原　著

臨床検査学教育 Vol.18, No.2 p.161 ～ 168, 2026.

専門知識向上のための学生指導における機械学習活用方法の検討

河　月　　稔 *1§　岩　田　浩　明 *1　森　　徹　自 *1　臼　井　真　一 *2

要　旨　本研究では、学生指導の判断に教員視点による評価と人工知能を併用する価値を検討した。2022 年度から 2024 年度の卒業生 114 人を対象として 3 年次までの必修専門科目の成績と 4 年次の国家試験模擬試験の結果を収集した。分析対象として 27 人を選出し、機械学習による模擬試験点数予測を行うとともに、教員 9 人に指導必要性の判定を実施してもらった。個々の教員が指導不要と判定した場合、模擬試験の点数が上位だった者の割合は 42.3% と高くなかったが、指導が必要と判定した場合、模擬試験の点数が下位だった者の割合は 75.0% と比較的高かった。一方、機械学習による点数予測は、模擬試験の点数と中程度の相関を認めた（$r = 0.631, p < 0.01$）。また、予測結果が 101 点以上だった場合、模擬試験の結果が 80 点以下だった人はいなかった。機械学習の結果は、特に模擬試験の結果が良くない人の見逃しを減らすための補助資料として、教員の従来の判断に併用して活用する価値があると考えられた。

キーワード　教員の視点、機械学習、模擬試験結果予測

I. 緒　　言

臨床検査技師養成校の学生教育において、国家試験に合格できるように指導を行うことは重要である。また、国家試験に向けた勉強によって知識量を増加させることは、その先の適切な検査技術の習得にもつながる。さらに、採用試験で専門科目に関して出題する医療関係施設は多いため、就職活動においても有利に働く。したがって、実践力を身につける臨地実習や就職活動が始まる前に学生へ適切な指導を行っていくことが有益と考えられる。この考えを著者らが所属する鳥取大学の検査技術科学専攻に当てはめると、4 年次から臨地実習が始まり、大半の学生は並行して就職活動を行うため、3 年終了時が指導に適した時期となる。

早い段階から学生に介入して成績を向上させる取り組みとして、全国の臨床検査技師養成校では創意工夫がなされているが、最近は人工知能（AI）の活用が注目されている。例えば、ディープラーニングにより模擬試験の結果から国家試験の点数を予測するモデルにおいて、かなりの相関を得ることができたと報告されている[1)]。また、同様の取り組みは他の医療系職種の学生教育でも行われており、機械学習に基づく国家試験合否予測を検討した研究がいくつも報告されている[2)～6)]。国家試験の合格率を上げるといった学生教育の質の向

*1 鳥取大学医学部保健学科生体制御学講座　§ kouzuki@tottori-u.ac.jp
*2 鳥取大学医学部保健学科病態検査学講座

上のためにAIを活用できる可能性が示唆されているが、AIの学習に用いるデータが不正確であったり、多様なデータを網羅できていなかったりする場合、十分な予測精度を得ることは難しい。予測精度が不十分であると教員の感覚をはるかに凌駕するものではないと考察している研究もある[4]。

以上のことより、人の感覚や経験値を重視しつつ、AIの評価結果を併用することが重要であると考えられる。しかし、AIを教育に活かした取り組みの多くは、教員の判断との比較や併用可能性については十分検討されていない。また、多くの取り組みは国家試験の点数や合否を予測することに焦点を当てている。しかし、就職活動に活かすことを考慮すると、教員の介入も入り、十分に勉強した成果の国家試験ではなく、自主的な学習姿勢が強く反映される早い時期の模擬試験の結果を予測し、指導に活用することが望ましいと考える。

そこで、本研究では、3年次までの成績を用いてAIで解析を行い、それとは別に3年次までの成績を見て積極的介入が必要かどうかを教員の視点で判定してもらい、それぞれの結果を4年次に実施した国家試験模擬試験の結果と比較した。そして、指導必要性の判断に際して、教員の視点とAIを併用することの有用性を検討した。

II. 対象と方法

研究デザイン

本研究は鳥取大学医学部倫理審査委員会の承認（No. 25A058）を得て実施した。同意取得について、過去の情報を用いる対象者へは、あらかじめ情報を鳥取大学医学部附属病院のホームページに公開して拒否できる機会を保証する方法（オプトアウト）とした。一方、新たに情報を取得する調査の対象者には、事前に研究の内容を説明したうえで、調査実施時に使用したアンケート回答用Googleフォーム内の同意確認欄へのチェックによって同意を得て実施した。

