Towards Fair Assessments: A Machine Learning-based Approach for Detecting Cheating in Online Assessments

Manika Garg*
The Hague University of Applied Sciences
The Hague, Netherlands
Department of Computer Science
University of Delhi
Delhi, India
manikagarg2007@gmail.com

Anita Goel
Department of Computer Science
Dyal Singh College, University of Delhi
Delhi, Delhi, India
goel.anita@gmail.com

Abstract

Academic cheating poses a significant challenge to conducting fair online assessments. One common way is collusion, where students unethically share answers during the assessment. While several researchers proposed solutions, there is lack of clarity regarding the specific types they target among the different types of collusion. Researchers have used statistical techniques to analyze basic attributes collected by the platforms, for collusion detection. Only few works have used machine learning, considering two or three attributes only; the use of limited features leading to reduced accuracy and increased risk of false accusations.

In this work, we focus on In-Parallel Collusion, where students simultaneously work together on an assessment. For data collection, a quiz tool is improvised to capture clickstream data at a finer level of granularity. We use feature engineering to derive seven features and create a machine learning model for collusion detection. The results show: 1) Random Forest exhibits the best accuracy (98.8%), and 2) In contrast to less features as used in earlier works, the full feature set provides the best result; showing that considering multiple facets of similarity enhance the model accuracy. The findings provide platform designers and teachers with insights into optimizing quiz platforms and creating cheat-proof assessments.

Keywords

Academic dishonesty, Cheating, Online education, Machine learning, Feature engineering, Integrity

ACM Reference Format:
Manika Garg and Anita Goel. 2025. Towards Fair Assessments: A Machine Learning-based Approach for Detecting Cheating in Online Assessments. In *LAK25: The 15th International Learning Analytics and Knowledge Conference (LAK 2025), March 03–07, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3706468.3706482

*Corresponding author.

LAK 2025, March 03–07, 2025, Dublin, Ireland

ACM ISBN 979-8-4007-0701-8/25/03
https://doi.org/10.1145/3706468.3706482

1 Introduction

In recent years, online education has received considerable attention from the global academic community. After the COVID-19 pandemic, there has been a significant shift towards online education, with many reputed institutions now offering online or blended courses as part of their main curriculum [9][15]. As part of this educational shift, online assessments have emerged as a crucial tool for evaluating student learning. However, as online assessments become more prevalent, there is a parallel increase in instances of academic cheating. Students, driven by personal goals or external pressures, frequently engage in unethical practices to attain high grades and excel academically [40]. The rising cheating practices pose a significant challenge to conducting fair assessments and even raise concerns about the credibility of online achieved credits [11].

Our current research expands upon our previous research about Academic Dishonesty Mitigation Plan (ADMP) [17] for managing cheating in online assessments. In the ADMP, different forms of dishonest practices observed in online assessments and their potential solutions are summarized. In this work, we focus on collusion – one of the most common forms of cheating observed in online assessments. Collusion refers to a cheating method where a student works together or copies from another student in an assessment [13]. Just like traditional examinations, it is expected that students complete their online assessments individually, but it is observed that many students unethically collaborate with each other. Students exploit the unmonitored settings of online environments to discreetly collaborate by exchanging answers in person, through phone calls, or via messaging applications such as email and WhatsApp [9, 36]. Previous research [16, 30] indicates the benefits of collaborative learning, however, unethical collaborative behaviors like collusion are always condemned. Collusion is a major concern for online educators as they strive to find solutions to maintain integrity [10].

Several researchers [27, 38] have proposed solutions to handle collusion. Some [1, 34] suggest change in assessment format to oral and subjective assessments, as they pose greater challenges for collusion compared to more objective formats. Another approach is using traditional proctoring method, which involves monitoring students during assessments through video streaming [4]. These methods, while effective, can be resource-intensive and laborious, making their implementation challenging. The balance between the need for anti-collusion measures and the practical constraint of implementation remains a key consideration in this area of research.

While many research studies[2, 6, 24] have proposed automated solutions for detecting collusion, there is a lack of clarity regarding the specific types of collusion they focus onamong the five possible types of collusion that we identified. Moreover, existing studies [6, 24] primarily apply statistical techniques to analyze basic attributes collected by the online platforms. Some research studies [13][39] leverage Machine Learning (ML) methods, but they consider two or three features only. The use of limited features can often lead to reduced accuracy and increased risk of false accusations; this poses a significant challenge for collusion detection [14]. Establishing ground truth for performance measurement remains another major challenge. Existing methods for establishing ground truth involve simulations and student interviews. These methods are prone to bias and may not provide a reliable foundation for identifying collusion [21].

In this paper, we identify five possible methods of collusion (explained in Section 2.1 in more detail). We focus on a specific type of collusion known as *In-Parallel Collusion* where students simultaneously work together on an assessment. We use feature engineering and machine learning to create a model for collusion detection. The basis of our research is data collected from several online assessments. The assessments were conducted using the online quiz tool, namely *iQuiz* [47], that facilitates the administration of online assessments. For this work, we upgraded *iQuiz* to capture a detailed clickstream of student interactions with the platform. From the collected data, we derive seven features at different levels of granularity. We conduct experiments with five classifiers applied to full feature set and feature subsets, to identify the most effective model for collusion detection. Finally, we analyze the performance of individual features and uncover the test-taking patterns of colluders. In summary, the main contributions of our work are:

- We improvise an online quiz tool to collect detailed clickstream data. From the collected data, we derive seven features at different levels of granularity, using feature engineering methods.
- We experiment with five classifiers applied to full features and feature subsets, and evaluate them to get the most effective model for collusion detection.
- We apply the best model on the unlabeled empirical data, in addition to the experimental data. The model predictions for unlabeled data were validated through the integration of a recognized personality test and spatial data analysis.
- We analyze the performance of individual features and discuss the test-taking patterns of colluders. Based on our findings, we highlight strategies for optimizing online quiz platforms and creating cheat-proof assessments.

The rest of the paper is organized as follows: Section II provides a brief background about the study. Section III describes the materials and methods of our study. Section IV provides experimental results and Section VIIdiscusses the findings and concludes the paper.

2 Background

2.1 Types of Collusion

Students can collude in different ways during the online assessment. We identified five possible methods of collusion. Identifying various types of collusion methods facilitates the ability to systematically measure and proactively address these violations. Table 1 provides a summary of category constructs and operational definitions of collusion. In this work, we focus on the In-Parallel collusion.

2.2 Existing Solutions

The prevalence of collusion in online assessments has prompted extensive research to develop effective integrity solutions. A commonly employed solution to deter cheating in educational settings is the implementation of ***honor code***. The honor code serves as a formal document outlining the definitions of cheating and the corresponding consequences for violations. Numerous research studies [27, 41] delve into strategies for the creation and adoption of honor codes. However, due to the inconsistent enforcement of honor codes by the institutions and their lack of awareness among students, these codes remain ineffective [32].

