https://doi.org/10.53360/2788-7995-2026-3(23)-24

IRSTI: 28.23.15; 28.23.29

G.N. Kazbekova*, Y.U. Serdaliyev
Khoja Akhmet Yassawi International Kazakh-Turkish University,
29 Bekzat Sattarkhanov Avenue. Turkestan, Kazakhstan
*e-mail: gulnur.kazbekova@ayu.edu.kz

LEAKAGE-FREE CAUSAL CONTEXT REPRESENTATION FOR POSE-BASED EXERCISE PHASE RECOGNITION

Abstract: *Pose-based exercise recognition provides a computationally efficient approach for analyzing human movement; however, classification based solely on the current pose may fail to capture short-term motion context. This study investigates whether leakage-free causal context can improve exercise-phase recognition without requiring a complex temporal neural architecture. A time-series exercise dataset containing 1,372 ordered pose observations from five exercises and ten exercise-phase classes was analyzed using seven pose representations derived from body landmarks, joint angles, and spatial distances. To reduce the risk of optimistic evaluation caused by neighboring observations appearing across training and testing partitions, a contiguous block-aware validation protocol was employed. A leakage-free causal mean context representation was constructed using only the current and preceding observations, without future samples or ground-truth labels. History-length ablation showed that a five-observation causal history achieved the best performance. The proposed representation increased mean Macro-F1 from 0.8358 for current-pose classification to 0.8973, corresponding to a 7.36% relative improvement, and improved performance in all five evaluation blocks. A causal Temporal Convolutional Network achieved a Macro-F1 of 0.8699, indicating that the simpler causal aggregation was more effective under the evaluated dataset conditions. Paired analysis yielded a mean Macro-F1 improvement of 0.0615 with a 95% confidence interval of [0.0308, 0.0921] and a large observed effect size (Cohen's $d_z = 2.49$); however, exact two-sided tests did not reach the conventional significance threshold ($p = 0.0625$) because only five paired blocks were available. These findings suggest that short causal pose context can provide useful information for exercise-phase recognition while remaining compatible with real-time processing.*

Key words: *human pose estimation, exercise recognition, causal context, pose representation, temporal modeling, block-aware validation, TCN.*

1. Introduction

Human action recognition has become an important research area in computer vision due to its applications in intelligent monitoring, human–computer interaction, rehabilitation, sports analysis, and automated exercise assessment. Conventional video-based approaches can extract rich visual information, but their performance may be affected by background appearance, illumination, clothing, and other environmental factors. Skeleton-based representations provide a compact alternative by describing human motion through body-joint coordinates and their spatial relationships, allowing recognition models to focus primarily on movement structure [1, 2]. Recent reviews demonstrate that skeleton-based human action recognition has evolved from handcrafted geometric descriptors toward deep spatial-temporal representation learning using graph convolutional networks, temporal architectures, and Transformer-based models [3, 4].

A central challenge in skeleton-based recognition is the effective representation of both body configuration and movement dynamics. Spatial coordinates of body landmarks provide detailed pose information, while joint angles and inter-joint distances can encode geometric relationships that are less directly represented by raw coordinates. Modern approaches increasingly learn these relationships using graph-based architectures such as ST-GCN, CTR-GCN, InfoGCN, and related models [5-7]. Other studies have investigated efficient skeleton representations and combinations of handcrafted geometric features with learned spatial-temporal features [8, 9]. These developments indicate that recognition performance depends not only on classifier complexity but also on how pose information is represented before classification.

Temporal information is particularly relevant when visually similar body configurations may correspond to different stages of an action. Modeling an observation independently ignores the

©G.N. Kazbekova, Y.U. Serdaliyev, 2026

movement trajectory that precedes it. Temporal Convolutional Networks (TCNs) provide one approach for learning such dependencies through convolutional sequence modeling [10, 11], while more recent skeleton-based methods employ graph convolution, attention, and Transformer architectures to capture increasingly complex spatial-temporal relationships [12, 13]. However, greater architectural complexity does not necessarily guarantee superior performance, particularly when the available dataset is limited. For practical exercise-monitoring systems, a useful temporal representation should therefore improve recognition while maintaining low computational complexity and causal operation suitable for real-time inference.

Evaluation methodology is another important consideration when working with ordered observations. If neighboring or strongly related samples are randomly distributed between training and testing partitions, a model may be evaluated on observations that are highly similar to samples already encountered during training. Such partitioning can produce optimistic performance estimates and obscure the model's ability to generalize to separated portions of the data. Data-leakage studies in human action recognition have demonstrated that evaluation protocols can substantially influence reported recognition performance [14]. Consequently, ordered pose observations require particular care when constructing validation partitions and temporal context.

In this study, we investigate whether short causal context can improve pose-based exercise-phase recognition without relying on a complex temporal neural architecture. Seven pose representations based on landmarks, joint angles, three-dimensional distances, coordinate-wise distances, and their combinations are first evaluated. A contiguous block-aware protocol is then used to reduce the influence of neighboring observations across evaluation partitions. Based on this protocol, we introduce a Leakage-Free Causal Mean Context Representation (LF-CMCR) that combines the current pose representation with an aggregated representation of strictly preceding observations. Context construction uses neither future observations nor ground-truth class labels and does not cross evaluation-partition boundaries. The method is additionally compared with a causal TCN using the same ordered-context principle.

The main contributions of this study are threefold. First, we systematically evaluate seven pose representations under both random and block-aware validation to examine representation robustness under a more restrictive evaluation protocol. Second, we propose LF-CMCR, a lightweight causal representation that incorporates short-term ordered context without future observations or label information, and analyze the effect of different history lengths. Third, we compare the proposed representation with both current-pose classification and a learnable causal TCN using paired block-wise evaluation and statistical analysis. The resulting framework is intended to clarify whether simple causal aggregation can provide a practical alternative to more complex temporal models for pose-based exercise-phase recognition.

2. Materials and Methods

2.1. Dataset and Pose Representation

The experiments were conducted using the Physical Exercise Recognition Time-Series dataset, which contains pose-based representations of human exercise movements. The dataset consists of 1,372 pose samples distributed across 10 exercise-state classes, representing the upward and downward phases of five common physical exercises: jumping jacks, pull-ups, push-ups, sit-ups, and squats. The class distribution is moderately imbalanced, with the number of samples per class ranging from 99 to 189.