2-1　対　　象

本研究の対象者は、鳥取大学医学部保健学科検査技術科学専攻の卒業生と教員である。卒業生の選択基準は、2022年度から2024年度に卒業した人とした。一方、教員については、選択基準を常勤かつ3年以上在籍していることとし、除外基準は研究の実施に関わった人とした。

2-2　手　　順

2025年7月までに、卒業生の3年次までの必修専門科目（49科目）の成績、および4年次に実施した国家試験模擬試験（医歯薬出版株式会社）の結果を、学内に保管してある情報から収集した（n=115）。なお、既修得単位として認定されている科目には評点がついていないため、該当の科目がある者の全てのデータは分析に用いないこととした（1人が該当）。

次に、例年9月に受験している第1回目の模擬試験の結果をもとに評価対象者を選出した。まず、3クラス（2022年度、2023年度、2024年度の3つの年度）それぞれについて、3分位法で上位、中位、下位の3群に分類した。そして、模擬試験の点数が各群の中央値付近に位置していた3人をそれぞれの群から選び出し、各クラス9名、計27名を評価対象として確定させた。なお、点数に基づく分類方法も検討したが、点数分類について科学的妥当性を検証した先行研究を調べることができなかったため、恣意的にならない分類方法として3分位法による分類を採用した。この評価対象者のデータを使用して、2025年9月までに機械学習による分析を行い、それとは別に指導必要性の判定を教員に実施してもらった。

2-2-1　機械学習による分析

卒業生の3年次までに履修した必修専門科目の成績を説明変数とし、国家試験模擬試験の成績を目的変数として予測する機械学習モデルを構築した。機械学習アルゴリズムには、分類・回帰問題において高い汎化性能が報告され、広く利用されているランダムフォレスト法を採用した[7]。データセットは、評価対象外の87人を学習データとしてモデル構築に用い、評価対象とした27人をテストデータとして予測性能の評価を行った。

2-2-2　教員による指導必要性の判定

評価対象とした27人にID番号を付して、個人を特定できる情報（氏名、学生番号、卒業年等の情報）を削除し、科目名、成績（各科目の得点）、

本研究で扱えるデータから算出した各科目の平均点を記載したデータセットをPDFファイルで作成し、研究対象者となる教員に配信した。教員には、成績を見て27名それぞれに対して「個別指導は必要ない」、「個別指導を要する可能性あり」、「個別指導が必要」の3段階で評価し、判定結果をエクセルに入力するようにお願いした。なお、判定については3段階で評価してもらうことのみを説明し、判定の根拠となる具体的な基準は各教員の考えに委ねることとした。また、「○○学という科目に重きを置いて評価した」といったように、どのような視点で成績判定を行ったのかについて、任意での回答を依頼した。個人が特定できないように全て無記名で収集を行うために、エクセルはオンラインストレージであるProselfから提出してもらい、その他の項目はGoogleフォームから回答してもらうこととした。

2-3 統計解析

教員による判定結果や機械学習による模擬試験点数予測の妥当性を評価するために、模擬試験の結果との相関分析や一致度を算出した。

相関分析については、数値として扱える機械学習の予測結果と模擬試験の結果の比較についてのみ実施した。統計ソフト（IBM SPSS Statistics version 27）を使用して、正規分布の評価にはShapiro-Wilk検定を用い、その後Pearsonの相関係数を算出した。有意水準は0.05として評価した。また、研究計画書に基づかない事後解析として、機械学習モデルの予測性能評価のために決定係数を算出した。

