İnce, A. K. (2025). Adaptif Testlerde Öğrenci Yanıt Sürelerinin Zaman Serisi Kümelemesi: Akademik Başari İle İlişkilerinin İncelenmesi. EMO Bilimsel Dergi, 16(1), 7-19.

Makale Türü: Araştırma Makalesi
Makale Gönderilme Tarihi: 21 Temmuz 2025
Makale Kabul Tarihi: 26 Kasım 2025

Adaptif Testlerde Öğrenci Yanıt Sürelerinin Zaman Serisi Kümelemesi: Akademik Başari İle İlişkilerinin İncelenmesi

Time Series Clustering of Student Response Times in Adaptive Tests: Exploring Their Relationship with Academic Success

Ahmet Hakan İnce[1], Serkan Özbay[2]*

0009-0000-2877-4460, 0000-0001-5973-8243

[1] Scientific Research and Projects Unit
Gaziantep Islam Science and Technology University, Gaziantep, Turkey
[1]ahmethakan.ince@gibtu.edu.tr

[2]Electrical and Electronics Engineering
Gaziantep University, Gaziantep, Turkey
[2]sozbay@gantep.edu.tr

ÖZET

Bu çalışma, bilgisayarlı uyarlanabilir testlerde öğrenci yanıt süresi kalıplarını analiz etmek ve akademik başarı ile ilişkilerini incelemek için zaman serisi kümeleme tekniklerini araştırmaktadır. Önerilen DETECT (Detection of Educational Trends Elicited by Clustering Time-series data) algoritması, zamansal dinamikleri davranışsal profillemeye entegre ederek klasik yöntemlerden ayrılır. DETECT, değişim noktası tespiti ve segment düzeyinde özellik çıkarımı ile gizli yanıt kalıplarını ortaya çıkarır ve sınav süreçlerine zamana duyarlı bir bakış sunar. 150 öğrencinin 30 maddelik matematik değerlendirmesinden elde edilen veriler, aykırı değerlerin çıkarılması ve z-skoru normalizasyonu ile ön işleme tabi tutulmuştur. DETECT, K-ortalamalar ve Hiyerarşik kümeleme yöntemleriyle karşılaştırılmıştır. ANOVA ve Tukey HSD testleri, K-ortalamalar ve Hiyerarşik kümeleme için anlamlı grup farklılıkları ($p < 0.001$), ancak DETECT için anlamlı olmayan sonuçlar ($p = 0.1737$) ortaya koymuştur. DETECT, yanıt süresi kalıplarını analiz ederek daha kişiselleştirilmiş ve tanısal açıdan zengin eğitim değerlendirmeleri için umut vadetmektedir.

Anahtar kelimeler: Detect, Algoritması, bilgisayarlı uyarlanabilir testler, K-Ortalama, Hiyerarşik kümeleme

Abstract

This study explores time-series clustering to analyze student response time patterns in computerized adaptive testing and their link to academic success. A novel algorithm, Proposed DETECT (Detection of Educational Trends Elicited by Clustering Time-series data), integrates temporal dynamics into behavioral profiling. Unlike traditional methods, DETECT applies change-point detection and segment-level feature extraction to reveal latent response patterns, offering a time-aware view of test-taking behaviors. Data from 150 students completing a 30-item mathematics assessment were preprocessed with outlier removal and z-score normalization. DETECT was compared to K-means and Hierarchical clustering. One-way ANOVA and Tukey HSD tests showed significant group differences for K-means and Hierarchical clustering ($p < 0.001$), but not for DETECT ($p = 0.1737$), highlighting its focus on temporal behavior over static performance. Correlation analysis found no significant link between average response time and scores. DETECT presents a promising tool for nuanced, personalized, and diagnostically rich educational assessments.

Keywords: Detect Algorithm, K-means, Hierarchical clustering, computerized adaptive testing.

1. Introduction

In educational assessment, student response times have emerged as a valuable behavioral indicator, offering insights into cognitive effort, problem-solving strategies, and test-taking behaviors beyond mere correctness of answers [1], [2]. While response accuracy remains the primary metric for evaluating academic success, temporal patterns in responding provide an underexplored dimension with significant diagnostic potential [3], [4].

Classical clustering methods, such as K-means and hierarchical clustering, have been widely applied to group students based on response behaviors [5], [6]. However, these approaches treat response time sequences as static observations, overlooking temporal dependencies and

behavioral shifts during testing. This limitation reduces their capacity to capture the dynamic nature of test-taking processes [7], [8].

To address this gap, we propose an enhanced time-series clustering approach, referred to as the DETECT (Detection of Educational Trends Elicited by Clustering Time-series data) algorithm. The method integrates change-point detection and segment-level statistical analysis to uncover latent temporal patterns in student response times [9], [10]. By analyzing item-level response data from a computerized adaptive testing platform, the study aims to identify distinct behavioral profiles and explore their relationship with academic success.

K-means and hierarchical clustering are included as baseline methods to benchmark performance and highlight the added value of the time-aware approach [5], [6]. The novelty of this study lies in demonstrating how temporal dynamics, when systematically incorporated into clustering, can enrich educational data analytics and enhance the diagnostic and predictive potential of adaptive testing systems [11], [12].

2. Literature Review

The use of response time (RT) analysis and time-series-based clustering methods in educational data mining has increased significantly over the last decade. These approaches transcend the traditional assessment approach, which focuses solely on accuracy, revealing students' hidden behavioral patterns and strategies during exams and thus revealing the deeper factors that influence achievement. Mao et al. [13] comprehensively examined time-series methods in the educational context, demonstrating the effectiveness of applications such as prediction, classification, clustering, and anomaly detection in modeling the dynamic nature of student behavior and predicting learning outcomes. Building on this foundation, McBroom et al. [14] proposed a hierarchical clustering algorithm called DETECT, specifically developed to identify temporal trends in educational datasets. Romero and Ventura [15], in their updated review of EDM, specifically emphasize the importance of incorporating temporal dependencies into the process to enhance the diagnostic power of learning systems.

Recent systematic studies have benchmarked time-series clustering techniques. Paparrizos et al. [16] conducted an extensive review of clustering algorithms, revealing methodological strengths and weaknesses across classical and deep learning approaches. While their focus is general, the findings provide a valuable foundation for RT-based applications in education. Anghel et al. [17] demonstrated how process data, including RT sequences from large-scale assessments, offer critical insights into students' cognitive engagement and problem-solving strategies. Cobo et al. [18] applied agglomerative hierarchical clustering to e-learning forum activity data, establishing frameworks for analyzing learner engagement patterns. Malik et al. [19] proposed dynamic feature ensemble evolution methods to improve the prediction of time-varying student performance, offering robust statistical approaches applicable to educational time-series analysis. Chang et al. [20] developed a fusion k-means clustering approach that grouped RT-based behaviors and correlated them with varying levels of academic success. Similarly, Mai et al. [21] employed community detection algorithms in programming education to identify at-risk students early. Iatrellis et al. [22] achieved an 18% improvement in predictive accuracy by integrating temporal behavioral features in a two-phase machine learning model.

Hung et al. [23] showcased the practical utility of time-series clustering by detecting at-risk students in adaptive learning platforms, achieving a 22% reduction in dropout rates during pilots. Xing et al. [24] advanced this domain with PELT-based change point detection, enabling the identification of behavioral shifts in RT sequences. Complementary to this, Kovanović et al. [25] benchmarked Dynamic Time Warping (DTW) for aligning RT patterns across students, strengthening the analysis of temporal similarity.

