TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 01/01 pp1–13
DOI: 10.26599/TST.2025.9010039
Volume x, Number x, xxxx xxxx

MOOC Dropout Prediction with Machine Learning Techniques: A Systematic Review and Meta-Analysis

Jorge Tenorio-Berrio*, Jorge Pérez-Martín, and Emilio Letón

Abstract: Massive Open Online Courses (MOOCs) have gained popularity as an accessible form of education, attracting a diverse and widespread student base. Despite their potential, MOOCs face a significant challenge: high dropout rates, which undermine their effectiveness and impact. The increasing interest in addressing this problem led to numerous studies developing new models to predict dropouts early and automatically, many of which use Machine Learning (ML) approaches. This research performs a quantitative synthesis of the performance of ML techniques for early dropout prediction in MOOCs. Following PRISMA guidelines, we perform a systematic review and meta-analysis. To analyze the overall performance, we use a random-effects model for a meta-analysis of proportions, analyzing two metrics: sensitivity and specificity. We have also studied the relationship between some of the studies' characteristics and the performance obtained by means of subgroup analysis. The results indicate that ML systems are capable of accurately detecting a significant percentage of potential dropouts. However, the performance of these systems varies depending on the dataset and the definition of dropout used in each study. Despite the promising findings, the high heterogeneity observed across studies suggests that these results should be interpreted with caution.

Key words: Massive Open Online Courses (MOOCs); meta-analysis; Machine Learning (ML); dropout; systematic review; Deep Learning (DL)

1 Introduction

Over the last few decades, the distance learning industry has grown considerably[1], giving rise to large platforms that offer flexible learning opportunities on a global scale. These companies have promoted a type of product that has been associated with their growth. These are Massive Open Online Courses (MOOCs), unregulated courses characterized by flexibility and low entry requirements. They offer a flexible schedule and allow students to explore their interests at a minimal cost, as they generally have few enrollment requirements. These characteristics make them a perfect educational supplement for students and employees.

However, MOOCs have a big problem: the percentage of students who complete these courses is very low, with the dropout rate around 90%[2]. Although Onah et al.[3] pointed out that some dropouts are practically unavoidable, mainly caused by practically non-existent enrollment requirements, this problem has attracted the community's interest by increasing the number of scientific papers addressing it. While some have tried to explore the causes of dropout[4], most have focused on developing systems for early detection and intervention. This task is much more complex than face-to-face education due to its

• Jorge Tenorio-Berrio, Jorge Pérez-Martín, and Emilio Letón are with Department of Artificial Intelligence, Universidad Nacional de Educación a Distancia (UNED), Madrid 28040, Spain. E-mail: jtenorio9@alumno.uned.es; jperezmartin@dia.uned.es; emilio.leton@dia.uned.es.
∗ To whom correspondence should be addressed.
Manuscript received: 2024-08-31; revised: 2024-12-15; accepted: 2025-03-18

distance and massive nature. Machine Learning (ML) provides techniques to deal with this complex problem, the most common approach is to pose it as a classification problem between dropouts and retentions[5, 6].

Multiple authors study this problem with a variety of approaches and techniques. Al-Shabandar et al.[7] used data from Harvard and MIT MOOCs on edX, using decision trees on aggregated behavioural data. Jiang and Li[8] worked with activity logs from MOOCs of the XuetangX platform. Other ML models have also been used, such as Random forest[9] or SVM[10]. Deep Learning (DL) has also been applied in this field, taking advantage of the temporal nature of behavioural data. In Ref. [11], raw data are transformed using a time-window-based algorithm to train a convolutional network. He et al.[12] used recurrent networks on the time series of behavioural data, in this case from the open university. More advanced models, such as those proposed by Fu et al.[13], combine convolutional layers with LSTM and attention mechanisms.

Moreno-Marcos et al.[14] performed a systematic literature review giving a survey of the state of the art on prediction in MOOC, and suggested that future research should be focused on heterogenous context giving by different platforms, thematic areas, or duration. This remark is aligned with Huang et al.[15] who in their review found out several factors that affect MOOC dropout: course attributes, social status, cognitive ability, emotional factor and learning behavior. Moreover, it is also important, as Basnet et al.[16] pointed out, to consider the impact of fakelearners on MOOCs as they biased the investigation on this topic[17].

Given the large number of publications in this research area, the need for methods for synthesizing results and conclusions arises. Several studies have been carried out in this area. All of them are included in the so-called reviews, of which Borenstein[18] distinguished two types:

- Narrative reviews, in which an expert in the field summarizes an arbitrary number of articles in the same line and presents his conclusions, as in the case of Dalipi et al.[19] who reviewed 25 studies or Prenkaj et al.[20] who presented a narrative review of dropout prediction in various types of online courses, including MOOCs.
- Systematic reviews, that eliminate the inherent subjectivity of classical reviews by defining a reproducible method of obtaining studies. Several systematic literature reviews have been performed in this area[14, 15, 21–29]. Two of the most relevant are Moreno-Marcos et al.[14], who analyzed 88 articles on MOOC prediction (not just dropout), and Alhothali et al.[25], a systematic review on predicting outcomes in different types of courses.

Unlike reviews, the meta-analysis technique tries to obtain a global value for a specific measure. In other words, it is a quantitative synthesis. To our knowledge, no meta-analysis has been carried out, although Alhothali et al.[25] mistakenly refered to the systematic literature review performed by Moreno-Marcos et al.[14] as a meta-analysis. Our work propose to use this technique to review and evaluate the performance of current dropout prediction systems in MOOCs. These objectives are innovative for two main reasons: (1) No previous work has obtained a quantitative synthesis of the performance of dropout prediction systems; (2) Although the meta-analysis technique is widely used in medicine[30, 31], its application in the field of data science and ML is an almost unexplored line of work.

We can find an example in Azeem et al.[32], where they analyzed 15 studies in code smell detection. It is a similar problem to dropout prediction since it is also a classification problem. We find other examples using meta-analysis in regression problems like Junior et al.[33], for performance synthesis in market prediction, and Ahmadi et al.[34] for groundwater monitoring/prediction.

Thus, this work tries to address two main Research Questions (RQ):

(1) RQ1: Is there a quantitative synthesis of the performance of ML techniques for dropout prediction in MOOCs?

(2) RQ2: How do the characteristics of the studies affect the performance of the models?

To answer these two questions, we structured the paper as follows: Up to this point, we have presented the problem, novelty, and objectives of this work. Section 2 describes the procedure used to prepare the systematic review and the subsequent meta-analysis. Section 3 presents the results of the systematic review and the synthesis of the effect measures. Section 4 introduces the key findings as well as the assumptions and limitations of the study. Finally, Section 5 presents the answers to the research questions and the lines of future work that arise from this work.

2 Materials and methodology

The PRISMA framework is used to carry out this work's systematic review and meta-analysis. First, we will present how the search strategy is approached, then the synthesis methods, and finally, the subgroups that are considered when performing the meta-analysis and the quantitative synthesis.

2.1 Search strategy

We have included all studies predicting dropout in MOOCs found in three scientific databases: Scopus‡, Web of Science (WoS)§, and IEEE Xplore¶. The queries were carried out on November 14th, 2024, so later articles are not included (see Fig. 1).