一致度については、算出に際して次の基準に従って模擬試験の結果や機械学習の予測結果を分類した。模擬試験の結果は、①前述の3分位法で3群に分類、②80点以下、81～100点、101点以上の3群に分類、という2つの方法で群分けを行い、点数が低い群から「C：個別指導が必要な群」、「B：個別指導を要する可能性がある群」、「A：個別指導は必要ない群」とした。なお、②の分類方法は科学的な根拠に基づくものではなく、研究関係者（本論文の著者）の経験に基づく分け方であるが、研究開始前に予め設定していた基準である。本研究の結果を多面的に評価するために、評価対象者を選出した①の分類方法に加えて、②の分類方法でも検討を行うこととした。一方、機械学習の予測結果は、上記②の点数による分類方法のみを用いて群分けを行った。この理由としては、上記①のように3分位法で分類すると、仮に全員が高得点や低得点と予測された場合において誤った分類になると考えられたためである。以上のように分類を行い、教員の判定結果や機械学習の予測結果と模擬試験の結果との一致度を算出した。なお、教員の判定結果は、個人の結果以外にも、各教員の判断基準の差異により見解が分かれる場合を考慮して全教員の判定で最も多かった結果を採用するといった多数決の結果での分析も実施した。一部、判定結果が同数の場合もあったが、保守的に考えて指導が必要とする判定結果を採用した。

教員の評価視点に関する自由記述式の質問については、回答の意味が変わらないように配慮し、さらに同じような趣旨の回答があった場合は統合して要旨を記述した。

III. 結　　果

3-1 研究対象者の概要

卒業生の情報については、115人のデータを収集し、成績情報に不足が無かった114人のデータを分析に使用した。一方、成績の判定調査については、9人の教員から研究への同意を得て、判定結果を収集した。なお、成績の判定調査対象となった卒業生の模擬試験の平均点数は、上位群で125点、中位群で97点、下位群で91点だった。

3-2 機械学習の予測結果と模擬試験の結果との相関分析

ランダムフォレスト法を用いて機械学習モデルを構築し、テストデータにおける予測性能を評価した。その結果、決定係数は$R^2 = 0.398$であった。また、機械学習で予測した点数と模擬試験の実際の点数の比較を行った結果、有意な中程度の相関関係を認めた（$r = 0.631$, $p < 0.01$）（**図1**）。実際の模擬試験の点数よりも低い得点を付けている場合が多かったが、模擬試験の結果が70点だった人を95点、89点だった人を124点といったように、

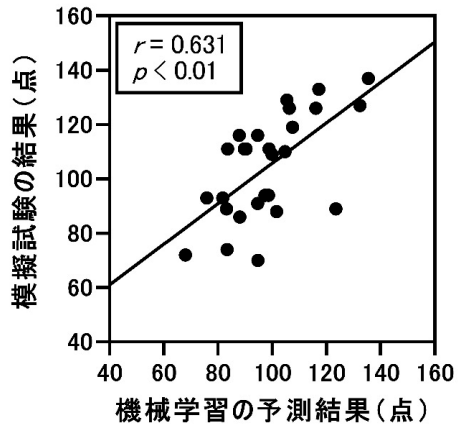


図 1　機械学習の予測結果と模擬試験の結果との相関分析

実際の点数よりも大幅に高い予測結果になっていた例もあった。

3-3　教員の判定結果や機械学習の予測結果と模擬試験の結果との一致度

3 分位法で分類した模擬試験の結果と教員の判定結果との比較を**表 1** に示す。教員個々の判定結果と教員多数決の判定結果の全体の一致度はそれぞれ 47.3% と 51.9% であり、大きな違いはみられなかった。各判定結果の一致度別の差は、A 判定で 2.7%、B 判定で 18.9%、C 判定で 8.3% と B 判定では若干の乖離があったものの、A 判定や C 判定では大きく乖離していなかった。教員は、「個別指導は必要ない」と判定することが多い傾向であったが、模擬試験の分類結果との一致度は教員個々の判定で 42.3%、教員多数決の判定で 45.0% と高くはなく、実際の模擬試験の結果が下位に位置していた場合もあった。一方、教員が「個別指導が必要」と感じたときは、模擬試験の分類結果との一致度が教員個々の判定で 75.0%、教員多数決の判定で 66.7% と比較的高く、概ね模擬試験の結果が下位に位置していた人を判定できていた。なお、機械学習では模擬試験の点数を予測したため、模擬試験の結果を点数ではなく 3 分位法で順位に基づいて分類した結果との比較を行うことは不適当と判断したことから、当該分析は実施していない。

