On online courses, such as MOOCs, students spend their time in mainly two ways: either by reviewing the course material or by solving exercises. However, there are cases in which the average student behavior differs and tends to become focused only on solving exercises alone or with another student, which can indicate illegal behavior in a learning context. By using a time difference analysis, course material interaction rate and multiple student vs student interactions, clustering techniques are being used to classify student accounts as fake or collaborative, in this project.

As explained on [1], fraudulent students use many techniques to cheat on MOOCs. The one revised on the paper is the Copying Answers using Multiple Existences Online (CAMEO) strategy, which consists in using a "harvester" account in order to solve questions and copy the correct answer to a real account. They show that some courses use techniques to prevent CAMEO, such as in-person assessments taken for a fee or withholding the answer until the problems are graded, but all have a downside. As for results, they found out that from 103 thousand people with only one certificate, 657 of them (1%) have obtained it though CAMEO. However, the presence of CAMEO users rises as the number of certificates in unique accounts rises: on users with 20 certificates, in 73 users, 18 (25%) used the CAMEO strategy, while for users with 40 certificates, in 3 of them, 2 used CAMEO (67%).

In [2] work, a MOOC log file is analyzed to model the student's path through the course and predict his final grade. This log file is comprised of a series of events for each student, which depicts everything that was done, all with a timestamp. The authors created features based on the student's event transitions and tried using many clustering methods, such as: Mean Shift, Hierarchical, Gaussian Mixture Model and k-means. The best and most interpretable was k-means, coupled with BIC. A second experiment was performed to test whether the student's grade was able to predict his grade, using a Multilayer Neural Network and a Random Forest regressor. The academic performance could be predicted with around 10% of mean absolute error.

On [3], the same CAMEO technique is analyzed through a machine learning approach, using a Random Forest model and 15 features to detect submissions as fraudulent. The model achieved a sensitivity level of 0.966 and a specificity level of 0.966. The dataset is from a single introductory physics MOOC called 8.MReV, from edX. From 13500 students enrolled, 502 obtained a certificate. From 502 users with a certificate, 65 (12.9%) of them were considered CAMEO users, while 84 (7.7%) that did not earn a certificate were also considered CAMEO users.

The work of Alexandron et al. [4] also tries to classify users by the CAMEO technique, but also includes collaboration as another method of cheating, were students solve exercises together. They show that a time-based anomaly-detection classifier can generalize from one type of cheating to another with an AUC mean of 0.85. However, their database is also limited to one course.

Our take on this analysis is that the majority of studies along this subject is in its infancy and also highly focused on single courses, with non-generalized models. We intend to combine the works of Alexandron et al. [4] and Valiente et al. [3] in a time-based, anomaly-detection CAMEO and collaboration fraud detector to create a more general classifier through clustering techniques and try to test it on more than one course database.

[1] C. G. Northcutt, A. D. Ho, and I. L. Chuang, “Detecting and preventing ‘multiple-account’ cheating in massive open online courses,” *Comput. Educ.*, vol. 100, pp. 71–80, Sep. 2016.

[2] Á. Pérez-Lemonche, G. Martínez-Muñoz, and E. Pulido-Cañabate, “Analysing event transitions to discover student roles and predict grades in MOOCs,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017, vol. 10614 LNCS, pp. 224–232.

[3] 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,” *IEEE Trans. Learn. Technol.*, vol. 12, no. 1, pp. 112–122, Jan. 2019.

[4] G. Alexandron, J. A. Ruipérez-Valiente, and D. E. Pritchard, “Towards a General Purpose Anomaly Detection Method to Identify Cheaters in Massive Open Online Courses.”