The queries executed in the different databases have defined the inclusion criteria. The query refines the proposal of Alhothali et al.[25] and includes terms related to three concepts: dropout, MOOC, and prediction. The search for these terms is performed in the title and abstract of each article, except in the case of IEEE Xplore, which has been searched in all article metadata. The queries performed in the three databases are presented at the end of this section.

Although we acknowledge the high quality of top-tier conferences, such as ACM LAK, AIED, EDM, ITS, and EC-TEL, we have chosen not to include them in our meta-analysis. This decision is based on three key reasons: (1) databases lack filters for high-quality conferences, making systematic inclusion difficult; (2) conference papers often lack the necessary detail for rigorous analysis; and (3) many are later expanded into journal articles, ensuring their inclusion. Our approach aligns with common practices, prioritizing peer-reviewed journal publications for robust and reliable evidence.

Once the articles are obtained, duplicates are discarded. Subsequently, the titles and abstracts of the remaining articles are reviewed, and those not related to the topic are discarded. A full-text version of the remaining references is then downloaded for a complete review. Only those written entirely in English have been taken into account. Next, articles that do not present an original model for dropout prediction in MOOC courses are excluded. Finally, we gather the

```
TITLE-ABS ((
("Dropout" OR "Retention" OR "Completion" OR "Attrition" OR "Withdrawal")
AND
("Online learning" OR "MOOC" OR "Online course" OR "Online Education")
AND
("Classification" OR "Prediction" OR " Machine Learning" OR "Predictive model" OR "Deep learning")
))
```

(a) Query 1: Scopus

```
(
    (TI=("Dropout" OR "Retention" OR " Completion" OR "Attrition" OR "Withdrawal ")
    OR
    AB=( "Dropout" OR "Retention" OR " Completion" OR "Attrition" OR "Withdrawal" ))
AND
    (TI=("Online learning" OR "MOOC" OR " Online course" OR "Online Education")
    OR
    AB=("Online learning" OR "MOOC" OR "Online course" OR "Online Education"))
AND
    (TI=( "Classification" OR "Prediction" OR "Machine Learning" OR "Predictive model" OR "Deep learning" )
    OR
    AB=( "Classification" OR "Prediction" OR " Machine Learning" OR "Predictive model" OR "Deep learning" ))
)
```

(b) Query 2: WoS

```
(
("All Metadata":"Dropout" OR "All Metadata ":"Retention" OR "All Metadata":" Completion" OR "All Metadata":"Attrition" OR "All Metadata":"Withdrawal" )
AND
("All Metadata":"Online learning" OR "All Metadata":"MOOC" OR "All Metadata":"Online course" OR "All Metadata":"Online Education" ) AND
("All Metadata":"Classification" OR "All Metadata":"Prediction" OR "All Metadata":" Machine Learning" OR "All Metadata":" Predictive model" OR "All Metadata":"Deep learning" ))
```

(c) Query 3: IEEE Xplore

Fig. 1 Queries executed in different databases.

necessary data from each article to conduct the meta-analysis. Articles that do not publish sufficient data are also excluded. We have extracted specific characteristics from each study, along with measures of model performance.

We consider as study features:

(1) Dropout definition. We use the categorization of Nagrecha et al.[35], which classifies them into two groups: Those based on lack of participation in the course (e.g., not accessing to the platform in a certain period, not submitting a specific assignment, etc.) and those based on unmet learning objectives (e.g., failing a particular assignment, not finishing a certain number of modules, etc.).

(2) Dataset employed. Many studies use the KDD Cup 015 dataset[25], so we will check for differences between the articles using this dataset and the rest.

‡ https://www.scopus.com/
§ https://www.webofscience.com/
¶ https://ieeexplore.ieee.org/

(3) Algorithm used. We distinguish between the studies that use ML algorithms and those that use DL algorithms.

It is very important to select the appropriate metric in these type of studies. Unfortunately there is no universally valid metric in general[14, 36–38]. In this work, we have used two measures of effect: sensitivity and specificity. We have opted for these two metrics to consider performance in both positive and negative cases. In a binary classification problem, a confusion matrix is given by four values: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). Given the confusion matrix, sensitivity and specificity are defined as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}},$$
$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}}.$$

We have to address two issues at this stage:

- For meta-analysis, each measure must be accompanied by its associated error, but ML studies do not usually publish it. When a study does not present the confusion matrix or the error, but does present other data that allow its calculation, these have been computed.
- Most of the articles reviewed perform several experiments. The main differences between them depend on the technique used or the time of the course at which the prediction is made. To avoid possible correlation between results within the same study, we have chosen to obtain only one result from each article. We use the "best among the earliest" criterion to select the results. In other words, we first chose the experiments that use the least amount of temporal data. When a study presents more than one model, it is necessary to choose one for meta-analysis. Choosing the one with the best sensitivity and the one with the best specificity will be unrealistic due to the trade-off between these metrics. This way, we select the model with the best F1 score, trying to select the model that performs better on these two dimensions. Although we initially consider the Area Under the Curve (AUC), we find the F1 score to be more appropriate. Given that not all published articles include AUC or F1, we believe that F1 allows us to keep a larger number of articles in the meta-analysis, since F1 can be calculated from the confusion matrix, unlike AUC. Moreover, as the MOOC dropout prediction datasets are often unbalanced (mostly because of the high dropout rate at MOOCs), the F1 score effectively balances precision and sensitivity, being the harmonic mean (the inverse of the means of the inverses) of precision and sensitivity. F1 can be computed as[29, 36]

$$\text{F1} = \frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}}.$$

Once the model is selected, the meta-analysis will focus on the ability of the models presented in the literature to predict the real dropout (sensitivity) and the attainment or retention (specificity).

2.2 Synthesis methods

From a formal point of view, a meta-analysis is intended to estimate an overall effect measure θ based on the observed measures of each study $\hat{\theta}_k$. We use the random effects model to calculate it, which takes into account differences between study characteristics, assuming that there is no single true effect size but rather a distribution of them. Thus, the random effects model assumes that there are two sources of variability on the one hand, the sampling error ϵ_k, associated with the measure provided by the study, and on the other ζ, which takes into account the variability between studies resulting from the differences between them.

Both sensitivity and specificity are proportions between two variables, so a meta-analysis of proportions is necessary. We use the binomial-normal model with logit transformation using GLMM[39]. We believe that the assumptions of this model best fit the problem definition. Thus, the model used is given by

$$\hat{y}_k \sim \text{Bin}\,(n_k, \theta_k),$$
$$\text{logit}\,(\theta_k) \sim \mathcal{N}(\theta_{\text{LO}}, \tau^2_{\text{LO}}) \qquad (1)$$

where, for the k-th study, $\hat{y}_k$ is the number of events or numerator of the proportion, n_k is the sample size or denominator of the proportion, and θ_k the observed value of the effect measure. The overall effect measure is given by θ_{LO}, and its variance τ^2_{LO}. The subscript "LO" indicates that they are in logit scale.

It is common for there to be differences between the measures of effect of different studies. As we have mentioned, a meta-analysis assumes the existence of two sources of error. On the one hand, sampling error is the result of variations due to chance within a study and is inversely proportional to the sample size. On the other hand, heterogeneity measures all variability not caused by this error, which is usually caused by characteristics not considered during the analysis.

