The Accuracy of AI-Based Automatic Proctoring in Online Exams

Adiy Tweissi, Wael Al Etaiwi and Dalia Al Eisawi
Princess Sumaya University for Technology (PSUT), Jordan
a.tweissi@psut.edu.jo
w.etaiwi@psut.edu.jo
d.aleisawi@psut.edu.jo

Abstract: This study technically analyses one of the online exam supervision technologies, namely the Artificial Intelligence-based Auto Proctoring (AiAP). This technology has been heavily presented to the academic sectors around the globe. Proctoring technologies are developed to provide oversight and analysis of students' behavior in online exams using AI, and sometimes with the supervision of human proctors to maintain academic integrity. Manual Testing methodology was used to do a software testing on AiAP for verification of any possible incorrect red flags or detections. The study took place in a Middle Eastern university by conducting online exams for 14 different courses, with a total of 244 students. The results were then verified by 5 human proctors in terms of monitoring measurements: screen violation, sound of speech, different faces, multiple faces, and eye movement detection. The proctoring decision was computed by averaging all monitoring measurements and then compared between the human proctors' and the AiAP decisions, to ultimately set the AiAP against a benchmark (human proctoring) and hence to be viable for use. The decision represented the number of violations to the exam conditions, and the result showed a significant difference between Human Decision (average 25.95%) and AiAP Decision (average 35.61%), and the total number of incorrect decisions made by AiAP was 74 out of 244 exam attempts, concluding that AiAP needed some improvements and updates to meet the human level. The researchers provided some technical limitations, privacy concerns, and recommendations to carefully review before deploying and governing such proctoring technologies at institutional level. This paper contributes to the field of educational technology by providing an evidence-based accuracy test on an automatic proctoring software, and the results demand institutional provision to better establish an appropriate online exam experience for higher educational institutions.

Keywords: AI-based proctoring, automatic proctoring, online exams, software accuracy, academic integrity

1. Introduction

Remote examination and proctoring have gained significant importance in recent years. The increasing importance is due to the need to accommodate comfort in education, increase security, and accessibility. Besides, the current COVID-19 pandemic has left learning institutions without an option but to adopt virtual learning methods to promote safety and adhere to social-distancing protocols. The increased adoption of remote examination has been linked with other factors besides the importance. For instance, online exams are assumed to help Massive Open Online Courses (MOOCs) and credit-based certifications to achieve credibility. Therefore, instead of students taking examinations in traditional classrooms, emphasis is now on promoting comfort-based learning and effective verification using digital proctoring.

The United Nations Educational, Scientific and Cultural Organization (UNESCO) identified the Coronavirus disease as one of the reasons for the high adoption of remote learning. According to UNESCO, the pandemic disrupted educational activities (UNESCO, 2021). As part of response strategies, governments globally had to shut down learning institutions, and most schools shifted their learning activities online. According to UNICEF (2020) report, an analysis showed high percentages of inability to reach students around the world unless proper educational policies and infrastructure were implemented. Although online learning has shown to work effectively, concerns have emerged about the models that learning institutions have been using for proctoring and the level of analysis used to ensure that students do not engage in malpractices. The concerns are valid because the virtual environment is complex. Therefore, the models used must have a high analogy of suspicious movement detection and the ability to flag and eliminate any aspects that affect accuracy or, at higher level, academic integrity. This type of software invites various detection mechanisms, which include face and noise detection, eyeball movement detection, and device detection. The software can also detect a change of tabs. With this, such a software can add credibility and integrity in examinations to minimize cases of academic dishonesty.

To evaluate, anticipate, and offer ways to defy the infection of COVID-19, new technologies such as Artificial Intelligence (AI), Big Data, and Machine Learning are needed. Artificial intelligence is a branch of computer science that can process large amounts of data and perform functions similar to those of the human brain. AI

technology is rapidly evolving, and its applications span almost every part of our lives, including medical, language processing, business forecasts, the environment, and even our smartphones. AI can also have a positive impact on eLearning processes and serve as a powerful enabler for improvement.

Digital proctoring has been compared with proctoring in traditional examination settings. Arguments have emerged providing various reasons why online courses could be susceptible to academic dishonesty. Kraglund-Gauthier and Young (2012) argued that remote assessments happen in an unsupervised environment. Therefore, it is difficult to determine the correct identity of the individual taking a test. Additionally, online test-takers can have access to materials that are not required in an assessment. It is also speculated that a virtual environment, which is often marked by the absence of an instructor, could encourage collaborative responses from groups of learners (Moore, Head, and Griffin, 2017). Although growing empirical evidence indicates that online examinations are prone to academic dishonesty than other methods, the debate is not yet settled. Therefore, there is a need for experiments to determine the accuracy of online proctoring to eliminate academic dishonesty.

1.1 Research gap

The value of this study paper is derived from filling the gap of testing and assuring the accuracy of auto-proctoring software. The intended contribution of this paper was to examine an AI-based automatic proctoring software and compare it to the human level of proctoring in online examinations. This comparison can be interpreted to better shape the understanding, hence the policy making of online exams' regulation and setup. In addition, this study highlights some of the limitations that may exist in AI-based proctoring software and adds more to the scope of coding improvement for developers (individuals or companies). Moreover, based on the research done by Nigam, et al. (2021) that reviewed 43 papers that were published between the years 2015 and 2021, the focus was mainly on the architecture of AI-based proctoring software in a technical aspect. This study brings human proctoring back to the formula by making the comparison between human decision on online exams and the AiAP decision. The study took place in a Middle Eastern university, which can extend the scope of research to have an international perspective for such technological advancement.

1.2 Research questions

The reviewed literature for this study focused on the improvement of proctoring technology in online exams, and the scope was determined to test the accuracy of one of the proctoring technology types, which is AI-based automatic proctoring (based on a proposed AiAP model). In order to explore the accuracy in terms of the number of cheating cases detected (i.e. number of violations), both Human and AiAP decisions need to be measured, compared, and then the difference between them should be computed to check the variation between a tested model (AiAP) and a verifying benchmark (Human Proctors). Therefore, the study sought to find answers for the following research questions:
RQ 1: Is there a significant difference between the Human Proctor Decision and the AiAP Decision?
RQ 2: Is the AiAP efficient in analyzing and verifying the cases of cheating in online exams?

2. Background review

Despite their different approach to learning, electronic forms of higher education have one objective, which is to help learners achieve the fundamental goals of a traditional university or college setting, such as research. Virtual education relies on the benefits of technology and the skills from campus-based teaching to create efficient designs and delivery. Moreover, it is suggested that an online educational institution must deliver all the aspects of a conventional learning institution, including teaching, staffing, aid, assessments, and evaluations.

2.1 AI-based solutions to maintain academic integrity

Several studies have investigated the accuracy of AI-based online proctoring using experimental approaches. For instance, Alessio, et al. (2017) compared online exam results from proctored and non-proctored online tests. One hundred and forty-seven students enrolled in different sections of a virtual course were assessed using linear mixed effects models. Half of these students were not proctored, while the rest used an online proctoring software. Students who used an online proctoring software required less time for online tests and scored less points. Similarly, another study compared physical processes and online exam evaluation tools.

Another study attempted to revive the unsettled debate concerning academic integrity in online courses. Dendir and Maxwell (2020) used a quasi-experiment where a webcam software was used in online proctoring to

evaluate high stake exams. The courses assessed in the study did not change their structural components even after online proctoring was introduced. Findings from this experiment show that exam scores reduced significantly after online proctoring was introduced in the courses. These results were interpreted as evidence which indicates that cheating happened in online courses before proctoring software solutions were introduced. This implies that online proctoring software is an effective tool for promoting academic integrity in online learning.

Online evaluation is an important part of universities. Evaluation is vital as it helps fulfill the purpose of education, which is to transfer knowledge and skills to students, and award correct credentials. According to Ismail, Safieddine, and Jaradat (2019), the credentials are based on reliable examinations and assessments of specific learning outcomes. Educators make significant efforts to identify and use testing ways to help fulfill the objective of education, while minimizing any form of inconvenient interruption in the process. Considering that most learning processes are conducted virtually, online evaluation relies on technology that can improve the proctoring and review process. AI-based proctoring has been suggested as one of the approaches that can be used to prevent malpractices that reduce the effectiveness of virtual assessments.

