Deep learning in the marking of medical student short answer question examinations: Student perceptions and pilot accuracy assessment

L. Hollis-Sando[1], C. Pugh[1], K. Franke[1], T. Zerner[1], Y. Tan[1, 2], G. Carneiro[1], A. van den Hengel[1], I. Symonds[1], P. Duggan[1] & S. Bacchi[1, 2]

Abstract

Introduction: Machine learning has previously been applied to text analysis. There is limited data regarding the acceptability or accuracy of such applications in medical education. This project examined medical student opinion regarding computer-based marking and evaluated the accuracy of deep learning (DL), a subtype of machine learning, in the scoring of medical short answer questions (SAQs).

Methods: Fourth- and fifth-year medical students undertook an anonymised online examination. Prior to the examination, students completed a survey gauging their opinion on computer-based marking. Questions were marked by humans, and then a DL analysis was conducted using convolutional neural networks. In the DL analysis, following preprocessing, data were split into a training dataset (on which models were developed using 10-fold cross-validation) and a test dataset (on which performance analysis was conducted).

Results: One hundred and eighty-one students completed the examination (participation rate 59.0%). While students expressed concern regarding the accuracy of computer-based marking, the majority of students agreed that computer marking would be more objective than human marking (67.0%) and reported they would not object to computer-based marking (55.5%). Regarding automated marking of SAQs, for 1-mark questions, there were consistently high classification accuracies (mean accuracy 0.98). For more complex 2-mark and 3-mark SAQs, in which multiclass classification was required, accuracy was lower (mean 0.65 and 0.59, respectively).

Conclusions: Medical students may be supportive of computer-based marking due to its objectivity. DL has the potential to provide accurate marking of written questions, however further research into DL marking of medical examinations is required.

Keywords: deep learning; natural language processing; automation; medical education

[1] University of Adelaide, Adelaide, South Australia, Australia
[2] Royal Adelaide Hospital, Adelaide, South Australia, Australia

Correspondence: Dr Stephen Bacchi stephen.bacchi@sa.gov.au

Introduction

Machine learning (ML) is the process of computers learning from data (Deo, 2015). Not only is it now important for medical students to learn about ML so that they may interpret the results produced by ML algorithms (James et al., 2021), ML may also aid in teaching medical students. One way in which ML may assist with medical education is through assessments of performance (Dias et al., 2018). ML has previously been applied to the grading of short answer questions (SAQs) in various fields outside of medicine (Burrows et al., 2015; Nadkarni et al., 2011).

SAQs may have multiple types, including 1-mark very short answer questions (VSAQs), which have been proposed as an alternative to standard multiple-choice questions (Puthiaparampil & Rahman, 2020). There is evidence that ML results in similar scores on 1-mark questions with a large sample size (n = 8,007) when compared to human markers (Latifi et al., 2016). ML has also been applied to the scoring of written notes for the United States Medical Licensing Examination (Salt et al., 2018).

Natural language processing (NLP) is one aspect of ML relevant to this topic. It involves the interaction between human languages (either written or spoken) and computers. At its most basic level, NLP could be viewed to include keyword searches. Sophisticated NLP analyses involve much more than this type of analysis, including looking at the context and order in which words appear (Locke et al., 2021). Proprietary software that utilises types of NLP to facilitate marking of SAQs is currently in use. However, open-source software that may be applicable to this task is also available. Open-source software includes algorithms, such as deep learning (DL) algorithms, and new content, which is constantly developing (Sidey-Gibbons & Sidey-Gibbons, 2019). Unlike earlier methods of NLP, more recent methods of ML text analysis, such as DL algorithms, including convolutional neural networks (CNN), have the potential to interpret the semantic context of the text that is analysed (Lai et al., 2015). These methods may increase the accuracy of the text analysis. There is limited research available on the accuracy of such new DL methods of NLP, including those implemented with open-source software, in the scoring of medical student examinations.

Currently, there are mixed opinions regarding the role of ML in marking of written assessments, and lack of acceptance may be a key barrier to implementation. At this time, there is limited evidence available on the acceptability to medical students of NLP in examination marking.