Estimating it is one of the objectives of meta-analysis since it is directly related to the validation of the evidence. There are several methods for this purpose. In this work, we have used three indices: Cochran's Q, the I^2 statistic, and the inter-study variance τ^2[40].

2.3 Subgroup analysis

The subgroup analysis[18] allows analyzing the relationship of each study's characteristics with the value of the effect measure while exploring the possible causes of heterogeneity. We use the same model as the global analysis, considering the three variables mentioned in Section 2.1: dropout definition, dataset used, and algorithms applied.

3 Result

This section presents the results of the systematic review and the meta-analysis.

3.1 Systematic review

The queries executed on the three databases return 1203 articles. Of these, 554 come from Scopus, 342 from WoS, and 307 from IEEE Xplore. Of this set, 404 are duplicated entries, and 526 articles are discarded because they are not from peer-reviewed journals.

Of the remaining 273 articles, the article selection process concludes by excluding 251 for the following reasons:

- 131 articles are discarded because they are unrelated to the topic of meta-analysis.
- Seven articles considered could not be obtained.
- Six articles are not written in English.
- 51 articles are discarded because they do not perform dropout prediction or the prediction of dropout is made only in non-MOOC courses.
- 56 articles are discarded because they do not provide the necessary information to obtain the confusion matrix and calculate the error.

Figure 2 shows the PRISMA flow diagram, which summarizes the results of applying the inclusion and exclusion criteria, where n is the number of articles.

3.2 Study characteristics

Table 1 shows the confusion matrix extracted from each article and the effect measures calculated. Sensitivity values are generally high, ranging from 0.6957[41] to 0.9951[42]. Specificity values are more variable, ranging from 0.0877[9] to 1.0000[8]. It is important to mention that this last study has many fewer cases than the rest.

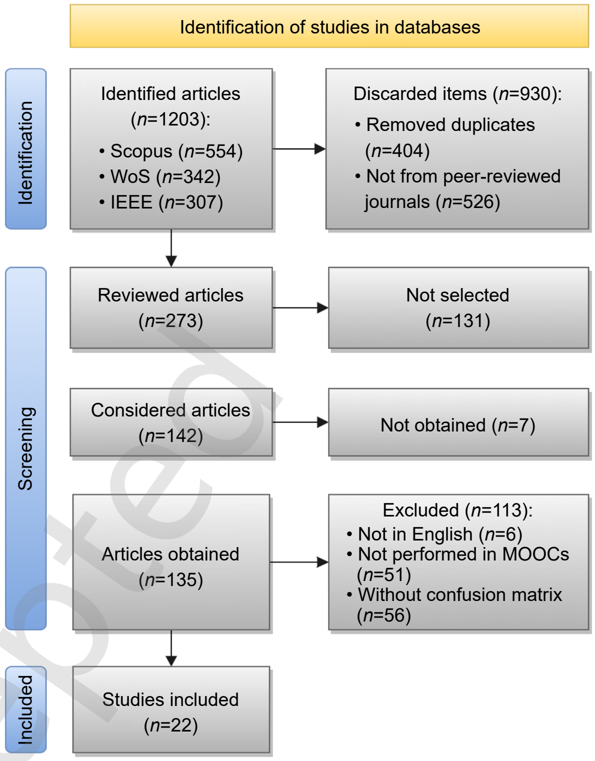


Fig. 2 PRISMA flow diagram for inclusion and exclusion of studies.

Regarding the characteristics of the studies, we have found that most of them base their definition of dropout on criteria related to lack of participation[7–11, 13, 42–45, 48, 49, 51, 52], and eight studies use definitions based on unmet learning objectives[41, 46, 47, 50, 53–56].

Eleven of the studies use the KDD Cup 2015 dataset[8–11, 13, 43, 45, 48, 49, 51, 52]. The remaining eleven studies use different datasets: Standford[42], OULAD dataset[7, 46, 53, 55, 56], HarvardX[44, 50], and from privates datasets[41, 47, 54].

As for the type of model used, ML and DL models are almost equally distributed. On the one hand, fourteen studies use ML. Random forest is the most common model used[9, 47, 49, 50, 56]. Gradient boosting is also frequently used[7, 41, 53]. Linear regression[43], SVM[10], rought sets[48], and ensemble techniques[8] are also used. On the other hand, eight studies use DL for their models. The most commonly used models are convolutional neural networks[11, 45], and the combination of these with LSTM neural networks and attention layers[13, 51, 52, 55]. Two other studies mention using neural networks but do not specify the

Table 1 Confusion matrix and measures of effect of selected studies selected.

Study	TN	FP	FN	TP	Sensitivity	Specificity
Reference [7]	2076	324	292	2957	0.9101	0.8650
Reference [8]	55	0	16	169	0.9135	1.0000
Reference [9]	438	4555	361	18 755	0.9811	0.0877
Reference [10]	811	1769	672	7497	0.9177	0.3143
Reference [11]	1123	1373	1309	8250	0.8631	0.4499
Reference [13]	2608	2384	2581	16 536	0.8650	0.5224
Reference [41]	67	4	7	16	0.6957	0.9437
Reference [42]	599	128	14	2844	0.9951	0.8239
Reference [43]	1123	1373	1309	8250	0.8631	0.4499
Reference [44]	1246	153	117	501	0.8107	0.8906
Reference [45]	1408	1088	402	9156	0.9579	0.5641
Reference [46]	2133	439	47	2270	0.9797	0.8293
Reference [47]	8296	1587	889	8993	0.9100	0.8394
Reference [48]	3449	551	588	3412	0.8530	0.8623
Reference [49]	146	9	9	166	0.9486	0.9419
Reference [50]	361	385	114	5183	0.9785	0.4839
Reference [51]	2748	2073	914	18 373	0.9526	0.5700
Reference [52]	2071	1644	614	13 770	0.9573	0.5575
Reference [53]	2539	556	847	2577	0.7526	0.8204
Reference [54]	95 205	8251	4552	98 904	0.9560	0.9202
Reference [55]	1030	41	45	436	0.9064	0.9617
Reference [56]	3850	293	54	3952	0.9865	0.9293

architecture employed[42, 46].

Table 2 shows the values of the variables of the 11 studies included in the meta-analysis.

3.3 Risk of bias

In this work, we have qualitatively analyzed the risk of publication bias. Following the recommendation of Barker et al.[57], this is the most appropriate method in a meta-analysis of proportions since there is no quantitative method to assess publication bias and the assumptions of classical methods are not necessarily met.

Our research process involves a careful consideration of excluding articles not published in peer-reviewed journals, a decision that was not made lightly. While this may introduce some publication bias in the data, we make this assumption in order to ensure the inclusion of higher-quality articles in our analysis.

In addition, as previously shown, many studies have been excluded because the data necessary to perform the analysis could not be obtained. Therefore, this work is limited to those studies that publish sufficient data to calculate the confusion matrix.

Another bias introduced in this analysis is the

Table 2 Characteristics of the studies selected for meta-analysis.