2.2 Types of proctoring in online exams: Automated and live

There have been several research papers that categorized online proctored exams in different ways. For example, Nigam, et al. (2021) categorized the proctoring software in general into Live Proctoring, Recorded Proctoring, and Automated Proctoring. Nie, et al. (2020) recommended that live proctoring is more accurate for online proctoring because it does not limit the invigilation process to remote monitoring. In live proctoring, the proctor can view all candidates and their environment through a screen. Live proctoring involves instructors or invigilators assigned the role of monitoring exams and viewing examination candidates in real-time. The proctors review the students assigned to them and confirm their presence. Although the proctor might not always know the identity of all students assigned to them, the device's camera often captures the image of learners holding their identity cards. The proctors monitor students through the process to ensure that there are no cases of impersonation during the exam and to detect any suspicious behaviors.

Live proctored exams are different from automated proctored assessments. Raman, et al. (2021) notes that the former examinees must wait for the specific scheduled time before they can take their exams. This means that many students must undertake an exam at the same time. The method has various challenges that limit its accuracy in online proctoring. One study identified that live proctoring depends on factors like student availability, access to strong Wi-Fi or other networks, and the availability of resources, such as tablets and mobile computers (Kaiiali, et al., 2016). In some systems, the school's preferred environment does not have access to the internet and important applications. This limits examination efficiency.

Recorded proctoring usually involves video recording the student during examination along with other log details. Human proctoring in this case is limited but required. Nevertheless, it can be a time-consuming and costly process (Nigam, et al, 2021). On the other hand, automated proctoring allows examinees to take an examination at their preferred time. The exam can still be monitored despite the time that an examinee takes it. Raman, et al. (2021) report that automated proctoring is fast growing as many educational institutions are adopting it because it saves time and allows individual candidate supervision. Additionally, automated proctoring provides a detailed report about any exam malpractices and fraud. Therefore, device applications like cameras help in monitoring students as they take exams virtually. Automated proctoring allows learners to take their exams without location restrictions. Raman, et al. (2021) found that automated proctoring remains the most effective way of conducting remote examinations.

2.3 Accuracy of AI-based proctoring software

In comparison to the decisions of human proctor versus AiAP, Prathish, Narayanan, and Bijlani (2016) did a study where a fully automated, multimodal proctoring software was used with no requirements to sophisticated or expensive hardware on the students, and without the need of live proctor presence during the exam. A comparison was made between malpractices detected using real (human) proctor and proposed software, and the recorded exam attempts were segmented to 14 timeslots. The human proctor and proposed software had a true false and true negative in terms of decision on malpractice in 11 out of 14 time slots. Their results also showed an accuracy of 80% for the automated proctor software in detecting the active window.

In another study, it was reported by Hussein, et al. (2020) that 12% of the students who took the automatically proctored exam in one of the campuses found difficulties in navigating through the questions. In another campus, there were 14% of students who reported an inability to complete the exam successfully. These are percentages that may require more attention to the technical reasons (if any) that prevented the students from having a seamless exam experience.

False flags are possible in automatic proctoring. In a comprehensive review of facial recognition methods, Anwarul and Dahiya (2020) specified 2 main categories of factors that affect the accuracy of facial recognition algorithms: intrinsic and extrinsic factors. Both types of factors are strongly related to AI-based automatic proctoring because the intrinsic components include the physical characteristics of the human face such as expressions, age, skin tone, and many others. These can affect the accuracy due to their variety and to the limitation of training data (i.e. face samples need to be large enough and representative). Whereas the extrinsic factors can influence the proctoring system because they are responsible for changing the appearance of the face like low resolution, low-quality connection, illumination, occlusion, and pose variation.

Concerning the facial recognition of test-takers, machine learning algorithms that are usually utilized in automatic proctoring are well-trained using thousands of images, nevertheless, they do receive a notable amount of criticism for inaccurate and sometimes biased results against colored skins (Coghlan et al, 2021).

In order to test the accuracy of AI-based automatic proctoring, a method with proper design is needed.

3. Methodology

The study relied on experiments as an effective method to determine the accuracy of AI-based automatic proctoring method. The experiment on 244 online exam sessions were done using Manual Testing methodology, which is a software testing technique used for verification of software to spot and resolve any possible errors or flaws (Anwar and Kar, 2019). Previous experiments used multiple features such as a deep learning neural network called Long Short-Term Memory (LSTM) to compare the proposed software performance with respect to human judgment (Nigam et al, 2021). LSTM are artificial recurrent neural networks (RRN), which is an architecture usually used in several applications, including but not limited to time series prediction, natural language processing, sentiment analysis, image and video captioning, text recognition, and sound detection (Van Houdt, Mosquera, and Nápoles, 2020). In this experiment, it was emphasized that violations of proctoring measurements are assessed collectively rather than individually. Despite this, some essential temporal patterns cannot be identified in original 3D pose. Therefore, LSTMs are important in learning the temporal patterns as they learn them directly from the sequences.

It was assumed that cheating students are likely to look around often, place their hands in their pocket, and keep checking if anyone is looking at them. These students also look at their phones and tablets many times more than individuals in natural conditions. However, proctors should not entirely rely on these findings as some violators are aware of the cues they portray. Therefore, they might intentionally display controlled movements or the opposite of what is expected (Randhavane, et al., 2019). Despite the challenges, universities globally may invest in machine learning technologies to ensure their software can distinguish between deceptive and real cues even as some students keep changing their malpractice tricks. The approach, however, has some limitations, primarily because it is based on information gathered from a lab setting in an educational institution. More accurate studies can be acquired from participants who know little about the context of the study.

3.1 Features

The proposed model of this study is called AI-based Automatic Proctoring (AiAP). The method in this model consists of 5 main measurements: screen violation detection, speech violation detection, different face detection, multi faces detection, and eye movement detection. As illustrated in Figure 1, the features that have been used were: LSTM network is to detect any human sound, the screen detection was performed using a modified version of Safe Exam Browser (SEB), whereas the image processing for facial recognition, eyes movement and identification were performed using Google Vision API.

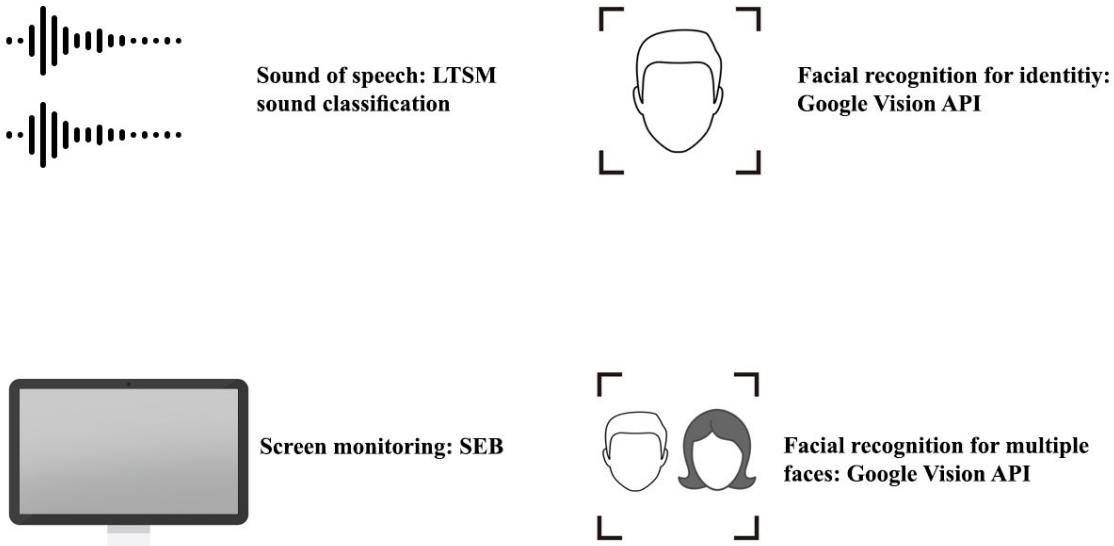


Figure 1: Five measurements taken by the features of AiAP

As previously mentioned, LSTM is an RRN network that is suitable for sound detection. The SEB is an open-source lockdown browser that can detect opened tabs, opened applications in the local computer, Virtual Machines (VM), and internet search or use attempts (SEB, 2021). This lockdown browser can also be integrated into multiple Learning Management Systems (LMS) and other exam solutions. The Google Vision API is cloud-based system that allows developers to integrate multiple features within applications, and these features include optical character recognition (OCR) face and landmark detection (Google Codelabs, 2021).

