Using Machine Learning to Understand Assessment Practices of Capstone Projects in Engineering

Herman Kamper
Dept. E&E Engineering
Stellenbosch University
South Africa
kamperh@sun.ac.za

Carla Niehaus
Dept. Mathematical Sciences
Stellenbosch University
South Africa
carla.niehaus@gmail.com

Karin Wolff
Fact. Engineering
Stellenbosch University
South Africa
wolffk@sun.ac.za

***Abstract*—Capstone projects are used as a final assessment in undergraduate engineering degrees to ensure alignment with international standards and profession-specific competencies. Accurate and consistent assessment of these projects is therefore crucial. Variability between assessors can also have direct consequences for a student's future career path. In light of previous work that has already indicated that variability is often high, we ask whether recent advancements in artificial intelligence and machine learning (ML) could help us gain a better understanding of what is going on in current assessment practices of capstone projects within engineering. To do this, we collect a dataset of previous capstone project reports with the marks awarded from a single department. We then have a new set of examiners re-examine the reports. We also train ML models to predict a mark given a report as input. Quantitatively, our analysis reveals a large discrepancy of roughly 12% between examiners. Qualitatively, the ML models show that marks are most affected by report length and the research area of the project. This supports previous work showing high inter-assessor variability, and showcases how ML can be used to start to uncover reasons for the differences between examiners. We also shared these findings within the department, leading to initial discussions to improve assessment practices and consistency.**

***Index Terms*—assessment, capstone projects, machine learning**

I. Introduction

Professional qualifications, globally, are regulated by standards-generating committees aligned to the relevant professional councils. These councils establish criteria for the demonstration of profession-specific competencies as agreed by a broad range of expert stakeholders in the associated community of practice. The competency criteria in engineering, for example, are framed by the International Engineering Alliance (IEA), "a global not-for-profit organisation, which comprises members from 41 jurisdictions within 29 countries, across seven international agreements. These international agreements govern the recognition of engineering educational qualifications and professional competence."[1] In the face of increased mobility and globalisation, the aim of such centralised and standardised criteria is to enable mutual recognition of qualifications across national borders.

While the professional competency criteria may broadly frame the design and delivery of a particular qualification curriculum, the implementation and interpretation at a pedagogical level remain the preserve of engineering educators appointed as academics in tertiary institutions. The Bachelor's in Engineering (BEng) qualification at South African universities is governed by a particular standard drawn up by our national engineering council (the Engineering Council of South Africa), a signatory to the Washington Accord. The BEng standard specifies specific minimum knowledge area credits over a four-year programme, and has recently included a revised range of 11 graduate attributes (GAs) aligned to the IEA competency profiles. The GAs specify competency levels, such as the ability to draw on natural, mathematical and engineering science knowledge, along with appropriate tools and techniques in addressing context-specific complex problems. While some of the GAs are summatively assessed in different knowledge-area subjects around the final year, it is common practice to use a final-year *capstone project* to assess the achievement of multiple GAs. Although the capstone assessment process differs across institutions, it usually entails at least two different examiners, one of whom is potentially external to the institution in question.

Despite the existence of ostensibly standardised criteria against which to assess a BEng capstone project report, the last two decades have seen a significant amount of literature dedicated to addressing the question of assessor variability [1]. In her early work, Shay [2], [3] examined the interpretations of engineering academics at South Africa's top research-intensive institution. She argued that "multiple subjectivities ... shape assessors' interpretations of student performance" and that "the contextually complex, communal character of professional judgement" as exercised by assessors entails a "double truth": professional decision-making is always "relational, situational, pragmatic and value-based" [3, p. 677].

Assessor variability in the case of the final hurdle in qualification achievement can mean make or break for a student. The process to achieve consensus in a summative grade, however, can also offer members of a community of practice the opportunity to engage dialogically [4] with questions of values, interpretation and validity. In so doing, a particular community may be more likely to align their assessment practices. However, it is simply not practical to expand the dialogue across the globalised world in which, say, thousands of engineering qualifications are on offer, all of

[1] https://www.ieagreements.org/

which are intended to be aligned with the IEA competency profiles. It is for this very reason that the broader professional community has entrusted the standards alignment role to the various professional bodies.