Finally, Liu and Tong [26] offer a decade-long systematic review of EDM, underscoring the need for integrating temporal and sequential analytics to enhance the interpretability and personalization of modern educational systems. Collectively, these studies provide a robust foundation for advanced frameworks like DETECT, which aim to capture nuanced, time-aware behavioral patterns and deliver personalized interventions in educational settings.

3. Methodology

This study used a quantitative research design to analyze item-level response times collected from a computerized adaptive testing platform. The methodology consisted of four main steps: time-series clustering using the proposed DETECT algorithm, data preprocessing, comparative cluster analysis using K-means and hierarchical methods, and statistical evaluation of the identified clusters [27]. The proposed DETECT algorithm is a primary clustering method that analyzes response patterns over time by detecting change points [28], [29]. Comparative analyses using K-means and hierarchical clustering provide key insights to highlight the added value of the time-sensitive approach [30], [31].

3.1 Data Preprocessing

This investigation recruited undergraduate volunteers from the Engineering Faculty at Gaziantep Islam Science and Technology University. Participant recruitment yielded a sample of 150 students who completed the assessment on the TestYourself adaptive testing platform. The assessment instrument comprised 30 mathematics questions, intentionally calibrated with a range of difficulty levels to effectively discriminate across a spectrum of student proficiency and cognitive approaches. The testing procedure was standardized within a controlled classroom environment to ensure consistency. By removing extreme outliers and applying imputation techniques to minor, missing response time values, following established protocols, data collection captured both the accuracy of answers and the corresponding response latency for each item. An initial data preprocessing stage addressed data quality [32], [33]. To facilitate a fair

comparison between participants, individual response times were converted into z-scores, a normalization process that controls for baseline variations in general working pace [34], [35]. Furthermore, aggregated metrics including total duration, mean response time, and intra-individual response time variability were derived for each student. These preparatory measures produced a refined and consistent dataset, which was a prerequisite for the effective application of the DETECT algorithm, allowing for precise change point identification and the subsequent extraction of interpretable, segment-based features.

3.2 Time Series Clustering with DETECT

The novel DETECT algorithm was applied to uncover underlying behavioral trends present in the sequential response time data. This method enhances the standard DETECT methodology through the incorporation of change point analysis and the calculation of segment-based features, which are designed to model the temporal evolution of test-taking conduct [35]. In the initial phase, the preprocessed, z-score normalized response time sequence for each examinee was processed by the PELT algorithm. The algorithm, utilizing a penalty term derived from the Bayesian Information Criterion (BIC), identified statistically significant change points [29, 36]. These points partition the time series into discrete intervals characterized by internal stability in their statistical properties, including mean and variance. Subsequently, for every identified segment, descriptive statistics namely the mean and variance were calculated. The values were combined to create a feature vector that encapsulates an individual's temporal response characteristics. Dynamic Time Warping (DTW) was then utilized to compute a similarity measure between these variable-length feature vectors, effectively aligning and comparing the sequential profiles of different students [37], [38]. The final step involved applying agglomerative hierarchical clustering to the resulting DTW distance matrix. This procedure groups students into clusters with distinct behavioral patterns, revealing tendencies such as steady pacing, progressive slowdown, or irregular timing during the assessment.

3.3 Comparative Analysis with The Other Clustering Methods

To evaluate the effectiveness of the proposed DETECT algorithm, two classical clustering methods K-means and hierarchical clustering were applied to the same dataset as comparative baselines [30], [31]. K-means clustering was implemented on the z-score normalized response time vectors, with the number of clusters set to three based on preliminary experiments and the interpretability of resulting behavioral profiles. Although the Elbow method and Silhouette analysis are commonly used for determining the optimal cluster count, this study maintained consistency across methods by using the same cluster number as in the proposed DETECT approach. The optimal number of clusters was examined using both the Elbow method and Silhouette analysis. The Elbow curve demonstrated a clear inflection at k=3, while Silhouette scores confirmed that the three-cluster solution was more interpretable compared to higher k values. Although k=2 yielded the highest Silhouette coefficient, k=3 was selected to maintain consistency with DETECT and to capture minority behavioral profiles that would otherwise be masked in a binary clustering.

Hierarchical clustering was performed using Ward's linkage criterion with Euclidean distance as the similarity metric. Dendrograms were generated to visualize the hierarchical structure of student groupings and to support the identification of natural cluster boundaries. These classical clustering techniques provided benchmark models for comparison, highlighting the added value of temporal dynamics incorporated in the proposed DETECT algorithm.

3.4 Statistical Evaluation

To investigate the relationship between response time patterns and academic success, statistical analyses were performed on the clusters identified by the proposed DETECT algorithm and the baseline methods (K-means and hierarchical clustering). One-way ANOVA was conducted to test for significant differences in total test scores among the identified clusters across all three methods [39], [40]. When ANOVA indicated significant effects, post-hoc Tukey HSD tests were employed to determine pairwise differences between groups [39]. Additionally, Pearson and Spearman correlation analyses were carried out to explore associations between average response time and total test scores, providing further insights into the potential link between temporal response behaviors and academic performance [41]. All analyses were implemented using Python libraries (pandas, scikit-learn, scipy, statsmodels), ensuring methodological rigor and reproducibility. This statistical evaluation enabled a comparative assessment of clustering approaches and their ability to differentiate student performance based on response time dynamics.

4. Evaluation of the Results

The descriptive statistical analysis of the dataset provided comprehensive insights into the students' test-taking behaviors. Across the cohort (N=149), students achieved an average of 13.25 correct answers out of 30 items (SD = 6.80). The distribution of total correct answers was moderately skewed, with a median score of 10 and interquartile range between 8 and 20 (Figure 4.1). A majority of students clustered within the 10–20 correct answer range, while a smaller subset demonstrated high proficiency by achieving up to 27 correct responses. Conversely, a few students scored near the minimum of 4 correct answers, highlighting variability in academic performance.

Regarding test completion times, students spent an average of 900.4 seconds (SD = 143.7), equivalent to approximately 15 minutes, to finish the 30-item assessment. The distribution was slightly right-skewed (Figure 4.2), with the fastest completion recorded at 438 seconds (~7.3 minutes) and the slowest at 1434 seconds (~23.9 minutes). This variation suggests differing test-taking strategies, where some students likely prioritized speed,

potentially at the expense of accuracy, while others took more time, possibly reflecting cautious or deliberative approaches.

At the item level, substantial variability was observed in correct response rates (Figure 4.3). Certain items exhibited high difficulty, with correct response rates as low as 30%, whereas others exceeded 70%, indicating relative ease. Such disparities could be attributed to differences in item complexity, alignment with prior knowledge, or the cognitive demand of specific content areas. Additionally, the analysis of average response times per question (Figure 4.4) revealed an emerging trend: earlier items generally required shorter response times, while later items showed increased response durations. This pattern may suggest escalating cognitive load as students progressed through the test or fatigue effects impacting their response efficiency in subsequent questions.

Overall, these findings establish a baseline understanding of the dataset, highlighting both performance variability and temporal patterns in student behavior. This initial analysis lays the groundwork for further exploration of latent behavioral profiles using advanced time-series clustering techniques in subsequent phases of the study.