Each sample is associated with a unique pose_id and a corresponding exercise-state label. Examination of the sample ordering showed that pose_id values form a strictly consecutive sequence from 0 to 1,371. Moreover, samples belonging to each class occupy a contiguous range in this ordering. Since the dataset does not provide original video, subject, repetition, or verified frame-sequence identifiers, the pose_id ordering is treated in this study as an ordered pose sequence rather than as confirmed temporal video frames. This distinction is important for preventing unsupported assumptions regarding the temporal structure of the dataset.

Four groups of pose descriptors were available: 3D body landmarks, joint angles, three-dimensional inter-joint distances, and axis-wise XYZ distance components. After alignment by pose_id, all descriptor groups contained the same 1,372 samples, with no missing identifiers, duplicated identifiers, or missing values. The resulting master dataset contained 170 numerical input features for each pose sample.

To investigate how different pose representations affect exercise recognition, seven feature configurations were constructed. R1 contained 99 raw 3D landmark coordinates, while R2 consisted of seven joint-angle features. R3 represented the pose using 16 three-dimensional inter-joint distance features, and R4 contained 48 axis-wise XYZ distance components. Two combined representations were additionally evaluated: R5, consisting of the 99 landmark coordinates and seven joint angles (106 features), and R6, containing the angle and distance-based geometric descriptors (71 features). Finally, R7 combined all available descriptors into a 170-dimensional representation.

Formally, the complete pose representation for an ordered observation t can be expressed as

$$\mathbf{x}_t = \left[\mathbf{L}_t, \mathbf{A}_t, \mathbf{D}_t, \mathbf{D}_t^{xyz}\right], \quad (1)$$

where $\mathbf{L}_t \in \mathbb{R}^{99}$ denotes the 3D landmark coordinates, $\mathbf{A}_t \in \mathbb{R}^7$ represents joint angles, $\mathbf{D}_t \in \mathbb{R}^{16}$ contains three-dimensional distance features, and $\mathbf{D}_t^{xyz} \in \mathbb{R}^{48}$ represents axis-wise distance components. Thus, the complete representation has dimensionality

$$99 + 7 + 16 + 48 = 170.$$

The seven representations were evaluated independently to determine whether compact geometric descriptors could preserve the discriminative information of the complete feature set. This representation analysis also provided the basis for selecting the feature space used in the subsequent ordered-context experiments (Table 1).

Table 1 – Pose representations evaluated in the study

Representation	Description	Features
R1	3D landmarks	99
R2	Joint angles	7
R3	3D inter-joint distances	16
R4	XYZ distance components	48
R5	Landmarks + angles	106
R6	Geometric descriptors	71
R7	All available features	170

2.2. Experimental Protocol and Evaluation

A rigorous evaluation protocol was designed to account for the ordered structure observed in the dataset. Initial inspection revealed that samples belonging to each exercise-state class were stored in contiguous pose_id ranges. Consequently, conventional random stratified splitting may place neighboring observations into both the training and test sets, potentially producing overly optimistic estimates of generalization performance.

To quantify this effect, two validation strategies were compared. The first employed five-fold random stratified validation, preserving the class distribution across folds. The second used a contiguous block-aware protocol. For each of the 10 classes, its ordered samples were divided into five approximately equal contiguous blocks. During each evaluation fold, one block from every class was assigned to the test partition, while the remaining four blocks were used for training. This produced five evaluation folds while preserving representation of all exercise-state classes in each test set.

The importance of this protocol was confirmed empirically. Under random stratified validation, the complete R7 representation achieved a mean Macro-F1 of 0.8704, whereas block-aware evaluation reduced the score to 0.8358 ± 0.0274. This corresponds to an absolute decrease of 0.0346 and a relative decrease of approximately 3.97%. Furthermore, approximately 96.57% of randomly selected test observations had an immediate ordered neighbor in the corresponding training data. These findings indicate that random splitting can benefit from strong local similarity between neighboring observations.

The block-aware protocol was subsequently applied consistently when comparing the seven pose representations. Under this evaluation, R7 achieved the highest Macro-F1 (0.8358), followed by R1 (0.8263) and R5 (0.8246). Interestingly, the compact R4 representation, containing only 48 features, achieved 0.8200 Macro-F1, retaining approximately 98.10% of the performance of the full 170-dimensional representation while reducing the feature dimensionality by 71.76%.

Because the blocks were constructed from the available pose_id ordering rather than original sequence metadata, they should not be interpreted as independent subjects, videos, or repetitions. Instead, the block-aware procedure serves as a more conservative evaluation mechanism that reduces direct adjacency between training and test observations while respecting the structure available in the dataset.

For all subsequent ordered-context experiments, feature construction was performed independently within each evaluation partition. Context vectors were therefore prevented from crossing training, validation, or test boundaries. In addition, future observations and ground-truth class labels were not used when constructing causal context features.

For the causal-context experiments, the complete R7 representation was used as the base feature space because it achieved the highest mean Macro-F1 under block-aware evaluation. The proposed CausalMean representation was evaluated with history lengths $h \in \{1,3,5,10\}$. For a history length h, each current observation was combined with a causal mean calculated from the available current and preceding observations within the same evaluation partition. The history length was treated as an experimental parameter, and $h = 5$ was selected as the primary configuration based on the history-ablation results.

A causal Temporal Convolutional Network (TCN) was additionally implemented as a learnable temporal baseline. To ensure a fair causal comparison, the TCN used the same history setting of $h = 5$, corresponding to a maximum input sequence length of six observations including the current pose. Only past and current observations were supplied to the network. Sequence windows were generated independently within the training, validation, and test partitions and were not allowed to cross partition boundaries. Feature standardization for the TCN was performed using a StandardScaler fitted exclusively on the training partition and subsequently applied to the corresponding validation and test data.

Macro-averaged F1 score was used as the primary evaluation metric because the 10 exercise-phase classes contained different numbers of samples. Classification accuracy and weighted F1 score were additionally calculated, while inference latency per sample was recorded to assess computational efficiency. Performance was reported across the five contiguous evaluation blocks using the mean and standard deviation of the block-wise scores.

To assess whether the observed improvements were consistent across evaluation blocks, paired statistical comparisons were performed using the Macro-F1 scores obtained on the same five blocks. The analysis included the mean paired difference, a 95% confidence interval for the mean difference, Cohen's d_z as a paired-sample effect-size measure, the Wilcoxon signed-rank test, an exact sign-flip permutation test, and an exact sign test. Because only five paired blocks were available, inferential p -values were interpreted cautiously and were considered together with the direction of the block-wise differences, effect size, confidence interval, and magnitude of the observed performance change.