Given that we still have evidence of variable assessor interpretation today, in 2023, and we have made great strides in developing artificial intelligence (AI) systems, the question behind the study in this paper is: Can we use a machine learning (ML) approach to see what is going on in current capstone assessment practices in an engineering faculty?

The paper presents a collaborative study between researching academics at a research-intensive institution in South Africa. The faculty in question is engaged in ongoing programme renewal processes, with the goal of improving the student learning experience and implementing more efficient systems in a large-class, Global South context. Methodologically, the study focuses on a single department in an engineering faculty and entailed collecting a dataset of 516 capstone project reports completed between 2017 and 2021, together with the marks that were assigned. To quantify assessor variability, a new set of examiners were asked to re-examine the reports. The findings reveal a large discrepancy (roughly 12%) between the new and original examiner assessments. To contextualise this, we train an ML assessment system that takes a project report and predicts a mark based on a set of extracted features. The ML model shows that marks are most affected by report length and the research area of the project. By communicating these results to the lecturing staff, we hope to sensitise engineering academics to the reality of subjectivity in the assessment process and stimulate discussions to improve consistency.

II. Conceptual Framework

In her sociologically empirical work on higher education assessment practices, Shay [2], [3], [5] argues that the process privileges an objective interpretation of reality while downplaying its subjective counterpart. She demonstrates that there is a "double truth" to assessment in the ostensible alignment to explicit standardised criteria. Firstly, assessment criteria in the professions are socioculturally created by specific communities of practice who attempt to reach a consensus about "what matters". The interpretation of these criteria, however, plays out in different contexts and involves "tacit understandings of expertise". Shay highlights the process versus product stages in the creation of an assessment artefact, such as a capstone project report or performance. Empirical studies in engineering capstone assessment demonstrate significantly divergent assessor variations [2], [4], [6]–[8]. While "assessor variability is often seen as unwanted bias or error" [1], and there have been multiple studies on the attempts to develop more rigorous assessment rubrics, the reality is that contextual, sociocultural interpretations of assessment criteria will always exist.

Sociologists such as Bourdieu and Bernstein are concerned with the question of social structures through which differentiated power relations play out [9], [10]. Shay [3] elaborates on the subjectivity of assessors, describing "professional judgement as inescapably (in part) an embodiment of the assessor". The assessment process is relational in that it is "a communicative exchange between the assessor and the assessed" in which the assessor holds the power. Furthermore, she qualifies the subjectivity of assessment processes as being context-dependent and pragmatic. While Shay's work focused on the implications for students, and the fact that the assessment process is not as objective as a student might think, the increasing call to international alignment of professional standards to enable mobility suggests that the capstone assessment process warrants further research. De Jonge et al. [11], in their examination of assessment practices in the medical profession, identify key features upon which aligned assessment can occur. Although their study focuses on alignment between the student and assessing supervisor, their concept of mutuality can be extended to any multi-stakeholder assessment process: when assessment is embedded in the learning process, and "constructive collaboration" with feedback enables "competency development towards professional standards" [11].

The emergence of AI in education [12] offers a number of possibilities. Several studies on educational techniques and the development of prototypes in statistical reasoning, data visualisation, and learning analytics have already emerged, extending the educational arenas that AI is impacting [13], [14]. Given the ubiquitous use of plagiarism technologies such as TurnItIn, and the more recent ChatGPT capabilities [15], ML techniques have advanced sufficiently to allow for their use in addressing the challenge of assessment interpretation. The present study intersects with several of the areas in higher education that Alam [12] identifies as being likely to be impacted by ML, including automatic grading, teaching evaluation, and supporting learning and teaching design.

We set out to examine the nature of capstone assessment practices as evident in a single engineering department at a research-intensive institution in the Global South. Our longer-term goal is to determine whether an ML approach can enable assessors to interrogate assumptions about the ostensible objectivity of standards, and in so doing achieve improved consensus and alignment in their future practices.

