Preserving the integrity of study behaviour in online retrieval practice using quantified learner dynamics

Maarten van der Velde
MemoryLab
maarten@memorylab.nl

Malte Krambeer
MemoryLab
malte@memorylab.nl

Hedderik van Rijn
MemoryLab
hedderik@memorylab.nl

ABSTRACT

Ensuring the integrity of results in online learning and assessment tools is a challenge, due to the lack of direct supervision increasing the risk of fraud. We propose and evaluate a machine learning-based method for detecting anomalous behaviour in an online retrieval practice task, using an XGBoost classifier trained on keystroke dynamics and task performance features to distinguish between genuine and fraudulent responses. The classifier requires only a modest amount of training data—approximately 100 short-answer responses, typically collected within 10 minutes of practice—and maintains good performance when not all feature types are available. This method enhances the reliability of online learning and assessment by identifying anomalous response behaviour in a way that preserves learners' privacy.

Keywords

Keystroke dynamics, learner authentication, online assessment, retrieval practice, integrity

1. INTRODUCTION

Online tools for learning and assessment offer benefits like adaptivity and immediate feedback, but also create new challenges for an educational sector that is rapidly integrating these tools into its curricula. A key issue is that of academic dishonesty: when learners are completing exercises outside the classroom, and are therefore not under the direct supervision of an instructor, how can we ensure that they are not looking up answers, engaging in plagiarism, or receiving help from others?

Preserving the integrity of online learning and assessment requires a cohesive set of methods to both prevent cheating in the first place, and to detect cheating when it does happen [13]. Systems can make specific forms of cheating more difficult through technological measures, such as authenticating a learner's identity through a password or biometrics, or monitoring their behaviour using proctoring tools [1]. Methods to preserve integrity can also include understanding why students engage in undesirable behaviour (e.g., because there are usability issues, or because the material is seen as low-quality, irrelevant, or too easy or difficult) and changing the learning environment accordingly [2, 28]. In addition, students may be explicitly dissuaded from dishonest behaviour, for instance by warning them about the consequences of being caught [9]. Within this wider framework, we focus on a specific aspect of preserving integrity in the context of an online retrieval practice environment: authenticating a learner's identity *during* the task through distinct patterns in their typing behaviour on short-form text answers and in their practice performance.

Maarten van der Velde, Malte Krambeer, and Hedderik van Rijn. Preserving the integrity of study behaviour in online retrieval practice using quantified learner dynamics. In Caitlin Mills, Giora Alexandron, Davide Taibi, Giosuè Lo Bosco, and Luc Paquette (eds.) Proceedings of the 18th International Conference on Educational Data Mining, Palermo, Italy, July, 2025, pp. 680–687. International Educational Data Mining Society (2025).

https://doi.org/10.5281/zenodo.15870147

1.1 Related work

Previous work has demonstrated how authenticating a learner's identity can extend to behavioural or biometric authentication, typically using a machine learning classifier trained on measurements of the learner's behavioural traits during the online task. Such behavioural biometrics reflect traits that are inherent to learners, and therefore difficult to fake by others. In tasks that involve typing, the frequency and timing information pertaining to particular (sets of) keystrokes, *keystroke dynamics*, can be very informative for distinguishing between learners. Keystroke dynamics have been used to determine whether text input was typed by the same person or not, both with specific phrases [25, 27] and with longer free-written texts [11, 6]. In addition, keystroke features can be used to identify particular task behaviours, like subprocesses in an essay writing task [34]. As such, they can also provide insight into other forms of academic dishonesty, for instance by detecting copying or reproduction from another source [18, 10]. For this kind of application of keystroke dynamics, typing behaviour in a trial or session is generally captured in a vector of features that serves as input for a classifier. Classification by machine learning methods, like neural networks and decision tree-based methods [6, 27, 18, 10], tends to outperform computationally simpler methods (e.g., cosine correlation [11]).

The current study explores the practical application of a method for biometric authentication in a specific context: short-answer online retrieval practice of declarative knowledge. This context poses several practical constraints. In retrieval practice, learners typically provide short answers of one or two words in response to retrieval cues that differ from trial to trial, such as vocabulary in a second language, or the name of a place indicated on a map. No other text input

is collected from learners. As such, there is relatively little keystroke data by which a learner might be identified. In addition, to protect learners' privacy, no additional privacy-sensitive information (e.g., audiovisual data [20]) is available. The retrieval practice task also provides a different kind of higher-level behavioural data that may be used for identifying learners: measures of retrieval performance. There are reliable individual differences in memory performance that can be measured through such a task [30], and in other contexts, performance-related features have been successfully used in detecting abnormal response behaviour [24].

We demonstrate a machine learning approach for learner authentication in the retrieval practice task, using a combination of keystroke features and task performance features (Figure 1), evaluating its performance in a realistic setting.

