

ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ INTELLIGENT SYSTEMS

УДК 004.8:519.876.5
<https://doi.org/10.37661/1816-0301-2026-23-3-92-106>

Поступила в редакцию | Received 01.06.2026
Подписана в печать | Accepted 22.06.2026
Опубликована | Published 30.09.2026

Вероятностный трансформер для предиктивного мониторинга редакционного процесса

Д. А. Сьянов
E-mail: syanov@bsu.by

*Белорусский государственный университет,
пр. Независимости, 4, Минск, 220030, Беларусь*

Аннотация

Цели. Целью исследования является разработка подхода к прогнозированию сроков наступления нескольких ключевых точек в многоэтапных процессах.

Методы. Предложен метод, основанный на обработке последовательностей событий с учетом стадий, промежутков времени и сведений об участниках процесса, а также на согласованном учете взаимосвязей между прогнозируемыми этапами.

Результаты. Проведенные эксперименты на данных редакционного документооборота показали повышение точности по сравнению с рекуррентными нейронными сетями и существенное преимущество на ранних этапах по сравнению с моделями градиентного бустинга.

Заключение. Метод позволяет получать согласованные интервальные прогнозы для нескольких этапов процесса и может использоваться для оперативного обновления оценок без применения специализированного оборудования.

Ключевые слова: предиктивный мониторинг процессов, вероятностный трансформер, прогнозирование задержек, квантильная регрессия, машинное обучение

Для цитирования. Сьянов, Д. А. Вероятностный трансформер для предиктивного мониторинга редакционного процесса / Д. А. Сьянов // Информатика. – 2026. – Т. 23, № 3. – С. 92–106. – <https://doi.org/10.37661/1816-0301-2026-23-3-92-106>.

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.

A probabilistic transformer for predictive monitoring of editorial process

Dmitry A. Syanov

E-mail: syanov@bsu.by

Belarusian State University

av. Nezavisimosti 4, Minsk, 220030, Belarus

Abstract

Objectives. The aim of this study is to develop an approach for predicting the timing of several key stages in multi-stage processes.

Methods. A method is proposed based on processing event sequences taking into account stages, time intervals, and information about process participants, as well as a consistent consideration of the relationships between the predicted stages.

Results. Experiments conducted on editorial document flow data demonstrated improved accuracy compared to recurrent neural networks and a significant advantage in the early stages compared to gradient boosting models.

Conclusion. The method allows obtaining of consistent interval forecasts for several stages of a process and can be used to quickly update the estimates without the use of specialized equipment.

Keywords: predictive process monitoring, probabilistic transformer, delay forecasting, quantile regression, machine learning

For citation. Syanov D. A. *A probabilistic transformer for predictive monitoring of editorial process*. Informatika [Informatics], 2026, vol. 23, no. 3, pp. 92–106 (In Russ.). <https://doi.org/10.37661/1816-0301-2026-23-3-92-106>.

Conflict of interests. The author declares of no conflict of interest.

Введение

Редакционный документооборот в системах управления журналами Open Journals System (OJS) реализуется как дискретно-событийный процесс, в рамках которого рукопись последовательно проходит этапы подачи материала, назначения рецензентов, рецензирования, вынесения редакторского решения и публикации. Согласно подходам процессной аналитики [1, 2] каждая такая последовательность событий рассматривается как экземпляр бизнес-процесса, а совокупность экземпляров формирует журнал событий, содержащий информацию о времени обработки, задержках и траекториях прохождения рукописей. При этом структура рассматриваемых процессов отличается высокой стохастичностью, так как длительности стадий существенно варьируются (от часов до месяцев), загрузка редакторов и рецензентов изменчива, а входящий поток обладает выраженной нерегулярностью. Данные факторы формируют сложную динамику очередей и задержек, что затрудняет использование традиционных методов моделирования и требует внедрения подходов, способных корректно описывать и прогнозировать эволюцию многомерного стохастического процесса в различных режимах нагрузки.