次に、予め設定していた点数で分類した模擬試験の結果と教員の判定結果や機械学習の予測結果との比較を**表 2** に示す。模擬試験の分類結果との全体の一致度は、教員個々の判定で 54.3%、教員多数決の判定で 55.6%、機械学習の予測結果で 55.6% であり、大きな差は無かった。教員の判定結果については、**表 1** の結果と比較して「個別指導が必要」と感じたときの一致度が大きく低下していた。一方、機械学習の予測結果は、101 点以上の点数を付けた場合、模擬試験の点数が 80 点以下だった人はおらず、模擬試験の分類結果との一致度も 80.0% と高かった。また、機械学習の予測結果が 80 点以下だった場合、模擬試験の分類結果との一致度は 50.0% と高くはなかったが、模擬試験の点数が 101 点以上だった人はおらず、大きく予測が外れてはいなかった。最後に、教員の判定結果と機械学習の予測結果を併用することで模擬試験の結果との一致度が向上するかについて検討した。教員の判定結果と機械学習の予測結果に乖離があった場合、保守的に考えて「個別指導が必要」、あるいは「個別指導を要する可能性がある」という悪い判定をしていた方を採用したが、模擬試験の分類結果との一致度は 55.6% であり、教員の判定結果と機械学習の予測結果を個別に用いた場合との差異は認められなかった。

3-4　教員の評価視点

どのような視点で評価したのかについて、教員 9 人中 8 人から回答を得た（1 人は無回答）。内容は「再試験を受験した科目の割合から判断」、「各科目の平均点からの差を考慮」、「上級学年の成績を優先して判断」、「学年が上がるにつれて成績が良くなっているかを考慮」、「血液学、生理学、免疫学の成績を重視して評価」に集約されると解釈した。

Ⅳ. 考　　察

教員は各自の基準で重みづけをしながら指導の必要性を評価していた。特に、国家試験の点数配分が大きい試験科目に関連性が高い講義が増えてくる上級学年の成績を重視していた傾向であった。なお、個々の教員の判定結果と多数決で一つの判

表 1　3 分位法で分類した模擬試験の結果と教員の判定結果との比較

	教員個々の結果；一致度 =115（47.3）[ab]			教員多数決の結果；一致度 =14（51.9）[ac]		
	A	B	C	A	B	C
模擬試験の結果						
A; 上位群	77（42.3）	4（9.8）	0（0）	9（45.0）	0（0）	0（0）
B; 中位群	53（29.1）	23（56.1）	5（25.0）	5（25.0）	3（75.0）	1（33.3）
C; 下位群	52（28.6）	14（34.1）	15（75.0）	6（30.0）	1（25.0）	2（66.7）

結果は人数（%）で示す。
[a] A; 個別指導は必要ない、B; 個別指導を要する可能性あり、C; 個別指導が必要
[b] 教員の回答を全て反映した結果
[c] 教員の回答で最も多かった判定を採用した結果

表 2　点数で分類した模擬試験の結果と教員の判定結果や機械学習の予測結果との比較

	教員個々の結果；一致度 =132（54.3）[ab]			教員多数決の結果；一致度 =15（55.6）[ac]		
	A	B	C	A	B	C
模擬試験の結果						
A; 101 点以上	108（59.3）	22（53.7）	5（25.0）	12（60.0）	2（50.0）	1（33.3）
B; 81 ～ 100 点	58（31.9）	16（39.0）	7（35.0）	6（30.0）	2（50.0）	1（33.3）
C; 80 点以下	16（8.8）	3（7.3）	8（40.0）	2（10.0）	0（0）	1（33.3）

	機械学習の結果；一致度 =15（55.6）[d]			併用した結果；一致度 =15（55.6）[e]		
	A	B	C	A	B	C
模擬試験の結果						
A; 101 点以上	8（80.0）	7（46.7）	0（0）	8（80.0）	6（42.9）	1（33.3）
B; 81 ～ 100 点	2（20.0）	6（40.0）	1（50.0）	2（20.0）	6（42.9）	1（33.3）
C; 80 点以下	0（0）	2（13.3）	1（50.0）	0（0）	2（14.3）	1（33.3）

結果は人数（%）で示す。
[a] A; 個別指導は必要ない、B; 個別指導を要する可能性あり、C; 個別指導が必要
[b] 教員の回答を全て反映した結果
[c] 教員の回答で最も多かった判定を採用した結果
[d] A; 101 点以上、B; 81 ～ 100 点、C; 80 点以下
[e] 教員多数決の結果と機械学習の結果のうち、判定結果が悪かった方を採用（A が最も良く、C が最も悪い判定と判断）