3.2 Exam structure

The preparations and techniques used in this experiment are summarized in the Figure 2. For more details on each step, the table in Appendix A shows the steps of each part. Screenshots of the online exam within AiAP are included in Appendix B.

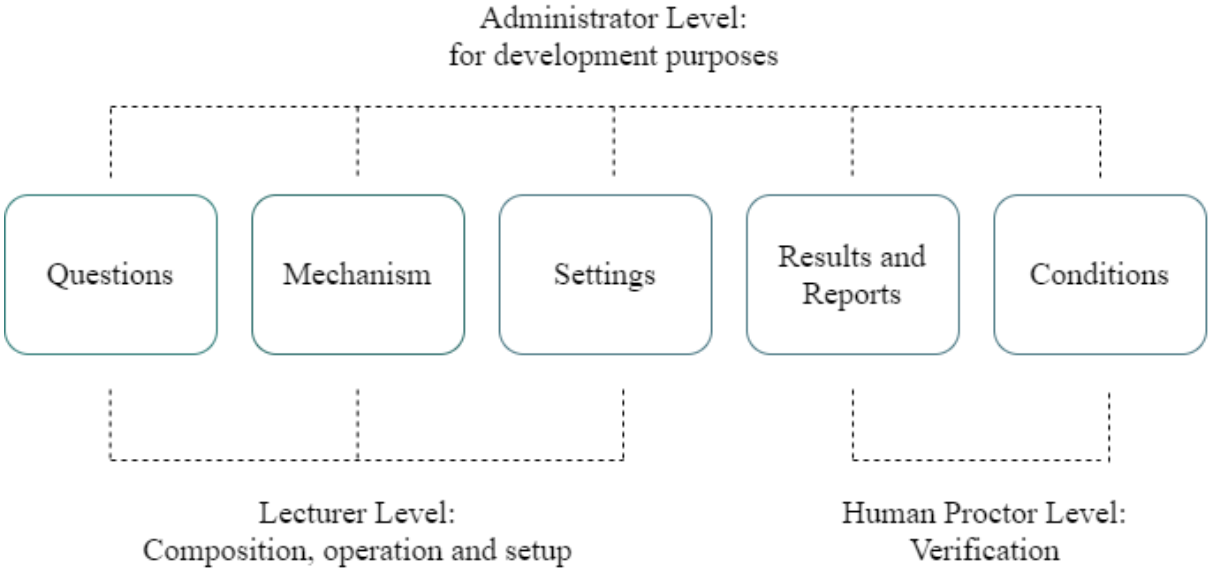


Figure 2: The online exam structure

The online exam consists of 5 parts: Questions, mechanism, settings, results and reports, and conditions. The questions are composed in multiple formats and types depending on the course subject, and they can be put into a specific classification or sub-classification for grouping purposes, in addition to being saved in a question bank. The mechanism consists of the ways the questions behave; for example: how the navigation works (one-way or multi-way), the time counter position and visibility, question shuffling, and live chat box for technical support. The settings are generally about adding and managing the courses and setting up the accounts and

privileges of lecturers and students. Results and reports are basically functions to show the recorded violations, and these violations can be sorted by student or by course. The conditions consist of a list of rules and regulations that students should follow before and during the exam.

The administrator level (i.e. role) has the privilege to control and/or edit all parts for development purposes, the lecturer level can control questions, mechanism and settings of the exam, and the human proctor level is responsible for verification of results and matching them with the conditions. The human proctors are fully aware of the proctoring standards and exam regulations at the university in which the study was conducted.

3.3 Model

The AiAP model is an in-house application designed and developed to help in auto-proctoring the online exams conducted for this experiment. The AiAP proposed model is summarized in Figure 3.

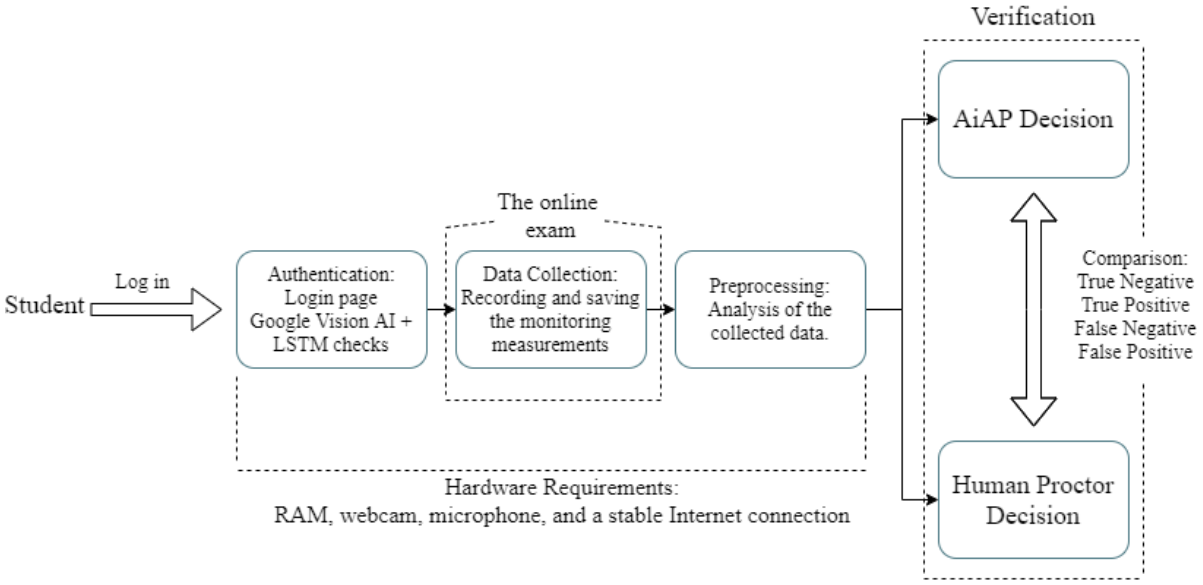


Figure 3: The AiAP proposed model

The AiAP shows the student -at the beginning- an authentication phase that consists of a login page (asking for username and password). Then, once logged in, the authentication continues to another page that asks the student to take a photo, say the word 'Hello' multiple times using the microphone, and then take a photo of their ID card using the webcam. By using Vision API and LSTM features, the software tracks the face, records the voice, scans the ID number out of the card, and finally matches all this information with an external database collected from the student information system (SIS) used at the same university. This process is done for authentication purposes before starting the exam.

The online exam begins, and hence the data collection phase starts. In this phase, the Vision API generates regions of interest from the live videos collected from each student (i.e. object or person), and it recognizes and extracts any suspicious behavior (i.e. the behavior that violates conditions). The pre-processing phase is then done by the software to generate a data of images, screenshots, and voice recordings to collectively form the set of monitoring measurements, namely: eye movement, different face, multiple faces, sound of speech, and screen activity. Lastly, the verification phase (i.e. testing phase 1) is done by AiAP to decide, based on the monitoring measurement results, whether a student has violated the online exam conditions (Positive) or has not violated them (Negative). The same verification is done by the human proctors (i.e. testing phase 2) who labeled the results of each student as positive and negative, in the form of software static (manual) testing methodology. A high number of violations can indicate an attempt to cheat in the online exam. The benchmark to decide whether a cheating case happened or not, was set to a decision of 25% and above. Meaning, if the average of monitoring measurements provided by AiAP or Human proctor was larger than 25% then the student committed a relatively-high number of violations to the conditions and, hence, most likely cheated in the exam.

It should be noted that the software requires some hardware resources, which can be restricted to the availability in every student's device: Random Access Memory (RAM) of a minimum 2GB, a webcam that supports HD aspect ratio, a high-quality microphone, and an internet card with a stable connection. At earlier

development stages, the software was put on training by conducting multiple mock exams to recognize the monitoring measurements (see samples in Appendix B).