III. Context

The Faculty of Engineering at a research-intensive South African institution has seen significant growth in student numbers over the past decade, and increasingly diverse cohorts. Nationally, attrition rates in STEM programmes are high (around 50% [6]). In order to improve retention, there is significant funding dedicated to both student support and staff capacity building. The faculty is actively engaged in programme renewal projects designed to address our challenges. One such University Capacity Development Grant (UCDG) project, the Recommended Engineering Education Project (REEP), enables academics and academic development staff to design, implement, evaluate, and research curricular and pedagogical practices. This paper reports on a REEP initiative to examine current capstone assessment practices using ML techniques.

This paper looks specifically at the capstone project at the end of the BEng degree in Electrical & Electronic Engineering.

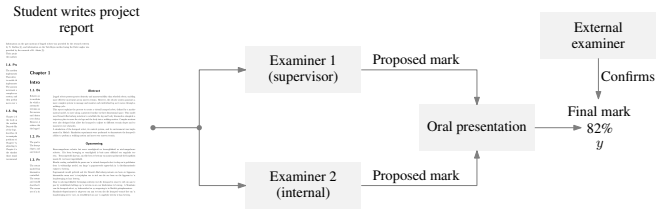

Fig. 1. Marks of a capstone project are awarded based on a process involving two examiners, a convenor, and an external examiner.

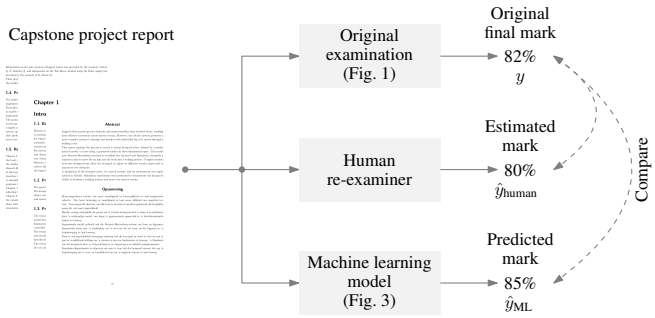

Fig. 2. The approach used for comparing the originally assigned marks to those obtained from re-examination or from the machine learning model.

In this degree, there are roughly 100 final-year students each year. Each student completes a unique open-ended project under the supervision of one of the lecturers in the department. The project is meant to assess eight of the eleven GAs (see Sec I): problem-solving; application of scientific and engineering knowledge; engineering design; investigations, experiments and data analysis; engineering methods, skills and tools, including information technology; professional and technical communication; individual work; and independent learning ability. Students have four months to complete their project. After completing the design, implementation and experimental work, a student writes a detailed report. The project report is the main deliverable on which a final mark is based. The body of the report is limited to 40 pages, but penalties aren't strictly enforced, so students regularly overrun.

After reports are handed in, the examination process proceeds as in Fig. 1. Every supervisor also acts as an examiner for the projects that they supervised (Examiner 1). An additional examiner from the department is appointed (Examiner 2). Reports are first evaluated separately by the two examiners. Each examiner awards a preliminary mark or gives a range (e.g. 60% to 65%). After the examiners complete their initial assessments based solely on the reports, an oral is scheduled, where the student presents their work and answers questions from the examiners under the guidance of a third lecturer (the convenor). After the presentation, the convenor facilitates a discussion between the two examiners to arrive at a combined mark. This mark is not necessarily the average of the two individual marks—this will depend on the discussion. The last step in the entire process is for an external moderator from outside of the university to confirm the mark that was assigned at the oral. It is rare for moderators to adjust marks—there will typically only be isolated cases in a particular year group.

Our main question is to see what is going on in the assessment practices of these capstone projects. Building on the growing body of work on using AI and ML for automatic assessment [16]–[18], our goals are to quantify how consistent lecturers are and to see what features in a report correlate with marks. E.g. do lecturers award lower marks for reports with many spelling mistakes? Or is the number of equations in a report a better indicator of a mark? Our bigger goal is to improve assessment quality and alignment between lecturers.

IV. Research Design

This paper presents a quantitative research study while being cognisant of the qualitative nature of the problem and the analysis methods that we employ. The structure of the capstone projects has already been described above, with Fig. 1 giving an overview of the examination process. This section details the rest of the research methodology.