Finally, three principal configurations were compared under the same block-aware framework: classification using the current R7 pose representation alone, the proposed leakage-free CausalMean representation with $h = 5$, and the causal TCN with $h = 5$. This paired design enabled direct assessment of whether ordered causal context improved recognition and whether a learnable temporal architecture provided an advantage over the proposed fixed causal aggregation.

2.3. Causal Ordered-Context Framework

To incorporate local ordered information without introducing future-sample leakage, a leakage-free causal ordered-context representation was constructed. Unlike symmetric temporal windows, which may use observations occurring after the current sample, the proposed representation relies exclusively on the current observation and a fixed number of preceding observations. This design makes the representation compatible with causal and potentially real-time inference.

Let $\mathbf{x}_t \in \mathbb{R}^d$ denote the feature vector of the current ordered pose observation at position t, where $d = 170$ for the complete R7 representation. For a history length h, the available causal context is defined as

$$\mathcal{H}_t^{(h)} = \{\mathbf{x}_{\max(0,t-h)}, \dots, \mathbf{x}_{t-1}, \mathbf{x}_t\}, \quad (2)$$

Thus, the context contains at most $h+1$ observations and never includes a future observation $\mathbf{x}_{t+k}$, where $k > 0$. At the beginning of an evaluation partition, shorter histories are retained rather than being completed using observations from another partition.
The local context is summarized using a causal mean vector:

$$\boldsymbol{\mu}_t^{(h)} = \frac{1}{|\mathcal{H}_t^{(h)}|} \sum_{\mathbf{x}_i \in \mathcal{H}_t^{(h)}} \mathbf{x}_i, \quad (3)$$

The final representation is obtained by concatenating the current pose vector with its causal contextual mean:

$$\mathbf{z}_t^{(h)} = \left[\mathbf{x}_t \parallel \boldsymbol{\mu}_t^{(h)}\right], \quad (4)$$

where ‖ denotes vector concatenation. With the 170-dimensional R7 representation, the resulting CausalMean descriptor therefore contains $170 + 170 = 340$ features.

History lengths $h \in \{1,3,5,10\}$ were evaluated to determine the amount of preceding ordered context required for effective recognition. A history of $h = 5$ was selected as the primary configuration based on the validation experiments. It represents each observation using the current pose together with information aggregated from up to five preceding ordered observations.

A central requirement of the proposed procedure is partition-aware context construction. For every block-aware fold, causal histories were generated independently within the training, validation, and test partitions. Consequently, no context window was permitted to cross a partition boundary. The construction procedure also used neither ground-truth exercise labels nor future observations. Therefore, the contextual representation available for a sample was derived solely from feature observations accessible up to its current ordered position within the corresponding partition.

This distinction is particularly important because the dataset provides ordered pose_id values but does not contain verified original-video frame sequences. Accordingly, the proposed method is referred to as causal ordered-context representation rather than conventional video temporal modeling. The method exploits local continuity present in the available ordering while avoiding assumptions about unavailable subject, repetition, or original frame-sequence metadata.

3. Results

3.1. Pose Representation and Validation Results

The initial experiments examined the discriminative performance and robustness of the seven pose representations under both random stratified and contiguous block-aware evaluation. Under random stratified validation, the complete R7 representation achieved a mean Macro-F1 of 0.8704, followed closely by R1 landmarks (0.8655), R4 XYZ distances (0.8632), R5 landmarks and angles (0.8624), and R6 geometric descriptors (0.8594). In contrast, the compact R3 distance and R2 angle representations achieved substantially lower scores of 0.6749 and 0.5631, respectively.

When the more conservative contiguous block-aware protocol was applied, performance decreased for all seven representations. R7 remained the highest-performing representation with a mean Macro-F1 of 0.8358 ± 0.0274, corresponding to an absolute decrease of 0.0346 relative to random stratified validation. R1 achieved 0.8263 ± 0.0372, while R5, R4, and R6 obtained Macro-F1 scores of 0.8246, 0.8200, and 0.8142, respectively. The largest relative reductions were observed for R3 and R2, which achieved block-aware Macro-F1 scores of only 0.6170 and 0.5159 (Table 2, Figure 1).

Table 2 – Performance of pose representations under random and block-aware evaluation

Representation	Features	Random Macro-F1	Block Macro-F1	F1 drop	Relative drop
R7 All features	170	0.8704	0.8358 ± 0.0274	0.0346	3.97%
R1 Landmarks	99	0.8655	0.8263 ± 0.0372	0.0392	4.53%
R5 Landmarks + angles	106	0.8624	0.8246 ± 0.0437	0.0378	4.38%
R4 XYZ distances	48	0.8632	0.8200 ± 0.0456	0.0432	5.01%
R6 Geometric	71	0.8594	0.8142 ± 0.0375	0.0452	5.25%
R3 3D distances	16	0.6749	0.6170 ± 0.0248	0.0579	8.58%
R2 Angles	7	0.5631	0.5159 ± 0.0444	0.0473	8.39%

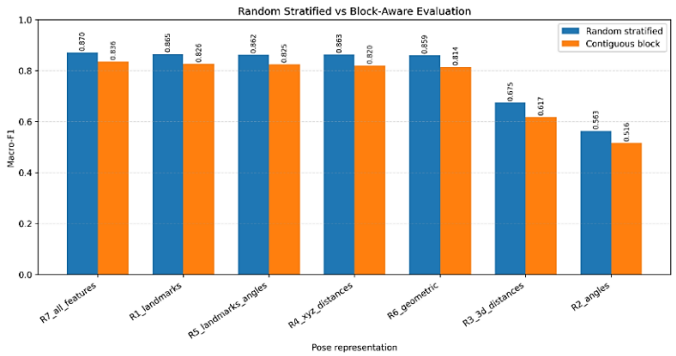


Figure 1 – Comparison of random stratified and contiguous block-aware Macro-F1 scores across the seven pose representations.