Many researchers [4, 38] propose the use of traditional ***human proctoring methods*** in online assessments where trained proctors monitor students in real-time through video streaming. However, implementing such methods can be costly and resource-intensive, requiring adequate internet bandwidth and webcams for all participants [8]. In many countries, video proctoring during assessments is not allowed because of recording regulation in personal spaces.

Different ***assessment design approaches***, such as oral and subjective assessments [6, 8], have been proposed to address the issue of collusion by introducing elements that are harder to manipulate collaboratively. While these approaches enhance the integrity of assessments, they may have practical challenges, such as resource requirements for oral assessments or increased grading time for subjective assessments [26].

While the methods described above demand some form of manual labor and supplementary resources, data analysis research has leveraged automated methods to identify collusion. With the advancement in technology, it is now possible to capture process data representing student behavior with the assessment. The existing research studies predominantly measure the degree of similarity between pairs of students to identify patterns of collusion. These studies consider basic attributes collected by the online platforms like responses to questions and associated timing information. In response-based analysis, researchers commonly utilize methods such as Simple matching coefficient [24], Jaccard similarity [8], and correlation [35] to assess the similarity between pairs of students based on their responses to questions. Conversely, in time-based analysis, the focus is on evaluating the correlation between the timelines of the assessment process [36][5][35][37].

Table 2 displays various studies addressing collusion through data analysis. We observe that the use of statistical methods is more common than ML methods. Moreover, features, like response similarity and exam submission time are frequently employed by previous studies. There exist some studies [39][8] that have considered question-specific features; however, these studies are limited to considering only two or three features. The use of limited features poses a challenge in the detection of collusion. Another major challenge is the identification of ground truth. Ground truth in this context refers to the accurate knowledge of which students engaged in collusion. In previous studies [22, 24], the determination

Table 1: Category Constructs and Operational Definitions

Category	Method	Reference
In-Parallel	Two students simultaneously work together on the assessment	[7]
Leader-Follower	One student attempt to solve the questions while systematically transferring the answers to another student.	[21, 23]
Key-Copy	One student shares the entire answer key with other student, who then simply copies all the answers without any attempt to solve the questions	[5, 27]
Divide-Conquer	Two students divide the quiz questions among themselves, complete their sections independently, and then collaborate to share their answers	[8]
Final-Key-Match	Two students work autonomously on the assessment and review their solutions collectively after finishing	[12]

Table 2: Previous Studies addressing Collusion using Data Analysis

Paper	Features	Method	Participants
[24]	Response, Exam submission time	Simple matching coefficient, Mean absolute deviation	192
[39]	Exam score, Question submission time	SVM	187
[8]	Response (subjective), Question starting and submission time	Absolute deviation, Clustering	-
[28]	Response, Exam starting and submission time, Question submission time	Euclidean distance	148
[33]	Response, Response time, Revision	Classification tree	304
[14]	Response, Question submission time, IP address	Correlation	78
[36]	Exam submission time	Mean absolute deviation	7518
[5]	Exam starting and submission time	Graphs	123
[6]	Exam submission time, Exam score	Graphs	132
[35]	Response	Correlation	63
[22]	Response, Exam starting and submission time	Distance metric	233
[13]	Response	Cluster analysis	1992

of ground truth has been constrained by methods like conducting student interviews or employing simulation studies. Student interviews, however, are susceptible to bias. Students are often reluctant to confess to academic misconduct, introducing an inherent limitation to the accuracy of the ground truth obtained through this method. On the other hand, simulation studies typically rely on artificially generated data to mimic real-world scenarios. However, this artificial data may not accurately reflect the complexities present in actual student interactions during assessments. As a result, the findings drawn from simulation studies may have limited generalizability to real-world settings. To address these limitations and obtain reliable evidence of collusion, it is necessary to adopt a more comprehensive approach; our research study is focused in this direction.

3 Materials and Methods

3.1 Dataset

We conducted several online assessments on the undergraduate students at a large public university in India during the academic year 2022-23. Each online assessment comprises twenty MCQs based on the fundamentals of computer science and information technology. The questions consist of four response options of which only one is correct. The student can select a response option or may leave it unanswered. The students were also allowed to return and revise already answered questions. The questions were presented in the same order for all students and a maximum time limit of twenty minutes was given to complete the assessment. The assessments were conducted in a university lecture hall where the students used their mobile phones to take the test.

The assessments were hosted on an online quiz tool – *iQuiz*, integrated with the Moodle learning management system. *iQuiz* allows users to create and administer quizzes, while capturing student interactions with the platform. We improvised *iQuiz* to present each question on a separate web page. This modification facilitated the collection of fine-grained data, capturing not only overall exam patterns but also detailed interactions at the question level. The process data for each student on every question is captured and stored in a downloadable CSV log file. For this study, we used the following entries of the log file:

- **Student** – User identification number
- **Question** – Numeric variable describing the question number.

Student	Question	Response	Score	Start	Revisions	Time
Student 13	1	option_2	1	["07-02-2023 07:57:02", "07-02-2023 07:57:55"]	[["option_1", "07-02-2023 07:57:49"], ["option_2", "07-02-2023 07:59:01"]]	113
Student 13	2	option_4	0	["07-02-2023 07:57:50", "07-02-2023 07:59:01"]	[["option_4", "07-02-2023 07:59:54"]]	58
Student 13	3	option_1	0	["07-02-2023 07:59:54"]	[["option_1", "07-02-2023 08:01:19"]]	85
Student 13	4	option_1	1	["07-02-2023 08:01:20"]	[["option_1", "07-02-2023 08:01:35"]]	15
Student 13	5	option_2	1	["07-02-2023 08:01:35"]	[["option_2", "07-02-2023 08:02:19"]]	43
Student 13	6	option_1	0	["07-02-2023 08:02:19"]	[["option_1", "07-02-2023 08:02:43"]]	24
Student 13	7	option_4	1	["07-02-2023 08:02:44"]	[["option_4", "07-02-2023 08:02:55"]]	12
Student 13	8	option_4	0	["07-02-2023 08:02:55"]	[["option_4", "07-02-2023 08:03:14"]]	17

Figure 1: A sample dataset used in this study.

- **Response** - Final response (option 1/2/3/4 or unanswered) submitted to the question
- **Score** - Numeric score (0/1) to the question. It is 0 if the response is incorrect/unanswered and 1 if correct.
- **Start** - List of timestamps when the question appears on the student's screen. The students can attempt questions multiple times.
- **Revisions** - List of submitted responses and the respective timestamps of submission to the question. The students can revise the previously submitted response or resubmit the same response multiple times (proofread).
- **Time** - The total time spent (seconds) on the question.

A sample dataset is shown in Fig 1. To enhance the usability and reproducibility of our work, we have open-sourced the iQuiz tool and made the collected data publicly available.