В этой связи особую значимость приобретают методы предиктивного мониторинга, обеспечивающие оценку будущих характеристик текущего экземпляра на основе наблюдаемого префикса журнала событий [3]. Возможность своевременно прогнозировать

вать длительности стадий, будущие контрольные точки и вероятность задержек имеет ключевое значение для оперативного принятия решений, планирования редакционных ресурсов и параметризации имитационных моделей. Однако обзорные исследования [4] показывают, что существующие методы остаются ограниченными, так как они обычно дают точечные оценки единственной скалярной характеристики (например, оставшегося времени), используют несопоставимые экспериментальные настройки и не рассматривают совместный вероятностный прогноз нескольких будущих событий. Кроме того, современные методы кодирования временной структуры журналов событий практически не исследованы.

Обзор смежных работ

Предиктивный мониторинг бизнес-процессов охватывает широкий спектр задач, отличающихся по типу прогнозируемой характеристики и требуемой глубине анализа. В частности, Тейнемаа и др. [3] выделяют класс задач прогноза исхода, где целевая переменная представляет собой бинарную или категориальную метку заверченного экземпляра процесса. Продолжает этот вектор исследований работа Маджи и др. [5], в которой задача выявления возможных нарушений нормативных требований формализована через LTL-предикаты, применяемые к наблюдаемому префиксу последовательности событий. Проблематика прогнозирования нежелательных исходов также изучалась с использованием как структурированных, так и неструктурированных атрибутов экземпляра процесса [6]. В более позднем сравнительном исследовании Крача и др. [7] были проанализированы классические (Random Forest, XGBoost) и нейросетевые (LSTM, CNN) модели, показавшие сопоставимую эффективность при наличии достаточного объема обучающих данных.

Также задача прогнозирования следующего действия рассматривалась Эверманном и др. [8] с использованием рекуррентных нейронных сетей, а Камарго и др. [9] предложили расширенную LSTM-модель, одновременно предсказывающую следующую активность и соответствующую временную метку. Этот класс методов позволяет моделировать динамику поведения процесса, однако не охватывает оценку длительностей выполнения и фокусируется преимущественно на последовательностных аспектах.

Наиболее близкой к настоящему исследованию является задача прогноза остаточного времени цикла. В систематическом обзоре Веренича и др. [4] подчеркивается отсутствие единообразия в экспериментальных настройках: в разных работах используются разнородные наборы данных, метрики и схемы разбиения, что существенно осложняет объективное сравнение методов. Такс и др. [10] применили LSTM для прогноза остаточного времени и последовательности будущих действий, сосредоточившись на оценке единственного значения остаточного времени для каждого экземпляра процесса. В последующих работах Наварина и др. [11] и Полато и др. [12] подход был расширен за счет включения атрибутов экземпляра процесса, что позволило учитывать дополнительный контекст. Сендерович и др. [13] показали, что качество прогноза существенно зависит как от внутриэкземплярных, так и межэкземплярных признаков, влияющих на временную структуру процесса.

Адекватное представление временных характеристик событий является ключевым фактором качества моделей предиктивного мониторинга процессов. Каземи и др. [14]

предложили метод Time2Vec, реализующий дифференцируемое отображение скалярного значения времени в вектор, состоящий из одного линейного и нескольких синусоидальных компонентов с обучаемыми частотами и фазами. Данный метод обеспечивает инвариантность к сдвигу и корректно захватывает периодические паттерны, однако не отражает особенности межсобытийных интервалов со значительно варьирующейся длительностью, характерных для процессов с диапазоном задержек от минут до месяцев.

Циклические преобразования ($\sin/\cos$) часа суток и дня недели широко применяются в прогнозировании временных рядов [2, 15], однако отражают лишь ограниченный аспект темпоральной структуры событий. В то же время функции экспоненциального затухания, моделирующие убывающую «память» системы о предшествующих событиях, активно используются в моделях точечных процессов. Тем не менее в контексте моделей предиктивного мониторинга процессов и, в частности, при построении квантильных прогнозов такие представления временной структуры пока не получили систематического рассмотрения.

Лим и др. [15] в архитектуре TFT предложили поэлементный пропускной механизм с сигмоидной активацией, обеспечивающий взвешенный отбор входных признаков. Однако известно, что такие механизмы подвержены деградации при насыщении сигмоидной функции, когда выходные веса стремятся к нулю. Это приводит к фактическому затуханию пропускного механизма, т. е. подавлению проходящего через него сигнала [16].