3.4 Sampling

The sample of this study was randomly collected from 14 courses in 3 different schools at a Middle Eastern university. These courses collectively consisted of a total 244 students (i.e. 244 online exam sessions). Each online exam was analyzed by the AiAP and given a percentage weight to each monitoring feature: screen violation, speech violation, different face violation, multi-faces violation, and eye movement violation. Afterward, the same online exam results (i.e. data) have been given to 5 human proctors to analyze the behavior of each student, and give a percentage weight for the same monitoring measurements after the exam has been done; which is a part of the verification phase. However, the weight percentages given by AiAP were hidden from the human proctors to avoid familiarity and bias.

4. Results

In this section, we present the results of the experiment done on 244 exam attempts, using the relevant paired sample tests, comparative results, and accuracy measurement. The experiment was based on two hypotheses to compare between both decisions: Human and AiAp.

Null Hypothesis: There is no difference between Human Decision and AiAP Decision (100%).

Alternative Hypothesis: There is a significant difference between Human Decision and AiAP Decision (100%).

4.1 Sample t-test results

A paired sample t-test (see Table 1) was used to compare the mean of Human Decision and AiAP Decision (100%). The t-test was statistically significant, with mean of AiAP Decision (100%) (M=35.6179, SD=12.23946) was significantly higher than Human Decision, (M=25.9552, SD=10.39789, t (245) =-16.146, p<.001, two-tailed.

Table 1: Mean difference of Human Decision and AiAP Decision (100%)

Decision	*n*	*M*	*SD*	*t*
Human Decision	244	25.9552	10.39789	-16.146
AiAP Decision (100%)	244	35.6179	12.23946	
p<0.05				

Therefore, the null hypothesis that there is no difference between Human Decision and AiAP Decision (100%) is rejected. It can be concluded that there is a significant difference between Human Decision and AiAP Decision (100%) in the population. In Table 2, the t-Test: Two-Sample results between Test Phase 1 and Test Phase 2 are included.

Table 2: t-Test: Two-sample assuming equal variances

Decision	Test Phase 1: AiAP Decision Average	Test Phase 2: Human Decision Average
Mean	35.6179	25.9552
Variance	150.90	108.85
Observations	244	244
Pooled Variance		129.88
Hypothesized Mean Difference		0
df		486
t Stat		-9.37
P(T<=t) one-tail		0
t Critical one-tail		1.65
P(T<=t) two-tail		0
t Critical two-tail		1.96

The two testing phases represent the verification process of human proctoring results over AiAP results, leading to a more definitely attributed differences that show the readiness of AiAP. According to Figure 4, the AiAP Decision seems to have detected and classified more violations during the online exams.

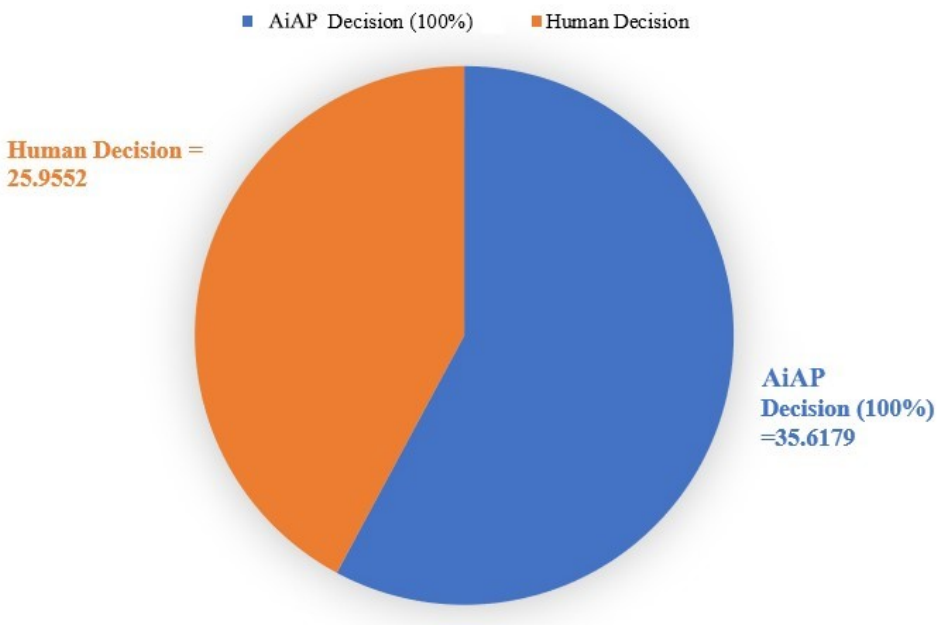


Figure 4: Percentage comparison between Human and AiAP Decisions

According to the AiAP model (review Figure 3), the data was initially collected for pre-processing by the AiAP software and presented in the form of images, recorded sounds, and screenshots. Then the same data was analyzed and verified by the AiAP in the verification phase by highlighting each image, sound, and screenshot that contained a violation with a red color. Each monitoring measurement is accordingly computed and given a specific score (100%). This data was then collected by the researchers and filled in multi-tabbed Excel sheet, sorted by a student sequence number (each student was given a sequence number) and the corresponding monitoring measurements were named after AiAP decision as follows: AI Speech, AI Screen, AI Different Face, AI Multi Face, AI Eye Detection, and finally AiAP Decision.

Then, the same structure was built in Excel and shared with 5 human proctors. The fields were empty, and the data file did not include any percentages/readings from AiAP decision to avoid similarity or bias. The human proctors were asked to perform the verification phase by going through each student record and filling the data under each field (monitoring measurement) as follows: Human Speech, Human Screen, Human Different Face, Human Multi-Face, Human Eye Detection, and finally Human Proctor Decision. The Human Decision is finally computed by finding the average of each monitoring measurement from all 5 human proctors. Two screenshots of the data are provided in Appendix C for reference.

4.2 AiAp accuracy results

In Figure 5, a comparison between the average violations detected by all human proctors and AiAP software is presented. The eye detection was the highest monitoring aspect in terms of the number of violations detected.

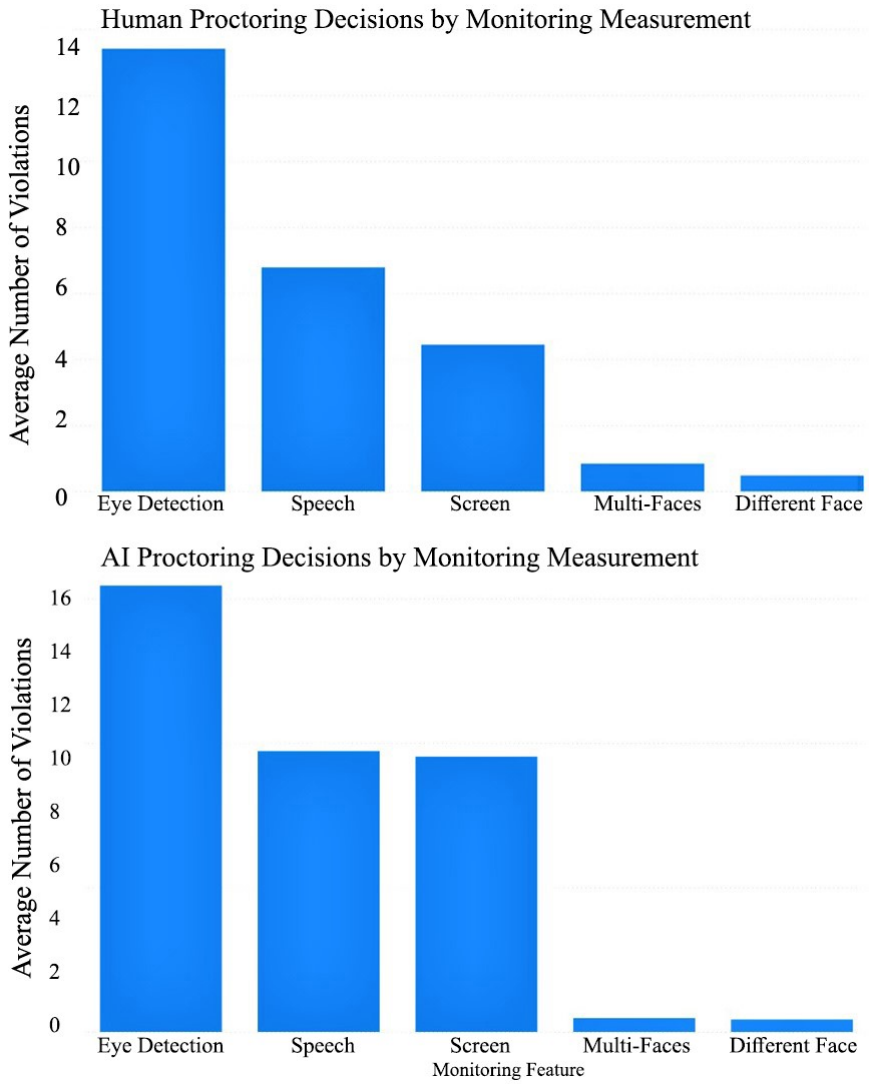


Figure 5: A comparison between the average violations detected by all human proctors and AiAP