Data collection took place in two different settings: experimental and empirical environments.

In the experimental setting, we created a cheating-induced environment to administer six online assessments. In total, 191 students participated in the assessments conducted in different sessions (approximately 30 students per session). The assessments were optional and non-graded formative assessments designed as practice quizzes. The high participation rate can be attributed to faculty encouragement and the value students perceived in practicing with quiz-based assessments. In each session, some pairs of students were randomly selected and directed to collaborate during the assessment, while the rest attempted the assessment individually. All participants were briefed that the assessment was part of a research study, and no credits would be awarded for the test. Student participation was voluntary, and their informed consent was obtained. Throughout the assessment, students were prohibited from communicating with others except for their selected partners. The assessment was proctored by two instructors. The student pairs that were instructed to collaborate were labelled as colluders in the dataset to be later used for the ML algorithm.

In the empirical setting, we administered a credit-based online assessment. However, unlike the assessments conducted in the experimental setting, no cheating was induced and the dataset obtained was purely empirical. In total, 30 students participated in the assessment. All students were directed to attempt the assessment individually. Upon entering the exam room, students followed the customary practice of choosing their own seats, and the seating positions of all students were recorded. The assessment was weakly proctored with only one invigilator. Students were given clear instructions prohibiting any form of discussion, with specified consequences for violations classified as cheating. Despite these instructions, we presume, based on previous research [46], factors such as academic credits and limited proctoring may serve as motivating factors for students to cheat. The dataset collected under these conditions is unlabeled, as it lacks information about which students engaged in collusion. Following the assessment, a professional psychologist administered a well-known personality test, namely NEO Five-Factor Inventory [19], to evaluate cheating-related personality traits among students. Students were informed about the personality test, and their consent was obtained. Using an offline questionnaire, students assessed the agreeableness and conscientiousness dimensions of their personality on a 5-point scale from very low (1) to very high (5). Agreeableness involves traits related to interpersonal relationships, such as helpfulness and trust, while conscientiousness pertains to being organized and conforming to societal norms [19]. The rationale for focusing on these two traits was due to their negative correlation with academic cheating [20, 44]. The psychologist reviewed and quantified both personality traits for each student. All the thirty students that took the assessment conducted in the empirical setting participated in the personality test.

A total of 4420 records were obtained from 221 students from all seven online assessments (six assessments under experimental and one assessment under empirical setting).

3.2 Feature Engineering

Feature engineering is a process of refining raw data to construct new features that can enhance the predictive performance of ML models [29]. Previous studies primarily focus on features like response similarity and overall exam timelines, for the detection of colluders. These studies are limited to considering only two or three features; leading to reduced accuracy and increased risk of false accusations. We assume that collusion is a multifaceted phenomenon involving nuanced similarities among students, often hidden when analyzing only broad-level features. While conventional features might overlook hidden patterns, question-specific details can offer a more granular perspective. Considering multiple and more detailed aspects of similarity may enhance the accuracy of results and reduce the likelihood of false positives.

In our study, we use feature engineering to derive seven distinct features based on both question and exam-level attributes of the log file (see Table 3). These features are derived through pairwise comparisons of students based on various aspects such as student

Table 3: Feature Information

S.no.	Level	Feature	Code
1	Question	Response Correctness	RC
2		Incorrect Response Similarity	IRS
3		Response Revision Similarity	RRS
4		View TimeStamp	VTS
5		Submit TimeStamp	STS
6	Exam	Time Difference	TD
7		Score Difference	SD

responses, their detailed timelines, revision activity, and scores. The features were developed based on previous research studies related to academic dishonesty and the research and teaching experience of co-authors.

3.2.1 Features. In this section, we describe the calculation of the proposed features. The following mathematical notations are used:

- Let s_i ($i = 1 \ldots N$) represent the list of students and k ($k = 1 \ldots Q$) represent the list of questions.
- Let r_{ik} represent the response and m_{ik} represent the score of the student s_i on question k.
- Let t_{ik} represent the time taken (in seconds) by the student s_i on question k.
- Let ts_{ik}^{view} be the question view timestamp (first view) and ts_{ik}^{sub} be the response submission timestamp (final submission) of the student s_i to question k.
- Let rev_{ik} (1 in case of revision, 0 otherwise) represent the act of revising a question k by a student s_i.
- Let I_i be the total number of incorrect responses from students s_i.

3.2.2 Features at Question Level: **•Response Correctness -** This feature represents the degree of similarity in the score vectors of any two students, s_i and s_j. The score vector is the list of scores across the questions of the assessment. It is calculated as follows:

$$RC_{i,j} = \sum_{k=1}^{Q} \begin{pmatrix} 1, & m_{ik} = m_{jk} \\ 0, & Otherwise \end{pmatrix} / Q \tag{1}$$

•Incorrect Response Similarity - This feature represents the relative frequency of identical incorrect responses of the student pair with their individual incorrect responses. We determine the total number of questions where both students select identical incorrect responses and then calculate the relative frequency. It is calculated as follows:

$$ir_{i,j} = \sum_{\substack{k=1 \\ m_{ik} = m_{jk} = 0}}^{Q} \begin{pmatrix} 1, & r_{ik} = r_{jk} \\ 0, & Otherwise \end{pmatrix} \tag{2}$$

$$IRS_{i,j} = \frac{ir_{i,j}}{2} \left(\frac{1}{I_i} + \frac{1}{I_j} \right) \tag{3}$$

•Response Revision Similarity - This feature represents the number of identical questions where both students performed revisions. It is calculated as follows:

$$RRS_{i,j} = \sum_{k=1}^{Q} \begin{pmatrix} 1, & rev_{ik} = rev_{jk} \\ 0, & Otherwise \end{pmatrix} / Q \tag{4}$$

•View TimeStamp - This feature represents the degree of similarity in the vectors of question view timestamps between any two students. It is computed using a common-sense threshold (th = 20 sec):

$$VTS_{i,j} = \sum_{k=1}^{Q} \begin{pmatrix} 1, & \left| ts_{ik}^{view} - ts_{jk}^{view} \right| < th \\ 0, & Otherwise \end{pmatrix} / Q \tag{5}$$

•Submission TimeStamp - This feature represents the degree of similarity in the vectors of response submission timestamps between any two students. It is calculated as follows:

$$STS_{i,j} = \sum_{k=1}^{Q} \begin{pmatrix} 1, & \left| ts_{ik}^{sub} - ts_{jk}^{sub} \right| < th \\ 0, & Otherwise \end{pmatrix} / Q \tag{6}$$

3.2.3 Features at Exam Level: **•Time Difference** - This feature represents the absolute difference in the total exam duration of any two students. It is calculated as follows:

$$TD_{i,j} = \left| \sum_{k=1}^{Q} t_{ik} - \sum_{k=1}^{Q} t_{jk} \right| \tag{7}$$