В области предиктивного мониторинга процессов трансформерные архитектуры [17] пока не получили широкого распространения: большинство исследований опираются на рекуррентные модели [9, 10] или ансамбли деревьев решений [4]. При этом остается открытой задача разработки моделей, способных формировать совместный вероятностный прогноз нескольких контрольных точек в едином пространстве состояний и учитывать их взаимные зависимости, что требует механизма согласованного внимания между такими точками. Формализация и решение данной задачи представляют собой важные шаги к построению более информативных и интерпретируемых предиктивных систем.

Сформулированная выше задача требует модели, одновременно способной корректно кодировать межсобытийные интервалы с существенно варьирующейся длительностью [14, 15], формировать согласованный вероятностный прогноз нескольких контрольных точек с учетом их взаимных зависимостей [4, 10] и последовательно обновлять контекст по мере поступления новых событий [8, 9].

Адаптивный вероятностный трансформер

Указанные ограничения обусловили разработку адаптивного вероятностного трансформера типа «последовательность-в-вектор» (Adaptive Probabilistic Sequence-to-Vector Transformer, APS2V)¹, предназначенного для предиктивного мониторинга редакционного процесса. В отличие от методов, предсказывающих единственную скалярную характеристику, APS2V должен формировать согласованный многомерный прогноз, задающий совместное вероятностное распределение времени достижения ключе-

¹Исходный код. – URL: https://github.com/DmitrijSyantov/OJS_timing (дата обращения: 10.03.2026).

Figure 1: Architecture of the proposed model. The diagram illustrates the flow from input data to the final quantized output. It starts with a 'rating_vals' input and a sequence of input features $[s_i, \Delta t_i, h_i, d_i, r_i, w_i, a_i]$ for $i=1..T$. These are projected into a 64-dimensional model space. The input is split into seven parallel processing blocks: Stage Embedding, Context Embedding, Load MLP, SimpleTimeEncoder, Age Linear, and Status Linear. These outputs are combined in a GatedVariableSelection block, followed by a TransformerEncoder. The output of the TransformerEncoder is then processed by a RatingMLP and a CrossAttentionLayer (Pre-LN). The CrossAttentionLayer output is split into two paths: one for a LastEvent block and another for a Fusion block. The LastEvent block output is combined with the Fusion block output in a Control Point-Context Attention (CPCA) block. The CPCA output is then processed by an Adaptive gating block and a Fusion block. The final output is processed by an Inter-Control-Point Attention (ICPA) block, which includes Pre-LN MHA, Pre-LN FFN, and an Interpolating gate. The final output is quantized ($\times 19$) to produce the final result.

Fig. 1. APS2V Architecture Graph

По наблюдаемому префиксу $\sigma_{1:t} = \langle e_1, e_2, \dots, e_n \rangle$ требуется предсказать абсолютные длительности до пяти контрольных точек редакционного цикла, такие как начало рецензирования (*rv*), завершение рецензирования (*rr*), решение редактора (*ed*), плани-

рование публикации (ps) и публикация (pub). Соответствующий целевой вектор обозначим

$$\mathbf{y} = (y_{rv}, y_{rr}, y_{ed}, y_{ps}, y_{pub})^T \in \mathbb{R}^{N_c}, \quad (1)$$

где $N_c = 5$ – число контрольных точек редакционного цикла.

Выход модели задается не одним значением, а набором условных квантилей

$$\hat{\mathbf{y}}(q) = (\hat{y}_{rv}(q), \hat{y}_{rr}(q), \hat{y}_{ed}(q), \hat{y}_{ps}(q), \hat{y}_{pub}(q)), \quad (2)$$

где q принадлежит фиксированному множеству Q из 19 квантильных уровней в диапазоне от 0,02 до 0,98. Для каждой контрольной точки из множества N_c модель оценивает 19 условных квантилей, формируя матрицу прогнозов размерности $N_c \times |Q| = 5 \times 19$. Соответственно, выход модели имеет размерность $\hat{\mathbf{Y}} \in \mathbb{R}^{B \times N_c \times |Q|}$. Такое представление формирует согласованный вероятностный прогноз нескольких зависимых контрольных точек по наблюдаемому префиксу.