The verification phase was done by AiAP to decide, based on the monitoring measurement results, whether a student has violated the online exam conditions (Positive) or has not violated them (Negative). The same verification is done by the human proctors who label the results of each student as positive or negative. A high number of violations can indicate an attempt of cheating in the online exam. As mentioned in Chapter 3, the benchmark to decide whether a cheating case happened or not was set to the average of monitoring measurements of 25% and above.

The final analysis of each decision was based on the following logic:

- If AiAP Decision = 'Negative' AND Human Proctor Decision = 'Positive', then the Result = 'False Negative'
- If AiAP Decision = 'Negative' AND Human Proctor Decision = 'Negative', then the Result = 'True Negative'
- If AiAP Decision = 'Positive' AND Human Proctor Decision = 'Negative', then the Result = 'False Positive'
- If AiAP Decision = 'Positive' AND Human Proctor Decision = 'Positive', then the Result = 'True Positive'

Based on this logic, and since AiAP was set to be compared with Human Proctor decision as a benchmark, the number of correct and incorrect decisions were as shown on Table 3.

Table 3: Number of correct decisions and incorrect decisions of AiAP.

Decision	*n*	AiAP Decision Result	Occurrence
False Negative	244	Incorrect decision	69
True Negative	244	Correct decision	47
False Positive	244	Incorrect decision	5
True Positive	244	Correct decision	123

From these results, the total number of incorrect decisions was 74 out of 244, which counts for approximately 30% of the online exams proctored that were analyzed incorrectly. This outcome leads for a further discussion in the next chapter.

5. Discussion

Debate about the effectiveness of remote online proctoring in online education is bound to continue, especially with the global circumstances that led many educational institutions to adopt distance learning. Despite some drawbacks, there are considerations for automated proctoring methods. It is vital to acknowledge that in-person proctoring has its weaknesses as proctors can miss a cheating incident, leading to ethical issues. Similarly, the analysis presented in this study shows that automated proctoring raises significant ethical concerns. Some of these concerns include cases of unfairness resulting from wrong artificial intelligence-informed judgment. There are also possibilities of intruding students' privacy, autonomy, and various psychological factors (Kharbat and Abu Daabes, 2021), especially in situations where proctoring relies on webcams to monitor learners' environments during assessments. Experiments presented in the study further show that online proctoring relied on data sets. Therefore, it can be speculated that monitoring software can contribute to privacy loss and reduced trust between learners and their instructors as the risk of unauthorized data mining is high. These ethical concerns can be justified by balancing the benefits and developing strategies to minimize risks.

This study sought to answer the following research questions:
RQ 1: Is there a significant difference between the Human Proctor Decision and AiAP Decision?
The answer is yes, there was a significant difference between both decisions based on a Paired sample t-Test. The AiAP made more decisions with 'Positive' cases than Human Proctors.
RQ 2: Is the AiAP efficient in analyzing and verifying the cases of cheating in online exams?

The answer is no, because 30% of the total online exams were incorrectly analyzed. Therefore, the software needs to be considered for further development and improvement in terms of Machine Learning, mechanism, and integration within features.

The analysis of false negative and false positive cases leads to a conclusion that AiAP can be taken to another model of testing and verification over several steps. It is important to notice that some studies found lower incorrectness rates in another AI-based proctoring software. For example, Raj, Narayanan, and Bijlani (2015) worked on an automatic proctoring software with a multi-modal method, through which the test-takers were scanned using image and audio processing, and PC monitoring techniques. The validation of detected malpractices during online exams was made for 12 video segments chosen to measure the accuracy of the software. Out of 12 segments, the software detected higher number of malpractices in 4 segments than the actual malpractices that were done by test takers, and in 6 segments the software detected lower number of malpractices. The overall results of this study showed that auto-proctoring software reached an 8% rate of false positive and 13% rate of true negative. The conclusion by Raj, Narayanan, and Bijlani (2015), however, was about lower necessity of having remote (human) proctors to supervise the online exams.

Based on the finding of incorrect flags, namely the false negative and false positive, some concerns about students' privacy and the accuracy of AI-based proctoring need to be addressed.

5.1 Consequences of using AI-based proctored examination

Any technological change applied in any setting can positively impact individuals and academic institutions regarding reduced time and increased accuracy in conducting online exams (Ilgaz and Afacan Adanir, 2020). However, and while technology is rapidly advancing, it has become imperative for universities to seek solutions for maintaining privacy issues while being connected to intelligent and smart systems.

In some cases, students can be cautious about giving out data that they suppose will not be stored and protected efficiently. Recorded videos, IP addresses, facial identity and ID card scans, and other identity elements are susceptible details that should be adequately preserved. Software solutions should not keep information and store data in a vulnerable way (Slusky, 2020). Moreover, developers of automated proctoring software can consider analyzing the test at any time (asynchronously) in case they suspect any malicious act.

Many universities and higher educational institutions are regularly acquiring novel and intelligent third-party technologies to prevent academic dishonesty aiming to reduce cheating in examinations, and that can replace human proctors. This research suggests that there might be a continuing need of a human presence to control and maintain the quality of examinations, but a robust, reliable, and efficient auto proctoring software can ensure that academic integrity is guaranteed and safeguarded.

5.2 Contribution

This study adds valuable empirical findings and implications targeted at the Middle Eastern region, targeting specific countries with a specific mindset in online examination systems. These countries faced immense examination integrity-related challenges, such as students' high cheating potential, random communication with others using multiple devices, and students who are freely browsing the internet while examining. The mentioned challenges were extremely apparent with the widespread and increasing use of online examinations at many higher education institutions in the Middle East (Tayan, 2017).

However, there are still some concerns about the accuracy of auto-proctoring software. Some studies were made to compare remote and auto-proctoring software with each other. For example, Hussein, et al. (2020) conducted a study of a four-phased method to test multiple auto-proctoring techniques and software for computerized exams at a large scale. The study made an evaluation matrix with the help of students' and professional live proctors' feedback. It is important to notice that software like AiAP requires high costs and efforts to assure the quality of service. Nevertheless, concerns about the accuracy of AI-based proctoring software have been raised, specifically in terms of accuracy of results and flagged behaviors. For example, a report from Clark (2021) showed the results from a well-known remote proctoring software in which there was no correct facial recognition of students from certain ethnicities, which led to putting a red flag on their exam attempt. Another research done by a proctoring company named ProctorU that heavily depends on the post-analysis of human being (similar to the phase 2 testing of this study), has found that no more than 10% of faculty members review the AI-based proctoring results, leaving the verification of flagged exam attempts under suspicion and question. Accordingly, ProctorU announced that their AI-only model would be stopped due to inaccuracy issues (ProctorU, 2021 a and b).

In this case, it is expected that more auto-proctoring development companies will consider adopting the hybrid model in which both AI-based proctoring software and human proctors work together to better shape and control the examination experience. As mentioned by Slusky (2020), ProctorU reported that more than 10% of 1.5 million exams required active intervention within a period of 12 months, which is a statistic that implies the need for this hybrid model where both live (human) and automatic (AI-based) proctors need to work together to assure quality and accuracy of online exam control. This argument was supported by Miguel, et al. (2015).