•Score Difference - This feature represents the absolute difference in the final scores of any two students. It is calculated as follows:

$$SD_{i,j} = \left| \sum_{k=1}^{Q} m_{ik} - \sum_{k=1}^{Q} m_{jk} \right| \tag{8}$$

3.2.4 Processed Datasets. For every proposed feature f_n ($n = 1$ to 7), we define $M \in R^{NxN}$ that represents a matrix of all student pairs where each $M_{i,j}$ entry represents the respective feature value between any two students s_i and s_j.

$$M_{f_n} = \begin{bmatrix} M_{0,0} & M_{0,1} & \cdots & M_{0,N} \\ M_{1,0} & M_{1,1} & \cdots & M_{0,N} \\ \vdots & \vdots & \cdots & \vdots \\ M_{N,0} & \cdots & \cdots & M_{N,N} \end{bmatrix} \tag{9}$$

These matrices collectively form a final dataset that comprises all possible student pairs, seven independent features and one dependent feature (1 represents collusion, 0 represents honest students). It should be noted that we have considered all possible combinations of student pairs, but to avoid redundancy, we retain only one pair of each symmetric pair, as they represent the same pairing.

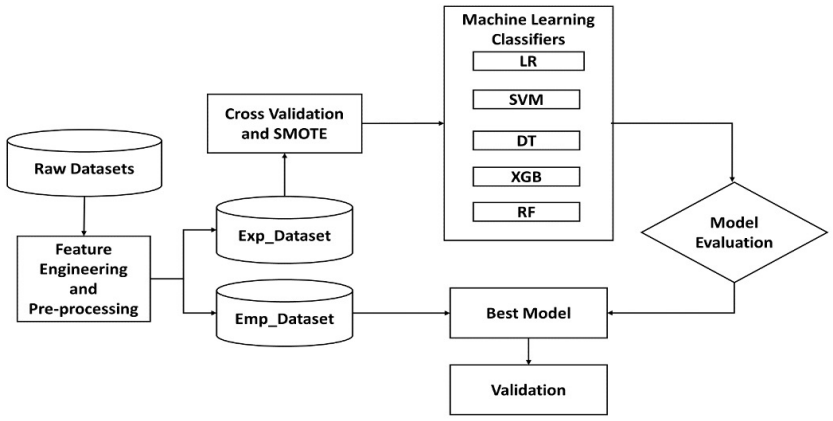

Figure 2: Building the ML model for Collusion Detection

For example, if we consider the pair (s_i, s_j), we have not included its symmetric counterpart (s_j, s_i). We repeat this process for each assessment. The processed datasets from six assessments in the experimental setting are merged into a comprehensive dataset referred to as *Exp_Dataset*. The dataset derived from the empirical setting is referred to as *Emp_Dataset*. The *Exp_Dataset* has a total of 3312 student pairs and the *Emp_Dataset* has a total of 435 student pairs.

3.3 Methods

In this section, we describe the ML model for collusion detection (Fig. 2). The detection of collusion among student pairs is a binary classification problem that involves categorizing each student pair into one of two classes: colluders or honest students. This classification task is performed based on the features derived through the feature engineering process. We used the datasets obtained in the previous section to construct the ML model. Both datasets underwent a pre-processing step to identify and address missing values. Additionally, standard scaling was applied to ensure a consistent scale across features in both datasets. The *Exp_Dataset* was employed for training and testing the ML model. Subsequently, the trained model was validated using the *Emp_Dataset*. The *Exp_Dataset* dataset is highly imbalanced (3245 honest student pairs: 67 colluding pairs), and can pose challenges for ML models. The data appears imbalanced because it is based on student pairs rather than individual students. For example, if we have four students—W, X, Y and Z—and only Y and Z collaborate, this results in just one colluding pair (Y, Z), while the other pairs (X, Y), (X, Z), (X, W), (Y, W), and (Z, W) are considered honest. Consequently, this setup leads to a higher number of honest pairs compared to colluding pairs, creating an imbalanced dataset. Using an imbalanced dataset without proper handling can lead to a biased model that performs poorly on the minority class. To address this issue, the Synthetic Minority Oversampling Technique (SMOTE) is used [45]. SMOTE is a popular data balancing technique that creates synthetic samples of the minority class.

In this study, five ML classifiers - Logistic Regression (LR), Support Vector Machine (SVM), Decision Trees (DT), Extreme Gradient Boost (XGB) and Random Forest (RF), are taken into consideration [42]. The rationale for choosing these classifiers is because they are commonly used for binary classification, and have shown to perform well on similar datasets [18].

Stratified five-fold cross-validation was used to estimate the performance of the model [3]. To achieve unbiased estimation, we apply SMOTE only across the training dataset, in every cross-validation fold separately. We use accuracy, precision, recall and F1-score for comparing the performance of classifiers. We also account for imbalanced data sets by using macro-averaging technique with all the metrics. The macro-averaging technique gives equal weight to each class, regardless of its size. It involves calculating the metric for each class separately, and then taking the average of these metrics across all classes.

4 Results

In this section, we present the results obtained using the proposed approach. The ML models were implemented using the *scikit-learn* library, and the coding was performed on *Google Collaboratory* using Python 3.10.12.

4.1 Results of Experimental Data

4.1.1 Performance of Classifiers. We compare the performance of five classifiers – LR, DT, SVM, RF and XGB, across four evaluation metrics (accuracy, precision, recall, and F1 score). Fig. 3 shows the performance comparison across different classifiers. We observe that both RF and XGB classifiers achieve the highest accuracy (98.8%), followed closely by DT (97.4%). The lowest accuracy is achieved by LR (93.8%). We find that RF has the highest precision (88.1%), followed by XGB (87.3%). The lowest precision is achieved by LR (61.8%). It is observed that LR has the highest recall (91%), followed by SVM (87.3%). The lowest recall is achieved by DT classifier (77.6%). We observe that RF has the highest F1 score (82.9%), followed closely by XGB (82.6%). The lowest F1 score is achieved by LR (67%). The comparative analysis of the results reveals that