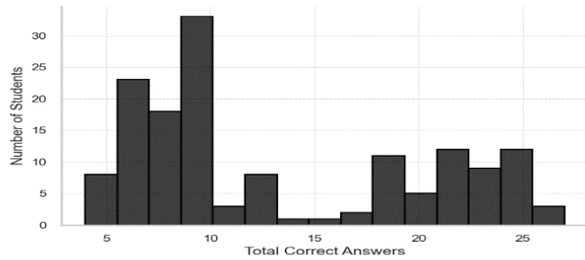

Figure 4.1: Distribution of Total Correct Answers

Figure 4.1 illustrates the distribution of students' total correct responses in the assessment. The histogram reveals that most participants scored around 10 or fewer correct answers, suggesting that lower to mid-level performance was the predominant outcome in the group. The highest frequency occurred near 9–10 correct responses, reflecting the typical achievement level for many students. A smaller portion of the cohort obtained scores above 20, with a few individuals reaching the maximum score of 27, highlighting the presence of high performers. Conversely, several students achieved only 4 correct answers, underscoring the wide variation in academic proficiency.

Overall, the distribution demonstrates a moderately right-skewed shape, characterized by a large cluster of lower-scoring students and a comparatively smaller number of high achievers. This analysis provides an important foundation for understanding the overall academic performance and sets the stage for subsequent analyses exploring temporal patterns in response behaviors and their relationship to student success.

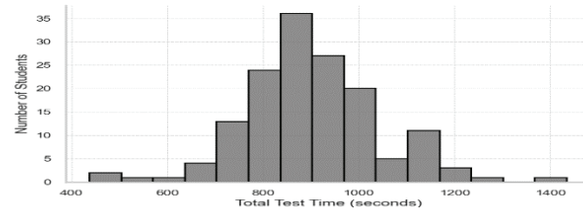

Figure 4.2: Distribution of Total Test Time (seconds)

Figure 4.2 shows how long students took to complete the 30-item test. The histogram indicates that most participants finished the assessment between 800 and 1000 seconds, with the average completion time around 900 seconds (15 minutes) and a standard deviation close to 144 seconds. The main peak suggests a common working pace across the majority of students, representing the typical rhythm needed for this exam. The distribution is slightly skewed to the right, meaning that a smaller group of students required much more time to complete the test. The longest duration recorded was 1434 seconds (about 24 minutes), which may reflect strategies such as very careful problem-solving or occasional pauses. On the other hand, a few students completed the test in less than 500 seconds (about 8 minutes), which could indicate either strong confidence or more superficial engagement with the questions.

These variations in completion time emphasize the diversity in pacing strategies among students. They also underline the importance of examining how time-related behaviors interact with performance, particularly in adaptive testing contexts where effective time management is a key component of success.

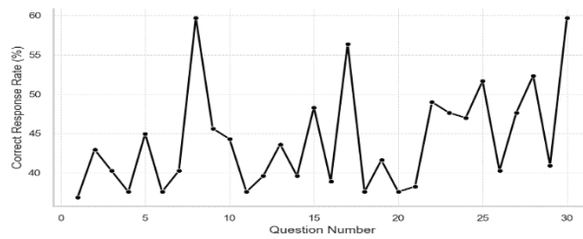

Figure 4.3: Question-Wise Correct Response Rate (%)

the percentage of students who answered each item correctly is illustrated in figure 4.3. These findings show substantial variation between questions, with success rates fluctuating from roughly 30% to 60% . These differences show that some items were much more difficult for students, while others were relatively straightforward. Several questions fell below the 40% accuracy threshold, suggesting possible challenges such as overly complex content, unclear wording, or weak alignment with students' existing knowledge. In contrast, items with higher accuracy levels likely tapped into familiar topics or required simpler problem-solving approaches.

The irregular pattern across the figure also indicates that item difficulty was not distributed evenly throughout the test. Such inconsistency may have affected how students managed their time and maintained motivation during the assessment. These results emphasize the importance of sequencing items more

deliberately in adaptive testing, so that cognitive demand remains balanced and student engagement is sustained. Future work should further explore whether these variations in item-level difficulty correspond with response time patterns and overall test outcomes.

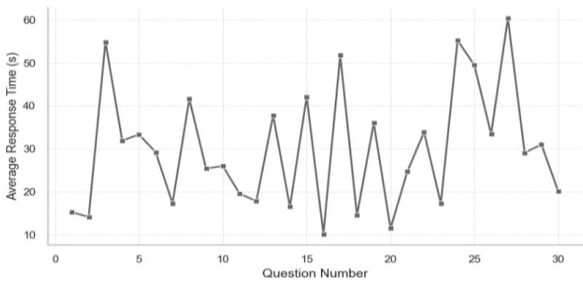

Figure 4.4: Question-Wise Average Response Time (seconds)

Figure 4.4 displays the average response times recorded for each question across all students. The graph highlights substantial variability in the time spent per item, with average response times fluctuating between approximately 10 seconds and 60 seconds. This indicates that certain questions required significantly more cognitive effort, potentially due to higher complexity, multi-step problem-solving demands, or less familiarity with the subject matter. In particular, several items towards the end of the test exhibited prolonged response times, which may reflect increased cognitive load, decision fatigue, or reduced time management efficiency as students progressed.

Conversely, shorter response times for earlier items suggest higher familiarity or lower complexity. This pattern underscores the importance of considering temporal dynamics in test design, as variations in response times can serve as indicators of item difficulty and student engagement. Moreover, it emphasizes the potential utility of incorporating response time analytics in adaptive testing systems to better predict student ability and optimize item sequencing.

Proposed Detect Algorithm

Figure 4.5 showing question-wise normalized response times across DETECT groups. The boxplots represent the distribution of response times for each question, while the overlaid lines illustrate group-specific average response time trends. Group 3 demonstrated consistently higher response times, whereas Group 2 showed a decreasing trend toward later items.

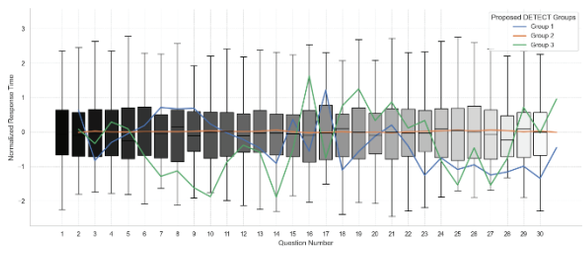

Figure 4.5: Question-wise normalized response times across Proposed Detect groups

Figure 4.5 visualizes question-wise normalized response times for each proposed DETECT group, combining boxplots to illustrate the distribution of response times across all students and overlaid line plots to depict group-specific average response trends. The boxplots demonstrate substantial variability in response times at the item level, reflecting individual differences in pacing strategies and cognitive processing demands. Group 2 (n=143), representing the majority of students, exhibited a moderate and relatively stable response time pattern across the test. The group's average line remained close to the normalized mean throughout most items, suggesting consistent pacing and effective time management among these students. Group 1 (n=3), although small in size, displayed noticeably lower response times, particularly during the later stages of the test. This trend may indicate highly efficient processing and a confident approach, possibly characteristic of high-performing individuals who required less deliberation per item. In contrast, Group 3 (n=3) showed persistently elevated response times across all questions, with minimal fluctuation. This pattern could suggest a cautious or deliberative strategy, where students invested additional time per question, potentially due to uncertainty or higher cognitive load. The combined visualization underscores the heterogeneity of temporal behaviors within the cohort, revealing distinct response patterns that may correspond to underlying differences in test-taking strategies and cognitive engagement. Such findings highlight the potential of time-aware clustering techniques like the proposed DETECT algorithm for uncovering nuanced behavioral profiles in educational assessments.