定に絞った結果において、模擬試験の結果との一致度は大きく変わらなかった。このことより、各教員が概ね同じ視点で判定していたと考えられた。教員の経験値によって学生評価の視点が異なることが報告されているが[8)9)]、本研究で選択基準とした 3 年以上同じ組織に在籍していれば、価値観の共有が行われ、考え方が収束していくのではないかと推察している。しかし、教員の判断基準が全く同じになることは無いため、教員間で意見交換を行い、学生指導にあたることが望ましいと考える。

教員が指導の必要性を感じた場合、概ね模擬試験の結果がクラスの中でも高くない傾向にあった。つまり、教員の評価基準は指導が必要であると判断した場合の正確性が高いと考えられた。ただし、模擬試験の結果を 3 分位法ではなく点数で分類した際は、80 点以下の人を判定する正確性が低くなっていた。教員が点数を予測しながら判断していた可能性はあるものの、本研究ではあくまで指導の必要性の判断を求めたことが要因と考える。

5 段階評価のような判定は具体的な点数を予測するような判定よりも実際のテストの点数との相関が低かったことが示されており[10]、何を判断してもらうのかによって結果は変わると理解できる。したがって、実際の教育現場において本研究と同様の検証を行う際は、教員に何を判断してもらうのかを明確にし、何と比較するのかを十分に考慮しなければならない点は注意を要する。

一方、教員が指導は必要ないと感じた場合、模擬試験の分類結果との一致度は高くはなかった。模擬試験の結果が悪かった例もあり、過大評価をしている傾向にあった。学生への指導に際して最も回避したいのは、指導が必要ないと感じた学生の模擬試験の結果が悪く、介入開始が後手に回ってしまう状況である。教員が学生の成績を過大評価する要因に関する一つの考察として、より高い能力評価は教員が学生を信頼していることを示していると記載していた文献があった[11]。教員には成績以外の個人が特定できる情報は提供していなかったが、全般的に心理的要因が影響した可能性はあったかもしれない。以上のように、本研究の結果においては、教員の判断のみでは指導が必要な学生の見逃しが生じる可能性が示唆されたが、この見逃しを防ぐために、機械学習の結果を併用することが有用であると考えられた。機械学習の予測結果は模擬試験の結果と中程度の相関を示しており、予測が 101 点以上の場合、模擬試験が 80 点以下だった人はおらず、一方で予測が 80 点以下の場合、模擬試験が 101 点以上だった人はいなかった。したがって、教員が指導は必要ないと感じたときに、機械学習の予測点数が高かった場合は本当に指導が必要ないと判断し、一方で予測点数が低かった場合は自分の判断を再考するといった活用方法が想定される。AI を用いた取り組みは、その予測精度の向上に焦点を当てることが多いが、一つの臨床検査技師養成校で使用できるデータ数には限界があるため、完璧な予測精度を得ることは難しい。そのため、教員の判断を重視しつつ、AI をどのように活用するのかを考えていくことが重要である。本研究の結果からは、教員が自信を持って学生指導にあたるため、および国家試験に関する専門知識の習得に後れを取る可能性が高い学生を看過しないために、AI を活用できると考えられた。今後も教育現場における適切な AI 活用法に関する情報が集積されていくことを期待する。

本研究の結果を踏まえた指導の在り方として、3 年次が終了した段階で、必修専門科目の成績や、機械学習の予測結果を提示しながら学生と個人面談を行うことが有用と考える。個人面談は労力を要するが、適切な時期に適切な対象者へ行うことは有効な支援手段となる[12]。学生にとっては就職活動や国家試験の勉強に向けての自己分析の機会となり、学習意欲の向上につながる可能性がある。また、成績を振り返ることで、客観的に自分の得意分野や苦手分野を把握することができ、就職試験の履歴書を書く際に役立つ可能性もある。具体的な機械学習の予測結果の示し方は検討中であるが、**図 2** のように総合評価を A、B、C 判定で示し、合わせて国家試験の試験科目（10 科目）毎の予測スコアをレーダーチャートで提示することを考えている。また、予測結果に対するそれぞれの特徴量（必修専門科目の成績）の貢献度を計算し、貢献度が高かったものを各試験科目の点数を上げるために勉強が必要な必修専門科目として提示することも検討している（**図 2**）。最終的には全ての科目を勉強しなければならないが、効率よく点数を伸ばすための端緒を提示することで、学生の満足度も高くなると推測する。今後は本研究の結果を踏まえた指導を実践し、その効果について検証していく必要があると考えている。