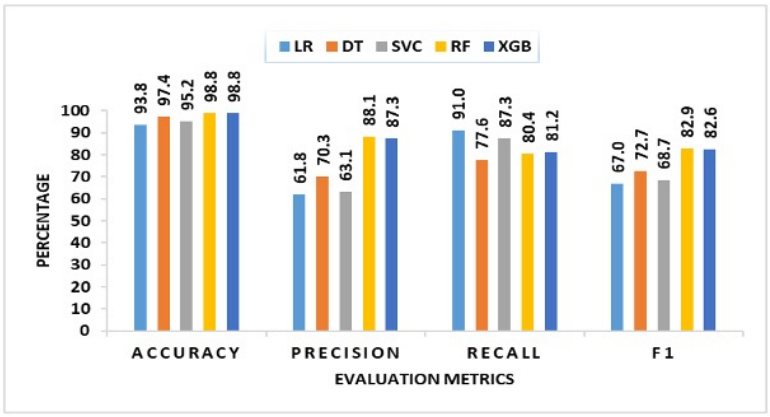

Figure 3: Performance Comparison of Classifiers

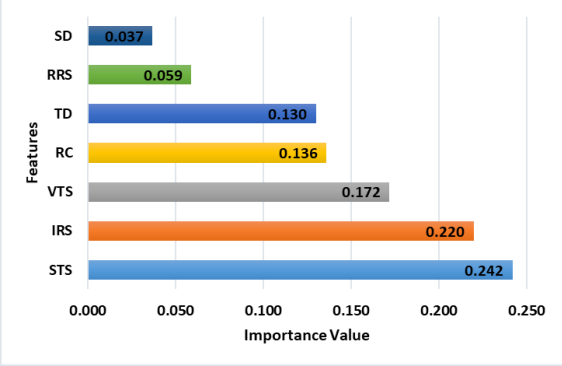

Figure 4: Feature Importance using RF classifier.

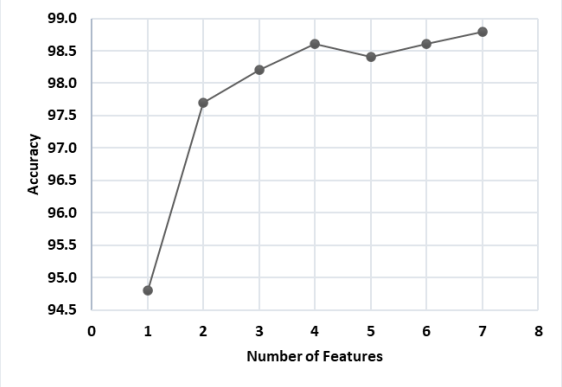

Figure 5: Model Accuracy corresponding to number of features

the RF classifier, in general, performs better than the rest of the classifiers. Therefore, we consider RF as the best classifier for our research problem.

4.1.2 Feature Importance. Feature importance refers to the process of determining the contribution of each feature in predicting the dependent variable. It helps to understand which features have the most impact on the model performance and provide insights into the underlying relationships between the features and the dependent variable.

We used the *rf.feature_importances* property of Random Forest classifier for this purpose. The *rf.feature_importances* returns an array of importance scores for each feature in the dataset. Figure 4 presents the feature importance values. It is noted that STS and IRS are the two most important features in the classifier, with importance scores of 0.242 and 0.220, respectively. The VTS, RC and TD features have scores above 0.1, indicating that they are also important for the ML model. The RRS and SD features have the lowest importance score of 0.059 and 0.037 respectively, suggesting that they may be less important than the other features.

We observed that while some features showed limited contribution individually, their collective inclusion, alongside other features, enhanced the overall performance of the model. Fig. 5. shows the model accuracy corresponding to the number of features. The graph was generated by iteratively incorporating one feature at a time, prioritizing them based on their decreasing importance as obtained in Fig. 4. We observed that the maximum accuracy was achieved when utilizing the complete feature set. Therefore, we consider the full feature set for our research problem.

4.2 Results of Empirical Data

We validated the trained ML model (RF classifier with seven features) on the *Emp_Dataset*. A total of 30 students were assessed by the model, out of which 11 students were predicted as colluders, eight of them colluded in pairs and three students colluded in a trio. It is to be noted that the trio of students was identified by analyzing the common students in the predicted pairs (e.g., {19, 20}, {20, 21} and {19, 21} form a trio {19, 20, 21}). As this dataset was unlabeled, we used the following indicators to support the model predictions.

First, all the predicted colluding students were found to be row-wise seat neighbors during the assessment. The exam room setup and the predicted honest and colluding students are shown in Fig. 6. The seating pattern of students was of the form 'student–empty seat–student'. Given the room setup, colluding with row-wise

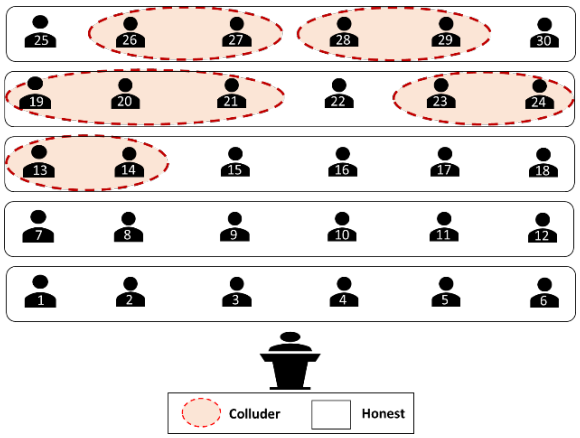

Figure 6: The exam-room setup and the model predictions

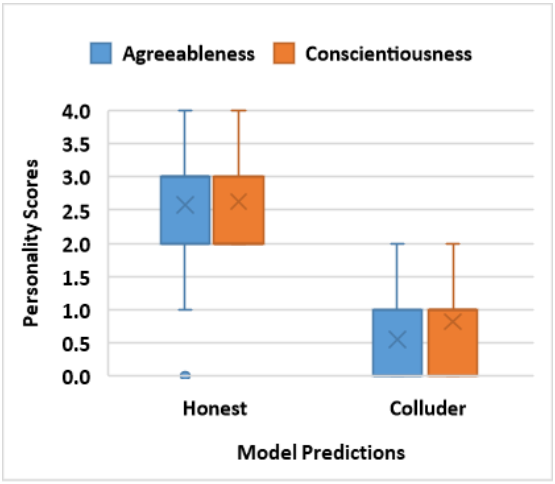

Figure 7: Personality scores corresponding to model predictions

neighbors was easier compared to front-back neighbors. The spatial patterns are self-evident showing that these groups of students can be working together.

Second, all the predicted colluders congregated near the back of the exam room. We assume that under weak baseline monitoring, sitting at the back of the exam room can provide opportunities to discreetly collude. Previous research studies [31, 43] supports this hypothesis that collusion activities are frequent among students who generally sit in the back seats. As the students selected their seats voluntarily, this hypothesis is highly possible.

Third, the students identified as potential colluders exhibited notably low scores in both agreeableness and conscientiousness dimensions assessed in the personality test (see Fig. 7, the "X" in the box plot denotes the median). Prior research [20, 44] has consistently demonstrated a correlation between lower scores in these traits and increased likelihood of academic dishonesty. This alignment with the predictions of our ML model further strengthens the model predictions.

Fourth, the suspected colluders exhibit a high degree of response and temporal similarity. These correlating patterns cannot be attributed to chance, otherwise, the predicted honest students should have captured these effects. This is a positive confirmation that our approach is working correctly.