A. Data

We compile a dataset of all the capstone projects submitted in Electrical & Electronic Engineering between 2017 and 2021. Reports are anonymised: each student is assigned a unique random ID and the front page and acknowledgements are manually removed from each report PDF. We also capture the identity of the supervising and examining lecturers (Examiners 1 and 2), but also anonymise this information using random IDs. The complete dataset consists of 516 report PDFs with the final marks that were awarded (following the process in Fig. 1). Reports range from 20 to 114 pages.

B. Measuring Assessor Consistency

To measure inter-assessor assessment consistency, we select 30 reports from 2021 for remarking. Six lecturers in the department were asked to remark the reports, each receiving five reports from their respective research areas. In almost all the cases, the re-examiners were different from the original examiners (but see comments about this in Sec. V-A). The task set to the re-examiners was to estimate the mark that was originally assigned to a report.

As illustrated in Fig. 2, we compare the difference between the originally assigned mark and the mark from re-examination. Formally, denote the original awarded mark for the nth report as $y^{(n)}$. Denote the mark from the re-examination as $\hat{y}^{(n)}_{\text{human}}$. To quantify the difference between the two marks, we calculate the root mean square error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(y^{(n)} - \hat{y}^{(n)}_{\text{human}}\right)^2} \tag{1}$$

A lower RMSE is better. Intuitively, if a model achieves an RMSE of 9%, this indicates that for most of the reports, the original and re-examination marks are within 9% of each other.

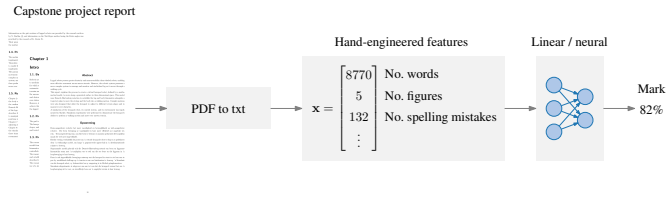

Fig. 3. Text is extracted from a project PDF, features are then extracted using a customised pipeline, the features are used as input to a linear or neural network regression model, and a predicted mark is obtained.

To situate the results, we repeat the above analysis but use six non-expert re-examiners. These non-experts are all postgraduate students from science and engineering who have not assessed capstone projects before. Again, we calculate an RMSE for these non-expert examiners, comparing their marks to those originally awarded by lecturers.

C. Machine Learning for Automatic Project Report Assessment

We consider several ML approaches that take in a report and predict a mark. In all cases the model uses reports from between 2017 and 2020 as its training data. The model is then presented with the same 30 reports from 2021 used for the human re-examination test. Denoting the ML model's prediction as $\hat{y}_{\text{ML}}^{(n)}$, we calculate the RMSE between the originally assigned mark and the prediction from the model following (1) as illustrated at the bottom of Fig. 2.

For our ML models, we follow the methodology illustrated in Fig. 3. Each project report is originally in PDF format. These are converted to text using PDF2GO.[2] Python's regular expression library[3] is then used to extract several features that we think would be indicative of the assigned mark, as listed in Table I. Apart from supervisor and examiner IDs, each lecturer is also assigned a category ID related to their main area of research and expertise: *control systems*, *energy*, *electromagnetics*, or *informatics*. The top-ten words are the word types used most often across all reports, ignoring stop words ("the", "if", etc.).

Not all features might be meaningful. We therefore use a feature selection approach called lasso regression [19, Sec. 6.2.2] to choose a subset of the most useful features. We feed these features to a neural network to give the predicted mark $\hat{y}_{\text{ML}}^{(n)}$. We call this approach *Neural: Lasso features* in the experiments below. A second ML approach that we found to work well is to use a single feature, word count, as input to a simple linear regression model; we refer to this as *Linear: Word count* below. As our most simple ML model, we use a *null model* [19, Sec. 3.2] that simply predicts the same mark for any project, irrespective of what is contained in the input; the mark is set to the average of the marks of the reports in the training data (2017 to 2020). Clearly, this is a very naive approach, but it is useful: if a more advanced ML approach doesn't beat this method, we know that it is essentially not meaningful.