The difference between the two validation strategies was further supported by the neighborhood analysis. Under random stratified splitting, an average of 96.57% of test observations had an immediate ordered neighbor present in the corresponding training set. For R7, random validation produced a Macro-F1 of 0.8704 compared with 0.8358 under contiguous block evaluation, representing a relative reduction of approximately 3.97%. This indicates that evaluation methodology has a measurable influence on the reported recognition performance when observations exhibit strong local ordering.

Despite the overall advantage of R7, several lower-dimensional representations preserved most of its block-aware performance. R1 retained 98.86% of R7 performance while reducing the feature dimensionality by 41.76%. More notably, R4 retained approximately 98.10% of R7 performance using only 48 features, corresponding to a 71.76% reduction in dimensionality. This suggests that the axis-wise geometric distance representation contains substantial discriminative information despite its considerably smaller feature space.

Paired comparison across the five contiguous blocks showed that R7 outperformed R1, R5, and R4 in four of the five blocks. The mean Macro-F1 differences were 0.0095, 0.0113, and 0.0158, respectively. However, after correction for multiple comparisons, these differences were not statistically significant. Given that only five paired blocks were available, the absence of statistical significance should not be interpreted as evidence of equivalence between the representations.

Overall, R7 provided the highest block-aware Macro-F1 and was therefore selected as the base representation for the subsequent causal ordered-context experiments. At the same time, the results demonstrate that substantially more compact pose descriptors, particularly R4, can preserve a large proportion of the discriminative performance of the complete representation.

3.2. Causal Context and History Analysis

After selecting R7 as the base pose representation, the effect of causal ordered context was investigated using different history lengths. The leakage-free CausalMean representation was evaluated with $h \in \{1,3,5,10\}$, where only the current observation and preceding observations available within the same evaluation partition were used. No future samples or ground-truth labels were involved in context construction.

The results demonstrate that incorporating a short causal history consistently improved recognition performance compared with using the current pose alone. The baseline R7 representation achieved a mean Macro-F1 of 0.8358 ± 0.0274. Using a single preceding observation ($h = 1$) increased Macro-F1 to 0.8540 ± 0.0239. Extending the history to three observations further improved performance to 0.8871 ± 0.0371, while the best result was obtained with $h = 5$, reaching a Macro-F1 of 0.8973 ± 0.0329 and an accuracy of 0.8966. Increasing the history to $h = 10$, however, reduced Macro-F1 to 0.8473 ± 0.0284 (Table 3).

Table 3 – Effect of causal history length on recognition performance

History	Macro-F1	Std.	Accuracy	Gain vs. current
Current (h=0)	0.8358	0.0274	0.8339	—
(h=1)	0.8540	0.0239	0.8565	+0.0182
(h=3)	0.8871	0.0371	0.8851	+0.0512
(h=5)	0.8973	0.0329	0.8966	+0.0615
(h=10)	0.8473	0.0284	0.8566	+0.0115

The optimal $h = 5$ configuration produced an absolute Macro-F1 improvement of 0.0615, corresponding to a relative gain of 7.36% over the current-pose baseline. Importantly, the improvement was observed in all five evaluation blocks. The block-wise Macro-F1 increased from 0.8515 to 0.8803, 0.8136 to 0.8866, 0.7994 to 0.8635, 0.8555 to 0.9492, and 0.8592 to 0.9070 for blocks 1-5, respectively.

Paired statistical analysis further showed a mean Macro-F1 difference of +0.0615, with a 95% confidence interval of [0.0308, 0.0921]. The observed paired effect size was large (Cohen's $d_z = 2.49$), and CausalMean outperformed the baseline in all five paired blocks. However, the Wilcoxon signed-rank test, exact sign-flip permutation test, and exact sign test each yielded $p = 0.0625$. Thus, the conventional $p < 0.05$ threshold was not reached (Figure 2).

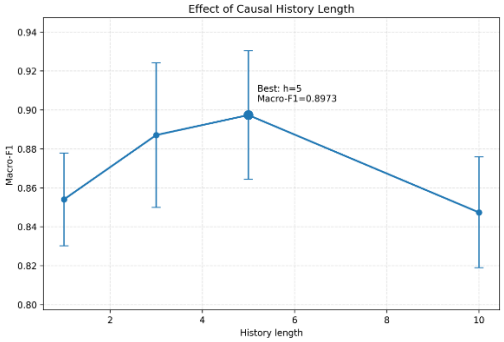


Figure 2 – Macro-F1 performance of the leakage-free CausalMean representation across different causal history lengths

This result should be interpreted in light of the small number of paired evaluation blocks ($n = 5$). With five consistently positive paired differences, the exact two-sided tests are inherently limited in attainable significance. Therefore, the statistical evidence is better characterized by the consistent 5/5 improvement direction, the large observed effect size, the positive confidence interval, and the magnitude of the performance gain rather than by the p-value alone.

The history ablation also indicates that additional context is not necessarily beneficial indefinitely. Performance increased substantially from $h = 1$ to $h = 5$, but declined at $h = 10$. This suggests that a relatively short local context captures useful ordered variation, whereas a longer history may incorporate less relevant observations and dilute information associated with the current pose.

3.3. Comparison with Causal TCN

To determine whether the observed benefit of causal ordered context required a learnable temporal model, the proposed CausalMean representation was compared with a causal Temporal Convolutional Network (TCN). Both approaches used the same history length of $h = 5$, corresponding to a maximum context of five preceding observations plus the current pose. The comparison was performed using the same five block-aware evaluation folds.

The causal TCN improved the mean Macro-F1 from 0.8358 ± 0.0274 for the current-pose baseline to 0.8699 ± 0.0195, corresponding to an absolute gain of 0.0340 and a relative improvement of 4.07%. However, the proposed CausalMean representation achieved the highest performance, with a mean Macro-F1 of 0.8973 ± 0.0329. This represents an absolute improvement of 0.0615 (7.36%) over the current-pose baseline and an improvement of 0.0275 over the causal TCN (Table 4).

Table 4 – Final comparison of current-pose, causal TCN, and CausalMean approaches

Method	Macro-F1	Std.	Gain vs. current	Relative gain
Current R7	0.8358	0.0274	–	–
Causal TCN (h=5)	0.8699	0.0195	+0.0340	+4.07%
CausalMean (h=5)	0.8973	0.0329	+0.0615	+7.36%

The paired block-wise comparison showed a consistent ranking. Causal TCN outperformed the current-pose baseline in all five blocks, with a mean Macro-F1 difference of +0.0340. CausalMean likewise outperformed the current representation in all five blocks and, importantly, also

exceeded the causal TCN in all five paired blocks. The mean difference between CausalMean and causal TCN was +0.0275, with a 95% confidence interval of [0.0021, 0.0529] and Cohen's $d_z = 1.34$.