5 Discussion and Conclusion

In this paper, we focused on collusion, a common cheating practice in online assessments. We used feature engineering and ML algorithms to create a model for the detection of collusion. We further validated the model on an unlabeled empirical dataset, whereby the model predictions were supported through spatial data analysis and a recognized personality test. The positive performance on both experimental and empirical datasets shows the general applicability of our approach.

The feature analysis also presents some useful insights about the test-taking pattern of students. We find that STS and VTS features were among the most relevant features of our ML model (Fig. 4). The box plot distribution in Fig. 8 shows that the colluding student pairs exhibit high values of STS and VTS, while the honest student pairs show very low values. This describes the correlation in the navigational patterns of colluders, confirming that they systematically move ahead with the assessments; synchronously view the questions and submit their responses. To avoid this behavior, we support randomization in the order of appearance of questions. This strategy can help limit collusion by disrupting the synchronized movement of colluders as they are not likely to get the same questions in the same order.

Other relevant features were IRS and RC (Fig. 4). We observe that the values of IRS and RC are high for colluders and low for honest student pairs (Fig. 8). This shows that colluding pairs were not only answering the same questions incorrectly but also, selecting the same incorrect response for those questions. We encourage the idea of implementing randomized response options for each question. This can possibly make it difficult for students to share answers as the response options will vary.

Other features, namely RRS, TD, and SD, exhibited limited impact on the model. We observe that due to the strictly timed nature of the assessment, instances of students revising their answers were relatively low. Despite that, we noted that colluding pairs displayed a correlated pattern of revisions on identical questions. Furthermore, we observed that colluding pairs exhibited similarities in exam duration and total score, in contrast to the significant variation observed in honest student pairs (Fig. 8).

We further evaluated the performance of various feature sets proposed in previous studies. The existing studies have employed different methods, including machine learning techniques or statistical analysis methods. To ensure fair comparisons, we assessed their features using the same classifier (RF) and evaluating them on the same dataset used in our study. This approach helps eliminate potential variations introduced by using different classifiers or datasets. Table 4 displays the accuracy and F1 score obtained from these evaluations. The results consistently demonstrate that the best performance is achieved when utilizing our comprehensive feature set. This finding reinforces our hypothesis regarding the

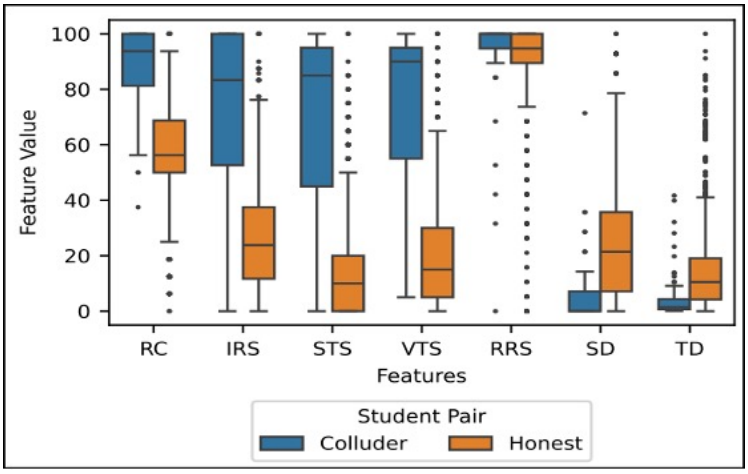

Figure 8: Box plot visualization of features

Table 4: Performance Comparison of Feature sets proposed in the Previous Studies

Paper	Original-Method	Features	Accuracy	F1
[35]	Statistics	Response	96.6%	72.0%
[13]	ML			
[36]	Statistics	Exam submission time	92.8%	56.7%
[24]	Statistics	Response, Exam submission time	96.1%	67.8%
[39]	ML	Exam score, Question submission time	98.1%	78.2%
[5]	Statistics	Exam starting and submission time	93.3%	56.6%
[6]	Statistics	Exam submission time, Exam score	92.7%	57.4%
[22]	Statistics	Response, Exam starting and submission time	97.3%	72.3%
[28]	Statistics	Response, Exam starting and submission time, Question submission time	98.4%	79.9%
Our approach	ML	Response Correctness, Incorrect Response Similarity, Revision Similarity, Question View Time, Question Submit Time, Exam Time Difference, Score Difference	98.8%	82.9%

significance of considering multiple facets of similarity to enhance collusion detection accuracy.

We also observed that the question-level features are more influential as compared to exam-level features for the detection of collusion. Most of the online quiz platforms collect basic data like scores and responses, but only few collect detailed clickstream data up to the question-specific level of the assessment. We recommend the improvisation of online quiz platforms for the collection and analysis of comprehensive clickstream data of students. To obtain question-level data, it is required for the platform to display one question per page. This feature can also be useful in preventing collusion where students take a screenshot of all questions and share them with others. Strategies like implementing a time limit for each question such that a student who knows the information has appropriate time to answer, and not too much time to help or collaborate with others, can also be used. We encourage platform providers and teachers to incorporate the above-discussed functionalities for the conduct of secure online assessments. While these techniques may not completely eliminate the potential for collusion, they can provide additional safeguards to prevent collusion.

We align our findings with existing research [19], highlighting a correlation between academic cheating and the personality traits of agreeableness and conscientiousness. The low values of these traits suggest a reduced concern for conflicts and a higher propensity for circumventing rules in the pursuit of success. Our model predictions identified colluders with low values in these traits. We recommend implementing value-based training programs that help individuals understand the importance of ethical behavior. For example, mindfulness training programs, emphasizing self-discipline, have demonstrated a positive impact on ethical decision-making [25]. Importantly, mindfulness is a skill that can be learned and developed through practice.

Despite the valuable insights, there are some limitations in the current study. The first limitation is the use of experimental datasets for training the ML model. However, unlike pure simulation studies, our approach was a near representation of the original scenario as the data was collected with real students. We further showed the generalizability of our approach with an empirical dataset, where we confirmed model predictions using a recognized personality test and spatial data analysis. This sets our study apart from previous

research that often relies on methods like student interviews to validate model predictions, which can be biased as students rarely confess to academic misconduct. Another limitation involves the potential for coincidental observations or false accusations. In contrast to existing studies that utilize only a few broad-level features, we incorporated seven features at both exam and question levels, to mitigate the risk of misinterpretation between collusion and coincidence. Nonetheless, we recommend that our approach is best suited for identifying suspicious cases that can be used as a filter for instructor review who may then make informed decisions. The current study can be integrated with other cheating mitigation strategies, such as eye-gaze tracking, audio recording, keystroke dynamics, etc., to enhance the overall detection process.