[2]https://www.pdf2go.com/

[3]https://docs.python.org/3/library/re.html

TABLE I
FEATURES EXTRACTED FROM PROJECT REPORTS

Number of pages	Average figure caption length
Word count	File size
Number of spelling mistakes	Supervisor ID
Number of spelling mistake types	Examiner ID
Number of figures	Supervisor domain ID
Number of tables	Examiner domain ID
Number of equations	Topic ID
Number of references	Top-ten words (true count)
Number of new line starts	Top-ten words (fraction of total words)

D. Methodology Recap

Taking the above together by referring to Fig. 2, we have a set of 30 reports that were originally assessed in 2021. We have a group of re-examiners remark these reports and assign new marks. We compare this to the marks originally awarded. In parallel, we repeat this with an automatic ML system.

V. ANALYSIS

A. Quantitative Results

Table II gives the quantitative results where ML models and human re-examiners were asked to predict the mark that was assigned to a capstone project. A lower RMSE is better.

The first observation is that the null model performs well. This model simply awards the average mark obtained between 2017 and 2020 to all the reports in this 2021 evaluation set, irrespective of the content of the report (this average is 67.06%). Although a low RMSE of 14.73% is therefore surprising, we found that this is due to the marks falling in a narrow band around the average; Fig. 4 shows a histogram of the marks in the training data: it is clear that the majority of reports are assigned marks between 60% and 80%. An average prediction for all reports therefore captures most of the assigned marks.

Looking at the performance of the human examiners in Table II, we see that the expert re-examiners are typically within 12.21% of the original examiners. Despite being better than the null model, the question of whether this is an acceptable margin of error is open for discussion (see Sec. VI). But it is clear that the re-examiners do not perfectly predict the same marks as

TABLE II
A COMPARISON OF HUMAN AND MACHINE LEARNING PERFORMANCE ON PREDICTING THE MARK AWARDED TO A CAPSTONE PROJECT

Model	RMSE
Null model	14.73
Linear: Word count	13.73
Neural: Lasso features	**11.22**
Human experts	12.21
Human non-experts	15.03

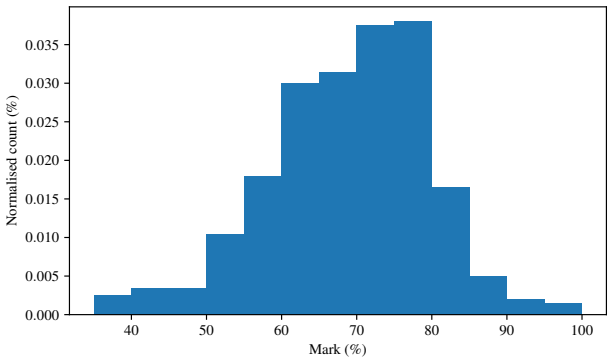

Fig. 4. A histogram of the marks assigned to reports in the training data (2017 to 2020).

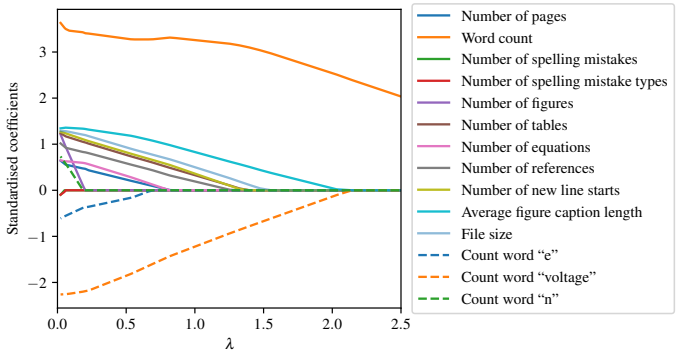

Fig. 5. Standardised model coefficients as the weight parameter in lasso regression is increased. Intuitively, the earlier that a feature coefficient snaps to zero, the less indicative it is of a report mark.

those originally awarded (which would give an RMSE of 0%). There is, fortunately, a bit of good news for the lecturers: the expert human examiners fare much better than the non-experts (the non-experts get a score worse than the null model).

We also looked at the performance of individual re-examiners. For the experts, there were big differences in individual performance, with a best RMSE of 1.84% and a worst RSME of 18.93%. For the non-experts, RMSEs were between 6.98% and 36.62%. As mentioned, in almost all cases, the expert re-examiners did not match examiners involved in the original examination. However, there were two cases where the examiners and re-examiners did actually overlap. In both cases the re-examiners indicated that they could not recall being an original examiner (despite this being the case), and, in both cases, the re-examiners awarded exactly the same mark as they did originally.