The exact two-sided tests yielded $p = 0.0625$ for each of the three principal paired comparisons. As discussed previously, this value reflects the limited number of independent paired blocks rather than evidence that the observed differences are absent. In particular, the consistent 5/5 direction of the differences and the positive effect sizes support the practical relevance of the observed improvements, while formal statistical significance at the conventional 0.05 threshold cannot be claimed.

The computational comparison also favored the fixed causal representation. CausalMean required approximately 0.066 ms per sample, whereas the causal TCN required approximately 0.963 ms per sample under the experimental implementation. Thus, the fixed causal aggregation achieved higher recognition performance while avoiding the additional training and inference complexity of the learnable temporal network.

Overall, the final experimental ranking was

$$\text{CausalMean}_{h=5} > \text{CausalTCN}_{h=5} > \text{Current R7}$$

These results indicate that the primary benefit was obtained from incorporating short-range causal ordered context itself, while increasing model complexity through a TCN did not provide an additional advantage on the present dataset.

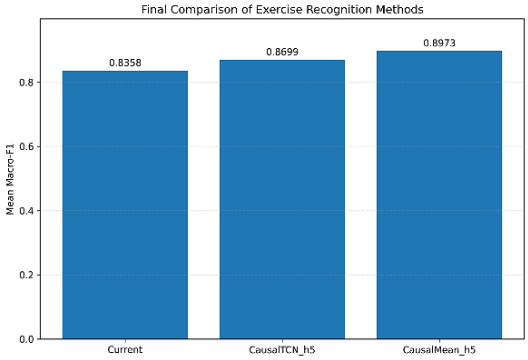


Figure 3 – Mean block-aware Macro-F1 comparison of the current-pose baseline, causal TCN, and the proposed CausalMean representation

4. Discussion

The experimental results demonstrate that both pose representation and evaluation protocol substantially influence exercise-recognition performance. Although the complete R7 representation achieved the highest overall block-aware performance, several compact representations retained most of its discriminative capability. In particular, the 48-dimensional R4 representation preserved approximately 98% of the performance of the full 170-dimensional representation. This observation supports previous studies showing that geometric relationships between skeletal joints can provide informative and computationally efficient descriptors for human action recognition [3, 4, 9]. Recent work has similarly demonstrated that handcrafted geometric information can complement skeletal representations while maintaining efficient inference [15, 16].

A second important finding concerns the evaluation protocol. Performance consistently decreased when random stratified validation was replaced by contiguous block-aware evaluation. The relatively high proportion of randomly selected test observations with immediate neighbors in the training set suggests that conventional random splitting can benefit from local similarity in ordered data. Therefore, randomly partitioned observations may provide an optimistic estimate of generalization when neighboring samples are strongly related. This finding is consistent with concerns raised in previous HAR research regarding inappropriate partitioning and data leakage [14]. More generally, recent evidence in human activity and exercise-recognition tasks confirms that inappropriate splitting can substantially overestimate generalization performance [14, 17]. Importantly, the present study does not claim that the derived blocks correspond to independent subjects or video sequences; rather, block-aware evaluation was introduced as a more conservative strategy given the metadata available in the dataset.

The central finding of this study is that short-range causal ordered context provided a clear advantage over isolated pose classification. The proposed CausalMean representation with $h = 5$ improved Macro-F1 from 0.8358 to 0.8973, with positive differences across all five evaluation blocks. This indicates that individual pose observations do not contain all discriminative information available in their local ordering. Aggregating recent observations can stabilize the representation and provide information about local movement evolution that is unavailable from a single pose. This observation is conceptually consistent with temporal action-recognition approaches, where relationships across observations are used to capture movement dynamics [5, 7, 10-13].

The history-ablation experiment provides additional insight into the scale of useful context. Increasing history from $h = 1$ to $h = 5$ progressively improved recognition, whereas extending the history to $h = 10$ reduced performance. A plausible interpretation is that short histories capture locally relevant changes associated with an exercise phase, while longer windows increasingly incorporate observations that are less representative of the current state. Thus, the results favor selective short-range context rather than simply maximizing temporal receptive field. Because original frame-level temporal metadata are unavailable, however, this interpretation should be understood in terms of ordered pose continuity rather than verified physical time.

An important result is that the fixed CausalMean representation outperformed the causal TCN under the same $h = 5$ setting. TCNs are designed to model temporal dependencies through hierarchical temporal convolutions and have demonstrated strong performance in action modeling [10], [11]. In the present experiments, the TCN improved substantially over current-pose classification, confirming the value of causal context, but it did not exceed the simpler aggregation approach. One explanation is the relatively limited dataset size of 1,372 observations. The additional parameters of a learnable temporal architecture may require substantially more independent sequential data to exploit their modeling capacity effectively, whereas CausalMean introduces no trainable temporal parameters and directly summarizes local ordered information. Therefore, increased architectural complexity did not translate into improved performance under the present experimental conditions.

From a practical perspective, this finding is relevant for lightweight and real-time exercise-recognition systems, where computational efficiency is an important consideration [16], [22], [23]. The proposed representation requires only the computation of a running local mean and feature concatenation before classification. In the experimental implementation, its inference latency was approximately 0.066 ms per sample compared with approximately 0.963 ms for the causal TCN. The combination of higher recognition performance and lower computational complexity suggests that fixed causal aggregation may be useful when low latency and limited computational resources are important.

The statistical results nevertheless require cautious interpretation. CausalMean improved Macro-F1 in all five paired blocks and produced a large observed effect size ($d_z = 2.49$) relative to the current-pose baseline. However, the exact two-sided tests yielded $p = 0.0625$, slightly above the conventional significance threshold. With only five paired evaluation blocks, statistical power is inherently limited. Consequently, the results support a consistent and practically meaningful observed improvement, but they should not be presented as definitive statistical evidence of superiority.

Several limitations should therefore be acknowledged. First, the dataset does not provide original video identifiers, subject identities, repetition boundaries, or verified frame-sequence metadata. Consequently, pose_id ordering cannot be interpreted as confirmed temporal video continuity. Second, the five contiguous blocks are derived evaluation groups rather than genuinely independent subject or repetition partitions. Third, the relatively small dataset limits both statistical power and the ability of higher-capacity temporal models such as TCNs to learn complex motion patterns. Finally, the present experiments address exercise-state recognition rather than complete exercise-technique assessment or error diagnosis.