2. METHODS

2.1 Data

We use data sampled from secondary school students in the Netherlands studying German vocabulary items in *Slim-Stampen*, an online retrieval practice application designed by MemoryLab and made available to students through the educational publisher Noordhoff. Each practice session consists of trials in which the learner sees a prompt on screen (e.g., a Dutch word), types a response into a textbox (e.g., the German translation), and then receives corrective feedback (Figure 1A). Practice sessions are variable in the number of vocabulary items and overall length, but on average contain about 20 trials. All students included in the sample had completed at least 3 sessions and typed at least 10 distinct bigrams (i.e., pairs of letters) in each session. In total, the data contained 131,819 trials across 5,765 sessions from 513 students. We operated under the assumption that there was no identity fraud in this sample, so that the student's unique user identifier could be taken as the ground truth. The data were fully anonymised and contained only a unique, non-identifiable user identifier. As a result, it was not possible to obtain informed consent from individual students. The analysis was conducted in accordance with a data usage agreement between Noordhoff and MemoryLab.

2.2 Features

We extracted two types of trial-level feature from the data: keystroke dynamics and task performance features (see Figure 1B and Figure 3).

2.2.1 *Keystroke dynamics*

From the timing information of individual keystrokes, we computed several common measures capturing aspects of typing speed, including duration or hold time, press-release or flight time, and press-press or inter-key delay (IKD). We also included 95th-percentile IKD as a measure of pausing, and mean absolute deviation in IKD as a measure of speed variability [21]. In addition, we computed bigram-specific median IKDs for the 40 most frequently occurring bigrams. Users may differ widely in their average typing speed for specific bigrams as a result of particular typing habits, such as hand positioning or strongly learned sequences. By capturing information on the timing of keystroke combinations, we therefore hope to allow for more user-specific adaptation. A similar approach has previously been used by e.g., [27],

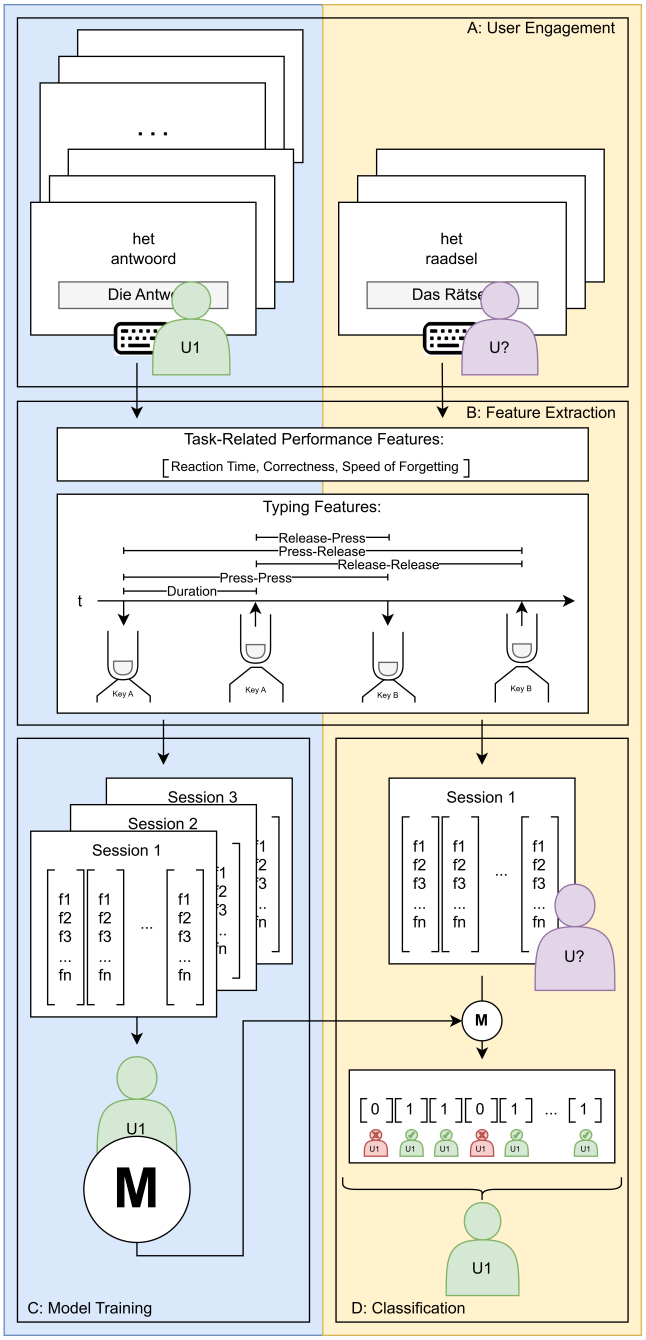

Figure 1: Diagram of the model training and classification process. A. Retrieval practice sessions consist of a sequence of trials in which learners type short answers to retrieval cues. B. Task-related and keystroke-related features are extracted from each response. C. A model is trained to identify the learner *U1* from their responses. D. After training, the model decides whether each new response is from learner *U1*. It also makes a session-level classification based on majority voting.