Therefore, this project was conducted with the aim of assessing: a) student opinion in regards to the acceptability of computer-based (machine learning) marking and b) a preliminary study of the accuracy of DL NLP implemented with open-source software in the scoring of medical student SAQs.

Methods

This study involved an anonymous survey and machine learning pilot performance analysis.

Data collection

Participants

A voluntary anonymous online formative examination was conducted for Year 4 and Year 5 University of Adelaide medical students in October 2018. The University of Adelaide Bachelor of Medicine Bachelor of Surgery program is a 6-year undergraduate degree.

Exam structure and marking

The examination comprised 100 questions, including 1-mark VSAQs (completed by only half the cohort), 2-mark SAQs, 3-mark SAQs and multiple-choice questions (MCQs) (completed by the entire cohort). The examination was time limited, with a time limit of 2.5 hours. The examination content covered topics including adult medicine, surgery and general practice. Prior to the examination, marking schemes were devised for the SAQs (see Table 1 for example questions and marking schemes). This marking rubric provided clear instructions on answers for which marks should be allocated. Following the completion of the examination, student written answers were marked by senior medical students (sixth-year students, who while on medical education rotations routinely engage in examination marking) and junior medical officers. A degree of leeway was afforded to examiners concerning answers that were considered reasonable but had not been explicitly discussed in the predefined marking criteria. In cases in which such an instance was encountered, the decision to allocate marks was achieved through group discussion. Similarly, when there was uncertainty regarding the number of marks that should be allocated to a given answer, consensus was achieved through group discussion.

Survey

Students completed a 6-point Likert-scale survey (see Appendix) prior to the examination, gauging opinion on computer-based marking. The questions for this survey were adapted from other machine learning and medical education literature (Pinto dos Santos et al., 2019). Answers ranged from "*very strongly disagree*" to "*very strongly agree*". Prior to administration, the survey was piloted on a convenience sample of senior medical students, with questions refined to improve clarity following this pilot. Students who were involved in this pilot survey did not partake in the study.

DL analysis

DL analysis was conducted using open-source Python libraries. Preprocessing of text (cleaning and simplifying the type of text present) involved removal of non-letter characters (such as punctuation) and stopwords (words that contain little value in the

Table 1

Example 1-, 2- and 3-Mark Short Answer Questions and Marking Schemes

Example Question	Example Marking Scheme
A 41-year-old female presents to the GP with epigastric pain and diarrhoea. An endoscopy is subsequently performed, which reveals a jejunal ulcer. Further investigations are conducted, and the patient is found to have an elevated fasting serum gastrin. What is the most likely diagnosis? (1 mark)	(1 mark) Zollinger-Ellison syndrome OR gastrinoma
You are a metropolitan GP. Mrs Brown, an 80-year-old female, is brought to see you by her adult children because they are worried about her memory. Over the past 3–6 months, they have observed that she has been increasingly dishevelled and is not keeping appointments. Mrs Brown reports feeling fatigued, constipated and having a reduced appetite. She has recently put on 5 kg. What is the most appropriate first line investigation? Justify your answer (2 marks).	(1 mark) Identifies hypothyroidism as most likely diagnosis OR demonstrates sound clinical reasoning. (1 mark) Identifies thyroid functions test (or thyroid stimulating hormone level) as most appropriate investigation.
You are a rural GP. Mrs Green, a 59-year-old female, presents with a tremor affecting both hands over the last 6 months. The tremor affects both hands equally and is most noticeable when she is using her hands (such as when lifting an item or trying to drink from a cup). She recalls that her mother also developed a similar tremor in her 50s. The tremor is bilateral, high frequency and flexion/extension in pattern at the wrists. Which medication is the first line pharmacological treatment for this condition? Justify your answer (3 marks).	(1 mark) Identifies essential tremor as most likely diagnosis. (1 mark) Identifies propranolol OR beta-blocker as first-line treatment. (1 mark) Demonstrates clinical reasoning to support diagnosis OR treatment choice.

meaning of a sentence, such as "the" and "is"), word stemming (shortening words to reduce the number of unique words present, such as shortening both "hypertensive" and "hypertension" to "hyperten"), tokenisation (replacing sub-word, word or word combinations with non-word values) and padding of sequence length (ensuring all sequences were of equal length with blank values). Further descriptions of these processes in a medical setting can be found in recent review articles (Locke et al., 2021). The final corpus was then split into a training set and test set (85%/15% split). The most frequently appearing 99% of words were included.