These limitations define a clear direction for future work. The proposed framework should be validated on larger datasets containing explicit subject, repetition, and frame-level temporal identifiers [21, 24, 25]. Subject-independent protocols such as leave-one-subject-out evaluation would enable stronger assessment of generalization, while genuine video sequences would permit direct comparison of causal aggregation with TCNs, recurrent architectures, graph-based temporal models, and transformer-based motion representations [19, 20]. Such validation would determine

whether the advantage observed for lightweight causal aggregation persists when richer and independently structured temporal data are available.

5. Conclusion

This study investigated pose-based exercise recognition under ordered-data conditions and proposed a leakage-free causal ordered-context representation (LF-CMCR) for incorporating short-range contextual information without using future observations or ground-truth labels. Seven pose representations were first evaluated using random stratified and contiguous block-aware validation. The complete R7 representation achieved the highest block-aware Macro-F1 of 0.8358, while several lower-dimensional representations retained a substantial proportion of its performance, demonstrating the potential of compact geometric pose descriptors.

The experiments further showed that evaluation protocol has a measurable influence on recognition performance. For R7, Macro-F1 decreased from 0.8704 under random stratified validation to 0.8358 under contiguous block-aware evaluation. This result highlights the importance of accounting for local ordering when neighboring observations may otherwise be distributed across training and test partitions.

The proposed causal representation substantially improved recognition performance. Among the evaluated history lengths, $h = 5$ achieved the best result, increasing mean Macro-F1 from 0.8358 to 0.8973, corresponding to an absolute gain of 0.0615 and a relative improvement of 7.36%. The improvement occurred across all five evaluation blocks and produced a large paired effect size ($d_z = 2.49$). However, because only five paired blocks were available, the exact two-sided tests yielded $p = 0.0625$; therefore, statistical significance at the conventional 0.05 level is not claimed.

Comparison with a causal TCN further demonstrated that the benefit was not dependent on a complex learnable temporal architecture. The causal TCN achieved a Macro-F1 of 0.8699, improving over the current-pose baseline but remaining below the proposed CausalMean representation. This finding suggests that, under the present dataset conditions, a short and computationally inexpensive causal aggregation can capture useful ordered information without the additional complexity of a trainable temporal model. TCNs nevertheless remain well-established sequence-modeling architectures, making this comparison an informative baseline rather than evidence against temporal convolutional modeling in general [10], [11].

The principal limitation of the study is that the available dataset does not provide verified original-video frame sequences, subject identities, or repetition boundaries. Consequently, the observed relationships should be interpreted as ordered-pose context rather than confirmed video-temporal dynamics. Future work will therefore validate the proposed framework using genuine sequential exercise datasets with explicit subject and repetition information, subject-independent evaluation protocols, and real-time exercise recordings.

Overall, the results indicate that carefully constructed short-range causal context can improve pose-based exercise recognition while preserving causal inference and low computational complexity, providing a practical foundation for future real-time exercise assessment systems.

References

1. Feng M. Skeleton Graph-Neural-Network-Based Human Action Recognition: A Survey / M. Feng, J. Meunier // Sensors. – 2022. – Vol. 22, № 6. – Art. 2091. https://doi.org/10.3390/s22062091.
2. Yue R. Action recognition based on RGB and skeleton data sets: A survey / R. Yue, Z.-Q. Tian, S. Du // Neurocomputing. – 2022. – Vol. 512. – P. 287-306. https://doi.org/10.1016/j.neucom.2022.09.071.
3. A Survey on 3D Skeleton-Based Action Recognition Using Learning Method / B. Ren et al // Cyborg and Bionic Systems. – 2024. – Vol. 5. – Art. 0100. https://doi.org/10.34133/cbsystems.0100.
4. Huang Q. A Review of Skeleton-Based Human Action Recognition / Q. Huang, J. Cui, C. Li // Journal of Computer-Aided Design & Computer Graphics. – 2024. – Vol. 36, № 2. – P. 173-194. https://doi.org/10.3724/SP.J.1089.2024.2023-00358.
5. Yan S. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition / S. Yan, Y. Xiong, D. Lin // Proceedings of the AAAI Conference on Artificial Intelligence. – 2018. – Vol. 32, № 1. https://doi.org/10.1609/aaai.v32i1.12328.
6. Channel-Wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition / Y. Chen et al // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). – 2021. – P. 13359-13368.