Study	Model	Dropout	Dataset
Reference [7]	ML	Participation	Others
Reference [8]	ML	Participation	KDD Cup
Reference [9]	ML	Participation	KDD Cup
Reference [10]	ML	Participation	KDD Cup
Reference [11]	DL	Participation	KDD Cup
Reference [13]	DL	Participation	KDD Cup
Reference [41]	ML	Learning objective	Others
Reference [42]	DL	Participation	Others
Reference [43]	ML	Participation	KDD Cup
Reference [44]	ML	Participation	Others
Reference [45]	DL	Participation	KDD Cup
Reference [46]	DL	Learning objective	Others
Reference [47]	ML	Learning objective	Others
Reference [48]	ML	Participation	KDD Cup
Reference [49]	ML	Participation	KDD Cup
Reference [50]	ML	Learning objective	Others
Reference [51]	DL	Participation	KDD Cup
Reference [52]	DL	Participation	KDD Cup
Reference [53]	ML	Learning objective	Others
Reference [54]	ML	Learning objective	Others
Reference [55]	DL	Learning objective	Others
Reference [56]	ML	Learning objective	Others

selection of a single result within each article. The best result among the earliest ones is used.

3.4 Results of syntheses

In this section, we present the results of the syntheses of the meta-analysis. For each of the two measures of effect, sensibility and specificity, we show the value of several statistics, such as I^2 (a percentage measuring the heterogeneity of the included studies), Q (a Chi-square test to assess the heterogeneity of studies included in a meta-analysis, in which the magnitude of the effect of each individual study is compared with the combined estimate), and τ^2 (the estimated standard deviation of the underlying effects across studies) as well as the graphic of a forest plot (the standard method for graphically representing the results of a meta-analysis in terms of the overall measure of effect). We also include results of subgroup analysis with the two measures, considering partioning by definition of dropout (groups $G_{\text{Participation}}$ and G_{Learning}), by type of dataset (groups G_{KDD} and G_{Others}) and by type of algorithm (groups G_{ML} and G_{DL}).

3.4.1 Meta-analysis for the sensibility

The sensitivity meta-analysis obtains a pooled effect of 0.9363 with a confidence interval of [0.9044; 0.9581]. The obtained forest plot can be seen in Fig. 3.

As can be seen, the heterogeneity of this model is very high. The value of I^2 is 99.7% and the p-value of the Q test is very close to 0. The value of τ^2 is 1.0930 (recall that this value is in logit scale).

Using the subgroup analysis technique, we verify the differences between the different groups, but they are not significant in any of the divisions. A summary of the results obtained can be seen in Fig. 4.

When analyzing the differences by **definition of dropout**, the pooled effect in $G_{\text{Participation}}$ is 0.9344 [0.8949; 0.9597], almost the same the 0.9392 [0.8722; 0.9722] obtained in G_{Learning}. The heterogeneity in G_{Learning} reach values for I^2 and τ^2 of 99.7% and 1.1622, respectively, compared to 99.6% and 0.9472 obtained in $G_{\text{Participation}}$. However, the differences between the two groups are not significant ($p = 0.8700$).

When analyze by the **dataset** used, G_{KDD} obtains a lower overall sensitivity than G_{Others}. The former group obtains 0.9284 [0.8956; 0.9514] and the latter 0.9431 [0.8831; 0.9733]. The heterogeneity values also vary in the case of τ^2. In G_{KDD} a value of 0.4727 and in G_{Others} 1.7379 are reached. However, both groups obtain an I^2 value of 99.7%. In this case, the differences between groups are not significant ($p = 0.5862$).

Finally, analyzing by model type also yields very similar measures. In the case of G_{ML}, we obtain a

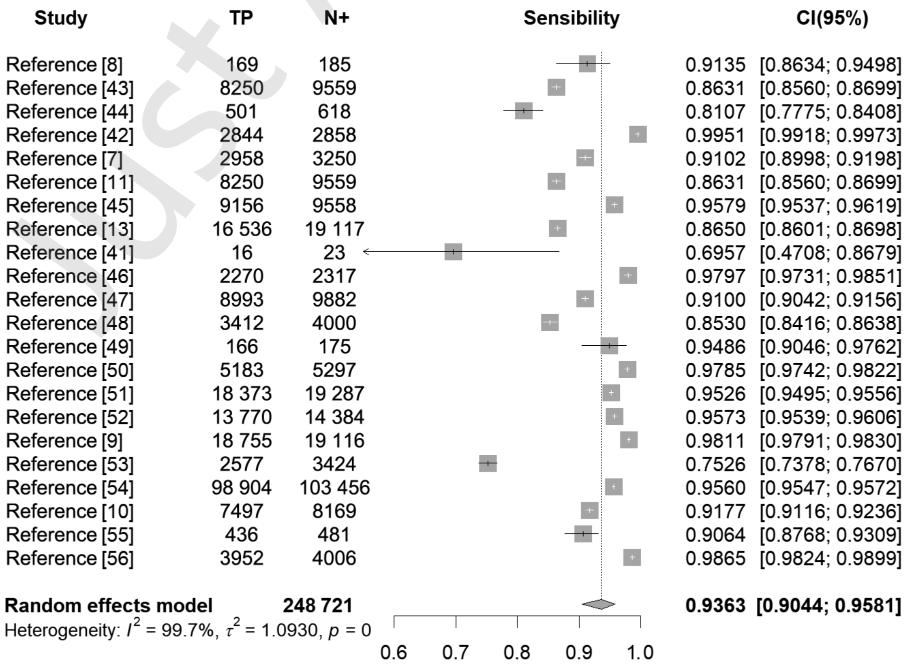


Fig. 3 Forest plot of the sensitivity meta-analysis. N+ stands for the number of total possitive cases and CI stands for the confidence intervals.

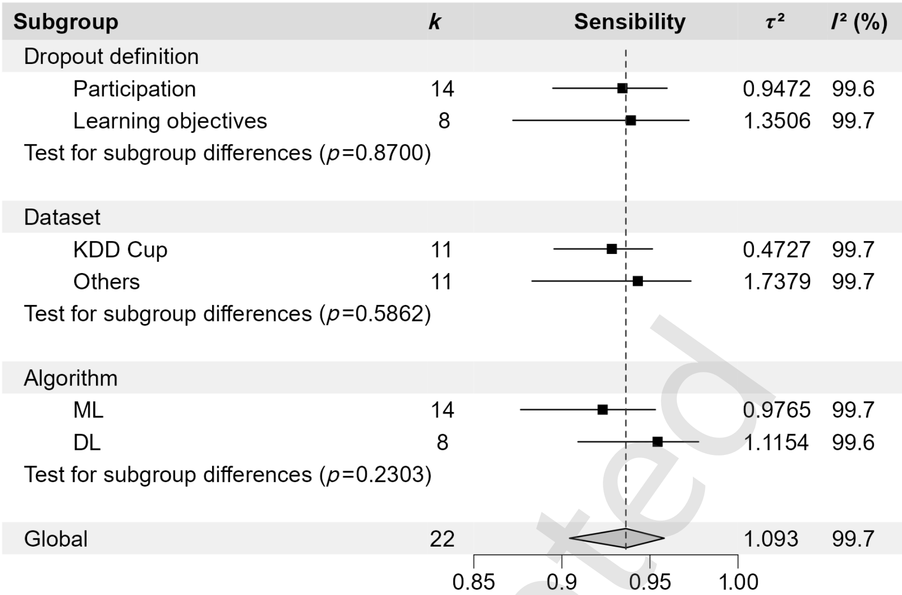


Fig. 4 Results of the subgroup analysis of the sensitivity meta-analysis.

sensitivity of 0.9232 [0.8766; 0.9531], while in G_{DL} it is 0.9544 [0.9092; 0.9776]. The heterogeneity is also very high in both groups. In G_{ML} values of 0.9765 and 99.7% are obtained for τ_g^2 and I^2, respectively, while G_{DL} are 1.1154 and 99.6%. In this case, we also do not obtain sufficient significance in the difference test ($p = 0.2303$).