A CNN structure was selected for classification in this study. The CNN was trained using supervised methods on the training dataset. This supervised training involved providing the models with the student answers in the training dataset (input) and the marks that each answer was allocated by the human scorer (output). The CNN was then provided with the opportunity to find the relevant associations between the input and the output such that it could predict the marks allocated to student answers that it had not seen previously (the test dataset for performance analysis). These training processes commenced with classification experiments conducted on the training set with a simple CNN structure prior to the addition of further hidden layers and nodes. Ultimately, the model used for all classification experiments had one embedding layer, one convolutional

layer and one maximum pooling layer. These were then followed by eight alternating dense layers (512 nodes) and dropout layers (dropout rate 0.2). The number of outputs in the output layer was defined by the number of categories possible in each question. For example, for a 3-mark question, there were four possible output classifications (0 marks, 1 mark, 2 marks or 3 marks).

Four 3-mark, four 2-mark and four 1-mark questions were randomly selected for cross validation experiments. For each question, 10-fold cross-validation experiments were conducted on the training set. This training set was used to tune the hyperparameters for each individual question. If > 90% accuracy was achieved on the training set, the size of the training set was then decreased to determine how low a number of training examples was required to maintain this accuracy. Following this process, the classifier developed on the training set was used on the hold-out test set to determine accuracy. Classification accuracy was calculated as the proportion of cases in which the correct number of marks was allocated—i.e., (true positives + true negative) / (true positives + true negative + false positive + false negative).

Ethical approval

This study received institutional ethics committee approval from The University of Adelaide Human Research Ethics Committee (approval number H-2018-083).

Results

Sample characteristics

One hundred and eighty-one students completed the practice examination (response rate 59.0%). Of these students, 100 (55.2%) were in the fifth year of the program, with the remaining students in the fourth year.

Student opinion regarding computer-based marking

In response to Likert-type options, students expressed mixed opinions regarding computer-based marking of written examinations (see Figure 1). Just over half of students reported feeling that computers would not be able to perform an adequately accurate job of marking written examinations (*agree*, *strongly agree* or *very strongly agree*: 56.6%). However, despite expressing this belief, the majority of students agreed that computer marking would be more objective than human marking (*agree*, *strongly agree* or *very strongly agree*: 67.0%), and over half of students reported that they would not object to having written examinations marked by a computer (55.5%). Conversely, 44.5% of students reported that they would object to having written examinations marked by a computer.

Figure 1

Medical Student Responses to Likert-Type Questions Regarding Their Opinion of Human and Computer-Based Marking of Examinations

