Cheating Detection in online exams using machine learning

كشف الغش في الامتحانات عبر الإنترنت باستخدام التعلم الآلي

Rehab Kareem kadthim **Zuhair Hussein Ali**

Department of Computer Science - College of Education

MustansiriyahUniversity

Abstract: Cheating reportedly affects online exams and might easily violate the educational process by providing an unfair competitive advantage to one student over others. Accordingly, several efforts have been made in the past years to find reliable and scalable approaches to solve this problem. Unfortunately, cheating behaviors are rather difficult to detect and existing approaches generally require human supervision. This paper proposes a model to automatically detect cheating in the final exam. The proposed model consists of two main phases. The first phase concern with prediction of students score using multiple linear regression. The second phase concern with applying three different classifier algorithms: Support Vector Machine (SVM), Random Forest (RF), and K-nearest neighbors (KNN) to classify students as a cheater or not. The experimental results show the good performance of the proposed method, especially for SVM when the accuracy reached 96%.

Keywords: online cheating, E-learning, Random forest, SVM, KNN.

الخلاصة

إن الغش في الامتحانات يؤثر على العملية التعليمية بسهولة من خلال توفير ميزة تنافسية غير عادلة لطالب على غيره. وبناءً على ذلك ، تم بذل العديد من الجهود في السنوات الماضية للعثور على مناهج موثوقة وقابلة للتطوير لحل هذه المشكلة. لسوء الحظ ، يصعب اكتشاف سلوكيات الغش ، وتتطلب الأساليب الحالية عمومًا إشرافًا بشريًا. تقترح هذه الورقة نموذجًا لاكتشاف الغش تلقائيًا في الاختبار النهائي. يتكون النموذج المقترح من مرحلتين رئيسيتين. تتعلق المرحلة الأولى بالتنبؤ بنتائج الطلاب باستخدام الانحدار الخطي المتعدد. تتعلق المرحلة الثانية بتطبيق ثلاث خوارزميات مختلفة للمصنفات: Support Vector Machine (SVM) و Random Forest (RF) وK-neatest neighbors لتصنيف الطلاب على أنهم غشاشون أم لا. أظهرت النتائج الت جريبية الأداء الجيد للطريقة المقترحة ، خاصة للطريقة SVM عندما وصلت الدقة إلى ٩٦٪.

1.INTRODUCTION

In the context of online examination, cheating refers to using artificial systems to gain a competitive advantage over other students. Unluckily, cheating is a widespread phenomenon among in online exams many students (especially after covid-19) that affects negatively the educational and academic reality. Accordingly, several efforts have been made in the past years to

find reliable and scalable solutions to this problem. Unfortunately, cheating behaviors are rather difficult to detect. Thus, the most successful and used approaches still heavily rely on Emerging Studies as well as on the monitoring activity of the competent authorities. Therefore, there must be ways to prevent or at least detect cheating. Various methods exist for the online cheating detection some of these are direct while others are indirect [1]. Direct detection required the existence of physical equipment, such as the use of online proctoring with a webcam or a camera designed with special techniques [2]. Also, the use of biometric methods, for example, the analysis of eye tracks or fingerprints [3]. Indirect methods try to prevent or detect cheating with less cost, such as analyzing or formulating the questions in a way to reduce cheating attempts [4].

While Machine Learning has become a cornerstone of information technology and, with it, a pretty essential, albeit sometimes hidden, part of our lives during the last two decades. With the availability of ever-increasing volumes of data, there is reason to expect that smart data analysis will become increasingly more prevalent as a crucial component of technological advancement.

This paper introduced an approach to automatically detect cheating in the final exams based on previous students grades. The proposed method consists of two main phases. The MLR was used in the first phase to predict a student's degree. While the second phase based on using three different classifiers SVM, RF, and KNN to classify each student as a cheater or not.

2.Related work

Recently, researchers have made many efforts to develop a realistic approach to identify ways to detect cheating in electronic exams. Studies are still expanding to provide the aspect of monitoring cheating cases. The past few decades present a set of studies related to approach.

Cavalcanti et al., 2012 [5] used a decision tree to build two classification models to detect cheating: one based on the cosine similarity, and the other based on the overlap coefficient. The results of both classifiers were compared against the results produced by a doexpert. The results proved that the decision tree with overlap coefficient is better to use for the purposed of cheating detection.

Li, 2019, [6], introduces a novel probabilistic strategy for detecting cheaters. The proposed method is based on analyzing the question-answering record and discovering that the incorrect answers have an extremely low chance under the condition of passing the test. The proposed model combines both probability estimation and feature bagging. Probability estimation is used to detect outlier detection, where every answer time exceeds an hour or the number of answers less than one hundred.