showing the informative value of bigram-specific data. We chose to include only the 40 most frequently occurring bigrams for considerations of model size and data sparsity. In addition to improving the model's chances of picking

up on user-specific patterns in relatively rare bigrams, we expect that including more bigram-specific timing features will strengthen performance of the model when applied to practice sessions of different languages and materials. For example, a fairly common English bigram like *U S* appears relatively infrequently in the current response data, but could become an important feature when the same Dutch student practising German also starts practising English.

2.2.2 Retrieval practice performance

Trial-level performance on the retrieval practice task was measured in terms of response accuracy (correct or incorrect), response time (duration from trial onset to the first keypress of the answer), and *Speed of Forgetting*, a continuous parameter estimated by the learning system that captures the difficulty of a specific item for a specific learner (details on how this parameter is estimated are provided in [33]). While the *Speed of Forgetting* parameter is specific to the current application, other adaptive learning systems include comparable parameters to capture individual differences in ability or difficulty that could similarly serve as performance-related features in a classifier.

2.3 Procedure

For each user, we trained an individual XGBoost model to classify new trials as being from the same user or from a different user [3]. The XGBoost model is based on the gradient tree boosting framework, which sequentially builds an ensemble of decision trees. It goes beyond methods such as random forests, while benefitting from the same low computational cost, interpretability, and ability to handle inhomogeneous features [12]. In binary classification, each new tree is trained to reduce the classification error by fitting to the gradient of the loss function, based on the predictions of the current ensemble. This process ensures that the model improves iteratively, focusing on hard-to-predict data. XGBoost is particularly suitable for this context, since it can handle missing values and complex, non-linear relationships in the data very well. Furthermore, XGBoost provides built-in regularization, helping to prevent overfitting, which is crucial when dealing with potentially noisy or sparse data [3].

A separate classifier was trained for each student, as illustrated in Figure 1. Retrieval practice trials were grouped by session and each session's trials were assigned to either the training set or the test set. To address class imbalance, the training set was downsampled to contain an even ratio of trials from the same user and trials from other users. For each trial, we extracted task-related performance features and keystroke-related features from the response of the learner (Figure 1B), resulting in a numeric vector of 51 feature values (3 performance features, 8 general keystroke features, and 40 bigram-specific keystroke features; see Figure 3). Since individual bigram-specific features are only available on trials in which the keystroke data contains those bigrams, the feature vector had missing values for absent bigrams.

Classifications were made on the test set at the individual trial level: each trial was classified as being from the same user or from a different user. Since in educational practice, we typically care about integrity at the session level, we also computed a session-level classification using majority voting: if more than half of all trials in a session are classified as being from the same user, the session as a whole is classified as being from the same user.

2.3.1 Performance evaluation

We evaluated classification performance on the test set through sensitivity, specificity, and ROC AUC. Sensitivity and specificity focus on positive or negative errors in classification, while ROC AUC serves as a measure of classification performance irrespective of a fixed decision threshold.

Sensitivity quantifies the probability that a new trial or session from the same learner is correctly identified as being from that learner. A high sensitivity means that the model is unlikely to incorrectly flag new data from the same user as coming from somebody else (i.e., false negatives). For the model to be practically useful, it is important that sensitivity is high, since falsely accusing learners of fraud is problematic.

The classifier's specificity describes the probability that a new trial or session from a different learner is correctly identified as being from a different learner. A high specificity means that the model is likely to catch impersonation of the learner by somebody else (i.e., false positives). In practice, the higher the specificity, the fewer cases of fraud will go unidentified. However, for our purposes, we would rather miss genuine fraud than make false accusations of fraud.

Sensitivity and specificity quantify performance given a default decision threshold of 0.5 (i.e., a higher than 50% estimated probability of a single response being from the same user, or a majority of responses within a session being classified as coming from the same user). The receiver operating characteristic (ROC) curve captures changes in sensitivity and specificity as the decision threshold changes. The area under this curve (AUC) therefore summarises the classifier's overall ability to discriminate between classes, independent of the specific threshold that is chosen.

2.3.2 Model variants

In addition to the full model, which included general and bigram-specific features along with features related to task performance, we also explored several variants with fewer features. To represent a scenario in which there is little to no overlap between specific bigrams in the training set and the testing set (e.g., because the learner is answering in a different language), we fitted a model using only general *bigram-agnostic* keystroke features along with task performance features. Furthermore, we assessed the added value of using learning task performance features by fitting a model with only keystroke features and *no performance* features. Finally, we also fitted a model to only learning task-related features, simulating a scenario in which the learner is answering multiple-choice questions and therefore generating *no keystrokes* at all.