最後に本研究の限界について述べる。1 つ目は、必修専門科目の点数のみで教員に指導の必要性を判定してもらった点である。例えば、学生の行動や性別の影響によって教員の評価に偏りが生じる可能性が示されている[13)～15)]。実際の教育現場において、点数だけでなく個々の学生の特性を加味して指導の必要性を評価する場合は、本研究の分析結果と乖離が生じる可能性があることに留意が必要である。2 つ目は、学習データが少なかった点である。機械学習は学習データが多くなるほど精度が上がるため、87 人というデータ数では予測

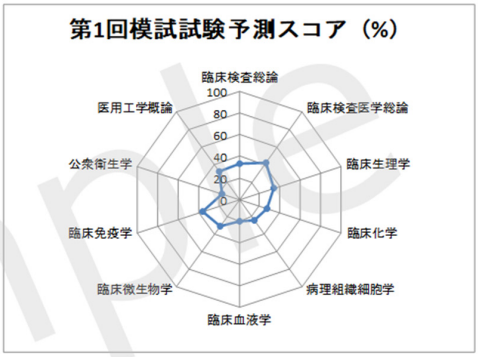
第1回模擬試験に向けた指導

学生番号　○○○○○○○○
氏名　○○ ○○

予測判定　C

A:上位1/3　B:中位1/3　C:下位1/3

重要視すべき専門科目を中心に勉学に励みましょう



重要視すべき専門科目

	臨床検査総論	臨床検査医学総論	臨床生理学	臨床化学	病理組織細胞学	臨床血液学	臨床微生物学	臨床免疫学	公衆衛生学	医用工学概論
1位	病態血液学	病原体検査学Ⅰ	基礎免疫輸血学	臨床化学検査学	疾病論Ⅰ	遺伝子検査学	臨床化学検査学	臨床化学検査学	疾病論Ⅰ	医用工学
2位	輸血・移植検査学実習	遺伝子検査学	医用工学	基礎免疫輸血学	臨床化学検査学	人体組織学	臨床病理学概論	病原体検査学Ⅰ	病態生理情報検査学実習	遺伝子検査学
3位	臨床病理学概論	人体の構造と機能	遺伝子検査学	病態血液学	人体組織学実習	病態血液学	病原体検査学Ⅰ	病態血液学実習	臨床化学検査学	疾病論Ⅱ

図 2　機械学習の分析結果を学生指導に活用する際の提示案

第 1 回模擬試験の予測点数がクラスの中で上位、中位、下位のどこに位置していたかの判定結果と、国家試験の全 10 科目毎に取得が見込まれる得点率を提示している。なお、予測点数といった数値情報は、過度に意識され、誤った解釈につながることが危惧されるため示していない。また、予測結果に対するそれぞれの特徴量（必修専門科目の成績）の貢献度を計算し、貢献度が高かった上位 3 つの必修専門科目を、各試験科目の点数を上げるために重要視すべき勉強科目として提示している。

の正確性に限界があったかもしれない。しかし、教員の入れ替わりや講義内容が更新されていくことを考慮すると、古すぎるデータを活用することは適当ではないと考える。そのため、本研究では 3 クラス分のデータで検証を行った。以上のように、本研究の結果を解釈するうえでいくつか考慮すべき点はあるが、これまでに十分検討されていなかった、教員の判断に AI を併用する有用性を調査した本研究の成果は、今後の学生教育における AI 活用の在り方を検討していくための貴重な情報源になると考えている。

V. 結　　論

機械学習の予測結果は、特に模擬試験の結果が良くない人の見逃しを減らすために活用できる可能性が示唆され、従来までの教員の判断に併用する価値はあると考えられた。しかし、機械学習の予測には一定の不確実性が伴うため、全ての結果を信じるのではなく、教員が自分の判断を見つめ直す、あるいは学生の学習意欲を向上させる補助資料として活用するのが望ましい。最終的には自身の教員としての経験を活かして学生指導にあたることが肝要である。AI は教員の代わりになることはなく、補完するものであるという考えも示されており[16)～18)]、適切な運用方法が求められる。