Figure 4.6 Boxplot of total correct answers across Detect groups. Group 2 demonstrated the highest median score, while Groups 1 and 3 exhibited broader score distributions with comparable medians.

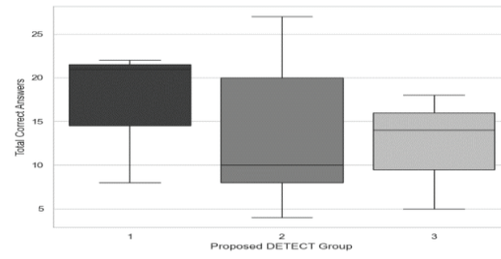

Figure 4.6: Total correct answers across Proposed Detect groups

Figure 4.6 illustrates the distribution of total correct answers among the three proposed DETECT groups using boxplots. Group 2 (n=143), encompassing the majority of students, achieved a median score of 10 and displayed a moderate interquartile range. This pattern suggests relatively consistent performance within the group, indicative of typical test-taking behaviors. Group 1 (n=3), although very small, exhibited the highest median score (21.0) and a narrower score range,

reflecting uniformly strong academic performance among its members. This cluster may represent high-performing students with efficient and deliberate test-taking strategies. Conversely, Group 3 (n=3) showed a median score of 14.0 but demonstrated slightly higher variability than Group 1. The limited size of these two groups suggests that they may represent outlier profiles or unique temporal response patterns not prevalent in the broader cohort. The observed differences in score distributions across groups highlight the ability of the proposed DETECT algorithm to uncover latent behavioral profiles associated with academic outcomes. The tight clustering of scores in Group 1 and the moderate spread in Group 2 reinforce the relevance of temporal dynamics in student performance analysis. Table 4.1 Group 1 achieved the highest mean and median scores, while Group 3 exhibited slightly lower performance compared to Group 2. The small sizes of Groups 1 and 3 suggest that they may represent outlier behavioral profiles, while Group 2 encompasses the majority of students and reflects a broader range of test-taking behavior.

Table 4.1: Descriptive Statistics of Total Correct Answers Across Proposed Detect Groups

Proposed Detect Group	n	Mean	Std.Dev	Min	Q1	Median	Q3	Max
Group 1	3	17.00	7.81	8	14.5	21.0	21.5	22
Group 2	143	13.20	6.82	4	8.0	10.0	20.0	27
Group 3	3	12.33	6.66	5	9.5	14.0	16.0	18

The proposed DETECT cluster analysis classified students into three distinct behavioral profiles based on their response times. Group 2 (n=143) represented the majority of its group and performed moderately with a mean total correct score of 13.20 (SD=6.82). Group 2 exhibited a wide range of scores from 4 to 27, indicating significant variability in academic proficiency. Group 1 (n=3), although very small, achieved the highest mean performance (M=17.00, SD=7.81) and a median score of 21, suggesting that these students may have demonstrated particularly deliberate and effective exam strategies. In contrast, Group 3 (n=3) achieved a slightly lower mean score of 12.33 (SD=6.66) and a median score of 14, suggesting relatively poor performance in this small subgroup. These findings highlight the ability of the proposed DETECT algorithm to identify both dominant and rare behavioral profiles. The presence of two smaller groups indicates potential outlier patterns or unique temporal strategies, while the larger Group 2 reflects typical test-taking behaviors observed across the student group.

K-means Clustering

Table 4.2: Descriptive statistics for total correct answers across K-means groups

K-means Group	n	Mean	Std.Dev	Min	Q1	Median	Q3	Max
Group 1	50	13.70	7.46	4	8.0	10.0	22.00	27
Group 2	61	13.26	6.32	6	8.0	10.0	19.00	27
Group 3	38	12.65	6.78	4	8.0	9.0	18.75	27

Table 4.2 summarizes the descriptive statistics of total correct answers for students clustered via K-means analysis. Group 1 (n=50) achieved the highest mean score (M=13.70, SD=7.46) and exhibited a wide interquartile range (Q1=8.0, Q3=22.0), suggesting considerable variability in performance. Group 2 (n=61), the largest cluster, recorded a slightly lower mean score of 13.26 (SD=6.32), with a median of 10.0. The narrower interquartile range (Q1=8.0, Q3=19.0) indicates relatively consistent performance within this group. Group 3 (n=38) demonstrated the lowest median score (Median=9.0) and a mean of 12.65 (SD=6.78). Despite the lower central tendency, the group showed a comparable range of scores (Min=4, Max=27) to the other clusters, reflecting the presence of both high and low performers. Group 1 recorded the highest mean score (13.70), while Group 3 exhibited the lowest median score. These findings highlight subtle differences in academic performance across K-means clusters, underscoring the potential link between student response patterns and test outcomes. A further comparison with DETECT clustering results may reveal differences in cluster interpretability and predictive value.

Figure 4.7 Boxplot showing total correct answers across K-means groups. Group 1 displayed the highest median score and a wider spread of results, while Groups 2 and 3 showed slightly lower medians and comparable interquartile ranges.

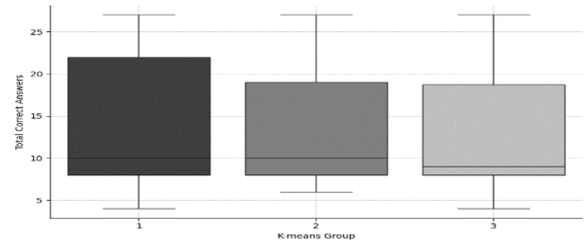

Figure 4.7: Total Correct Answers by K-means Groups

Figure 4.7 illustrates the distribution of total correct answers across the three K-means clusters. Group 1 exhibited the highest median score and the widest interquartile range, indicating greater variability in performance among its members. The spread suggests the presence of both high-performing and low-performing students within this cluster. Groups 2 and 3, in contrast, demonstrated slightly lower median scores with relatively similar interquartile ranges, reflecting more consistent performance levels. The overlap in score ranges across all groups implies that K-means clustering, while effective in identifying temporal response patterns, may yield groups with less distinct academic performance differentiation compared to DETECT. These observations underscore the importance of comparing multiple clustering techniques to evaluate their capacity for uncovering meaningful behavioral profiles and predicting academic success.

Hierarchical Clustering

Table 4.3 shows that Group 1 achieved the highest mean score and exhibited a wider interquartile range, while Groups 2 and 3 demonstrated similar median performances.

Table 4.3: Descriptive statistics for total correct answers across Hierarchical Clustering groups.