Comparing the human results to linear and neural ML approaches in Table II, we see that the neural network taking in features selected with lasso regression comes closest to the marks originally awarded (RMSE of 11.22%). Given that this is slightly better than the 12.21% from the human experts, this indicates that the ML methodology would be hard to improve, given the data which is based on the assessment practices currently followed in the department.

B. Qualitative Results: The Most Informative Features

The close performance between the best ML approach and the human experts motivates us to use the ML method to analyse the type of aspects that examiners look at when examining reports. To probe the ML model, we use lasso regression and look at its selection as we force it to select an increasingly smaller number of features. This is illustrated in Fig. 5. Intuitively, the idea is that features that snap to zero first (left on plot) are less important than those that snap to zero later (right on plot).

We see that the spelling mistake weights snap to zero first, indicating that these are least important, while features like the number of words and average figure caption length snap to zero last and are therefore deemed most important. Many lecturers indicated beforehand that spelling mistakes would be a very good feature, so it is surprising that the number of spelling mistakes seems to be far less useful compared to features measuring the length of a report.

The research area of the report is not included explicitly in the figure, but it is captured indirectly: The count of the word “voltage” snaps to zero very late in Fig. 5 (i.e. it is very informative). The word “voltage” is often associated with projects from the energy area, and the higher the count of this word, the lower the predicted mark, as indicated by the negative coefficient. Similarly the word “e” is often used as a symbol in energy research, and again the negative coefficient indicates that this negatively affects a mark. In contrast, “n” is often used to denote the number of items in a set in the machine learning research area; the more this word is used, the higher the mark.

Taking this analysis together qualitatively, lecturers give higher marks to longer reports; and the area in which a project falls also impacts the predicted mark, with areas like machine learning achieving higher marks than areas like energy.

It is important here to emphasise the difference between correlation and causation. It might be tempting to conclude that lecturers are unfairly biased towards shorter reports. But it could also be that better projects require longer reports, and then it is good that report length and marks are positively correlated. To state this differently, a long report doesn’t cause a better project (but it could be indicative of one).

VI. Discussion

Before setting out on this study, the authors informally asked lecturers in the department what they would deem as an acceptable margin of error. Most lecturers indicated an acceptable margin of between 5% and 10%. The best ML approach in Table II comes close to this, with an RMSE of just above 11%. During the study, we also asked the expert re-examiners how close they believed they were to the assigned mark, and most lecturers indicated that they would be within 5%. The scores from Table II indicate that the assessors are much worse than they anticipated.

Despite the questions of correlation and causation, the analysis above still raises a crucial question: Why do the examiners differ so much in their assessments? This should

be investigated in detail in follow-up work, but we briefly outline some potential reasons. It is clear that the research area influences interpretation, e.g. a newer research area like machine learning seems to result in higher marks than a more established area such as energy. So do perspectives on what counts change over time? Similarly, it could be argued that assessor perspectives shift and develop over time with increased assessment experience—the big range of individual RSMEs across individual examiners points in this direction. There are also differences in the original examination process (Fig. 1) compared to the re-examination process (Fig. 2). For instance, in re-examination, a mark is obtained from a single examiner without a discussion with another examiner. Moreover, in the original process, there is an oral examination, which can affect perceptions. In the original process, examiners are also aware of the student's identity, in contrast to the re-examination process where identifying information is removed. This can further influence perceptions, e.g. an examiner might be aware of a particular student's past performance.

The variability in assessments has a direct implication for a student. The difference between failing and passing has big financial implications. Similarly, the difference between achieving a distinction or not directly influences access to subsequent post-graduate studies. Effectively speaking, assessors hold the power of determining a student's future career. They are, indeed, entrusted to hold this power: they are appointed as experts in their fields and are ostensibly following the prescribed standards as internationally accepted.

After performing the qualitative analysis above, we also had a feedback session where we shared the findings with the department. Many lecturers indicated informally that they were unsurprised by the findings indicating poor consistency. Many lecturers also started discussing strategies for how consistency could be improved during and after the session, especially given the potential consequences for a student.