VI. 謝　　辞

成績の判定調査にご協力いただいた鳥取大学医学部保健学科検査技術科学専攻の先生方に深謝申し上げます。本研究は令和 7 年度 学長裁量経費 学長リーダーシップ経費の助成を受けて実施した。

VII. COI 状態

投稿論文に関連し、発表者らに開示すべき COI 関係にある企業などはありません。

※本稿の一部は、第 19 回日本臨床検査学教育学会学術大会（2025 年 8 月 22 日）において一般演題として発表した。

文　献

1）高橋祐司 , 沖野久美子 , 松尾淳司 , 吉田 繁 , 幸村 近 , 二瓶裕之 . Google Colaboratory を用いた国家試験予測 AI の開発（抄）. 臨床検査学教育（第 18 回日本臨床検査学教育学会学術大会）2024; 15（Supp）: 76.

2）清水典史 , 井上 寛 , 松延千春 , 高露恵理子 , 椿 友梨 , 白谷智宣 . サポートベクターマシンに基づく試験合否予測モデルの構築 . 薬学教育 2018; 2: 141-7.

3）清水典史 , 井上 寛 , 松延千春 , 椿 友梨 , 白谷智宣 . 薬剤師国家試験合否予測モデルを利用した学生の学修状況の把握 . 薬学教育 2019; 3: 111-5.

4）小牧祥太郎 , 戌亥啓一 , 松尾康弘 , 高吉 進 , 福元恵美 , 福永陽平 . 言語聴覚士国家資格取得を支援する機械学習モデルの有効性の検討－模擬試験解答データを用いた試験の合否予測より－ . 言語聴覚研究 2020; 17（4）: 309-17.

5）松延千春 , 井上 寛 , 窪田敏夫 , 白谷智宣 . ランダムフォレストに基づく薬剤師国家試験合否予測モデルの構築とその活用法に関する検討 . エコノミクス 2022; 27（1）: 1-8.

6）松延千春 , 白谷智宣 , 窪田敏夫 . 低学年次成績データを用いた薬剤師国家試験合否予測のためのランダムフォレストによるモデル構築と予測因子の探索・予測精度の評価 . 薬学教育 2025; 9: e09001.

7）Breiman L. Random Forests. Machine Learning 2001; 45: 5-32.

8）Kosel C, Bauer E, Seidel T. Where experience makes a difference: teachers' judgment accuracy and diagnostic reasoning regarding student learning characteristics. Front Psychol 2024; 15: 1278472.

9）Seidel T, Schnitzler K, Kosel C, Stürmer K, Holzberger D. Student Characteristics in the Eyes of Teachers: Differences Between Novice and Expert Teachers in Judgment Accuracy, Observed Behavioral Cues, and Gaze. Educ Psychol Rev 2021; 33: 69-89.

10）Hoge RD, Coladarci T. Teacher-Based Judgments of Academic Achievement: A Review of Literature. Review of Educational Research 1989; 59（3）: 297-313.

11）Urhahne D, Wijnia L. A review on the accuracy of teacher judgments. Educational Research Review 2021; 32: 100374.

12）Hashimoto S, Saimaru H, Yamaki F, Kase Y. Analysis and examination of individual interviews with sixth-year pharmacy students on their performance. Japanese Journal of Pharmaceutical Education 2023; 7: 271-280.

13）Ferman B, Fontes LF. Assessing knowledge or classroom behavior? Evidence of teachers' grading bias. Journal of Public Economics 2022; 216: 104773.

14）Protivínský T, Münich D. Gender Bias in teachers' grading: What is in the grade. Studies in Educational Evaluation 2018; 59: 141-9.

15）Lavy V, Sand E. On the origins of gender gaps in human capital: Short- and long-term consequences of teachers' biases. Journal of Public Economics 2018; 167: 263-79.

16）Office of Educational Technology. Artificial intelligence and the future of teaching and learning: insights and recommendations. U.S. Department of Education, Washington, DC, 2023.

17）Fajardo-Ramos DC, Chiappe A, Mella-Norambuena J. Human-in-the-loop assessment with AI: implications for teacher education in Ibero-American universities. Front Educ 2025; 10: 1710992.

18）Devika KH. AI Teachers: Partners in Education, Not Replacements. International Journal of Latest Technology in Engineering Management & Applied Science 2025; 14（4）: 678-81.