Golden et al., 2021 [7], used a case study to assess the prevalence of potential cheating and propose preventive measures that may be implemented. They utilized cheating intelligence agents as a tool for identifying online cheating behaviors, which is made up of two key modules: IP detector and behavior detector. The intelligence agent observes the students' behavior and can prevent and detect any malicious conduct. It can be used to give randomized multiple-choice questions in a course examination and can be combined with online learning programs to monitor students' behavior. The proposed method was validated by testing it on a variety of datasets (mid-term and final-term exams).

3.Proposed methodology for cheating detecting

The proposed model for cheating detection based on two main phases: MLR and machine learning algorithms. The MLR phase is used to predict the final student score, based on many exam scores. Where there are five exams before the final exam arranged in the following sequence (First Quiz (Q1), Second Quiz (Q2), Midterm (M), Third Quiz (Q3), and Fourth Quiz (Q4)). The output of this phase helps to put (0) or (1) labels for each student record. Where (0) refers to not cheat and (1) means cheating. The second phase is based on using different machine learning algorithms to classify students as either cheaters or not.

3.1 Prediction Phase using multiple linear regression

In this phase, each student's scores are entered, and the mathematical equation of the MLR approach is implemented to determine the weight of each exam. That helps in turn to predict the final exam score. Figure (1) shows the main steps of this phase. MLR is used to establish the relation between dependent and independent variables. There is one dependent variable, whereas the independent variables must be more than one.

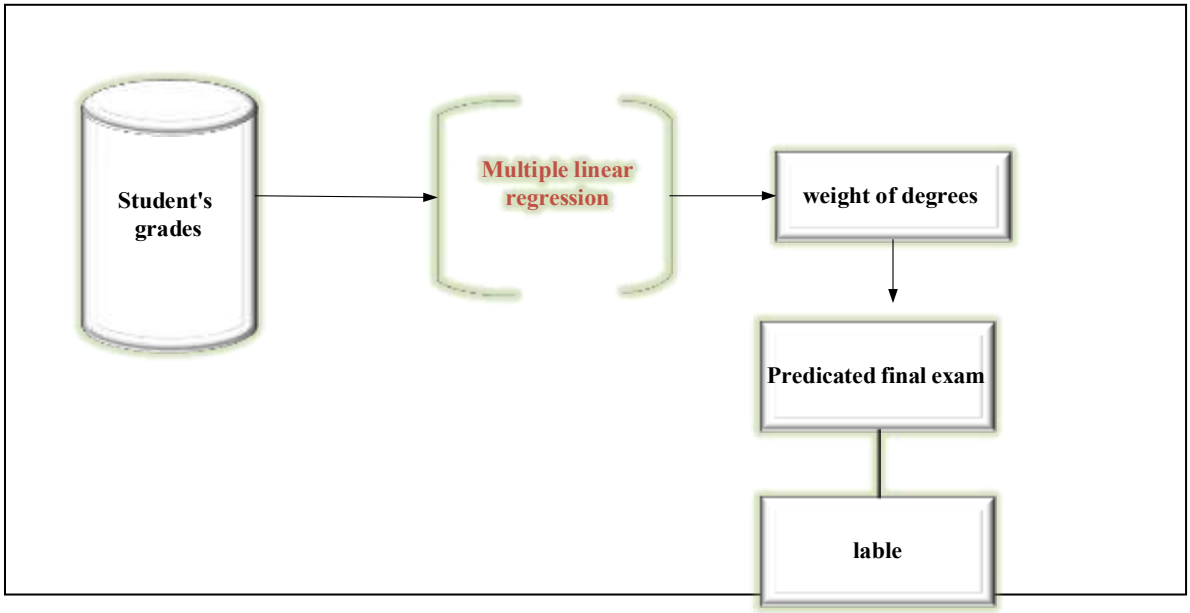

Figure (1): MLR Phase

The general form of MLR is represented in Eq.(1).

$$Y=W_0+W_1Q1+W_2Q2+W_3M+W_4Q3+W_5Q4$$

Where

W_i: represent the exam weight

Q_1, M: represents student quizzes and midterm exam respectively.

3.2 Machine Learning Algorithms for Prediction

In this subsection, a brief descrption of the three machine learning algorithms was given.

3.2.1 Random forest

Random Forest is a combination of a series of tree structure classifiers. It's also the most customizable and user-friendly algorithm. A forest is made up of trees. The strength of a forest is supposed to be proportionate to the number of trees in it. The random forest includes several decision trees in various subsets of a particular dataset. To increase predicting accuracy from that dataset, the average is used [9]. Every tree has a divide prediction by category and the class with the most votes become the subtree.