Для каждого события формируются шесть представлений размерности $d_{model} = 64$: стадии, контекста издания, системной загрузки, времени, статуса и возраста рукописи. Эти группы объединяются механизмом взвешенного отбора признаков, близким к блоку Gated Variable Selection из архитектуры TFT [15], но применяющим нормировку softmax [17] вместо независимых поэлементных пропускных механизмов с сигмоидной активацией:

$$\mathbf{w} = \text{softmax}(W_{gate} [\mathbf{g}^{(1)}; \dots; \mathbf{g}^{(6)}] + \mathbf{b}_{gate}) \in \mathbb{R}^6, \mathbf{h}_{base} = \sum_j w_j \mathbf{g}^{(j)}. \quad (3)$$

После добавления позиционных эмбеддингов представление $\mathbf{h}_{base}$ подается в двухслойный трансформерный энкодер [18], формирующий матрицу событийного контекста $Z \in \mathbb{R}^{B \times T \times d_{model}}$.

Параллельно пара признаков рецензирования [$rating_val$, has_rating] проецируется многослойным персептроном (англ. Multilayer Perceptron, MLP) в последовательность представлений $R \in \mathbb{R}^{B \times T \times d_{model}}$. Для интеграции редакционных оценок с событийным контекстом используется блок перекрестного внимания $C_{attn} = \text{CrossAttentionLayer}(Q = Z, KV = R)$, построенный на основе архитектуры Pre-LN Transformer. Блок включает операции предварительной нормализации, многоголового внимания (англ. Multi-Head Attention, МНА), остаточных соединений и последующей обработки посредством полносвязной сети прямого распространения (англ. Feed-Forward Network, FFN).

В последующие модули передается итоговое контекстуализованное представление C_{attn} , сформированное блоком перекрестного внимания. При этом матрица коэффициентов внимания используется только на этапе вычисления представления и не входит в дальнейший вычислительный граф модели. Для адаптивного согласования событийного контекста с информацией о рецензировании вычисляется коэффициент

$$\alpha = \sigma(\text{Linear}(\text{MaskedMean}(R))), \quad (4)$$

после чего формируется уточненное представление префикса

$$Z_p = \text{LayerNorm}(Z + \alpha \cdot C_{\text{attn}}). \quad (5)$$

Такое уточнение позволяет согласовать событийный контекст с информацией о рецензировании и передать более информативное описание состояния рукописи в последующие блоки.

Далее последовательность событий преобразуется в глобальное представление рукописи $\mathbf{g}_d$, формируемое путем агрегирования матрицы уточненного префиксного контекста Z_p с последующим дополнением агрегатными статистиками префикса и представлением последнего наблюдаемого события. На следующем этапе формируется множество из N_c контрольных точек, соответствующих прогнозируемым временным рубежам редакционного процесса. Для каждой контрольной точки строится запрос $\mathbf{cp}_q \in \mathbb{R}^{B \times N_c \times d_{\text{model}}}$: обучаемые эмбединги контрольных точек дополняются позиционным кодированием и проецируются в пространство признаков размерности d_{model} . Для построения контекстно-зависимых представлений контрольных точек используется механизм перекрестного внимания:

$$\mathbf{cp}_{\text{ctx}} = \text{CrossAttn}(\text{query} = \mathbf{cp}_q, \text{keys} = Z_p, \text{values} = Z_p). \quad (6)$$

В выражении (6) запросами выступают представления контрольных точек, а ключами и значениями – элементы наблюдаемого префикса процесса. Такой механизм перекрестного внимания (англ. Control Point-Context Attention, CPCA) обеспечивает построение специализированных представлений для каждой контрольной точки за счет избирательного применения информации из событийного контекста. В результате модель получает контекстно-зависимые описания различных этапов редакционного процесса, используемые для последующего прогнозирования их временных характеристик.