7. InfoGCN: Representation Learning for Human Skeleton-Based Action Recognition / H.-g. Chi et al // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2022. – P. 20186-20196. https://doi.org/10.1109/CVPR52688.2022.01955.
8. Constructing Stronger and Faster Baselines for Skeleton-Based Action Recognition / Y.-F. Song et al // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2023. – Vol. 45, № 2. – P. 1474-1488. https://doi.org/10.1109/TPAMI.2022.3157033.
9. Nan M. Spatio-temporal neural network with handcrafted features for skeleton-based action recognition / M. Nan, M. Trăscău, A.-M. Florea // Neural Computing and Applications. – 2024. – Vol. 36. – P. 9221-9243. https://doi.org/10.1007/s00521-024-09559-4.
10. Temporal Convolutional Networks for Action Segmentation and Detection / C. Lea et al // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2017. – P. 156-165.
11. Bai S. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling / S. Bai, J. Z. Kolter, V. Koltun. – 2018. – arXiv:1803.01271.
12. Plizzari C. Skeleton-based action recognition via spatial and temporal transformer networks / C. Plizzari, M. Cannici, M. Matteucci // Computer Vision and Image Understanding. – 2021. – Vol. 208-209. – Art. 103219. https://doi.org/10.1016/j.cviu.2021.103219.
13. MotionBERT: A Unified Perspective on Learning Human Motion Representations / W. Zhu et al // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). – 2023. – P. 15085-15099.
14. Varga D. Critical Analysis of Data Leakage in WiFi CSI-Based Human Action Recognition Using CNNs / D. Varga // Sensors. – 2024. – Vol. 24, № 10. – Art. 3159. https://doi.org/10.3390/s24103159.
15. Recognition and Scoring Physical Exercises via Temporal and Relative Analysis of Skeleton Nodes Extracted from the Kinect Sensor / R. Esmaeeli et al // Sensors. – 2024. – Vol. 24, № 20. – Art. 6713. https://doi.org/10.3390/s24206713.
16. Efficient Skeleton-Based Action Recognition for Real-Time Embedded Systems / N. Noor et al // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). – 2024. – P. 5889-5897.
17. Varga D. Mitigating Data Leakage in a WiFi CSI Benchmark for Human Action Recognition / D. Varga // Sensors. – 2024. – Vol. 24, № 24. – Art. 8201. https://doi.org/10.3390/s24248201.
18. Liu J. Multi-Modality Co-Learning for Efficient Skeleton-based Action Recognition / J. Liu, C. Chen, M. Liu // Proceedings of the 32nd ACM International Conference on Multimedia. – 2024. – P. 4909-4918. https://doi.org/10.1145/3664647.3681015.
19. BlockGCN: Redefine Topology Awareness for Skeleton-Based Action Recognition / Y. Zhou et al // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2024. – P. 2049-2058. https://doi.org/10.1109/CVPR52733.2024.00200.
20. MGSAN: Multimodal Graph Self-Attention Network for Skeleton-Based Action Recognition / J. Wang et al // Multimedia Systems. – 2024. – Vol. 30. – Art. 369. https://doi.org/10.1007/s00530-024-01566-8.
21. FineRehab: A Multi-modality and Multi-task Dataset for Rehabilitation Analysis / J. Li et al // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). – 2024. – P. 3184-3193. https://doi.org/10.1109/CVPRW63382.2024.00324.
22. Towards Automating Personal Exercise Assessment and Guidance with Affordable Mobile Technology / M. Sideridou et al // Sensors. – 2024. – Vol. 24, № 7. – Art. 2037. https://doi.org/10.3390/s24072037.
23. Kwon H.-J. A novel machine learning-based shoulder exercise assistant program / H.-J. Kwon, S.-H. Choi // Discover Artificial Intelligence. – 2025. – Vol. 5. – Art. 37. https://doi.org/10.1007/s44163-025-00259-z.
24. M3GYM: A Large-Scale Multimodal Multi-view Multi-person Pose Dataset for Fitness Activity Understanding in Real-world Settings / Q. Xu et al // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2025. – P. 12289-12300.
25. A Knee Rehabilitation Exercises Dataset for Postural Assessment using Wearable Devices / P. Kasnesis et al // Scientific Data. – 2025. – Vol. 12. – Art. 610. https://doi.org/10.1038/s41597-025-04963-4.

Г.Н. Казбекова*, Е.У. Сердалиев
Қожа Ахмет Ясауи атындағы Халықаралық қазақ-түрік университеті,
Бекзат Саттарханов даңғылы, 29, Түркістан, Қазақстан
*e-mail: gulnur.kazbekova@ayu.edu.kz

ДЕРЕКТЕРДІҢ АҚПАРАТТЫҚ АҒЫП КЕТУІН БОЛДЫРМАЙТЫН СЕБЕПТІК КОНТЕКСТІК РЕПРЕЗЕНТАЦИЯ НЕГІЗІНДЕ ЖАТТЫҒУ ФАЗАЛАРЫН ТАНУ

Позаға негізделген жаттығуларды тану адам қозғалысын талдаудың есептеу тұрғысынан тиімді тәсілін ұсынады. Алайда тек ағымдағы позаға негізделген жіктеу қысқа мерзімді қозғалыс контекстін толық қамти алмауы мүмкін. Бұл зерттеуде күрделі уақыттық нейрондық архитектураны қолданбай-ақ, ағып кетусіз себептік контекстің жаттығу фазаларын тану сапасын арттыра алатындығы зерттеледі. Бес жаттығу және жаттығу фазаларының он класына жататын 1 372 реттелген позалық бақылаудан тұратын уақыттық қатарлар деректер жиыны дене landmark нүктелері, буын бұрыштары және кеңістіктік қашықтықтар негізінде алынған жеті позалық репрезентацияны пайдалана отырып зерттелді. Көршілес бақылаулардың оқыту және тестілеу бөліктеріне қатар түсуінен туындайтын бағалаудың оптимистік нәтижелер алу қаупін азайту мақсатында блоктардың реттілігін ескеретін валидация хаттамасы қолданылды.

Ағып кетусіз себептік орташа контекстік репрезентация тек ағымдағы және алдыңғы бақылауларды пайдалана отырып құрылды; бұл ретте болашақ үлгілер де, нақты эталондық белгілер де (ground-truth labels) қолданылған жоқ. Тарих ұзындығын абляциялық талдау бес бақылаудан тұратын себептік тарихтың ең жоғары нәтижені көрсеткенін анықтады. Ұсынылған репрезентация ағымдағы поза бойынша жіктеу кезінде алынған орташа Macro-F1 көрсеткішін 0,8358-ден 0,8973-ке дейін арттырды, бұл салыстырмалы түрде 7,36% жақсаруға сәйкес келеді және барлық бес бағалау блогында нәтижелілікті арттырды.

Себептік Temporal Convolutional Network (TCN) моделі 0,8699 Macro-F1 көрсеткішіне қол жеткізді, бұл зерттелген деректер жиыны жағдайында қарапайым себептік агрегацияның тиімдірек болғанын көрсетті. Жұптық талдау орташа Macro-F1 өсімінің 0,0615 болғанын, 95% сенімділік интервалы [0,0308; 0,0921] және байқалған әсер мөлшерінің үлкен болғанын (Cohen's d) көрсетті. Алайда қолда бар бес жұптық блок қана болғандықтан, дәл екіжақты статистикалық сынақтар дәстүрлі статистикалық мәнділік шегіне ($p < 0,05$) жеткен жоқ. Алынған нәтижелер қысқа мерзімді себептік позалық контекст нақты уақыт режимінде өңдеуге сәйкес бола отырып, жаттығу фазаларын тану үшін пайдалы ақпарат бере алатынын көрсетеді.