2.4 Code and data

Feature extraction and model fitting and evaluation were done in R [29]. We used the `xgboost` package [4] together with `tidymodels` [22] for modelling, and the `vip` package [15] to visualise variable importance. The R code and feature data are available at `https://github.com/SlimStampen/quantified-learner-dynamics`.

3. RESULTS

3.1 Classification performance

The classification performance of all full models is summarised in Figure 2A and in the first row of Table 1.

3.1.1 Sensitivity

As Figure 2A and Table 1 show, trial-level sensitivity is reasonably high, with a median of 82.8% and an interquartile range of 16.7 percentage points. This means that on average about 1 in every 5 responses will still be incorrectly marked as coming from a different user. Session-level sensitivity is substantially better, with a median of 100%: in the vast majority of cases, there are no false negatives at all.

3.1.2 Specificity

Trial-level specificity is quite good but not perfect: the median of 80.9% means that about 4 out of every 5 responses from other users are identified as such by the classifier. As we would expect, session-level specificity is better, with a median of 89.7%: about 9 out of every 10 sessions from other users are caught.

3.1.3 ROC AUC

The difference in sensitivity and specificity at both the trial-level and the session-level suggests that the classifier is somewhat conservative in both cases, and that a higher threshold might result in better overall performance. The ROC AUC is high when classifying individual trials (median: 89.6%) and even higher when classifying entire sessions (median: 97.1%).

3.1.4 Intermediate summary

Across metrics, we observe a consistent pattern: performance is higher when the classification is made for a session, rather than for an individual response. This aligns with our expectations, both when it comes to correctly identifying the learner and when it comes to detecting anomalous behaviour that could indicate fraud. Since individual responses are short, typically only consisting of one or two words, it is quite possible that a momentary distraction or slip causes the response as a whole to deviate from a learner's norm. Provided that such anomalies occur in a minority of responses within a session, the model can still correctly identify a learner at the session level. Similarly, while individual fraudulent responses might not be distinctive enough to cause suspicion, it is more difficult for entire sessions of anomalous responses to slip through undetected.

3.2 Effect of training set size

Since the training set for the model consists entirely of responses made in the retrieval practice task, training set size can be a limiting factor. We evaluated how the number of available trials from a learner in the training set impacted classification performance. Figure 2B shows that, as expected, performance generally improves as the training set size grows. For session-level classification, improvement appears to reach a plateau once there are around 100 training trials from the target learner. For most learners, that would mean that stable classification performance can already be reached within the first 10 minutes of practice.

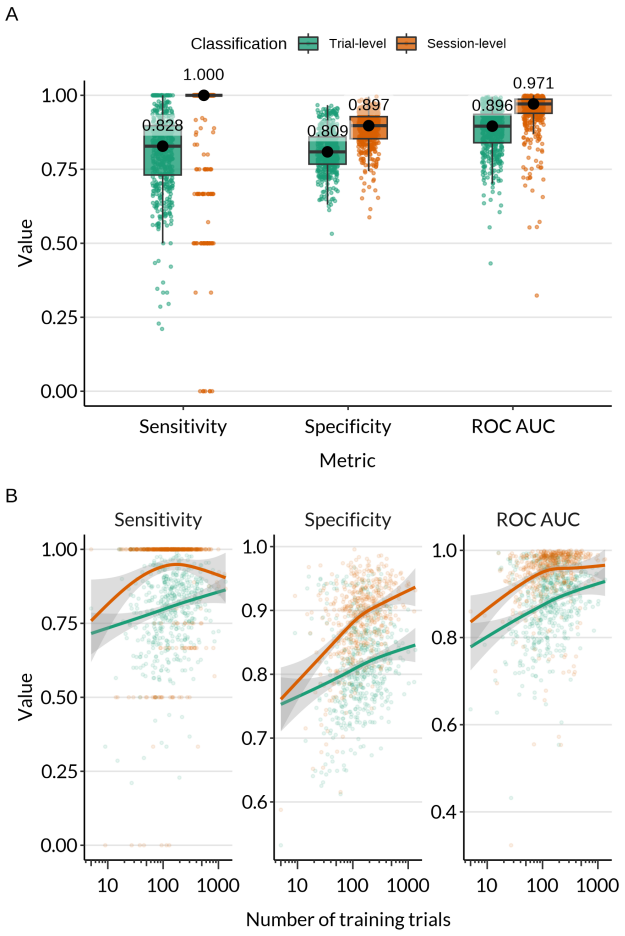

Figure 2: Performance of the XGBoost model, by classification level. Each point represents a separate model trained to identify a single learner. A. Overall performance. Labelled black points show the median. B. Performance as a function of the number of trials from the target learner in the training set. The fitted curves are generalised additive models (GAMs).