Представления, сформированные механизмом CPCA, дополнительно объединяются с исходными представлениями контрольных точек и глобальным дескриптором рукописи. Для этого контекстное представление после адаптивного взвешиванияconcatенируется с запросом контрольной точки и глобальным дескриптором $\mathbf{g}_d$, после чего применяется линейное преобразование, нелинейность GELU и нормализация слоя:

$$\hat{\mathbf{c}} = \text{LayerNorm}\left(\text{GELU}\left(\text{Linear}\left([\mathbf{cp}_q \parallel \mathbf{cp}_{\text{ctx}}^{\text{gated}} \parallel \mathbf{g}_d]\right)\right)\right) \in \mathbb{R}^{B \times T \times d_{\text{model}}}. \quad (7)$$

Таким образом, каждая контрольная точка получает собственное контекстно-зависимое представление, отражающее ее связь как с глобальным состоянием рукописи, так и с наблюдаемым префиксом событий.

Для моделирования взаимных зависимостей между контрольными точками используется межточечный трансформерный модуль (англ. Inter-Control-Point Attention, ICSPA), который последовательно применяет к представлениям $\hat{\mathbf{c}}$ механизм МНА с каузальной маской и предварительной нормализацией слоя (англ. Pre-Layer Normalization, Pre-LN), за которым следует полносвязный блок FFN. Каузальная маска обеспечивает передачу информации от более ранних контрольных точек к более поздним, что соответствует причинно-следственной структуре редакционного процесса.

Финальное обновление выполняется посредством адаптивного механизма взвешивания:

$$\mathbf{ica}_{gate} = \sigma(\text{Linear}(\hat{\mathbf{c}})), \hat{\mathbf{c}} \leftarrow \hat{\mathbf{c}} + \mathbf{ica}_{gate} \odot (\mathbf{c}_{post_ffn} - \hat{\mathbf{c}}). \quad (8)$$

Описанная схема обеспечивает согласованное моделирование взаимосвязей между ранними и поздними этапами процесса, позволяя модели учитывать порядок следования ключевых точек и усиливать релевантные связи.

Формирование вероятностного прогноза осуществляется на основе представлений контрольных точек, полученных после блока ICSPA. Выходной слой генерирует для каждой контрольной точки набор квантильных оценок, соответствующих уровням из множества Q и образующих согласованный вероятностный прогноз. Для обеспечения монотонного упорядочения квантилей используется специальная параметризация выходного слоя, предотвращающая пересечение соседних квантильных оценок.

Данные, обучение и интеграция

Модель обучалась на журнале событий, сформированном по данным редакционного документооборота систем OJS Белорусского государственного университета². До разбиения на выборки исходный набор включал 2568 экземпляров процесса (74 320 зарегистрированных событий). Медианная длина последовательности событий составляла 29 событий (при диапазоне от 3 до 131), что отражает высокую вариативность жизненного цикла рукописей. Все встречающиеся обозначения стадий были приведены к унифицированному словарю из 20 токенов. Наиболее частотными оказались «подача рукописи» (submissionSubmitted), «действия с файлами» (fileUploaded/Revised), «назначение рецензента» (reviewerAssigned) и «решение редактора» (log.editor.decision).

Обучение выполнялось посредством минимизации суммы квантильных потерь (англ. pinball loss), аппроксимирующих метрику непрерывной ранговой вероятностной меры (англ. Continuous Ranked Probability Score, CRPS) [18], с добавлением штрафного члена, обеспечивающего монотонность квантилей:

$$L_{total} = \frac{1}{N \cdot 5 \cdot K} \sum_{n,c,k} p_{q_k}(y_{n,c} - \hat{y}_{n,c}(q_k)) + \lambda \sum_{k=1}^{K-1} \max(0, \hat{y}(q_k) - \hat{y}(q_{k+1})). \quad (9)$$

Целевые метки для контрольных точек определялись по временным отметкам соответствующих редакционных действий. Эмпирические распределения длительностей

²Набор данных APS2V-OJS (dataset_inspect.csv). – URL: <https://w3.bsu.by/index.php/catalog/5> (дата обращения: 10.03.2026).