This study shed light on the need for a smart and state of the art technological proctoring system while conducting online exams and proposed a tested AI-based proctoring system. Thus, the paper similarly suggests that effectiveness of the online exams can be achieved by automated, valid, reliable, and most importantly secure proctoring software with the determination of reassuring and supporting accurate learning outcome production, and safeguarding alignment with constructive implementation. Additionally, the effectiveness and accuracy of such software demands institutional provision and support, as well as the establishment of appropriate atmospheres and surrounding inside most Middle Eastern higher educational institutions (Raza, et al., 2021). There are speculations that online proctoring technologies could lead to social trajectories or similar possibilities. These risks are not easy to assess. Despite this, they should not be ignored. Some colleges and universities have avoided AI-based technology as a way of avoiding such risks. Additionally, it is expected that more universities will make efforts to enhance security to encourage more students to embrace online learning and assessment methods. Besides their efforts, universities should consider reaffirming to stakeholders, including instructors and students that they will not become victims of governments and powerful corporations who might want information to deprive online learning participants of their freedom. This implies that it is the university's responsibility to uphold a culture of trust that has been there in traditional educational settings. Furthermore, higher educational institutions are encouraged, based on the findings of this study, to involve the support of human proctors as a backup, or as an extra verification engine, to seek a higher quality of automated proctoring. A 'Blended' type of proctoring may be proposed for future research.

5.3 Limitations

The proposed software has some technological limitations, however, such as the inability to share the webcam with two distinct applications, which prevents the proctor from utilizing any other online communication application to monitor exam candidates in real-time. This restriction also applies to microphones, which prevent candidates from asking questions or communicating with the human proctor, teacher, or advisor during the exam. Because of these constraints, the auto-proctoring software of this study can only work on its own, with no online (live) monitoring from a human proctor.

Furthermore, in terms of technical capabilities and logistics, the proposed software has limitations and several challenges. The main challenge that may affect the overall performance of the proposed software is the low quality of webcams and microphones (Usually, the auto-proctoring needs a full HD quality to have more accurate results), in addition to internet connection fluctuation and poor internet connection in some areas, especially in the country where this study was conducted. In addition, as proposed by Slusky (2020), there are multiple constraints that need to be considered in terms of cybersecurity for AI-based proctoring software in general, such as compliance with digital criminology and security laws nationally and internationally, lockdown of copying and pasting in the online exam, prevention of proxy in taking the exam (i.e. a student taking the exam on behalf of another), prevention of using virtual machines and remote access, and detection of a second monitor.

5.4 Recommendations and closure

Online proctoring becomes more accurate when the system is automated with less human intervention. Automation allows efficient data transfer and real-time monitoring, consequently detecting malpractice cases that human proctors cannot identify during assessments, especially in the case of large numbers of students. More studies should be conducted to test the AI-based automatic proctoring software in terms of compatibility with Internet browsers, variation of course types, more question types and extra detection techniques.

In terms of methodology, more human proctors can join future studies and participate in accuracy and accuracy testing of auto-proctoring. Testing can be done by taking samples of recorded videos of the students, rather than having to screen and review a full recording. This can be done by choosing random samples from the recorded -proctored- material, which might be labeled as 'violating' and 'non-violating' samples to focus the efforts of accuracy testing.

The primary objective for this study was to evaluate the accuracy of AI-based automated proctoring in enhancing academic integrity. An analysis of various proctoring methods showed that all invigilation approaches have some weaknesses. Therefore, it is recommended that educational institutions should employ a multiple-approach model to improve exam monitoring accuracy. Despite this, delivering and monitoring examinations using an online, automated software have the potential to be more effective and accurate than the traditional approach.

References

Alessio, H.M., Malay, N., Maurer, K., Bailer, A.J. and Rubin, B., 2017. Examining the effect of proctoring on online test scores. *Online Learning*, 21(1), pp.146-161. http://doi.org/10.24059/olj.v21i1.885.

Anwar, N. and Kar, S., 2019. Review paper on various software testing techniques & strategies. *Global Journal of Computer Science and Technology*, 19(2), pp.43-49. Available at: <https://computerresearch.org/index.php/computer/article/view/1873> [Accessed 11 June 2021].

Anwarul, S., Dahiya, S. 2020. A comprehensive review on face recognition methods and factors affecting facial recognition accuracy. In: Singh, P., Kar, A., Singh, Y., Kolekar, M., Tanwar, S. (eds) Proceedings of ICRIC 2019. *Lecture Notes in Electrical Engineering*, vol 597. Springer, Cham. https://doi.org/10.1007/978-3-030-29407-6_36

Clark, M., 2021. Students of color are getting flagged to their teachers because testing software can't see them. [online] Available at: <https://www.theverge.com/2021/4/8/22374386/proctorio-racial-bias-issues-opencv-facial-detection-schools-tests-remote-learning> [Accessed 12 December 2021].

Coghlan, S., Miller, T. and Paterson, J. 2021. Good proctor or "big brother"? ethics of online exam supervision technologies. *Philosophy & Technology*. vol 34, pp.1581–1606.

Dendir, S. and Maxwell, R.S., 2020. Cheating in online courses: Evidence from online proctoring. *Computers in Human Behavior Reports 2*. [Advance online publication]. https://doi.org/10.1016/j.chbr.2020.100033.

Google Codelabs., 2021. Using the Vision API with Python. [online] Available at: <https://codelabs.developers.google.com/codelabs/cloud-vision-api-python#0> [Accessed 10 February 2022].

Hussein, M.J., Yusuf, J., Deb, A.S., Fong, L. and Naidu, S., 2020. An evaluation of online proctoring tools. *Open Praxis*, 12(4), pp.509-525. https://doi.org/10.3316/informit.620366163696963.

Ismail, R., Safieddine, F. and Jaradat, A., 2019. E-university delivery model: Handling the evaluation process. *Business Process Management Journal*, 25(7), pp.1633-1646. https://doi.org/10.1108/BPMJ-10-2018-0281.

Ilgaz, H. and Afacan Adanır, G., 2020. Providing online exams for online learners: Does it really matter for them? *Education and Information Technologies*, 25(2), pp.1255-1269. https://doi.org/10.1007/s10639-019-10020-6.

Kaiiali, M., Ozkaya, A., Altun, H., Haddad, H. and Alier, M., 2016. Designing a secure exam management system (SEMS) for M-learning environments. *IEEE Transactions on Learning Technologies*, 9(3), pp.258-271. https://doi.org/10.1109/TLT.2016.2524570.

Kharbat, F.F. and Abu Daabes, S.A., 2021. E-proctored exams during the COVID-19 pandemic: A close understanding. *Education and Information Technologies*. [Advance online publication]. https://doi.org/10.1007/s10639-021-10458-7.

Kraglund-Gauthier, W.L. and Young, D.C., 2012. Will the real John Doe stand up? Verifying the identity of online students. In L. A. Wankel and C. Wankel (Eds.), *Misbehavior online in higher education* (Vol. 5, pp. 355-377). England: Emerald Group Publishing Limited. https://doi.org/10.1108/S2044-9968(2012)0000005019.

Miguel, J., Caballé, S., Xhafa, F. and Prieto, J., 2014. Security in online learning assessment towards an effective trustworthiness approach to support E-learning teams. *Proceedings of 2014 IEEE 28th International Conference on Advanced Information Networking and Applications*: *IEEE AINA 2014* (pp. 123-130), 13-16 May 2014, University of Victoria, Victoria, Canada. IEEE. http://doi.org/10.1109/aina.2014.106.

Moore, H.P., Head, J.D. and Griffin, R.B. 2017. Impeding students' efforts to cheat in online classes. *Journal of Learning in Higher Education*, 13(1), pp.9-23. [online] Available at: <https://files.eric.ed.gov/fulltext/EJ1139692.pdf> [Accessed 10 February 2022].

Nie, D., Panfilova, E., Samusenkov, V. and Mikhaylov, A., 2020. E-learning financing models in Russia for sustainable development. *Sustainability*, 12(11), pp.4412-4426. http://doi.org/10.3390/su12114412.