3.3.2 K-Nearest Neighbor's Algorithm

K-Nearest Neighbors is a one of the simplest techniques that gives very good results. It is a non-parametric and instance-based learning algorithm. KNN is used in classification mode to obtain the class, of the new unlabeled data belongs. Such that, several neighbors' elements are gathered as a 'k' and will be decided generally odd. The distance between the object and the nearest data points is calculated by the methods like Euclidian's distance, Hamming distance, Manhattan distance or Makowski distance. The following Figure (2) illustrates an example of binary classification.

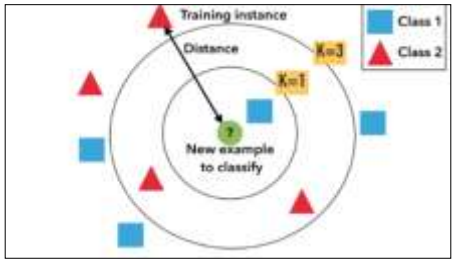

Figure (2): KNN with Two Classes [10].

3.3.3 Support Vector Machine

Support Vector Machine (SVM) is a supervised learning technique that can be used to classify the training data efficiently into two categories. learning the SVM is very time-consuming for a large scale of data. Select these hyperplanes so that there are no points between them and then try on maximizing the distance between the hyperplane. We have found the distance between the hyperplane as 2/|x| as shown in Figure (3). To minimize, we need to ensure for all (i) either using Equation (3) [11].

$$w.x_i - b \geq 1 \text{ or } w.x_i - b \leq -1 \qquad (3)$$

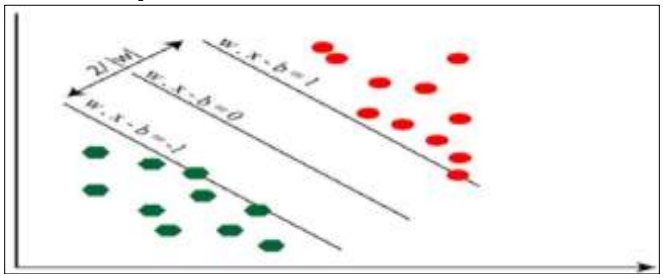

Figure (3): SVM with Samples from Two Classes.

4. Result and Discussion

In this section, The performance evaluation of the proposed method is presented and discussed.

4.1 Dataset collection

Due to the difficulty of obtaining actual exam scores for students, this research uses an idea of a hypothetical dataset for a group of students. The basic idea of creating the dataset is based on the paper [12]. The dataset consists of student course assessments, including quizzes, the midterm exam, and the final exam. The assumption is based on detecting abnormal scores by comparing the results before the final exam to the final exam score. A student with a substantial difference between his or her two scores is more likely to have cheated. The dataset consists of 110 students' records. Each student record includes four quizzes, a midterm exam, and a final exam.

4.2 Result of multiple leaner regression phases

The MLR basically used to assign labels to the student record. This label is either cheater or not cheater. This process is very important before starting the machine learning technique. The MLR based on computing weight for each exam. These weights with exams scores help to predict the final score. Table (1) show the weight of each exam.

Table (1): Exams Weight

Features	Weights of Feature
Quiz1	0.15580391
Quiz2	0.13213351
M	0.41318684
Quiz3	0.12006481
Quiz4	0.12585633

4.3 RESULT OF machine learning algorithms Experiment 1

This experiment evaluates the performance of the system classifiers (SVM, KNN, and Rf) where the difference between the predicted score and the final score exceeds 15. The number of students is 110. The number of students indicated as cheater resulted from MLR is 20, while the remaining 90 indicated as not a cheater. Table (2) shows the performance of SVM, KNN and RF.

Methods	TP	FN	TN	FP	P	R	F- measure	ACC.
SVM	14	6	82	8	0.63	0.7	0.66	0.87
KNN	14	6	81	9	0.60	0.7	0.64	0.86
RF	12	8	71	19	0.38	0.6	0.47	0.75

Table (2): Performance of System Classifiers When Threshold 15

4.4 RESULT OF machine learning algorithms Experiment 2

This experiment evaluates the performance of the system classifiers (SVM, KNN, and Rf) where the difference between the predicted score and the final score exceeds 25. The number of students indicated as cheaters resulted from MLR is 10, while the remaining 100 indicated as not a cheater. Table (4) shows the performance of SVM, KNN, and RF.

Table (3): performance of System Classifiers When Threshold 25.