3.3 Variable importance

To identify the features with the strongest contribution to the classification, we computed variable importance per feature, based on the total gain of splits in the decision trees associated with that feature. Figure 3 shows how each feature's variable importance is distributed across all of the full models. Many of the most important features are general keystroke features, such as the median hold time and flight time.

In addition, several features related to the retrieval practice task rate quite highly, particularly the *Speed of Forgetting* and reaction time. The *Speed of Forgetting* summarises a learner's overall memory performance and is typically quite a consistent and distinctive feature of a learner [30, 33]. The relatively high importance of this feature in the current classification confirms it to be a useful feature by which to identify a learner in this context, too. Similarly, reaction time can capture individual differences in processing speed (e.g., the time needed to read the retrieval cue on the screen [32]), which makes it a relevant feature in the classification.

Table 1: Median and interquartile range (in parentheses) of performance metrics of the model variants.

Model	Sensitivity		Specificity		ROC AUC	
	Trial-level	Session-level	Trial-level	Session-level	Trial-level	Session-level
Full	.828 (.167)	1.000 (.000)	.809 (.092)	.897 (.075)	.896 (.096)	.971 (.048)
Bigram-agnostic	.795 (.192)	1.000 (.000)	.760 (.114)	.848 (.109)	.848 (.128)	.951 (.069)
No performance	.828 (.163)	1.000 (.000)	.804 (.092)	.894 (.075)	.894 (.098)	.970 (.046)
No keystrokes	.557 (.175)	.600 (.667)	.582 (.075)	.680 (.170)	.597 (.117)	.709 (.286)

Finally, a number of bigram-specific features have relatively high importance, such as the median IKD for the *Space Shift* bigram and for the *E R* and *E N* bigrams. Unlike features based on general keystroke dynamics and task performance, which are available for every response, bigram-specific features are only available if the response happens to contain a given bigram. Higher-importance bigram-specific features tend to be associated with higher-frequency bigrams. While specific low-frequency bigrams might be highly informative in individual cases, as the scattered points in Figure 3 indicate, their overall importance is limited.

3.4 Model variants

The analysis of variable importance indicated that general keystroke features, bigram-specific features and task performance features all contributed to the classification of the full model. Figure 4 compares performance of the full model with performance of models with fewer features.

A model that does not include bigram-specific information still maintains most of the performance of the full model, suggesting that having overlap in the specific bigrams available in the training data and those present in new practice sessions is not critical for learner authentication. Linear mixed-effects regression models fitted to the session-level metrics indicated that, compared to the full model, there was no significant drop in sensitivity ($\beta = -0.027$, $t = -2.039$, $p > .05$), a small but significant drop in specificity ($\beta = -0.048$, $t = -11.474$, $p < .001$), and a small but significant drop in ROC AUC ($\beta = -0.026$, $t = -3.867$, $p < .001$).

Similarly, a model that does not have access to task performance features achieves similar performance to the full model, which suggests that good classification is still possible if performance features are not available (e.g., in cold-start situations when *Speed of Forgetting* estimates for new materials are not yet available [33], or in cases when response times are uninterpretable due to distraction). Linear mixed-effects regression models indicated no significant drop in session-level sensitivity ($\beta = -0.003$; $t = -0.248$, $p > .05$), specificity ($\beta = -0.005$; $t = -1.252$, $p > .05$), or ROC AUC ($\beta = -0.002$; $t = -0.308$, $p > .05$), compared to the full model. While *Speed of Forgetting* and response time were found to be relatively important features in the full model, other features appear to be able to compensate for their omission here.

Finally, a model that does not have access to any keystroke features but bases its classification on learning performance features alone, performs much worse, although still above chance level. Linear mixed-effects regression models confirmed a significant decrease relative to the full model in session-level sensitivity ($\beta = -0.337$; $t = -25.448$, $p < .001$), specificity ($\beta = -0.207$; $t = -50.007$, $p < .001$), and ROC AUC ($\beta = -0.269$; $t = -40.528$, $p < .001$).

Together, these analyses suggest that while all features contain information that is helpful for classification, there is some redundancy among them. Among the different types of features, general keystroke dynamics appear to provide the most important information for learner authentication.