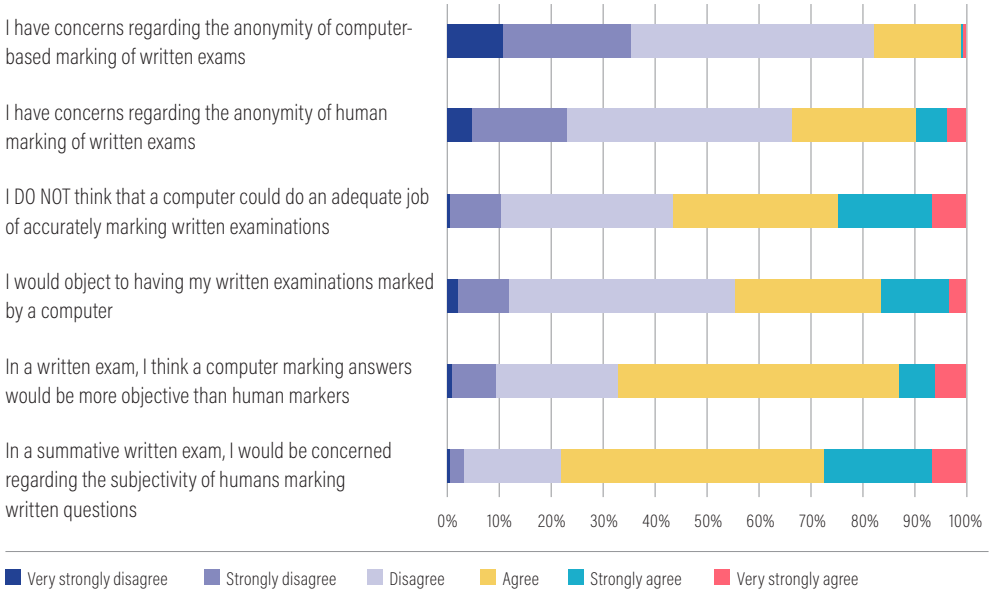

Table 2

Accuracy of Classification Into Grade Bands by DL Marking

Question Topic	Train/Test Split (%)	Maximum Possible Marks	Training Set Mean Accuracy (±SD)	Test Set Mean Accuracy	Test Set Maximum Proportion of a Single Mark Category
Zollinger-Ellison syndrome	20/80	1	0.86 (0.24)	0.98	0.74
Kidney stones	20/80	1	0.8 (0.21)	0.94	0.59
Conversion disorder	20/80	1	1.0 (0)	1.0	0.59
Compartment syndrome	20/80	1	1.0 (0)	0.98	0.66
Neurovascular injury	85/15	2	0.61 (0.13)	0.71	0.65
Acute angle-closure glaucoma	85/15	2	0.59 (0.14)	0.55	0.46
Thyroid cancer	85/15	2	0.46 (0.16)	0.56	0.44
Hypothyroidism	85/15	2	0.67 (0.28)	0.76	0.64
Acute pulmonary oedema	85/15	3	0.62 (0.18)	0.61	0.39
Anorexia nervosa	85/15	3	0.54 (0.11)	0.49	0.36
Essential tremor	85/15	3	0.38 (0.11)	0.5	0.41
Acute pancreatitis	85/15	3	0.77 (0.12)	0.76	0.48

DL analysis

For questions that required binary classification (1-mark questions, VSAQ), there were consistently high classification accuracies (> 90%) on the training set, with the original split (85%/15% train/test split). Subsequently, classification experiments were trialled with fewer samples in the training set. Ultimately, high accuracies on the training set (see Table 2) were able to be achieved with 30 training examples (20%/80% train/test split). The accuracies on the test set were similar to that of the training set. Across the four 1-mark questions, the mean accuracy on the test set was 0.98.

For questions in which multiclass classification was required (2-mark, 3 classes and 3-mark, 4 classes), accuracy was significantly lower. Even with the original 85%/15% train/test split, it was not possible to achieve accuracies over 0.80. The mean accuracy on the test set among the four 2-mark questions and four 3-mark questions was 0.65 and 0.59, respectively.

Discussion

Our results highlight that this DL algorithm was able to achieve high accuracy scoring 1-mark VSAQs using a small number of training questions. Regarding 2-mark and 3-mark questions, lower accuracies were achieved, in particular for 3-mark questions. However, it should be noted that the sample size in this study was small. Additionally, the results of this study highlight that while students would support a more objective marking process, there is some trepidation regarding the accuracy of ML marking systems.

This study has demonstrated that this method of machine learning requires small training datasets to be effective for 1-mark SAQs (VSAQs). This method of marking retained a high accuracy when using a small number of training examples. These results support the findings of previous studies that demonstrate that machine learning algorithms demonstrate a high reliability when marking VSAQs (Sam et al., 2018). Although medical student examinations comprise many formats, this strategy presents a viable system for marking 1-mark short answer questions in the assessment of medical students. Recent research has supported that VSAQs are an authentic means of assessment, with good discriminative value, and may be an alternative to standard multiple-choice questions (Sam et al., 2019). It is also important to note that employing a combination of both machine learning and human marking may be a viable strategy, one that is in use in some centres and in proprietary software.