Nigam, A., Pasricha, R., Singh, T. and Churi, P., 2021. A systematic review on ai-based proctoring systems: Past, present and future. *Education and Information Technologies*, 26(5), pp.6421-6445. https://doi.org/10.1007/s10639-021-10597-x.

Prathish, S. Narayanan, S.A. and Bijlani, K., 2016. An intelligent system for online exam monitoring. *Proceedings of the 2016 International Conference on Information Science (ICIS)* (pp. 138-143), 12-13 August 2016, Kochi, India. IEEE. https://doi.org/10.1109/infosci.2016.7845315.

ProctorU., 2021a. A human-centered proctoring policy. [online] Available at: <https://www.proctoru.com/human-centered-proctoring> [Accessed 10 February 2022].

ProctorU., 2021b. ProctorU to discontinue exam integrity services that rely exclusively on AI. [online] Available at: <https://www.proctoru.com/industry-news-and-notes/proctoru-to-discontinue-exam-integrity-services-that-rely-exclusively-on-ai> [Accessed 10 February 2022].

Raj, R.V., Narayanan, S.A. and Bijlani, K., 2015. Heuristic-based automatic online proctoring system. *Proceedings of the 2015 IEEE 15th International Conference on Advanced Learning Technologies* (pp. 458-459), 6-9 July 2015, Hualien, Taiwan. IEEE. https://doi.org/10.1109/ICALT.2015.127.

Raman, R., Sairam, B., Veena, G., Vachharajani, H. and Nedungadi, P., 2021. Adoption of online proctored examinations by university students during COVID-19: Innovation diffusion study. *Education and Information Technologies*. [Advance online publication]. https://doi.org/10.1007/s10639-021-10581-5.

Randhavane, T., Bhattacharya, U., Kapsaskis, K., Gray, K., Bera, A. and Manocha, D. 2019. The liar's walk: Detecting deception with gait and gesture. *ArXiv Preprint*. [online] Available at: <https://arxiv.org/pdf/1912.06874.pdf> [Accessed 10 February 2022].

Raza, S.A., Qazi, W., Khan, K.A. and Salam, J., 2021. Social isolation and acceptance of the learning management system (LMS) in the time of COVID-19 pandemic: An expansion of the UTAUT model. *Journal of Educational Computing Research*, 59(2), pp.183-208. https://doi.org/10.1177/0735633120960421.

Safe Exam Browser (SEB)., 2021. Developer documentation - Integration. [online] Available at: <https://safeexambrowser.org/developer/seb-integration.html> [Accessed 10 February 2022].

Slusky, L., 2020. Cybersecurity of online proctoring systems. *Journal of International Technology and Information Management*, 29(1), pp.56-83. [online] Available at: <https://scholarworks.lib.csusb.edu/jitim/vol29/iss1/3> [Accessed 10 February 2022].

Tayan, B.M., 2017. Academic misconduct: An investigation into male students' perceptions, experiences & attitudes towards cheating and plagiarism in a Middle Eastern university context. *Journal of Education and Learning*, 6(1), pp.158-166. http://doi.org/10.5539/jel.v6n1p158.

UNESCO., 2021. One year into COVID-19 education disruption: Where do we stand? [online] Available at: <https://en.unesco.org/news/one-year-covid-19-education-disruption-where-do-we-stand> [Accessed 10 February 2022].

UNICEF., 2020. COVID-19: Are children able to continue learning during school closures?: A global analysis of the potential reach of remote learning policies. [online] Available at: <https://data.unicef.org/resources/remote-learning-reachability-factsheet/> [Accessed 10 February 2022].

Van Houdt, G., Mosquera, C. and Nápoles, G., 2020. A review on the long short-term memory model. *Artificial Intelligence Review*, 53(8), pp.5929-5955. https://doi.org/10.1007/s10462-020-09838-1.

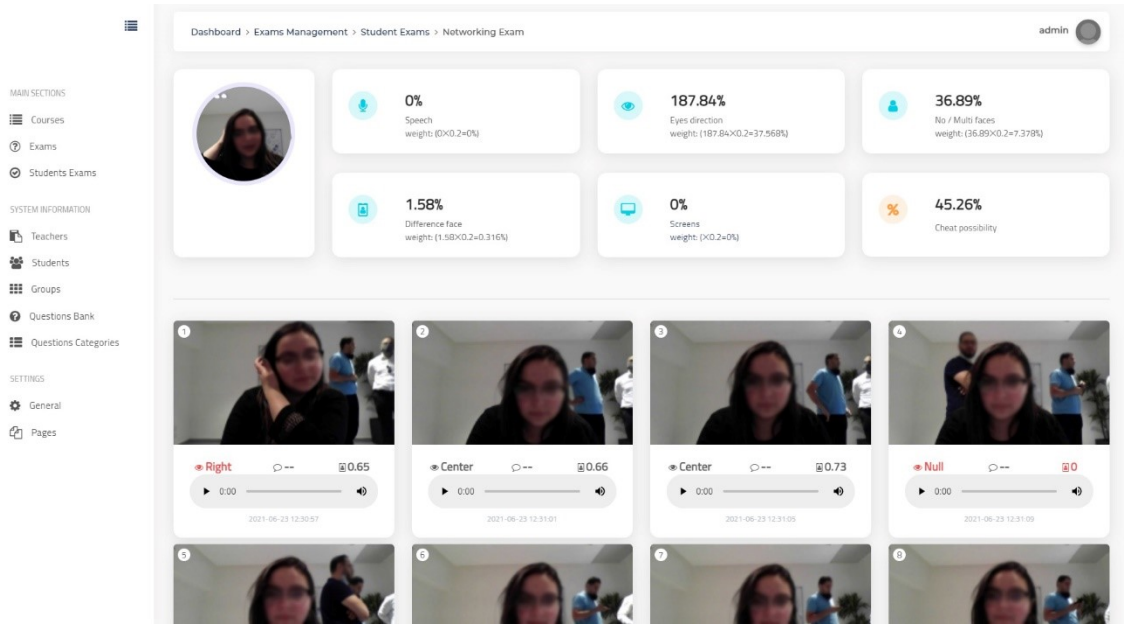
Appendix A: Exam mechanism

Examinations	Examinations are administered as follows:
	1. Adding a classification and sub-classification of questions 2. Choose type of question: a. Multiple choice b. True/False c. Write the correct answer (to be graded manually) 3. Create question bank Add a question a. Name of question b. Question text c. Type of answer d. Add answers if using MCQs e. The question-and-answer text is added using Text Editor to add images and other font formats.
Administration	The system is managed according to the following:
	System General Manager: 1. Managing lecturers 2. Management of teaching materials 3. Student management 4. Managing the question bank 5. Viewing the results 6. Creating an exam 7. Viewing reports Lecturers 1. Managing his/her own teaching materials 2. Viewing examinations 3. Administering of exams 4. Managing the question bank 5. Confirming results 6. Viewing the students
Exam mechanism	The exam has a specific mechanism, as follows:
	1. The exam can be one-way: Each question is allocated a certain time to answer, and the student can/cannot return to the previous question after answering. 2. The exam can be two-way, where students can move between the questions. 3. At the end of the exam, a specific time will be allowed to review the answers. 4. The questions are selected so that they cannot be repeated for any student in the same exam. 5. At the beginning of the exam, the following are checked: a. The student's photo b. Screen sharing c. Sound sharing which is checked by saying "Hello" 6. The exam is shown on-screen with three separate areas: a. Question area b. Answer area c. Online observer area d. Also the clock, the calculator icon, and the chat text. 7. The exam may only be taken on a desktop device, using the Chrome browser. 8. The exam cannot be interrupted. In the event that the exam is disconnected, it is possible to return to the exam at the same point of disconnection.
Obtaining the result	The results are obtained by following criteria:
	1. The number of monitoring violations made by the student is extracted so that the human proctor or the system administrator can see the errors and then either reject or approve the examination and the result. 2. Violations include:

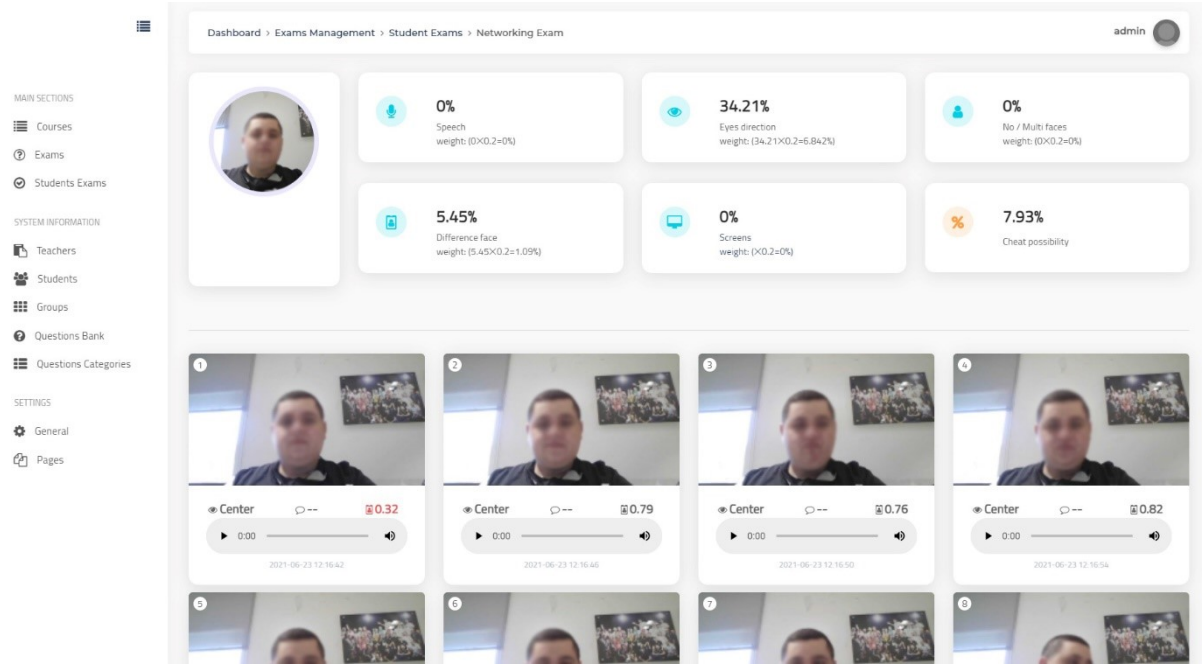
	a. A different face or two faces on the screen, or a face that does not look at the screen or is covered. b. The student cannot use a still image or a static or pre-made video as it is verified during the exam. c. Moving the eyes or fixing them in a place other than the middle of the screen. d. The student speaks. e. The student opens any page on his/her computer that violates the exam interface. f. Use of any still image or pre-made video.
Conditions placed on the student	The following conditions are placed on students:
	1. Opening the exam in the Chrome browser is preferred. 2. The presence of a camera on the device in the middle of the upper part of the screen. 3. Documenting the student file by filming the student during the exam. 4. A microphone that is working properly and is tested prior to the exam by the system. 5. Allow student screen sharing. 6. Take the exam in a quiet room and avoid making human speech sounds. 7. The exam cannot start if one of the control applications is running on the student's device (Zoom - AnyDesk - TeamViewer - Remote Desktop).
Reports	The reports are generated by following:
	The lecturer receives a page showing the names of the students who took the exam in addition to the total number of errors (violations) and the progress of the examination. Pictures of all students who took the exam and pictures of all exam screens.
Technical support	Technical Support includes the following:
	Live chat system between the student and the lecturer in case of inquiries.

Appendix B: Screenshots from the online exam within AiAP, showing a sample of the visual dataset

Faces sample 1:

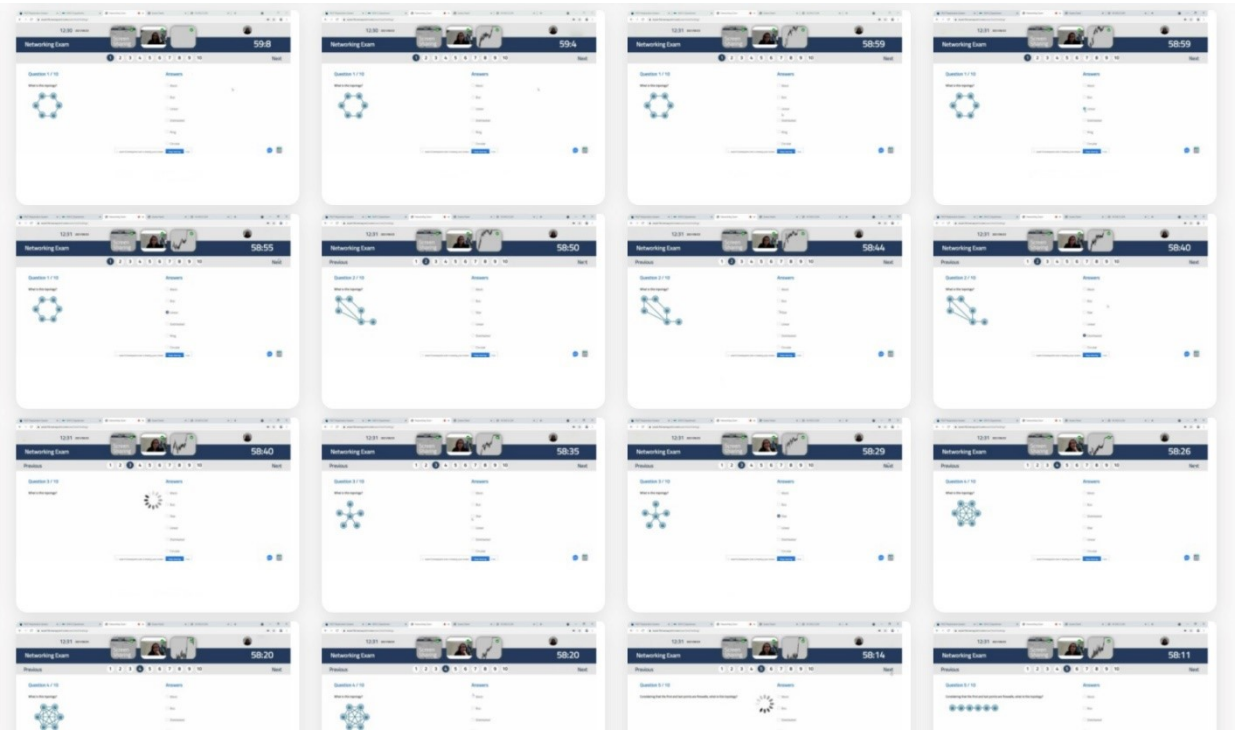


Faces sample 2:



Screenshots from the online exam within AiAP

Screen monitoring sample:



Appendix C: Screenshots showing a sample of the collected data

Student Sequence Number	AI Speech (20%)	AI Screen (20%)	AI Different Face (15%)	AI Multi Face (15%)	AI Eye Detection (30%)	AI Proctor Decision (100%)
1	15	8	0	0	14	37
2	14	8	15	0	29	66
3	19	17	0	11	1	48
4	0	6	0	0	9	15
5	0	2	0	0	14	16
6	8	4	0	0	10	22
7	3	4	0	12	25	44
8	11	9	0	0	14	34
9	18	14	0	0	14	46
10	18	1	0	0	5	24
11	10	18	0	0	12	40
12	4	15	0	0	10	29
13	8	16	0	0	0	24
14	17	6	0	0	4	27
15	14	11	15	0	19	59

Student Sequence Numb	Human 1 Speech (20%)	Human 1 Screen (20%)	Human 1 Different Face (15%)	Human 1 Multi Face (15%)	Human 1 Eye Detection (30%)	Human 1 Decision (1
1	10	5	0	0	11	26
2	12	0	15	0	26	53
3	10	20	0	15	0	45
4	2	3	0	0	6	11
5	2	0	0	0	11	13
6	6	1	0	0	7	14
7	1	1	0	15	22	39
8	9	6	0	0	11	26
9	0	11	0	0	11	22
10	0	0	0	0	2	2
11	8	0	0	0	9	17
12	2	0	0	0	7	9
13	6	0	0	0	0	6
14	15	3	0	0	1	19
15						