Methods	TP	FN	TN	FP	P	R	F- measure	ACC.
SVM	8	2	98	2	0.8	0.8	0.8	0.96
KNN	8	2	97	3	0.72	0.80	0.76	0.95
RF	7	3	86	14	0.33	0.70	0.45	0.84

It's clear from the two experiments that the results of SVM and KNN are much closed to each other and outperform the results of RF. Two aspects must be considered before discussing the results. The first aspect is the size of the dataset. As it is known, all machine learning methods work well for big data set and give better results. This is due to the fact that larger data for training, provides an opportunity for learning and more data diversity. The Second aspect the dataset balancing. This point is very sensitive when dealing with data classification. Machine learning requires creating a robust training dataset since the training performs as the seed for subsequent model evaluation. If the training data are unbalanced, the model will perform badly as its accuracy will drop. Data classification is a thriving sector for machine learning where a balanced training dataset is extremely important.

As it's shown in table (2) and table (3) the effect of the two aspects especially the performance of RF. It's clear from the results that SVM and KNN can perform well for a small dataset and unbalanced data, while the performance is less for RF. The used dataset contains six attributes. These six attributes are few for RF, where the depth of the decision tree is low, therefore the efficiency of the results is low.

5. Conclusions

the methodology of e-learning has been applied in the whole world due to the flexibility, accessibility, and ease of use of e-learning, but the main challenge remains in education via the Internet It is the evaluation in exams because the phenomenon of cheating has become increasing. In this search, a solution is proposed to discover cheating by the student's score in the final test was also evaluated according to the analysis of his previous scores during the semester and using machine learning techniques. As a result of the proposed system, it was found that the SVM algorithm gives a high rate in the classification of cheating among students according to the tests that have been applied. As it is the best classification with (0.96) % based on the results of experiments and the accuracy of classification. Also, the number of features affects the performance of the Random Forest when high performance is achieved by a Random Forest with a large number of features and the performance decreased with less number of features. The MLR gives good results with such applications of prediction.

6. REFERENCES

[1] K. Curran, G. Middleton, and C. Doherty, “Cheating in exams with technology,” 2011. *Int. J. Cyber Ethics Educ.*, vol. 1, no. 2, pp. 54–62. 2011.

[2] A. M. Duhaim, S. O. Al-Memory, and M. S. Mahdi, “Cheating detection in online exams during Covid-19 pandemic using data mining techniques” 2021.

[3] R.Bawarith, Abdullah, D. Anas, and P. Dr, “E-exam cheating detection system” 2017. *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 4, pp. 176–181.

[4] L. Li, C. Qiu, and A. Fang, “Detection of test cheating by outlier analysis approach using

probability estimation" 2019. *Pervasive Health Pervasive Compute. Technol. Health*.

[5] A. W. Muzaffar, M. Tahir, M. W. Anwar, Q. Chaudry, S. R. Mir, and Y. Rasheed, "A systematic review of online exams solutions in e-learning: Techniques, tools, and global adoption," IEEE Access, vol. 9, pp. 32689– 32712, 2021.

[6] J. A. Ruiperez-Valiente, P. J. Munoz-Merino, G. Alexandron, and D. E. Pritchard, "Using machine learning to detect 'multiple-account' cheating and analyze the influence of student and problem features" 2019. *IEEE Trans. Learn. Technol*, vol. 12, no. 1, pp. 112–122.

[7] O. R. Harmon and J. Lambrinos, "Are online exams an invitation to cheat?" 2008. *J. Econ. Educ*, vol. 39, no. 2, pp. 116–125.

[8] A. Dahiya and P. Mittal, "Evaluation of Data Mining Techniques and Its Fusion with IoT Enabled Smart Technologies for Effective Prediction of Available Parking Space," *Int. J. Electr. Comput. Eng. Syst.*, vol. 12, no. 4, pp. 187–197, 2021.

[9] Foundations of Machine Learning Mehryar Mori, Afshin Rostami Zadeh, and A meet Talwalkar. The MIT Press Cambridge, Massachusetts London, England 2012.

[10] Daanouni, O., Cherradi, B., & Tmiri, A. (2019, October). Predicting diabetes diseases using mixed data and supervised machine learning algorithms. In Proceedings of the 4th International Conference on Smart City Applications (pp. 1-6).

[11] W. Wang, C. Men, and W. Lu, "Online prediction model based on support vector machine," *Neurocomputing*, vol. 71, no. 4–6, pp. 550–558, 2008.

[12] F. Kamalov, H. Sulieman, and D. S. Calongne, "Machine learning based approach to exam cheating detection," 2021. *PLoS One*, vol. 16, no. 8 August, pp. 1–15, 2021.