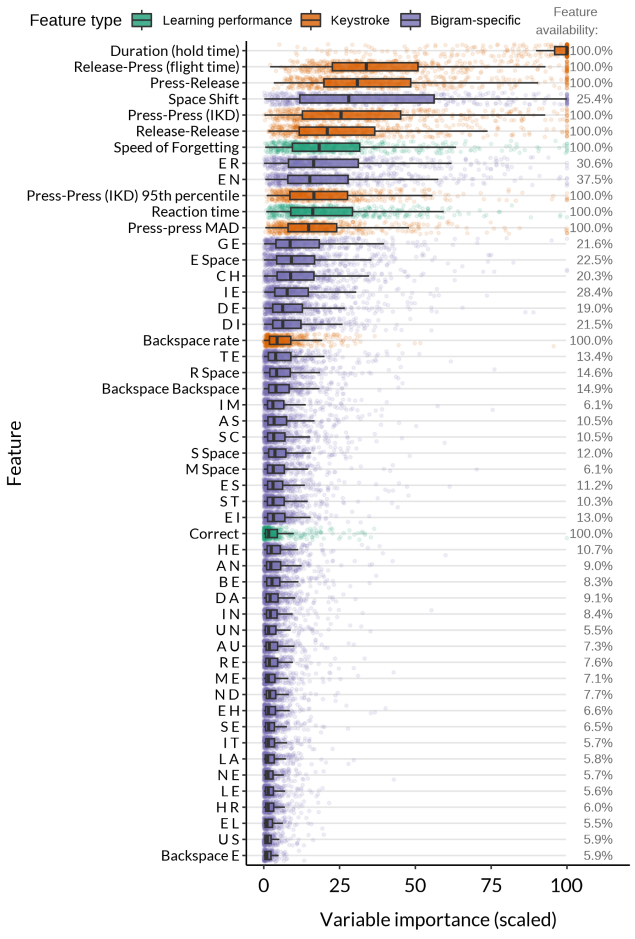

Figure 3: Variable importance across models, coloured by feature type. Each point represents a separate model trained to identify a single learner. The percentage of retrieval practice responses in which a feature is available is listed on the right.

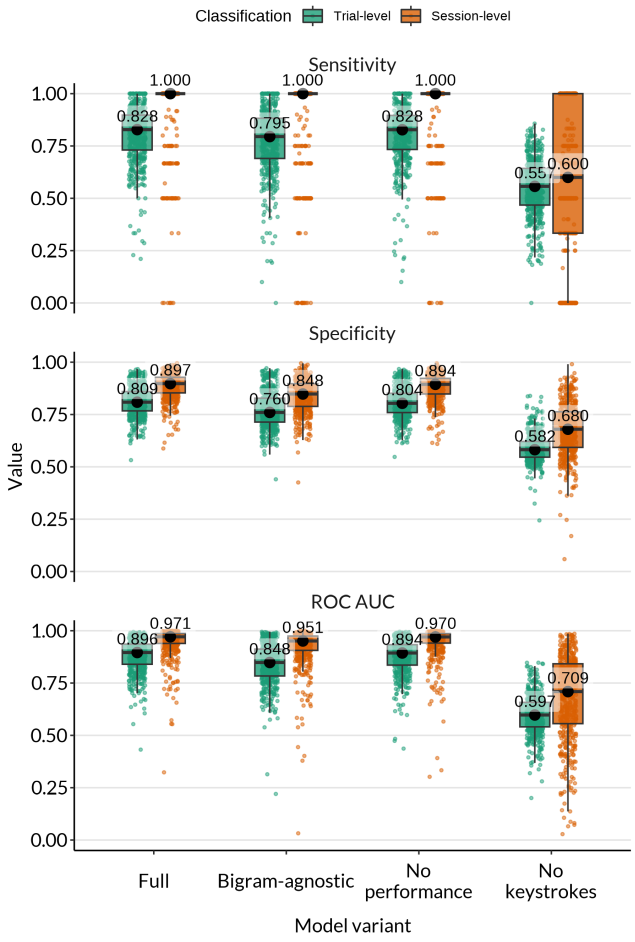

Figure 4: Performance of variants of the XGBoost model with different features, by classification level. Each point represents a separate model trained to identify a single learner. Labelled black points show the median. Full: all features; Bigram-agnostic: no bigram-specific features; No performance: only keystroke features; No keystrokes: only task performance features.

4. DISCUSSION

It is important to emphasise that this method for learner authentication focuses on addressing a particular component of academic fraud. For instance, it is still unclear to what extent a method like this could identify a learner reproducing an answer from another source. While keystroke-based classifiers have been shown to be able to identify such behaviour in longer-form text [18], it remains to be seen whether the same holds for the short-form answers in typical retrieval practice tasks. In general, ensuring integrity in an online learning tool requires a comprehensive approach towards prevention and detection of fraud, of which the current method can be an important component [13].