With the continued growth of technology and online education, the use of online assessments will likely become even more prevalent in higher education. It is essential to maintain the fairness of assessments to preserve their credibility. We presented an approach for the detection of a commonly observed academic dishonesty - In-Person collusion in online assessments. The approach helped in building a ML model that facilitates automated detection of students involved in collusion. The analysis of features helped in identifying various strategies for designing quiz platforms and online assessments that proactively eliminate opportunities for collusion. This study aligns with the recommendations of ADMP, which emphasizes addressing cheating through the combination of prevention and detection techniques. Ultimately, the future of education depends on our ability to uphold the highest standards of academic integrity.

Data Availability The data that support the findings of this study are available at https://doi.org/10.6084/m9.figshare.23744283.v1

References

[1] Alexandr Akimov and Mirela Malin. 2020. When old becomes new: a case study of oral examination as an online assessment tool. *Assessment and Evaluation in Higher Education* 45, 8 (2020), 1205–1221. https://doi.org/10.1080/02602938.2020.1730301

[2] Giora Alexandron, José A. Ruipérez-Valiente, and David E. Pritchard. 2019. Towards a general purpose anomaly detection method to identify cheaters in massive open online courses. *EDM 2019 - Proceedings of the 12th International Conference on Educational Data Mining* (2019), 480–483. https://doi.org/10.35542/osf.io/wuqv5

[3] Md L. Ali, Kutub Thakur, and Muath A. Obaidat. 2022. A Hybrid Method for Keystroke Biometric User Identification. *Electronics (Switzerland)* 11, 17 (2022). https://doi.org/10.3390/electronics11172782

[4] Yousef Atoum, Liping Chen, Alex X. Liu, Stephen D.H. Hsu, and Xiaoming Liu. 2017. Automated Online Exam Proctoring. *IEEE Transactions on Multimedia* 19, 7 (2017), 1609–1624. https://doi.org/10.1109/TMM.2017.2656064

[5] Antonio Balderas and Juan Antonio Caballero-Hernández. 2020. Analysis of Learning Records to Detect Student Cheating on Online Exams: Case Study during COVID-19 Pandemic. *ACM International Conference Proceeding Series* (2020), 752–757. https://doi.org/10.1145/3434780.3436662

[6] Antonio Balderas, Manuel Palomo-Duarte, Juan Antonio Caballero-Hernández, Mercedes Rodriguez-Garcia, and Juan Manuel Dodero. 2021. Learning Analytics to Detect Evidence of Fraudulent Behaviour in Online Examinations. *International Journal of Interactive Multimedia and Artificial Intelligence* 7, 2 (2021), 241. https://doi.org/10.9781/ijimai.2021.10.007

[7] Gregory J. Cizek and James A. Wollack. 2016. *Handbook of Quantitative Methods for Detecting Cheating on Tests*. Routledge. https://doi.org/10.4324/9781315743097

[8] Catherine Cleophas, Christoph Hönnige, Frank Meisel, and Philipp Meyer. 2023. Who's Cheating? Mining Patterns of Collusion from Text and Events in Online Exams. *INFORMS Transactions on Education* 23, 2 (January 2023), 84–94. https://doi.org/10.1287/ited.2021.0260

[9] Jamie Costley. 2019. Student Perceptions of Academic Dishonesty at a Cyber-University in South Korea. *Journal of Academic Ethics* 17, 2 (2019), 205–217. https://doi.org/10.1007/s10805-018-9318-1

[10] Charles Crook and Elizabeth Nixon. 2019. The social anatomy of 'collusion.' *British Educational Research Journal* 45, 2 (2019), 388–406. https://doi.org/10.1002/berj.3504

[11] Charles Crook and Elizabeth Nixon. 2021. How internet essay mill websites portray the student experience of higher education. *The Internet and Higher Education* 48, (January 2021), 100775. https://doi.org/10.1016/j.iheduc.2020.100775

[12] Jiameng Du, Yifan Song, Mingxiao An, Marshall An, Christopher Bogart, and Majd Sakr. 2022. Cheating Detection in Online Assessments via Timeline Analysis. *SIGCSE 2022 - Proceedings of the 53rd ACM Technical Symposium on Computer Science Education* 1, (2022), 98–104. https://doi.org/10.1145/3478431.3499368

[13] Carol Eckerly. 2021. Answer Similarity Analysis at the Group Level. *Applied Psychological Measurement* 45, 5 (July 2021), 299–314. https://doi.org/10.1177/01466216211013109

[14] A. Ercole, K. D. Whittlestone, D. G. Melvin, and J. Rashbass. 2002. Collusion detection in multiple choice examinations. *Medical Education* 36, 2 (2002), 166–172. https://doi.org/10.1046/j.1365-2923.2002.01068.x

[15] Kelum A.A. Gamage, Erandika K. de Silva, and Nanda Gunawardhana. 2020. Online delivery and assessment during COVID-19: Safeguarding academic integrity. *Education Sciences* 10, 11 (2020), 1–24. https://doi.org/10.3390/educsci10110301

[16] Manika Garg and Anita Goel. 2021. A Data-Driven Approach for Peer Recommendation to Reduce Dropouts in MOOC. *Lecture Notes in Electrical Engineering* 735 LNEE, (2021), 217–229. https://doi.org/10.1007/978-981-33-6977-1_18

[17] Manika Garg and Anita Goel. 2022. A systematic literature review on online assessment security: Current challenges and integrity strategies. *Computers and Security* 113, (2022), 102544. https://doi.org/10.1016/j.cose.2021.102544

[18] Manika Garg and Anita Goel. 2022. Preserving Integrity in Online Assessment using Feature Engineering and Machine Learning. (2022).

[19] Tamara L. Giluk and Bennett E. Postlethwaite. 2015. Big Five personality and academic dishonesty: A meta-analytic review. *Personality and Individual Differences* 72, (2015), 59–67. https://doi.org/10.1016/j.paid.2014.08.027

[20] William G. Graziano, Lauri A. Jensen-Campbell, and Elizabeth C. Hair. 1996. Perceiving interpersonal conflict and reacting to it: The case for agreeableness. *Journal of Personality and Social Psychology* 70, 4 (1996), 820–835. https://doi.org/10.1037//0022-3514.70.4.820

[21] Arto Hellas, Juho Leinonen, and Petri Ihantola. 2017. Plagiarism in Take-home Exams. In *Proceedings of the 2017 ACM Conference on Innovation and Technology in Computer Science Education*, June 28, 2017. ACM, New York, NY, USA, 238–243. https://doi.org/10.1145/3059009.3059065

[22] Arto Hellas, Juho Leinonen, and Petri Ihantola. 2017. Plagiarism in Take-home Exams. In *Proceedings of the 2017 ACM Conference on Innovation and Technology in Computer Science Education*, June 28, 2017. ACM, New York, NY, USA, 238–243. https://doi.org/10.1145/3059009.3059065

[23] Kenrie Hylton, Yair Levy, and Laurie P. Dringus. 2016. Utilizing webcam-based proctoring to deter misconduct in online exams. *Computers and Education* 92–93, (2016), 53–63. https://doi.org/10.1016/j.compedu.2015.10.002