Inconsistency between assessors is a known phenomenon—as shown by the studies summarised in Sec. II. Our hope is that this study would start a conversation about how ML technology can be used do identify and address these inconsistencies.

VII. Conclusion

This paper set out to quantitatively and qualitatively use machine learning (ML) to gain a better understanding of current assessment practices of capstone projects within an engineering department at a top research institution in the Global South. Our methodology involved collecting a dataset of previous project reports, and then comparing the originally assigned marks to those assigned by a set of re-examiners. We also looked at the performance of an ML model on predicting the original assigned mark. The analysis showed that the inter-assessor variation is large (roughly 12% in RMSE), which is also similar to our best ML approach. The ML model also showed that marks correlate most with features measuring report length and the research area of the report. We engaged with the stakeholders in order to start discussions about how assessment consistency can be improved. Future work will look particularly into whether an ML approach can be used to improve consistency across lecturers. E.g. an ML system could make a first-pass prediction. Or it could be used to highlight areas of bias for a particular lecturer. These are exciting avenues for future endeavours.

Acknowledgements

This project was supported by a South African University Capacity Development Grant (UCDG).

References

[1] C. Domínguez, A. Jaime, F. J. García-Izquierdo, and J. J. Olarte, "Factors considered in the assessment of computer science engineering capstone projects and their influence on discrepancies between assessors," *ACM Transactions on Computing Education*, 2020.

[2] S. Shay, "The assessment of complex performance: A socially situated interpretive act," *Harvard Educational Review*, 2004.

[3] ——, "The assessment of complex tasks: A double reading," *Studies in Higher Education*, 2005.

[4] K. Wolff and F. Hoffman, "'Knowledge and knowers' in engineering assessment," *Critical Studies in Teaching and Learning*, 2014.

[5] S. Shay, "Beyond social constructivist perspectives on assessment: The centring of knowledge," *Teaching in Higher Education*, 2008.

[6] K. Chan, "Statistical analysis of final year project marks in the computer engineering undergraduate program," *IEEE Transactions on Education*, 2001.

[7] C. Pathirage, R. Haigh, D. Amaratunga, and D. Baldry, "Enhancing the quality and consistency of undergraduate dissertation assessment: A case study," *Quality Assurance in Education*, 2007.

[8] A. Nyamapfene, "Involving supervisors in assessing undergraduate student projects: Is double marking robust?" *Engineering Education*, 2012.

[9] P. Bourdieu, *Outline of a Theory of Practice*. Cambridge University Press, 1977.

[10] B. Bernstein, *Pedagogy, Symbolic Control, and Identity: Theory, Research, Critique*. Rowman & Littlefield, 2000.

[11] L. P. de Jonge, A. A. Timmerman, M. J. Govaerts, J. W. Muris, A. M. Muijtjens, A. W. Kramer, and C. P. van der Vleuten, "Stakeholder perspectives on workplace-based performance assessment: Towards a better understanding of assessor behaviour," *Advances in Health Sciences Education*, 2017.

[12] A. Alam, "Possibilities and apprehensions in the landscape of artificial intelligence in education," in *ICCICA*, 2021.

[13] B. Whitby, *Artificial Intelligence: A Beginner's Guide*. The Rosen Publishing Group, 2009.

[14] V. Devedžić, "Web intelligence and artificial intelligence in education," *Journal of Educational Technology & Society*, 2004.

[15] OpenAI, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[16] X. Zhai, Y. Yin, J. W. Pellegrino, K. C. Haudek, and L. Shi, "Applying machine learning in science assessment: A systematic review," *Studies in Science Education*, 2020.

[17] P. Vittorini, S. Menini, and S. Tonelli, "An AI-based system for formative and summative assessment in data science courses," *International Journal of Artificial Intelligence in Education*, 2021.

[18] A. Botelho, S. Baral, J. A. Erickson, P. Benachamardi, and N. T. Heffernan, "Leveraging natural language processing to support automated assessment and feedback for student open responses in mathematics," *Journal of Computer Assisted Learning*, 2023.

[19] J. Gareth, W. Daniela, H. Trevor, and T. Robert, *An Introduction to Statistical Learning: With Applications in R*, 2nd ed. Springer, 2021.