3.4.2 Meta-analysis for the specificity

The pooled effect value in the specificity meta-analysis is 0.7746 [0.6455; 0.8644]. The obtained forest plot can be seen in Fig. 5. In this case, the measurements of the different studies are spread over the whole possible range of values.

As for heterogeneity, again, very high values are obtained. The value of I^2 is 99.9%, and in the case of τ^2 it is 2.2526. Similarly, the p-value of the Q test is very close to 0.

The results of the subgroups analysis are slightly different to the sensitivity meta-analysis, in this case significant differences are found in two of the

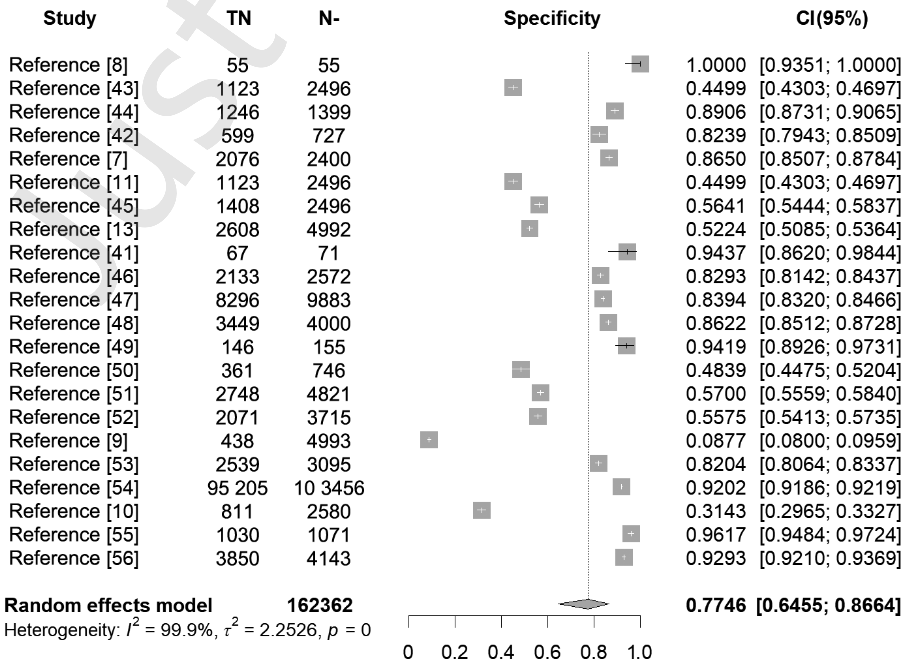


Fig. 5 Forest plot of the specificity meta-analysis. N− stands for the number of total negative cases.

divisions. A summary of these results can be seen in Fig. 6.

Partitioning by **definition of dropout** results in a pooled effect on $G_{\text{Participation}}$ of 0.6938 [0.4936; 0.8405], also below the 0.8758 [0.7817; 0.9329] of G_{Learning}. In terms of heterogeneity, on the one hand, $G_{\text{Participation}}$ obtains an I^2 value of 99.8% and τ^2 of 2.5175. G_{Learning}, on the other hand, obtains a 99.7% and 0.9265, respectively. The differences between the two groups are significant ($p = 0.0398$).

Although in this case the differences are significant ($p = 0.180$). In the case of G_{KDD}, an overall specificity of 0.6333 [0.3836; 0.8273] is obtained. In contrast, in G_{Others} it is 0.8715 [0.8054; 0.9175]. Differences also occur in heterogeneity. In G_{KDD} we obtain a value of I^2 of 99.8% and of 2.8714 for τ^2, and in G_{Others} 99.6% and 0.8225.

Dividing the articles by **model type** obtains different measures than in the analysis of sensitivity. A specificity of 0.8128 [0.6359; 0.9152] is obtained for G_{ML}, and 0.7050[0.5312; 0.8344] for G_{DL}. In G_{ML} we obtain a value of 2.9101 for τ_g^2, while G_{DL} is 1.1542. In the case of I^2 we obtain 99.9% and 99.5% for G_{ML} and G_{DL}, respectively. The test for differences between groups obtains a p-value of 0.3202.

4 Discussion

This work has carried out a quantitative synthesis in MOOC dropout prediction using ML techniques. We use PRISMA guidelines to carry it out, which is a differentiating element from other meta-analyses carried out in the field of data science.

This section will discuss the results and the main limitations of this work.

Concerning the first research question (RQ1), the results of the overall meta-analysis indicate that the ability of these systems to predict dropout is very good overall, in agreement with several systematic literature reviews (see Refs. [21, 26, 28, 29], among others). This is reflected in the high pooled sensitivity value. However, this performance has not been equally reflected in the prediction of retentions since the overall specificity obtained is barely close to 80%. In addition, the measures of each study are more variable in this second measure.

Two observations can be made on this fact. On the one hand, priority is being given to optimizing the performance of predicting dropouts at the cost of more errors in predicting retentions. This may be because the impact or cost associated with the mistake in retentions is lower than that of dropouts. On the other hand, high variability is a cause of unbalanced datasets. Typically, they have a lower proportion of negative examples, causing the specificity to decrease much faster with each error.

In the results of the subgroup analysis (RQ2), we have verified that there is some correlation between the dataset used in the study and its performance (in agreement with some systematic literature reviews, see Refs. [15, 29], among others), which is repeated in both

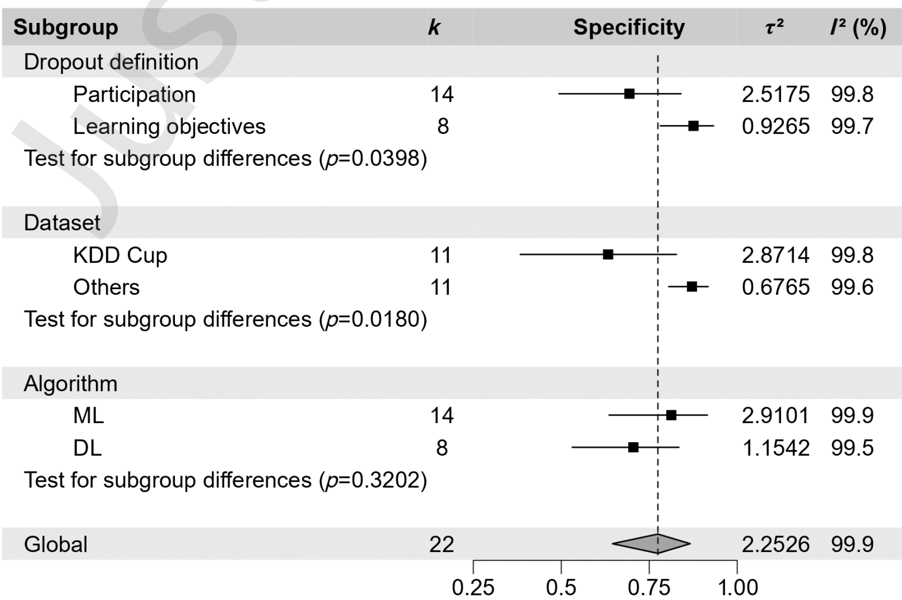


Fig. 6 Results of the subgroup analysis in the specificity meta-analysis.

measures of effect. The studies that use the KDD dataset obtain worse results than those that use a different one. This fact is not repeated when the dropout definition or type of algorithm is used, so we can conclude that there is some dependence between the data used and the prediction performance.