Similarly, a combination human and machine learning approach may be utilised for 2-mark and 3-mark SAQs. In this study, the level of performance on these types of questions would be insufficient (i.e., the models would not be sufficiently accurate) to use ML alone for marking. The performance on the 2-mark and 3-mark SAQs is salient in that previous research has suggested that this type of question may complement standard multiple-choice questions (Bird et al., 2019), although this point may be debated (Hift, 2014). The results of this study also suggest that questions with greater numbers of marks are more difficult

to classify accurately with machine learning. This finding is intuitive in the setting of multiclass classification tasks in other areas (such as image analysis) and of relevance when considering automated scoring of longer answers and essays (Gierl et al., 2014).

The sentiment expressed by medical students in this study demonstrated a conflict between support for a potentially more objective and efficient marking process. Further education of students in regards to the implementation of supervised machine learning methodologies that can assess word combinations (e.g., phrases) rather than simply the frequency of individual words may influence these opinions. It should be noted that the survey in this study was administered prior to the examination (so that the content and type of questions in the examination would not influence student responses to the survey).

A limitation of this study is the small sample size, although we note that the accuracy assessment was intended as a pilot study. A larger sample size may have resulted in significantly higher accuracies for the marking of 2- and 3-mark questions, as this change would have increased the number of data from which the algorithms could learn. The marking schemes employed in this study were deliberately structured to encompass complex concepts such as "demonstrates sound clinical reasoning" (see Table 1). If simpler marking schemes were employed for the 2-mark and 3-mark questions, this approach would also likely have resulted in a higher accuracy. It is possible that the phrasing of survey questions may influence participant responses. In this case, more neutral survey items may have provided more positive responses. Another limitation is that the examination was formative rather than summative, and students may not have been as detailed in their explanations as a result. Finally, this study was conducted at a single site in only the English language. Accordingly, the findings may not be applicable to content from other centres and in other languages. In addition, this study applied machine learning marking strategies to clinical questions. It should be noted that performance on other types of questions, such as questions regarding professionalism and ethics, may result in different performance.

Future research into methods to improve the accuracy of DL marking of 2-mark and 3-mark questions, for example with larger datasets and simpler marking schemes, may be beneficial. Studies may also examine the application of DL to different styles of written medical questions, such as modified essay questions. Research may also seek to examine the influence of structured machine learning education for medical students on medical student acceptance of computer-based marking (for example, highlighting the differences between simple "keyword" and DL marking programs).

Conclusion

Medical students may be supportive of computer-based (machine learning) marking, in part due to its perceived objectivity. DL has the potential to provide accurate marking of written questions, although multiclass classification tasks may require larger datasets. Further research into DL marking of medical examinations is required.

Conflicts of interest and funding

The authors declare that there are no conflicts of interest. This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

References

Bird, J. B., Olvet, D. M., Willey, J. M., & Brenner, J. (2019, December). Patients don't come with multiple choice options: Essay-based assessment in UME. *Medical Education Online*, *24*(1), Article 1649959. https://doi.org/10.1080/10872981.2019.1649959

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, *25*(1), 60–117. https://doi.org/10.1007/s40593-014-0026-8

Deo, R. (2015). Machine learning in medicine. *Circulation*, *132*(20), 1920–1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593

Dias, R., Gupta, A., & Yule, S. (2018). Using machine learning to assess physician competence: A systematic review. *Academic Medicine*, *94*(3), 427–439. https://doi.org/10.1097/ACM.0000000000002414

Gierl, M. J., Latifi, S., Lai, H., Boulais, A. P., & De Champlain, A. (2014, October). Automated essay scoring and the future of educational assessment in medical education. *Medical Education*, *48*(10), 950–962. https://doi.org/10.1111/medu.12517