характеризуются высокой вариабельностью и выраженной асимметрией: медианное время до начала рецензирования составляет около 196 ч (8 сут) при 90-м перцентиле 1756 ч, тогда как для публикации эти значения равны 2361 и 5960 ч соответственно. Значительный разброс временных интервалов делает задачу вероятностного прогнозирования более обоснованной по сравнению с использованием точечных оценок и приводит к высоким абсолютным значениям метрики CRPS, измеряемым сотнями часов. Для повышения численной устойчивости обучения целевые длительности масштабировались делением на 1000. Поскольку исходные интервалы измеряются в часах и находятся в диапазоне от 3 до 5960 ч, такое преобразование уменьшает диапазон целевой переменной приблизительно с $[0; 6000]$ до $[0; 6]$, сохраняя относительные различия между наблюдениями и упрощая оптимизацию квантильного регрессора.

Отфильтрованный набор данных был разделен на обучающую (2028 завершенных рукописей) и валидационную (507 рукописей) выборки в соотношении 80/20 со стратификацией по идентификатору издания при фиксированном значении генератора случайных чисел 42. Из каждой обучающей трассы длиной n событий формировались обучающие примеры методом выборки префиксов, длина которых равномерно распределялась на интервале $[1, n - 1]$. Такой подход обеспечивал обучение модели на различных стадиях редакционного процесса. В качестве методов регуляризации использовались событийное прореживание (вероятность 0,15) и случайное исключение нейронов (dropout) с вероятностью 0,2.

Размер предложенного варианта модели составляет $\approx 263\,000$ обучаемых параметров ($d_{model} = 64$). Компактность архитектуры позволила исключить необходимость применения GPU: обучение выполнялось на шестиядерном CPU (AMD Ryzen 5 7535HS, 32 ГБ ОЗУ) и занимало порядка 2 с на эпоху.

Обучение проводилось с использованием оптимизатора AdamW с шагом $2 \cdot 10^{-4}$ и весовым затуханием $5 \cdot 10^{-4}$ в течение 150 эпох (общее время ≈ 5 мин). Применялись косинусное расписание обучения (с разогревом пять эпох) и механизм ранней остановки (patience = 20) по метрике CRPS. В рассматриваемом эксперименте критерий ранней остановки не был достигнут и обучение продолжалось до установленного лимита в 150 эпох, а лучшая валидационная метрика была получена на последней эпохе. Для стабилизации весов использовалось экспоненциальное скользящее среднее (ЕМА) с коэффициентом 0,995. Динамика обучения показана на рис. 2. Полный состав входных признаков, используемых моделью, приведен в табл. 1.

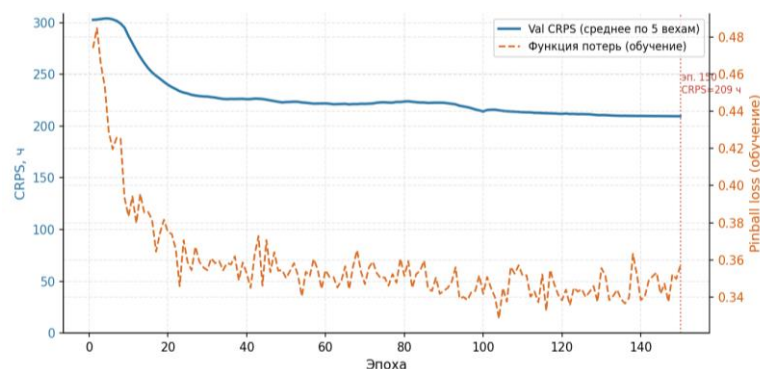


Рис. 2. Динамика обучения APS2V

Fig. 2. APS2V Learning Dynamics

Таблица 1
Входные признаки лога событийTable 1
Event log input features

Категория признаков <i>Category of features</i>	Признак <i>Sign</i>	Описание <i>Description</i>
Статусные	Идентификатор стадии	Категориальный признак события процесса
Статусные	Флаг состояния рукописи	Бинарный индикатор текущего статуса
Временные	Межсобытийная динамика	Логарифм межсобытийного интервала, циклические координаты часа суток и дня недели, а также набор экспоненциальных временных базисов
Редакционные	Оценка рецензента	Нормализованная оценка и бинарный индикатор ее наличия
Контекстные	Возраст рукописи	Логарифм суммарного времени с момента первичной подачи
Контекстные	Системная загрузка	Количество других активных рукописей в текущий момент
Контекстные	ID издания	Категориальный идентификатор журнала или конференции OJS