In accordance with Ref. [58], another interesting result of this analysis is that there is hardly any difference between the performance of ML and DL models. Although the trend nowadays is to use complex neural networks, this analysis shows that the differences with the more classical models are insignificant.

The main fact that characterizes this analysis is the high heterogeneity. In the overall meta-analyses, extremely high values of the I^2 index have been obtained, always close to 100%; and p-values very close to 0 in the Q test, despite the fact that τ^2 seems acceptable. This fact may indicate that there is some variability among the characteristics of the studies, and therefore subgroup analysis with three variables is used. However, this technique has not yielded very enlightening results. It has been found in all experiments that heterogeneity varies between subgroups. In all cases, one of the groups obtains a higher value with respect to the pooled value, while the other is lower. Beyond this fact we have not found any relevant patterns in this aspect.

From the point of view of Barker et al.[57], the high levels of heterogeneity are expected in a meta-analysis of proportions and should not be considered a drawback. This is mainly due to certain incompatibilities of the Q test and the I^2 index with this type of meta-analysis.

Even so, it is crucial to bear in mind that these types of studies have unique characteristics that have not been taken into account and can affect heterogeneity. This is the great challenge facing meta-analyses applied to studies in data science. In medicine, the vast majority of variables are controlled and published, but this is not the case in data science. Only the value of specific evaluation measures is usually published, but not all the model and dataset characteristics are published.

We have verified this in this article when retrieving specific data for the meta-analysis. Initially, we assum a particular publication bias by including only peer-reviewed journals. We consider that we would find articles of higher quality, thus facilitating results extraction. However, in most articles, retrieving the data to calculate the error associated with the evaluation measure has been impossible. This can be done using the confusion matrix or the characteristics of the set where the system is evaluated. We consider that these data are not only necessary to carry out a meta-analysis but also essential to ensure its reproducibility.

It is also important to mention that it is possible that some heterogeneity is caused by a certain correlation between measures not taken into account, caused by two reasons. On the one hand, it is known that the decision threshold chosen in the study affects the value of the evaluation metrics[59]. This is clearly seen in pairs of sensitivity-specificity type metrics, causing one to increase while reducing the other. On the other hand, as Deeva et al.[60] mentioned, there is some correlation between prediction moment and model performance. If the prediction moments are not the same, the measurements could become incomparable. Taking both aspects into account adds some complexity that we propose as future work.

Related to the above, the studies included in this meta-analysis often report predictions at multiple time points, which has required careful consideration for measure selection. Since each study uses a different temporal granularity (e.g. Mourdi et al.[42] used weeks, Adnan et al.[46] used days, and Dass et al.[47] used a percentage of the course), it is difficult to standardize the temporal granularity used, as it depends on the format and structure of the course. In line with the community's interest in making predictions as early as possible, we decide to select the earliest available measure for each study, trying to alleviate the heterogeneity between measurements. Another possibility that could be considered for future work will be taking into account the time of the prediction as another variable of the meta-analysis.

5 Conclusion and Future Work

Throughout this work, a quantitative analysis of the performance of ML and DL techniques for the prediction of student dropout in MOOCs by means of ML has been carried out. The meta-analysis technique on 22 articles and two evaluation metrics has been used to carry it out. The PRISMA guide is followed for all the work.

It is concluded that, despite the high levels of heterogeneity, these systems are particularly useful for dropout detection, since they are capable of detecting more than 90% of students who drop out in most of the studies carried out. However, we have found that some of these systems have a high false negative rate, leading to errors in predicting students who complete the course. Using the subgroup analysis technique, we have found that the data used have a greater influence on the performance of the system than the type of model used, in agreement with Gašević et al.[61]

To continue this work we propose some lines, all of which are aimed at reducing heterogeneity and publication bias:

● Extend the meta-analysis by taking into account the time of prediction as a variable. Despite our analysis take into account only the "best among the earliest" models, giving that early identification of at-risk learners could help to reduce dropout[62], it should be interesting to consider other approaches. One of these approaches would be to consider using meta-regression models.

● The correlation between measures could also be taken into account by means of a bivariate analysis. The so-called diagnostic meta-analysis takes into account this aspect among others.

● Extend the analysis by applying less restrictive selection criteria, incorporating the so-called gray literature.

References

[1] Research and Markets, Massive Open Online Courses-Global Strategic Business Report, https://www.researchandmarkets.com/reports/5140372/massive-open-online-courses-global-strategic, 2024.

[2] N. Floratos, T. Guasch, and A. Espasa, Recommendations on formative assessment and feedback practices for stronger engagement in MOOCs, *Open Praxis*, vol. 7, no. 2, pp. 141–152, 2015.

[3] D. F. O. Onah, J. Sinclair, and R. Boyatt, Dropout rates of massive open online courses: Behavioural patterns, in *Proc. 6th Int. Conf. Education and New Learning Technologies*, Barcelona, Spain, 2014, pp. 5825−5834.

[4] C. Gütl, R. H. Rizzardini, V. Chang, and M. Morales, Attrition in MOOC: Lessons learned from drop-out students, in *Proc. 3rd Int. Workshop on Learning Technology for Education in Cloud—MOOC and Big Data*, Santiago, Chile, 2014, pp. 37−48.

[5] P. M. Moreno-Marcos, P. J. Muñoz-Merino, J. Maldonado-Mahauad, M. Pérez-Sanagustín, C. Alario-Hoyos, and C. D. Kloos, Temporal analysis for dropout prediction using self-regulated learning strategies in self-paced MOOCs, *Comput. Educ.*, vol. 145, p. 103728, 2020.

[6] J. Zhang, M. Gao, and J. Zhang, The learning behaviours of dropouts in MOOCs: A collective attention network perspective, *Comput. Educ.*, vol. 167, p. 104189, 2021.

[7] R. Al-Shabandar, A. J. Hussain, P. Liatsis, and R. Keight, Detecting at-risk students with early interventions using machine learning techniques, *IEEE Access*, vol. 7, pp. 149464–149478, 2019.

[8] F. Jiang and W. Li, Who will be the next to drop out? Anticipating dropouts in MOOCs with multi-view features, *Int. J. Performability Eng.*, vol. 13, no. 2, pp. 201–210, 2017.

[9] F. Khoushehgir and S. Sulaimany, Negative link prediction to reduce dropout in Massive Open Online Courses, *Educ. Inf. Technol.*, vol. 28, no. 8, pp. 10385–10404, 2023.

[10] S. Al Amoudi, A. Alhothali, R. Mirza, H. Assalahi, and T. Aldosemani, Click-based representation learning framework of student navigational behavior in MOOCs, *IEEE Access*, vol. 12, pp. 121480–121494, 2024.