[24] Daniel Jaramillo-Morillo, José Ruipérez-Valiente, Mario F. Sarasty, and Gustavo Ramírez-Gonzalez. 2020. Identifying and characterizing students suspected of academic dishonesty in SPOCs for credit through learning analytics. *International Journal of Educational Technology in Higher Education* 17, 1 (December 2020), 45. https://doi.org/10.1186/s41239-020-00221-2

[25] Yaprak Kalafatoğlu and Tülay Turgut. 2017. Another benefit of mindfulness: Ethical behaviour. *International Journal of Social Sciences and Education Research* 3, 3 (2017), 772–772. https://doi.org/10.24289/ijsser.311367

[26] Sathiamoorthy Manoharan. 2017. Personalized Assessment as a Means to Mitigate Plagiarism. *IEEE Transactions on Education* 60, 2 (May 2017), 112–119. https://doi.org/10.1109/TE.2016.2604210

[27] Tony Mason, Ada Gavrilovska, and David A. Joyner. 2019. Collaboration Versus Cheating. (2019), 1004–1010. https://doi.org/10.1145/3287324.3287443

[28] Riccardo Mazza. 2021. A visual method to identify and characterise students suspected of collaboration during remote quizzes submissions in Learning Environments. In *2021 25th International Conference Information Visualisation (IV)*, July 13, 2021. IEEE, 255–260. https://doi.org/10.1109/IV53921.2021.00048

[29] Fatemeh Nargesian, Horst Samulowitz, Udayan Khurana, Elias B. Khalil, and Deepak Turaga. 2017. Learning feature engineering for classification. *IJCAI International Joint Conference on Artificial Intelligence* 0, August (2017), 2529–2535. https://doi.org/10.24963/ijcai.2017/352

[30] Lisa-Maria Norz, Verena Dornauer, Werner O. Hackl, and Elske Ammenwerth. 2023. Measuring social presence in online-based learning: An exploratory path analysis using log data and social network analysis. *The Internet and Higher Education* 56, (January 2023), 100894. https://doi.org/10.1016/j.iheduc.2022.100894

[31] Douglas Attoh Odongo, Eric Agyemang, and John Boulard Forkuor. 2021. Innovative Approaches to Cheating: An Exploration of Examination Cheating Techniques among Tertiary Students. *Education Research International* 2021, (March 2021), 1–7. https://doi.org/10.1155/2021/6639429

[32] Vidya Raman and Shaun Ramlogan. 2020. Academic integrity and the implementation of the honour code in the clinical training of undergraduate dental students. *International Journal for Educational Integrity* 16, 1 (2020), 1–20. https://doi.org/10.1007/s40979-020-00058-2

[33] Jochen Ranger, Nico Schmidt, and Anett Wolgast. 2020. The Detection of Cheating on E-Exams in Higher Education—The Performance of Several Old and Some New Indicators. *Frontiers in Psychology* 11, October (2020), 1–16. https://doi.org/10.3389/fpsyg.2020.568825

[34] Jake Renzella, Andrew Cain, and Jean Guy Schneider. 2022. Verifying student identity in oral assessments with deep speaker. *Computers and Education: Artificial Intelligence* 3, July 2021 (2022), 100044. https://doi.org/10.1016/j.caeai.2021.100044

[35] Peter Richmond and Bertrand M. Roehner. 2015. The detection of cheating in multiple choice examinations. *Physica A: Statistical Mechanics and its Applications* 436, (October 2015), 418–429. https://doi.org/10.1016/j.physa.2015.05.040

[36] José A. Ruipérez-Valiente, Dragan Gašević, Srećko Joksimović, Vitomir Kovanović, Pedro J. Muñoz-Merino, and Carlos Delgado Kloos. 2017. A data-driven method for the detection of close submitters in online learning environments. *26th International World Wide Web Conference 2017, WWW 2017 Companion* (2017), 361–368. https://doi.org/10.1145/3041021.3054161

[37] José A. Ruipérez-Valiente, Daniel Jaramillo-Morillo, Srećko Joksimović, Vitomir Kovanović, Pedro J. Muñoz-Merino, and Dragan Gašević. 2021. Data-driven detection and characterization of communities of accounts collaborating in MOOCs. *Future Generation Computer Systems* 125, (2021), 590–603. https://doi.org/10.1016/j.future.2021.07.003

[38] Yousef W. Sabbah. 2017. Security of Online Examinations. . 157–200. https://doi.org/10.1007/978-3-319-59439-2_6

[39] Vicenzo Abichequer Sangalli, Gonzalo Martinez-Munoz, and Estrella Pulido Canabate. 2020. Identifying cheating users in online courses. *IEEE Global Engineering Education Conference, EDUCON* 2020-April, (2020), 1168–1175. https://doi.org/10.1109/EDUCON45650.2020.9125252

[40] Sumit Sarkar. 2022. Collaborative cheating and group confession: Findings from a natural experiment. *Cogent Social Sciences* 8, 1 (December 2022). https://doi.org/10.1080/23311886.2022.2069909

[41] Brenda M. Stoesz and Anastassiya Yudintseva. 2018. Effectiveness of tutorials for promoting educational integrity: A synthesis paper. *International Journal for Educational Integrity* 14, 1 (2018). https://doi.org/10.1007/s40979-018-0030-0

[42] Nikola Tomasevic, Nikola Gvozdenovic, and Sanja Vranes. 2020. An overview and comparison of supervised data mining techniques for student exam performance prediction. *Computers & Education* 143, (January 2020), 103676. https://doi.org/10.1016/j.compedu.2019.103676

[43] Alexandru Topîrceanu. 2017. Breaking up friendships in exams: A case study for minimizing student cheating in higher education using social network analysis. *Computers & Education* 115, (December 2017), 171–187. https://doi.org/10.1016/j.compedu.2017.08.008

[44] Kevin M. Williams, Craig Nathanson, and Delroy L. Paulhus. 2010. Identifying and profiling scholastic cheaters: Their personality, cognitive ability, and motivation. *Journal of Experimental Psychology: Applied* 16, 3 (2010), 293–307. https://doi.org/10.1037/a0020773

[45] Marcin Woźniak, MichałWieczorek, and Jakub Siłka. 2023. BiLSTM deep neural network model for imbalanced medical data of IoT systems. *Future Generation Computer Systems* 141, (April 2023), 489–499. https://doi.org/10.1016/j.future.2022.12.004

[46] Hongwei Yu, Perry L. Glanzer, Byron R. Johnson, Rishi Sriram, and Brandon Moore. 2018. Why college students cheat: A conceptual model of five factors. *Review of Higher Education* 41, 4 (2018), 549–576. https://doi.org/10.1353/rhe.2018.0025

[47] iQuiz. Retrieved from https://github.com/anitagoel/iQuiz