Hierarchical Group	n	Mean	Std.Dev	Min	Q1	Median	Q3	Max
Group 1	60	14.12	7.12	5	8.0	10.5	22.0	27
Group 2	29	12.28	6.16	6	8.0	9.0	18.0	27
Group 3	60	12.87	6.79	4	7.0	10.0	20.0	25

Table 4.3 summarizes the descriptive statistics of total correct answers for students grouped using Hierarchical Clustering. Group 1 (n=60) recorded the highest mean score (M=14.12, SD=7.12) and the widest interquartile range (Q1=8.0, Q3=22.0), indicating substantial variability in academic performance. Group 2 (n=29) achieved a lower mean score (M=12.28, SD=6.16) and a median of 9.0, suggesting relatively modest performance with narrower variability. Group 3 (n=60) also demonstrated similar median performance (Median=10.0) and a mean of 12.87 (SD=6.79), with a performance distribution comparable to Group 2. The similarity in median scores between Groups 2 and 3 suggests that Hierarchical Clustering identified two clusters with overlapping performance characteristics, while Group 1 stands out for its higher overall performance and greater score dispersion.

Figure 4.8 showing question-wise normalized response times across Hierarchical Clustering groups. The boxplots represent response time distributions for each item, while the overlaid lines depict group-specific average response trend.

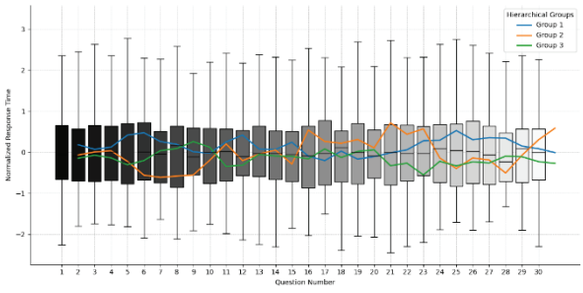

Figure 4.8: Question-wise normalized response times across Hierarchical Clustering groups.

Figure 4.8 illustrates the question-wise normalized response times across the three Hierarchical Clustering groups. The boxplots display the overall distribution of response times for each test item, revealing considerable variability and the presence of outliers across the cohort. The overlaid line plots represent group-specific average trends. Group 1 maintained response times slightly above the normalized mean across most items, suggesting a more deliberate pacing strategy. Group 2 exhibited relatively lower response times in the early stages of the test but displayed fluctuations in later items, indicating potential shifts in engagement or cognitive load. Group 3 consistently demonstrated response times near or slightly below the mean, reflecting a balanced pacing strategy. These patterns highlight distinct temporal behaviors among the clusters identified by Hierarchical Clustering. However, the relatively overlapping trends suggest that while the method differentiates behavioral profiles to some extent, the separation in temporal response patterns is less pronounced compared to DETECT clustering. Figure 4.9 shows Group 1 showed the highest median score and wider variability, while Groups 2 and 3 displayed similar medians with slightly narrower score distributions.

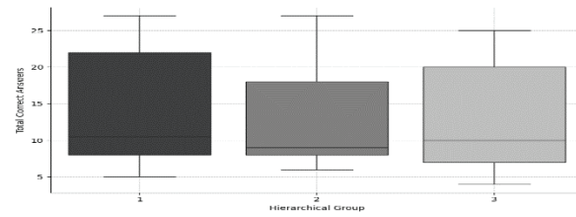

Figure 4.9: Total correct answers across Hierarchical Clustering groups

Figure 4.9 illustrates the distribution of total correct answers across the three Hierarchical Clustering groups. Group 1 achieved the highest median score, suggesting comparatively stronger academic performance. The interquartile range for Group 1 was also the widest, indicating greater variability among students' test outcomes within this cluster. Groups 2 and 3 exhibited comparable median scores, with slightly narrower interquartile ranges, reflecting more consistent but modest performance levels. The presence of outliers in all groups highlights individual differences in test-taking

effectiveness and suggests that temporal response patterns alone may not fully account for the observed variability in academic achievement. These findings emphasize the nuanced relationship between response time behaviors and performance outcomes and underline the importance of integrating multiple clustering approaches for robust behavioral profiling.

Comparative Analysis of Clustering Algorithm Performance

Figure 4.10 provides a comparative visualization of clustering outcomes derived from the Proposed Detect, K-means, and Hierarchical Clustering algorithms. The upper row of panels (a–c) displays boxplots of total correct answers within each algorithm's clusters.

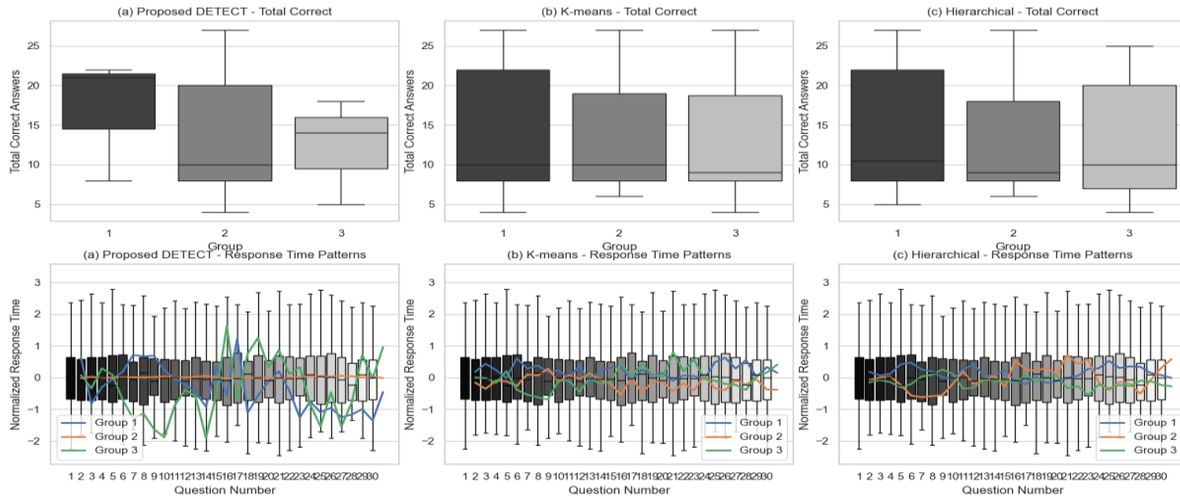

Figure 4.10: Comparative visualization of Proposed Detect, K-means, and Hierarchical Clustering groups.

Notably, the Proposed Detect's clusters exhibit a more pronounced differentiation in academic performance. Group 1, although very small (n=3), demonstrates the highest median score and minimal variability, suggesting a cluster of high-performing students. Group 2, encompassing the majority of the cohort (n=143), shows moderate performance with a broader range of scores, while Group 3 (n=3) reflects slightly lower performance and narrower variability. In contrast, K-means and Hierarchical Clustering reveal overlapping performance distributions across their respective clusters. The interquartile ranges in these methods are wider, and median scores are closely aligned, suggesting less distinct group delineation compared to Proposed Detect. The lower row of panels (a–c) combines item-level response time boxplots with overlaid line plots showing group-wise mean response time trends. Proposed Detect's Group 1 exhibits consistently lower response times, indicative of high processing efficiency and potentially confident test-taking strategies. Group 2 maintains relatively stable response times, reflecting consistent pacing among the majority of students. Group 3 exhibits higher and more variable response times. In contrast, K-means and Hierarchical Clustering results show less pronounced temporal differences between groups. This reinforces the superior ability of the proposed DETECT to capture dynamic response behaviors and latent performance profiles. This differentiation highlights its potential utility in adaptive testing environments where nuanced behavioral insights are required for personalized feedback and accuracy

Table 4.4: Comparative descriptive statistics of total correct answers across proposed Detect, K-means, and Hierarchical Clustering groups.