[11] L. Qiu, Y. Liu, Q. Hu, and Y. Liu, Student dropout prediction in massive open online courses by convolutional neural networks, *Soft Comput.*, vol. 23, no. 20, pp. 10287–10301, 2019.

[12] Y. He, R. Chen, X. Li, C. Hao, S. Liu, G. Zhang, and B. Jiang, Online at-risk student identification using RNN-GRU joint neural networks, *Information*, vol. 11, no. 10, p. 474, 2020.

[13] Q. Fu, Z. Gao, J. Zhou, and Y. Zheng, CLSA: A novel deep learning model for MOOC dropout prediction, *Comput. Electr. Eng*., vol. 94, p. 107315, 2021.

[14] P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Muñoz-Merino, and C. D. Kloos, Prediction in MOOCs: A review and future research directions, *IEEE Trans. Learn. Technol.*, vol. 12, no. 3, pp. 384–401, 2019.

[15] H. Huang, L. Jew, and D. Qi, Take an MOOC and then drop: A systematic review of MOOC engagement pattern and dropout factor, *Heliyon*, vol. 9, no. 4, p. e15220, 2023.

[16] R. B. Basnet, C. Johnson, and T. Doleck, Dropout prediction in Moocs using deep learning and machine learning, *Educ. Inf. Technol.*, vol. 27, no. 8, pp. 11499–11513, 2022.

[17] G. Alexandron, L. Y. Yoo, J. A. Ruipérez-Valiente, S. Lee, and D. E. Pritchard, Are MOOC learning analytics results trustworthy? With fake learners, they might not be! *Int. J. Artif. Intell. Educ*., vol. 29, no. 4, pp. 484−506, 2019.

[18] M. Borenstein, L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein, *Introduction to Meta-Analysis*. Hoboken, NJ, USA: John Wiley & Sons, Ltd., 2009.

[19] F. Dalipi, A. S. Imran, and Z. Kastrati, MOOC dropout prediction using machine learning techniques: Review and research challenges, in *Proc. 2018 IEEE Global Engineering Education Conf.*, Santa Cruz de Tenerife, Spain, 2018, pp. 1007−1014.

[20] B. Prenkaj, P. Velardi, G. Stilo, D. Distante, and S. Faralli, A survey of machine learning approaches for student dropout prediction in online courses, *ACM Comput. Surv.*, vol. 53, no. 3, p. 57, 2020.

[21] M. M. Tamada, J. F. de Magalhães Netto, and D. P. R. de Lima, Predicting and reducing dropout in virtual learning using machine learning techniques: A systematic review, in *Proc. 2019 IEEE Frontiers in Education Conf.*, Covington, KY, USA, 2019, pp. 1−9.

[22] B. Albreiki, N. Zaki, and H. Alashwal, A systematic literature review of student' performance prediction using machine learning techniques, *Educ. Sci.*, vol. 11, no. 9, p. 552, 2021.

[23] H. Luan and C. C. Tsai, A review of using machine learning approaches for precision education, *J. Educ. Technol. Soc.*, vol. 24, no. 1, pp. 250–266, 2021.

[24] H. V. Pant, M. C. Lohani, and J. Pande, Exploring the factors affecting learners' retention in MOOCs: A systematic literature review, *Int. J. Inf. Commun. Technol. Educ.*, vol. 17, no. 4, pp. 1–17, 2021.

[25] A. Alhothali, M. Albsisi, H. Assalahi, and T. Aldosemani, Predicting student outcomes in online courses using machine learning techniques: A review, *Sustainability*, vol. 14, no. 10, p. 6199, 2022.

[26] F. Ouyang, L. Zheng, and P. Jiao, Artificial intelligence in online higher education: A systematic review of empirical research from 2011 to 2020, *Educ. Inf. Technol.*, vol. 27, no. 6, pp. 7893–7925, 2022.

[27] S. Elibol and A. Bozkurt, Student dropout as a never-ending evergreen phenomenon of online distance education, *Eur. J. Investig. Health, Psychol. Educ.*, vol. 13, no. 5, pp. 906–918, 2023.

[28] J. Chen, B. Fang, H. Zhang, and X. Xue, A systematic review for MOOC dropout prediction from the perspective of machine learning, *Interact. Learn. Environ.*, vol. 32, no. 5, pp. 1642–1655, 2024.

[29] S. Alghamdi, B. Soh, and A. Li, A comprehensive review of dropout prediction methods based on multivariate analysed features of MOOC platforms, *Multimodal Technol. Interact.*, vol. 9, no. 1, p. 3, 2025.

[30] R. D. Nindrea, T. Aryandono, L. Lazuardi, and I. Dwiprahasto, Diagnostic accuracy of different machine learning algorithms for breast cancer risk calculation: A meta-analysis, *Asian Pac. J. Cancer Prev.*, vol. 19, no. 7, pp. 1747–1752, 2018.

[31] A. L. D. Araújo, M. C. Moraes, M. E. Pérez-de-Oliveira, V. M. da Silva, C. Saldivia-Siracusa, C. M. Pedroso, M. A. Lopes, P. A. Vargas, S. Kochanny, A. Pearson, et al., Machine learning for the prediction of toxicities from head and neck cancer treatment: A systematic review with meta-analysis, *Oral Oncol.*, vol. 140, p. 106386, 2023.

[32] M. I. Azeem, F. Palomba, L. Shi, and Q. Wang, Machine learning techniques for code smell detection: A systematic literature review and meta-analysis, *Inf. Software Technol.*, vol. 108, pp. 115–138, 2019.

[33] M. A. Junior, P. Appiahene, O. Appiah, and C. N. Bombie, Forex market forecasting using machine learning: Systematic literature review and meta-analysis, *J. Big Data*, vol. 10, no. 1, p. 9, 2023.

[34] A. Ahmadi, M. Olyaei, Z. Heydari, M. Emami, A. Zeynolabedin, A. Ghomlaghi, A. Daccache, G. E. Fogg, and M. Sadegh, Groundwater level modeling with machine learning: A systematic review and meta-analysis, *Water*, vol. 14, no. 6, p. 949, 2022.

[35] S. Nagrecha, J. Z. Dillon, and N. V. Chawla, MOOC dropout prediction: Lessons learned from making pipelines interpretable, in *Proc. 26th Int. Conf. World Wide Web Companion*, Perth, Australia, 2017, pp. 351−359.

[36] R. Pelanek, Metrics for evaluation of student models, *J. Educ. Data Min.*, vol. 7, no. 2, pp. 1–19, 2015.

[37] R. Pelánek, The details matter: Methodological nuances in the evaluation of student models, *User Model. User-Adapted Interaction*, vol. 28, no. 3, pp. 207–235, 2018.

[38] R. Pelánek, Measuring similarity of educational items: An overview, *IEEE Trans. Learn. Technol.*, vol. 13, no. 2, pp. 354–366, 2020.

[39] T. Stijnen, T. H. Hamza, and P. Özdemir, Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data, *Statist. Med.*, vol. 29, no. 29, pp. 3046–3067, 2010.

[40] R. DerSimonian and N. Laird, Meta-analysis in clinical trials, *Control. Clin. Trials*, vol. 7, no. 3, pp. 177–188, 1986.

[41] T. Panagiotakopoulos, S. Kotsiantis, G. Kostopoulos, O. Iatrellis, and A. Kameas, Early dropout prediction in MOOCs through supervised learning and hyperparameter optimization, *Electronics*, vol. 10, no. 14, p. 1701, 2021.