Hift, R. (2014). Should essays and other "open-ended"-type questions retain a place in written summative assessment in clinical medicine? *BMC Medical Education*, *14*, 249. https://doi.org/10.1186/s12909-014-0249-2

James, C. A., Wheelock, K. M., & Woolliscroft, J. O. (2021, July 1). Machine learning: The next paradigm shift in medical education. *Academic Medicine*, *96*(7), 954–957. https://doi.org/10.1097/ACM.0000000000003943

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent convolutional neural networks for text classification. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, *29*(1), 2267–2273. https://doi.org/10.1609/aaai.v29i1.9513

Latifi, S., Gierl, M., Boulais, A., & De Champlain, A. (2016). Using automated scoring to evaluate written responses in English and French on a high-stakes clinical competency examination. *Evaluation & the Health Professions*, *39*(1), 100–113. https://doi.org/10.1177/0163278715605358

Locke, S., Bashall, A., Al-Adely, S., Moore, J., Wilson, A., & Kitchen, G. B. (2021). Natural language processing in medicine: A review. *Trends in Anaesthesia and Critical Care*, *38*, 4–9. https://doi.org/10.1016/j.tacc.2021.02.007

Nadkarni, P., Ohno-Machado, L., & Chapman, W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544–551. https://doi.org/10.1136/amiajnl-2011-000464

Pinto dos Santos, D., Giese, D., Brodehl, S., Chon, S. H., Staab, W., Kleinert, R., Maintz, D., & Baessler, B. (2019, April). Medical students' attitude towards artificial intelligence: A multicentre survey. *European Radiology*, *29(*4), 1640–1646. https://doi.org/10.1007/s00330-018-5601-1

Puthiaparampil, T., & Rahman, M. M. (2020, May 6). Very short answer questions: A viable alternative to multiple choice questions. *BMC Medical Education*, *20*(1), Article 141. https://doi.org/10.1186/s12909-020-02057-w

Salt, J., Harik, P., & Barone, M. A. (2018, December 11). Leveraging natural language processing: Toward computer-assisted scoring of patient notes in the USMLE Step 2 clinical skills exam. *Academic Medicine*, *94*(3), 314–316. https://doi.org/10.1097/ACM.0000000000002558

Sam, A., Field, S., Collares, C., van der Vleuten, C. P., Wass, V., Melville, C., Harris, J., & Meeran, K. (2018). Very-short-answer questions: Reliability, discrimination and acceptability. *Medical Education*, *52*(4), 1–9. https://doi.org/10.1111/medu.13504

Sam, A. H., Westacott, R., Gurnell, M., Wilson, R., Meeran, K., & Brown, C. (2019, September 26). Comparing single-best-answer and very-short-answer questions for the assessment of applied medical knowledge in 20 UK medical schools: Cross-sectional study. *BMJ Open*, *9*(9), Article e032550. https://doi.org/10.1136/bmjopen-2019-032550

Sidey-Gibbons, J. A. M., & Sidey-Gibbons, C. J. (2019, March 19). Machine learning in medicine: A practical introduction. *BMC Medical Research Methodology*, *19*(1), Article 64. https://doi.org/10.1186/s12874-019-0681-4

Appendix

Pre-exam Survey

Please select the box that indicates how strongly you agree/disagree with each of the following statements:

	Very strongly disagree	Strongly disagree	Disagree	Agree	Strongly agree	Very strongly agree
In a summative written exam, I would be concerned regarding the subjectivity of humans marking written questions	○	○	○	○	○	○
In a written exam, I think a computer marking answers would be more objective than human markers	○	○	○	○	○	○
I would object to having my written examinations marked by a computer	○	○	○	○	○	○
I DO NOT think that a computer could do an adequate job of accurately marking written examinations	○	○	○	○	○	○
I have concerns regarding the anonymity of human marking of written exams	○	○	○	○	○	○
I have concerns regarding the anonymity of computer-based marking of written exams	○	○	○	○	○	○

If you would object to having your written exams marked by a computer, please explain why: