持っているため，これらのデータセットを利用する際には対象となるタスクに合わせて，適切に非独立性を排除する必要があると考えられる．

6. まとめと今後の課題

本研究では，データセットの非独立性がコーディング能力判定モデルの精度に与える影響を調査した．その結果，非独立性によってモデルの精度が過剰に高くなっており，非独立性を排除することでモデルの精度が約 30%減少することが確認できた．非独立性の中でも，特に同一のユーザが提出したソースコードが訓練・テストセットの両方に含まれることで，モデルの精度が大きく上がることがわかった．

今後の課題は，他のデータセットやモデルにおけるデータセットの非独立性の影響を調べることである．競技プログラミングから収集したデータセットは，5 節で紹介したもの以外にも多く存在しており，様々な研究で利用されている．それらのデータセットやタスクにおいて非独立性の影響を調査することで，本研究で得られた結果の一般性が明らかになると考えられる

[*4] https://codingcompetitions.withgoogle.com/codejam
[*5] https://open.kattis.com/

参考文献

[1] Allamanis, M.: The adverse effects of code duplication in machine learning models of code, *Proceedings of the 2019 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*, pp. 143–153 (2019).
[2] Alon, U., Brody, S., Levy, O. and Yahav, E.: code2seq: Generating Sequences from Structured Representations of Code, *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, (online), available from ⟨https://openreview.net/forum?id=H1gKYo09tX⟩ (2019).
[3] 服部文志，松下　誠，井上克郎：深層学習を用いたコーディング能力判定モデルの汎化性能調査，情報処理学会研究報告， Vol. 2022-SE-211, No. 15, pp. 1–7 (2022).
[4] Hendrycks, D., Basart, S., Kadavath, S., Mazeika, M., Arora, A., Guo, E., Burns, C., Puranik, S., He, H., Song,

D. and Steinhardt, J.: Measuring Coding Challenge Competence With APPS, *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, (online), available from ⟨https://openreview.net/forum?id=sD93GOzH3i5⟩ (2021).

[5] Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780 (1997).

[6] Laakmann McDowell, G.: *Cracking the coding interview: 189 programming questions and solutions*, CareerCup (2015).

[7] Li, Y., Choi, D., Chung, J., Kushman, N., Schrittwieser, J., Leblond, R., Eccles, T., Keeling, J., Gimeno, F., Dal Lago, A. et al.: Competition-level code generation with alphacode, *Science*, Vol. 378, No. 6624, pp. 1092–1097 (2022).

[8] Lili Mou, Ge Li, L. Z. T. W. and Jin, Z.: Convolutional neural networks over tree structures for programming language processing, *In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, Vol. AAAI'16, pp. 1287–1293 (2016).

[9] 槇原啓介，松下　誠，井上克郎：ソースコード特徴量を用いた機械学習によるソースコード品質の評価手法，電子情報通信学会技術研究報告，Vol. 119, No. 113, pp. 105–110 (2019).

[10] Maleki, F., Ovens, K., Gupta, R., Reinhold, C., Spatz, A. and Forghani, R.: Generalizability of machine learning models: quantitative evaluation of three methodological pitfalls, *Radiology: Artificial Intelligence*, Vol. 5, No. 1, p. e220028 (2022).

[11] 松井智寛，松下　誠，井上克郎：判定対象の拡大を目的とした 3 値分類によるソースコード品質の評価手法，情報処理学会研究報告，Vol. 2020-SE-205, No. 7, pp. 1–8 (2020).

[12] 松井智寛，松下　誠，井上克郎：ソースコードのグラフ表現を利用した深層学習によるコーディングの専門性の判定手法，情報処理学会研究報告，Vol. 2022-SE-210, No. 12, pp. 1–8 (2022).

[13] Puri, R., Kung, D. S., Janssen, G., Zhang, W., Domeniconi, G., Zolotov, V., Dolby, J., Chen, J., Choudhury, M., Decker, L. et al.: CodeNet: A large-scale AI for code dataset for learning a diversity of coding tasks, *arXiv preprint arXiv:2105.12655* (2021).

[14] Schlichtkrull, M., Kipf, T. N., Bloem, P., vanÂ den Berg, R., Titov, I. and Welling, M.: Modeling Relational Data with Graph Convolutional Networks, *European Semantic Web Conference*, Springer International Publishing, pp. 593–607 (2018).

[15] Srinivasan Iyer, Ioannis Konstas, A. C. and Zettlemoyer, L.: Mapping Language to Code in Programmatic Context, *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1643–1652 (2018).

[16] Tal Ben-Nun, A. S. J. and Hoefler, T.: Neural code comprehension: A learnable representation of code semantics, *Advances in Neural Information Processing Systems 31*, pp. 3588–3600 (2018).

[17] 堤　祥吾：プログラミングコンテスト初級者・上級者におけるソースコード特徴量の比較，大阪大学大学院情報科学研究科修士論文 (2018).

[18] Ye, F., Zhou, S., Venkat, A., Marcus, R., Tatbul, N., Tithi, J. J., Petersen, P., Mattson, T. G., Kraska, T., Dubey, P., Sarkar, V. and Gottschlich, J.: MISIM: An End-to-End Neural Code Similarity System, *CoRR*, Vol. abs/2006.05265 (online), available from ⟨https://arxiv.org/abs/2006.05265⟩ (2020).

[19] Zhang, X., Zhao, J. and LeCun, Y.: Character-Level Convolutional Networks for Text Classification, *Advances in Neural Information Processing Systems*, Vol. 1, pp. 649–657 (2015).

[20] Zhao, G. and J.Huang: DeepSim: Deep Learning Code Functional Similarity, *Proc. FSE 2018*, pp. 141–151 (2018).