[42] Y. Mourdi, M. Sadgal, H. El Kabtane, and W. B. Fathi, A machine learning-based methodology to predict learners' dropout, success or failure in MOOCs, *Int. J. Web Inf. Syst.*, vol. 15, no. 5, pp. 489–509, 2019.

[43] L. Qiu, Y. Liu, and Y. Liu, An integrated framework with feature selection for dropout prediction in massive open online courses, *IEEE Access*, vol. 6, pp. 71474–71484, 2018.

[44] R. Al-Shabandar, A. J. Hussain, P. Liatsis, and R. Keight, Analyzing learners behavior in MOOCs: An examination of performance and motivation using a data-driven approach, *IEEE Access*, vol. 6, pp. 73669–73685, 2018.

[45] Y. Wen, Y. Tian, B. Wen, Q. Zhou, G. Cai, and S. Liu, Consideration of the local correlation of learning behaviors to predict dropouts from MOOCs, *Tsinghua Science and Technology*, vol. 25, no. 3, pp. 336–347, 2020.

[46] M. Adnan, A. Habib, J. Ashraf, S. Mussadiq, A. A. Raza, M. Abid, M. Bashir, and S. U. Khan, Predicting at-risk students at different percentages of course length for early intervention using machine learning models, *IEEE Access*, vol. 9, pp. 7519–7539, 2021.

[47] S. Dass, K. Gary, and J. Cunningham, Predicting student dropout in self-paced MOOC course using random forest model, *Information*, vol. 12, no. 11, p. 476, 2021.

[48] C. Blundo, V. Loia, and F. Orciuoli, A time-aware approach for MOOC dropout prediction based on rule induction and sequential three-way decisions, *IEEE Access*, vol. 11, pp. 113189–113198, 2023.

[49] S. Nithya and S. Umarani, An identification of the prominent learner behavioral features to predict MOOC dropouts using hybrid algorithm, *J. Theor. Appl. Inf. Technol.*, vol. 101, no. 3, p. 1261, 2023.

[50] Z. Chi, S. Zhang, and L. Shi, Analysis and prediction of

MOOC learners' dropout behavior, *Appl. Sci.*, vol. 13, no. 2, p. 1068, 2023.

[51] K. Niu, G. Lu, X. Peng, Y. Zhou, J. Zeng, and K. Zhang, CNN autoencoders and LSTM-based reduced order model for student dropout prediction, *Neural Comput. Appl.*, vol. 35, no. 30, pp. 22341–22357, 2023.

[52] G. Kumar, A. Singh, and A. Sharma, Ensemble deep learning network model for dropout prediction in MOOCs, *Int. J. Electr. Comput. Eng. Syst.*, vol. 14, no. 2, pp. 187–196, 2023.

[53] E. M. Smaili, M. Daoudi, I. Oumaira, S. Azzouzi, and M. E. H. Charaf, Towards an adaptive learning model using optimal learning paths to prevent MOOC dropout, *Int. J. Eng. Ped.*, vol. 13, no. 7, pp. 128–144, 2023.

[54] X. Li and H. Tang, Analysis of factors influencing online learning using the decision tree method, *HighTech and Innovation Journal*, vol. 5, no. 2, pp. 479–488, 2024.

[55] X. Xia and W. Qi, Driving STEM learning effectiveness: Dropout prediction and intervention in MOOCs based on one novel behavioral data analysis approach, *Humanit. Soc. Sci. Commun.*, vol. 11, no. 1, p. 430, 2024.

[56] H. A. Althibyani, Predicting student success in MOOCs: A comprehensive analysis using machine learning models, *PeerJ Comput. Sci.*, vol. 10, p. e2221, 2024.

[57] T. H. Barker, C. B. Migliavaca, C. Stein, V. Colpani, M. Falavigna, E. Aromataris, and Z. Munn, Conducting proportional meta-analysis in different types of systematic reviews: A guide for synthesisers of evidence, *BMC Med. Res. Methodol.*, vol. 21, no. 1, p. 189, 2021.

[58] M. Vaarma and H. Li, Predicting student dropouts with machine learning: An empirical study in Finnish higher education, *Technol. Soc.*, vol. 76, p. 102474, 2024.

[59] K. W. Kim, J. Lee, S. H. Choi, J. Huh, and S. H. Park, Systematic review and meta-analysis of studies evaluating diagnostic test accuracy: A practical review for clinical researchers-Part I. General guidance and tips, *Korean J. Radiol.*, vol. 16, no. 6, pp. 1175–1187, 2015.

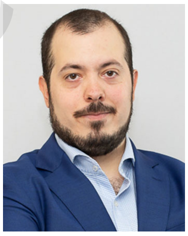
[60] G. Deeva, J. De Smedt, and J. De Weerdt, Educational sequence mining for dropout prediction in MOOCs: Model building, evaluation, and benchmarking, *IEEE Trans. Learn. Technol.*, vol. 15, no. 6, pp. 720–735, 2022.

[61] D. Gašević, S. Dawson, T. Rogers, and D. Gasevic, Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success, *Internet High. Educ.*, vol. 28, pp. 68–84, 2016.

[62] D. Bañeres, M. E. Rodríguez-González, A. E. Guerrero-Roldán, and P. Cortadas, An early warning system to identify and intervene online dropout learners, *Int. J. Educ. Technol. High. Educ.*, vol. 20, no. 1, p. 3, 2023.

Jorge Tenorio-Berrio received the BEng degree in computer engineering from Universidad Nacional de Educación a Distancia (UNED), Spain in 2021, and the MEng degree in data science from the same university in 2024. Nowadays he is researching in data analysis in the field of distance education.

Emilio Letón received the PhD degree in mathematics from Universidad Complutense de Madrid (UCM), Spain in 2002. He is currently a lecturer at Department of Artificial Intelligence, UNED, Spain, where he joined in 2009. Previously, he was a lecturer at Department of Statistics, Universidad Carlos III de Madrid (UC3M), Spain, for a period of five years. He has also worked for 15 years at Departments of Planning and Statistics in the banking sector and the pharmaceutical industry. He is the co-author of 15 articles reflected in his ORCID (https://orcid.org/0000-0001-6908-9590). He makes his transfer from the miniXmodular group (www.minixmodular.ia.uned.es) of which he has been coordinator since its creation in 2016, from the ONCE Foundation Channel at UNED in which he has collaborated since its creation in 2016 and from the UNESCO Chair in Education to Distance (CUED) where he joined in 2022 as a member of the academic council.



Jorge Pérez-Martín received the BEng degree in IT business management from Complutense University of Madrid, Spain in 2014. One year later he received the MEng degree in advanced artificial intelligence from Universidad Nacional de Educación a Distancia (UNED), Spain. From 2015 to 2019 he completed his PhD studies in intelligent systems thanks to a predoctoral grant (FPU14/02577) from the Spanish Ministry of Education. In October 2020, he started working at Department of Artificial Intelligence, UNED, Spain. He currently holds an associate professor position in this same department. He is a member of the "Research Center for Intelligent Decision-Support Systems" and "miniXmodular", a teaching innovation group. He has participated in several teaching innovation projects related to accessible interactive materials and in several research projects mainly related to the use of machine learning and deep learning models in medical contexts.