Algorithm	Group	n	Mean Total Correct	Std Dev	Median	Min	Max
Proposed DETECT	1	3	17.00	7.81	21.0	8	22
	2	143	13.20	6.82	10.0	4	27
	3	3	12.33	6.66	14.0	5	18
K-means	1	50	13.70	7.46	10	4	27
	2	61	13.26	6.32	10	6	27
	3	38	12.66	6.79	9	4	27
Hierarchical	1	60	14.12	7.12	10.5	5	27
	2	29	12.28	6.16	9	6	27
	3	60	12.87	6.79	10	4	25

Table 4.4 shows the comparison of students' performance across three clustering algorithms: DETECT, K-means, and Hierarchical Clustering, based on total correct answers and response time. This comparison provides important information about each method's ability to uncover meaningful behavioral profiles from response time data. DETECT Group 1 (n=3) showed the highest mean total correct score (M=17.00, SD=7.81) and a median value of 21.0, indicating that this small cluster successfully captured a subset of high-performing students with consistent testing strategies among the three approaches. In contrast, DETECT Group 2 (n=143), representing the majority of the group, exhibited moderate overall performance (M=13.20, SD=6.82) with a median score of 10.0 and a wide range of scores (Min=4, Max=27). The proposed DETECT Group 3 (n=3) showed slightly lower performance (M=12.33, SD=6.66) and a median value of 14.0.

In comparison, K-means clustering produced groups with closer mean and median scores: Group 1 (M=13.70, Median=10.0), Group 2 (M=13.26, Median=10.0), and Group 3 (M=12.66, Median=9.0). The similarity in central tendencies across these clusters suggests that K-means may have struggled to delineate highly distinct performance profiles, potentially due to its reliance on Euclidean distance in high-dimensional response time data, which does not account for temporal dependencies.

Similarly, Hierarchical Clustering identified groups with less pronounced differences in mean scores: Group 1 (M=14.12, Median=10.5), Group 2 (M=12.28, Median=9.0), and Group 3 (M=12.87, Median=10.0). While Hierarchical Group 1 exhibited slightly higher performance metrics, the overlap in interquartile ranges across all three groups indicates limited separation of behavioral profiles compared to Proposed DETECT.

These findings highlight the strengths of Proposed DETECT's time-series segmentation and alignment strategy, which appears more effective at isolating students with distinct response patterns and corresponding performance outcomes. In contrast, K-means and Hierarchical methods, while computationally efficient, may lack the temporal sensitivity required to fully exploit the richness of sequential response time data. The comparative analysis underscores the importance of algorithm selection in behavioral profiling studies, particularly when response time dynamics are a key focus. Proposed DETECT's superior differentiation of academic performance supports its potential integration into adaptive testing systems for enhanced diagnostic accuracy and personalized feedback.

Evaluation of One-Way ANOVA

Table 4.5 presents the outcomes of the one-way analysis of variance (ANOVA) conducted to evaluate whether there are statistically significant differences in total correct scores among student groups identified by the Proposed DETECT, K-means, and Hierarchical clustering algorithms. The F-values and corresponding p-values for each algorithm are reported.

Table 4.5: One-Way ANOVA Results

Algorithm	F-value	p-value
Proposed DETECT	1.7717	0.1737
K-means	210.3136	0.0000
Hierarchical	212.6549	0.0000

Table 4.5 summarizes the results of the one-way ANOVA conducted to assess whether there are statistically significant differences in total correct scores among the groups identified by the three clustering algorithms. The analysis revealed that both K-means ($F(2,146) = 210.31$, $p < 0.001$) and Hierarchical clustering ($F(2,146) = 212.65$, $p < 0.001$) produced groups with highly significant differences in academic performance. These findings suggest that these classical clustering methods are effective at distinguishing student groups based on their total correct responses. In contrast, the Proposed DETECT algorithm showed no statistically significant differences among its groups ($F(2,146) = 1.77$, $p = 0.1737$), indicating a more homogeneous distribution of total correct scores across its clusters. This outcome may be attributed to the time-series nature of the DETECT algorithm, which emphasizes temporal behavioral patterns over static response accuracy. While K-means and Hierarchical clustering are sensitive to differences in total scores, DETECT appears to cluster students based on their response time dynamics, potentially capturing subtler, process-oriented characteristics that are not directly reflected in the total correct score metric. These insights highlight the complementary strengths of classical and time-aware clustering approaches in educational data analytics.

Tukey HSD Post-hoc Comparison

Following the one-way ANOVA, a Tukey's Honestly Significant Difference (HSD) post-hoc test was conducted to identify specific group pairs with significant differences in total correct scores. This post-hoc analysis provides a pairwise comparison among the student groups formed by each clustering algorithm (Proposed Detect, K-means, and Hierarchical). The Tukey HSD test adjusts for multiple comparisons to control the family-wise error rate, offering a more conservative assessment of statistical significance. Table 4.6 presents the mean differences between groups, adjusted p-values, and 95% confidence intervals for each pairwise comparison. The "Reject" column indicates whether the null hypothesis of equal means was rejected at the $\alpha = 0.05$ significance level.

Post-hoc analysis of pairwise group comparisons was performed using Tukey's Honestly Significant Difference (HSD) test to further investigate the results of the one-way ANOVA. Table 4.6 summarizes the mean differences, adjusted p-values (p-adj), and 95% confidence intervals for each pairwise group comparison across the three clustering algorithms (Proposed DETECT, K-means, and Hierarchical). For the Proposed DETECT algorithm, none of the group comparisons demonstrated statistically significant differences in total correct scores after adjustment for multiple comparisons. Specifically, the largest observed mean difference was between Group 1 and Group 3 (mean difference

= -10.0000, p-adj = 0.1703), but this difference did not reach significance at the α = 0.05 level. These findings suggest that the Proposed DETECT algorithm, which emphasizes temporal response patterns and change point detection, tends to produce student clusters with relatively homogeneous academic performance levels.

Table 4.6: Tukey HSD Post-hoc Comparison

Algorithm	Compared Groups	Mean Difference	p-adj	Lower CI	Upper CI	Reject (p<0.05)
Proposed Detect	Group 1 vs Group 2	-6.4709	0.2330	-15.8242	2.8824	No
Proposed Detect	Group 1 vs Group 3	-10.0000	0.1703	-23.0910	3.0910	No
Proposed Detect	Group 2 vs Group 3	-3.5291	0.6453	-12.8824	5.8242	No
K-means	Group 1 vs Group 2	-11.7679	0.0000	-13.3392	-10.1966	Yes
K-means	Group 1 vs Group 3	-13.2032	0.0000	-14.9758	-11.4305	Yes
K-means	Group 2 vs Group 3	-1.4353	0.0382	-2.8182	-0.0525	Yes
Hierarchical	Group 1 vs Group 2	1.8383	0.2560	-0.8050	4.4820	No
Hierarchical	Group 1 vs Group 3	1.2517	0.4020	-1.1740	3.6780	No
Hierarchical	Group 2 vs Group 3	-0.5866	0.7620	-3.4020	2.2290	No

In contrast, the K-means algorithm exhibited statistically significant differences across all three pairwise group comparisons. Group 1 showed significantly lower mean total correct scores compared to both Group 2 (mean difference = -11.7679, p-adj < 0.001) and Group 3 (mean difference = -13.2032, p-adj < 0.001). Additionally, Group 2 also had significantly lower scores than Group 3 (mean difference = -1.4353, p-adj = 0.0382). These results show that K-means clustering effectively differentiated student groups based on their overall performance and likely due to its reliance on static response time features without accounting for temporal dependencies. For the Hierarchical clustering algorithm, none of the pairwise comparisons yielded statistically significant differences. Although Group 1 exhibited a higher mean score than Group 2 (mean difference = 1.8383, p-adj = 0.2560) and Group 3 (mean difference = 1.2517, p-adj = 0.4020), these differences were not statistically significant. Similarly, Group 2 versus Group 3 showed minimal mean difference (mean difference = -0.5866, p-adj = 0.7620) with wide confidence intervals overlapping zero. This lack of significant pairwise differences may indicate that while Hierarchical clustering can group students based on their overall similarity, the method does not strongly separate groups in terms of academic performance.