A factor that was not considered in the current evaluation of the method, but is worth exploring in future work, is that learners' behaviour can change over time. The performance of a classifier that was trained once on typing behaviour in an initial set of practice trials may deteriorate as a learner becomes a more proficient typist [26], or as their fluency in the target language increases [16]. In addition, metrics related to memory performance may also shift as a learner's knowledge of the material changes [8]. To ensure that a classifier maintains its performance over time, it can be periodically retrained as new practice data becomes available [19].

In addition, implementation of the method in practice requires careful consideration of how classifications are communicated to students and/or their instructors, in a way that includes instructors in any decisions informed by automated classifications, and that does not disempower students [7, 31]. While the decision-making process of this kind of model is complex, analysing the contribution of specific features to the classification process (through variable importance, as in Figure 3, or through a method like Shapley Additive Explanations [23]) can help instructors and students understand why a particular classification was made. It is also important to critically evaluate the fairness and possible biases of such an automated method [18], to ensure that application of the method does not disproportionally affect specific (groups of) students.

Our findings suggest that the current approach is insufficient for learning or assessment tasks with multiple-choice questions (MCQs). The comparison of model variants revealed that classification performance relies heavily on keystroke dynamics: when the only information available to the model is task performance-related, performance is substantially worse. Future work may explore whether it is possible to engineer additional features specific to a multiple-choice task, such as features related to specific error patterns or response preferences. Alternatively, occasional typing prompts could be inserted into an MCQ task to provide keystroke information [14].

In general, performance of the classifier might be further optimised by including additional features related to clickstream or navigation data throughout the application [17], or features related to cutting, pasting, edits involving jumps through the answer [6]. In addition, since there is a degree of collinearity between features, a dimensionality reduction technique like principal component analysis (PCA) may be used to increase training efficiency. Finally, a comparison to other classification algorithms (e.g., twin neural networks that learn embeddings per user and use a similarity function for classification [5]) would provide insight into the model's relative performance.

5. CONCLUSION

We have demonstrated a method for identifying anomalous behaviour in an online retrieval practice task that may indicate fraud, using an XGBoost classifier trained on both keystroke dynamics and performance features specific to the learning task. The method shows good performance, with high sensitivity and specificity, particularly when classifications are aggregated at the level of a practice session. The classifier requires only a modest amount of training data: about 100 short answer responses, which can normally be obtained from fewer than 10 minutes of practice. In addition, the method appears quite robust to missing features: good classification performance is maintained when bigram-specific features or learning performance features are not

available. These findings mean that the method is suitable for application in practice, enabling online learning and assessment tools in which learners provide short text answers to automatically flag irregular behaviour, while preserving learners' privacy.

6. REFERENCES

[1] A. Amigud, J. Arnedo-Moreno, T. Daradoumis, and A.-E. Guerrero-Roldan. Using Learning Analytics for Preserving Academic Integrity. *The International Review of Research in Open and Distributed Learning*, 18(5), Aug. 2017.

[2] R. Baker, W. J., H. Neil, R. Ido, C. Albert, and K. Kenneth. Why Students Engage in "Gaming the System" Behavior in Interactive Learning Environments. *Journal of Interactive Learning Research*, 19(2):185–224, 2008.

[3] T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, San Francisco California USA, Aug. 2016. ACM.

[4] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, Y. Li, and J. Yuan. Xgboost: Extreme Gradient Boosting. Comprehensive R Archive Network, 2024.

[5] D. Chicco. Siamese Neural Networks: An Overview. In *Artificial Neural Networks*, volume 2190 of *Methods in Molecular Biology*. Humana, New York, NY, 2021.

[6] I. Choi, J. Hao, P. Deane, and M. Zhang. Benchmark Keystroke Biometrics Accuracy From High-Stakes Writing Tasks. *ETS Research Report Series*, 2021(1):1–13, Dec. 2021.

[7] S. Coghlan, T. Miller, and J. Paterson. Good Proctor or "Big Brother"? Ethics of Online Exam Supervision Technologies. *Philosophy & Technology*, 34(4):1581–1606, Dec. 2021.

[8] M. G. Collins, F. Sense, M. Krusmark, and T. Jastrzembski. Modeling Change Points and Performance Variability in Large-Scale Naturalistic Data. In *Proceedings of Virtual MathPsych/ICCM 2023*, 2023.

[9] H. Corrigan-Gibbs, N. Gupta, C. Northcutt, E. Cutrell, and W. Thies. Deterring Cheating in Online Environments. *ACM Transactions on Computer-Human Interaction*, 22(6):1–23, Dec. 2015.

[10] S. Crossley, Y. Tian, J. S. Choi, L. Holmes, and W. Morris. Plagiarism Detection Using Keystroke Logs. In P. Benjamin and D. E. Carrie, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 476—483, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

[11] E. Flior and K. Kowalski. Continuous Biometric User Authentication in Online Examinations. In *2010 Seventh International Conference on Information Technology: New Generations*, pages 488–492, Las Vegas, NV, USA, 2010. IEEE.