Түйін сөздер: *адам позасын бағалау, жаттығуларды тану, себептік контекст, позалық репрезентация, уақыттық модельдеу, блоктарды ескеретін валидация, TCN.*

Г.Н. Казбекова*, Е.У. Сердалиев
[1]Международный казахско-турецкий университет имени Ходжи Ахмеда Ясави,
Республика Казахстан, Туркестан, проспект Бекзата Саттарханова, 29
*e-mail: gulnur.kazbekova@ayu.edu.kz

РАСПОЗНАВАНИЕ ФАЗ ОБУЧЕНИЯ НА ОСНОВЕ ПРИЧИННО-КОНТЕКСТНОЙ РЕПРЕЗЕНТАЦИИ, ПРЕДОТВРАЩАЮЩЕЙ УТЕЧКУ ИНФОРМАЦИИ ИЗ ДАННЫХ

Распознавание упражнений на основе поз представляет собой вычислительно эффективный подход к анализу движений человека. Однако классификация, основанная исключительно на текущей позе, может не учитывать краткосрочный контекст движения. В данном исследовании изучается, может ли свободный от утечки причинный контекст повысить качество распознавания фаз упражнений без необходимости использования сложной временной нейронной архитектуры. Набор временных рядов, содержащий 1 372 упорядоченных наблюдения поз, относящихся к пяти упражнениям и десяти классам фаз упражнений, был проанализирован с использованием семи позовых представлений, сформированных на основе ориентиров тела (landmark), углов суставов и пространственных расстояний. Для снижения риска получения оптимистичных результатов оценки вследствие попадания соседних наблюдений одновременно в обучающую и тестовую выборки был применён протокол валидации с учётом последовательности блоков.

Свободное от утечки причинное представление на основе среднего контекста было сформировано с использованием только текущего и предшествующих наблюдений, без использования будущих образцов или истинных меток классов (ground-truth labels). Абляционный анализ длины истории показал, что причинная история из пяти наблюдений обеспечивает наилучшие результаты. Предложенное представление повысило среднее значение Macro-F1 с 0,8358 для классификации по текущей позе до 0,8973, что соответствует относительному улучшению на 7,36%, а также обеспечило повышение качества во всех пяти оценочных блоках.

Причинная сеть Temporal Convolutional Network (TCN) достигла значения Macro-F1 0,8699, что указывает на более высокую эффективность простой причинной агрегации в условиях исследуемого набора данных. Парный анализ показал среднее улучшение Macro-F1 на 0,0615 с 95%-ным доверительным интервалом [0,0308; 0,0921] и большим наблюдаемым размером эффекта (Cohen's d). Однако точные двусторонние тесты не достигли традиционного порога статистической значимости ($p < 0,05$), поскольку в исследовании было доступно только пять парных блоков. Полученные результаты показывают, что краткосрочный причинный контекст поз может предоставлять полезную информацию для распознавания фаз упражнений, сохраняя при этом совместимость с обработкой данных в режиме реального времени.

Ключевые слова: *оценка позы человека, распознавание упражнений, причинный контекст, позовое представление, временное моделирование, блочно-ориентированная валидация, TCN.*

Information about the authors

Gulnur Nagimetovna Kazbekova* – Candidate of Technical Sciences, Associate Professor, International Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkestan, Kazakhstan; e-mail: gulnur.kazbekova@ayu.edu.kz. ORCID: https://orcid.org/0000-0002-2756-7926.

Yerlan Serdaliyev – PhD student in the educational program «Information systems», International Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkestan, Kazakhstan; e-mail: erlan.serdaliev@ayu.edu.kz. ORCID: https://orcid.org/0009-0007-5403-0423.

Авторлар туралы мәліметтер

Гулнур Нагиметовна Казбекова* – техника ғылымдарының кандидаты, қауымдастырылған профессор, Қожа Ахмет Ясауи атындағы Халықаралық қазақ-түрік университеті, Түркістан қ., Қазақстан; e-mail: gulnur.kazbekova@ayu.edu.kz. ORCID: https://orcid.org/0000-0002-2756-7926.

Ерлан Утегенулы Сердалиев – «Ақпараттық жүйелер» ББ докторанты, Қожа Ахмет Ясауи атындағы Халықаралық қазақ-түрік университеті, Түркістан қ., Қазақстан; e-mail: erlan.serdaliev@ayu.edu.kz. ORCID: https://orcid.org/0009-0007-5403-0423.

Сведения об авторах

Гулнур Нагиметовна Казбекова* – кандидат технических наук, ассоциированный профессор, Международный казахско-турецкий университет имени Ходжи Ахмеда Ясави, г. Туркестан, Казахстан; e-mail: gulnur.kazbekova@ayu.edu.kz. ORCID: https://orcid.org/0000-0002-2756-7926.

Ерлан Утегенулы Сердалиев – докторант ОП «Информационные системы», Международный казахско-турецкий университет имени Ходжи Ахмеда Ясави, г. Туркестан, Казахстан; e-mail: erlan.serdaliev@ayu.edu.kz. ORCID: https://orcid.org/0009-0007-5403-0423.

Received 18.06.2026
Revised 25.08.2026
Accepted 26.08.2026

https://doi.org/10.53360/2788-7995-2026-3(23)-25

МРНТИ: 20.53.19; 14.85.09

CC BY 4.0
Check for updates

А.К. Шайханова*, Г.А. Шангытбаева[2], Г.Б. Бекешова[1], Лили Нурлияна Абдулла[3]
[1]Евразийский национальный университет им. Л.Н. Гумилева,
10000, Республика Казахстан, г. Астана, ул. Сатпаева, 2
[2]Актюбинский региональный университет имени К. Жубанова,
030000, Актюбинская область, г. Актобе, пр. Алии Молдагуловой, ст-е 34
[3]Университет Путра Малайзия, Серданг, Селангор, Малайзия
*e-mail: aigul.shaikhanova@gmail.com

ИНФОРМАЦИОННО-АНАЛИТИЧЕСКАЯ МЕТОДИКА ОЦЕНКИ ЭФФЕКТИВНОСТИ ВНЕДРЕНИЯ ЦИФРОВЫХ ИНСТРУМЕНТОВ В ПРОФЕССИОНАЛЬНУЮ ДЕЯТЕЛЬНОСТЬ ПЕДАГОГОВ

Аннотация**:** *В статье представлена разработанная информационно-аналитическая методика комплексной оценки эффективности внедрения цифровых инструментов в профессиональную деятельность педагогов. Актуальность исследования определяется необходимостью перехода от описательной самооценки цифровых компетенций к*

©А.К. Шайханова, Г.А. Шангытбаева, Г.Б. Бекешова, Лили Нурлияна Абдулла, 2026