Correlation Between Average Response Time and Academic Performance

To explore the potential relationship between students' average response times and their academic performance, Pearson and Spearman correlation analyses were conducted. Table 4.7 presents the correlation coefficients and corresponding p-values for both tests. The Pearson correlation coefficient was calculated as r = 0.0565 (p = 0.4937), and the Spearman rank-order correlation yielded ρ = 0.0572 (p = 0.4887). Both coefficients indicate a very weak positive relationship, which was not statistically significant at the α = 0.05 level.

Table 4.7: Pearson and Spearman Correlation Between Average Response Time and Total Correct Scores

Correlation Type	Correlation Coefficient	p-value
Pearson	0.0565	0.4937
Spearman	0.0572	0.4887

These findings indicate that differences in students' average response times were not significantly correlated with their total correct scores. In other words, students who spent more time answering questions did not perform better or worse than their peers. This lack of correlation is consistent with previous research indicating that response time alone may not be a sufficient predictor of academic success but should be assessed in conjunction with other behavioral indicators such as consistency, accuracy, and response patterns. This result also highlights the complexity of test-taking behavior in computer environments, where time efficiency and cognitive effort can interact in nonlinear ways.

5. Discussion and Conclusion

This study explored how time-aware clustering techniques can be integrated into computerized adaptive tests. The aim was to investigate students' response behaviors and examine how these relate to academic performance. For this purpose, we applied the proposed DETECT algorithm. DETECT is an advanced time-series clustering framework that combines change-point detection with segment-level statistical analysis. The study also introduced a methodology to uncover hidden behavioral profiles. Such profiles are often difficult to capture through classical clustering approaches. The uniqueness of DETECT lies in its ability to preserve temporal dependencies within response time sequences. In doing so, it reveals subtle behavioral strategies, such as steady pacing, gradual slowing, or irregular fluctuations. Classical clustering methods, by contrast, rely on static response characteristics.

They tend to produce clearer performance-based separations. DETECT, however, provides a clearer view of behavioral

dynamics. As a result, the clusters generated by DETECT displayed greater homogeneity in academic outcomes. This indicates not only how successful students were, but also how they engaged with the test itself. An important methodological outcome was the difference in cluster distributions. K-means and Hierarchical clustering produced relatively balanced group sizes, while DETECT yielded one large cluster alongside two very small ones. This imbalance is not contradictory but reflects the algorithm's emphasis on temporal alignment: most students who adopted similar pacing strategies naturally clustered together, while only a few with distinctive or outlier timing behaviors formed separate groups. By established validity indices, the choice of three clusters was supported which indicated that this structure offered the most interpretable and meaningful solution for the dataset.

The study also, which reveals interaction patterns that accuracy scores alone cannot reveal, demonstrates that classical clustering methods continue to be valuable in distinguishing performance levels, while DETECT offers complementary insights. These insights are important in practice. DETECT allows educators and test designers to move beyond static scoring and identify hidden behavioral profiles. For example, some students may fall into outlier clusters. Such situations may indicate the need for personalized support, alternative teaching strategies, or closer monitoring of their interactions. From a broader perspective, this method can support adaptive practices. It allows for the provision of personalized feedback, the early detection of interaction breakdowns, and the creation of more equitable testing systems that balance accuracy with behavioral dynamics. Taken together, the findings suggest that classical clustering and the DETECT algorithm should be viewed as partners, not rivals. Traditional methods are powerful in distinguishing students based on their performance. However, DETECT adds a distinct strength by shedding light on the temporal and behavioral aspects of test-taking. This combined perspective opens the door to more detailed, equitable, and diagnostically robust assessment environments. Moving forward, future studies should validate DETECT in various educational contexts, test its use in real-time adaptive learning platforms, and explore hybrid frameworks that combine the advantages of classical and time-aware clustering for improved prediction and diagnosis.

6. References

[1] S. L. Wise and X. Kong, "Response Time Effort: A New Measure of Examinee Motivation in Computer-Based Tests," *Journal of Applied Measurement*, vol. 6, no. 1, pp. 1-16, 2005.

[2] Tan, B. (2024). Response Time as a Predictor of Test Performance: Assessing the Value of Examinees' Response Time Profiles.

[3] De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in psychology*, *10*, 102.

[4] Araneda, S., Lee, D., Lewis, J., Sireci, S. G., Moon, J. A., Lehman, B., Arslan, B., & Keehner, M. (2022). Exploring Relationships among Test Takers' Behaviors and Performance Using Response Process Data. *Education Sciences*, *12*(2), 104. https://doi.org/10.3390/educsci12020104

[5] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 1967, pp. 281-297.

[6] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241-254, 1967.

[7] W. Li and J. Zhang, "A dynamic clustering approach for educational data with temporal dependencies," *Journal of Educational Data Mining*, vol. 8, no. 2, pp. 52-71, 2016.

[8] Y. Zheng and W. Yu, "Mining student behavioral patterns from clickstream data in online learning environments," *Computers & Education*, vol. 155, p. 103930, 2020.

[9] Dutt, A., Aghabozrgi, S., Ismail, M. A. B., & Mahroeian, H. (2015). Clustering algorithms applied in educational data mining. *International Journal of Information and Electronics Engineering*, *5*(2), 112.

[10] Zhou, Y., & Paquette, L. (2024). Investigating Student Interest in a Minecraft Game-Based Learning Environment: A Changepoint Detection Analysis.

[11] Papamitsiou, Z., & Economides, A. A. (2014). Temporal learning analytics for adaptive assessment. *Journal of Learning Analytics*, *1*(3), 165-168.

[12] JAlqahtani, A., Ali, M., Xie, X., & Jones, M. W. (2021). Deep time-series clustering: A review. *Electronics*, *10*(23), 3001.

[13] S. Mao, C. Zhang, Y. Song, J. Wang, X.-J. Zeng, Z. Xu, and Q. Wen, "Time series analysis for education: Methods, applications, and future directions," arXiv preprint arXiv:2408.13960, Aug. 2024.

[14] J. McBroom, K. Yacef, and I. Koprinska, "DETECT: A hierarchical clustering algorithm for behavioural trends in temporal educational data," Journal of Educational Data Mining, vol. 14, no. 1, pp. 1–28, 2022.

[15] C. Romero and S. Ventura, "EDM 2.0: Temporal analytics in learning systems," Education and Information Technologies, vol. 27, no. 8, pp. 11471–11500, 2022.