[12] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), Oct. 2001.

[13] M. Garg and A. Goel. A systematic literature review on online assessment security: Current challenges and integrity strategies. *Computers & Security*, 113:102544, Feb. 2022.

[14] M. Garg and A. Goel. A comprehensive approach for mitigating impersonation in online assessment: Integrity policy and random authentication. *International Journal of Information Security*, 24(1):1, Feb. 2025.

[15] B. Greenwell, M. and B. Boehmke, C. Variable Importance Plots—An Introduction to the vip Package. *The R Journal*, 12(1):343, 2020.

[16] L. Haake, S. Wallot, M. Tschense, and J. Grabowski. Global temporal typing patterns in foreign language writing: Exploring language proficiency through recurrence quantification analysis (RQA). *Reading and Writing*, 37(2):385–417, Feb. 2024.

[17] J. Hao and M. Fauss. Test Security in Remote Testing Age: Perspectives from Process Data Analytics and AI, Nov. 2024.

[18] Y. Jiang, M. Zhang, J. Hao, P. Deane, and C. Li. Using Keystroke Behavior Patterns to Detect Nonauthentic Texts in Writing Assessments: Evaluating the Fairness of Predictive Models. *Journal of Educational Measurement*, 61(4):571–594, Dec. 2024.

[19] P. Kang, S.-s. Hwang, and S. Cho. Continual Retraining of Keystroke Dynamics Based Authenticator. In S.-W. Lee and S. Z. Li, editors, *Advances in Biometrics*, volume 4642, pages 1203–1211. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[20] F. F. Kharbat and A. S. Abu Daabes. E-proctored exams during the COVID-19 pandemic: A close understanding. *Education and Information Technologies*, 26(6):6589–6605, Nov. 2021.

[21] L. Knol, A. Nagpal, I. E. Leaning, E. Idda, F. Hussain, E. Ning, T. A. Eisenlohr-Moul, C. F. Beckmann, A. F. Marquand, and A. Leow. Smartphone keyboard dynamics predict affect in suicidal ideation. *npj Digital Medicine*, 7(1):54, Mar. 2024.

[22] M. Kuhn and H. Wickham. Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles., 2020.

[23] S. M. Lundberg and S.-I. Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774, Long Beach, CA, 2017.

[24] K. Man, J. R. Harring, Y. Ouyang, and S. L. Thomas. Response Time Based Nonparametric Kullback-Leibler Divergence Measure for Detecting Aberrant Test-Taking Behavior. *International Journal of Testing*, 18(2):155–177, Apr. 2018.

[25] P. K. Mungai and R. Huang. Using keystroke dynamics in a multi-level architecture to protect online examinations from impersonation. In *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, pages 622–627, Beijing, China, Mar. 2017. IEEE.

[26] B. Ngugi, B. K. Kahn, and M. Tremaine. Typing Biometrics: Impact of Human Learning on Performance Quality. *Journal of Data and Information Quality*, 2(2):1–21, Feb. 2011.

[27] K. Nova. Analyzing Keystroke Dynamics for User

Authentication: A Comparative Study of Feature Extractions and Machine Learning Models. *Sage Science Review of Applied Machine Learning*, 5(2):67–80, Nov. 2022.

[28] R. Pelánek and T. Effenberger. Improving Learning Environments: Avoiding Stupidity Perspective. *IEEE Transactions on Learning Technologies*, 15(1):64–77, 2022.

[29] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, 2020.

[30] F. Sense, F. Behrens, R. R. Meijer, and H. van Rijn. An Individual's Rate of Forgetting Is Stable Over Time but Differs Across Materials. *Topics in Cognitive Science*, 8(1):305–321, 2016.

[31] Z. Swiecki, H. Khosravi, G. Chen, R. Martinez-Maldonado, J. M. Lodge, S. Milligan, N. Selwyn, and D. Gašević. Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, 3:100075, 2022.

[32] M. van der Velde, F. Sense, J. P. Borst, L. van Maanen, and H. van Rijn. Capturing Dynamic Performance in a Cognitive Model: Estimating ACT-R Memory Parameters With the Linear Ballistic Accumulator. *Topics in Cognitive Science*, 14(4):889–903, 2022.

[33] M. van der Velde, F. Sense, J. P. Borst, and H. van Rijn. Large-scale evaluation of cold-start mitigation in adaptive fact learning: Knowing "what" matters more than knowing "who". *User Modeling and User-Adapted Interaction*, 34:1467–1491, June 2024.

[34] M. Zhang, H. Guo, and X. Liu. Using Keystroke Analytics to Understand Cognitive Processes during Writing. In *Proceedings of the 11th International Conference on Educational Data Mining*, pages 384–390, Paris, France, 2021.