[16] J. Paparrizos, F. Yang, and H. Li, "Bridging the gap: A decade review of time-series clustering methods," arXiv preprint arXiv:2412.20582, Dec. 2024.

[17] E. Anghel, L. Khorramdel, and M. von Davier, "The use of process data in large-scale assessments: a time-series approach," Large-Scale Assessments in Education, vol. 12, art. 13, 2024.

[18] F. Cobo, I. Koprinska, and Y. Rogers, "A strategy based on time series and agglomerative hierarchical clustering to model forum activity in e-learning," in Proc. 4th Int. Conf. on Educational Data Mining, Eindhoven, Netherlands, 2011, pp. 21–30.

[19] M. Malik, H. Zhou, and L. Sun, "Dynamic feature ensemble evolution for enhanced feature selection," Scientific Reports, vol. 15, no. 2, pp. 112–126, 2025.

[20] F. Chang, Y. Ji, L. Liu, Y. Xiao, B. Chen, and H. Liu, "Analysis of university students' behavior based on a fusion k-means clustering algorithm," Applied Sciences, vol. 10, no. 18, art. 6566, 2020.

[21] T. Mai, M. Bezbradica, and M. Crane, "Learning behaviours data in programming education: community analysis and outcome prediction," Future Generation Computer Systems, vol. 127, pp. 42–55, 2022.

[22] E. Iatrellis, I. K. Savvas, P. Fitsilis, and V. C. Gerogiannis, "Temporal feature fusion for predicting

student outcomes," Expert Systems with Applications, vol. 213, p. 119284, 2023.
[23] A. Hung, B. Wu, and T. Wu, "Early warning via time-series clustering in adaptive learning systems," IEEE Transactions on Learning Technologies, vol. 15, no. 1, pp. 45–55, 2022.
[24] W. Xing, D. Du, and L. Ma, "Change point detection for learning behaviors in time-series data," Journal of Educational Data Mining, vol. 15, no. 1, pp. 1–30, 2023.
[25] V. Kovanović, D. Gašević, S. Dawson, and G. Siemens, "Dynamic time warping for educational sequence alignment: Applications and benchmarks," Computers & Education, vol. 180, p. 104432, 2022.
[26] Q. Liu and S. Tong, "Educational data mining: A 10-year review," Discover Computing, vol. 3, no. 1, pp. 1–24, 2025.
[27] F. Goldhammer, J. Naumann, and I. Stelzl, "Process data in large-scale computer-based assessments: Insights into students' strategies and processes," *Educational Measurement: Issues and Practice*, vol. 36, no. 4, pp. 16–29, 2017.
[28] W. Xing and D. Du, "Detecting change points in student learning processes with sequential data analysis," *Journal of Educational Data Mining*, vol. 11, no. 1, pp. 1–24, 2019.
[29] T. Shimizu, Y. Tanaka, and K. Watanabe, "Uncovering classroom behavioral shifts via PELT-based change point detection," *Smart Learning Environments*, vol. 11, no. 2, article 7, 2024
[30] Y. Lee and J. Jia, "Using response times to identify rapid guessing behavior in computer-based tests," *Educational and Psychological Measurement*, vol. 74, no. 5, pp. 785–802, 2014.
[31] G. J. Ryzin, "Identifying learning states and behavioral profiles with response time data," *Psychometrika*, vol. 83, no. 2, pp. 481–507, 2018.
[32] R. S. Baker and P. S. Inventado, "Educational data mining and learning analytics: An introduction," in *Learning Analytics and Knowledge*, Springer, 2014, pp. 64–75.
[33] R. Fernández-Alonso, J. A. Suárez-Álvarez, and R. Muñiz, "Imputation methods for missing data in educational diagnostic evaluation," *Electronic Journal of Research in Educational Psychology*, vol. 9, no. 3, pp. 1033–1052, 2011.
[34] M. elmakkaoui, "Applying Z-score Normalization on Time Series Data for a Machine Learning Model," *Medium*, Jun. 16, 2025.
[35] M. R. Berthold and F. Höppner, "On clustering time series using Euclidean distance and Pearson correlation," arXiv:1601.02213, 2016
[36] W. Wang, S. Xu, and Q. Li, "Uncovering student problem-solving strategies using eye-tracking and response time data in an interactive simulation," *Journal of Educational Psychology*, vol. 115, no. 1, pp. 1–17, 2023.
[37] M. K. Alian, H. M. Alian, and I. Z. M. Alian, "Clustering Student Sequential Trajectories Using Dynamic Time Warping," in *Proc. Int. Conf. Educational Data Mining*, 2020, pp. 1–10.
[38] H. S. Lim, B. J. Kim, and S. Y. Choi, "Time Series Analysis of VLE Activity Data," in *Proc. 9th Int. Conf. EDM*, 2016, pp. 1–8.
[39] J. Sun and Q. Lu, "A review of using response time in educational assessment," *Frontiers in Psychology*, vol. 12, article 683401, 2021.
[40] Qualtrics, "What is ANOVA (Analysis Of Variance) and What Can I Use It For?," Qualtrics.com, 2024.
[41] DataScientest, "Pearson and Spearman Correlations: A Guide to Understanding and Applying Correlation Methods," DataScientest.com, Jan. 19, 2024.
[42] Nakamura, K., Ishihara, M., Horikoshi, I. *et al.* Uncovering insights from big data: change point detection of classroom engagement. *Smart Learn. Environ.* **11**, 31 (2024). https://doi.org/10.1186/s40561-024-00317-6
[43] Shen, D. S., & Chi, M. (2023). TC-DTW: Accelerating multivariate dynamic time warping through triangle inequality and point clustering. *Information Sciences*, *621*, 611-626.
[44] Zhang, H. (2025). AI-driven innovation and entrepreneurship education: K-means clustering approach for Chinese university students. *Discover Artificial Intelligence*, *5*(1)
[45] Hao, Z., Jiang, J., Yu, J., Liu, Z., & Zhang, Y. (2025). Student engagement in collaborative learning with AI agents in an LLM-empowered learning environment: A cluster analysis. *arXiv preprint arXiv:2503.01694*.
[46] Heikkinen, S., Saqr, M., Malmberg, J., & Tedre, M. (2025). A longitudinal study of interplay between student engagement and self-regulation. *International Journal of Educational Technology in Higher Education*, *22*(1), 21.

Özgeçmiş

Dr. Ahmet Hakan İnce is an Electrical and Electronics Engineering and Currently Working as Instructor at the Rectorate of Gaziantep Islam Science and Technology University, where he also Coordinator in the Scientific Research and Projects Coordination Unit. He received his M.Sc. and Ph.D. degrees in Electrical and Electronics Engineering from Gaziantep University. His research interests focus on artificial intelligence and its applications, particularly in educational data analytics, adaptive testing systems, and distance learning technologies.

Özgeçmiş

Dr. Serkan Özbay is an Associate Professor in the Department of Electrical and Electronics Engineering at Gaziantep University. He received his Ph.D. in Electrical and Electronics Engineering from Gaziantep University. His research interests include artificial intelligence, machine learning, signal processing, and their applications in engineering systems and educational technologies. Dr. Özbay has authored numerous scientific publications in peer-reviewed journals and international conferences. He has also supervised various graduate theses and projects focusing on AI-driven systems.