Точность прогнозирования

Для позиционирования APS2V относительно типичных методов предиктивного мониторинга было проведено сравнение с двумя базовыми моделями: рекуррентной нейронной сетью LSTM и ансамблем градиентного бустинга (GBM). Обе модели обучались и оценивались на одном и том же разбиении данных, описанном в предыдущем разделе. С целью обеспечения воспроизводимости результатов все эксперименты проводились при фиксированном значении начального состояния генератора псевдослучайных чисел ($\text{seed} = 42$). Для оценки устойчивости полученных результатов к стохастической составляющей процедуры обучения были дополнительно проведены контрольные эксперименты без фиксации генератора случайных чисел. Полученные в их рамках результаты демонстрируют соответствие основным выводам исследования и позволяют исключить случайный характер наблюдаемых различий. Оценка качества выполнялась по метрике CRPS с использованием единой сетки из 19 квантилей. Сравнение проводилось для нескольких долей наблюдаемой трассы (10, 40 и 90 %), что позволило проанализировать поведение моделей на ранних, промежуточных и поздних стадиях редакционного процесса.

Модель LSTM представляет собой двухслойную рекуррентную нейронную сеть со скрытым состоянием размерности 64 и многоквантильным выходным слоем. На вход модели подается тот же набор признаков, что и в случае предложенной архитектуры (стадия процесса, логарифмическая дельта, циклические признаки часа и дня недели, рейтинг, возраст рукописи, загрузка), однако в отличие от нее модель LSTM не содержит специализированного модуля временного кодирования и механизмов внимания. Процедура обучения предусматривала раннюю остановку по метрике CRPS на валидационной выборке при максимальном числе эпох, равном 150. Фактически обуче-

ние было завершено досрочно на 60-й эпохе ввиду отсутствия дальнейшего улучшения качества модели.

Модель GBM обучалась по схеме «отдельная модель на каждую контрольную точку × каждый квантиль» (95 моделей) на агрегированных признаках префикса: числе событий, статистиках дельт (среднее, дисперсия, максимум), возрасте рукописи, циклических признаках последнего события, счетчиках стадий и загрузке контекста. Использовался алгоритм градиентного бустинга для квантильной регрессии (200 деревьев, глубина четыре, скорость обучения 0,1). Результаты тестирования приведены в табл. 2.

Таблица 2

Сравнение моделей на валидационной выборке

Table 2

Comparison of models on the validation sample

Префикс Prefix, %	Модель Model	Начало рецензирования Start of review	Завершение рецензирования Completion of review	Решение редакции Editorial decision	План публикации Publication plan	Публикация Publication	Среднее Average
10	GBM	162,0	284,6	370,4	542,0	530,6	377,9
10	LSTM	175,5	306,6	419,8	589,7	578,2	414,0
10	APS2V	153,3	260,8	334,0	488,6	483,1	344,0
40	GBM	103,3	164,2	222,5	407,1	407,1	260,8
40	LSTM	78,2	260,5	160,2	515,5	502,1	303,3
40	APS2V	80,8	148,4	72,2	369,8	376,0	209,5
90	GBM	–	–	–	1094,0	276,2	685,1
90	LSTM	0,0	0,0	0,0	1373,7	708,9	1041,3
90	APS2V	0,0	0,0	0,0	588,0	764,2	676,1

Анализ вероятностных свойств

Результаты экспериментов показывают, что качество прогнозирования существенно зависит как от выбранной контрольной точки, так и от объема доступной информации о процессе. На ранних стадиях редакционного цикла наибольшей предсказуемостью характеризуются события начала и завершения рецензирования, тогда как для контрольных точек, связанных с планированием публикации и фактической публикацией, значения CRPS оказываются заметно выше. Данная закономерность согласуется с увеличением неопределенности по мере прохождения рукописью последовательности редакционных процедур и накопления влияния внешних факторов, не отраженных непосредственно в журнале событий.

Анализ вероятностных характеристик модели показал, что для широких прогнозных интервалов эмпирическое покрытие остается близким к номинальному уровню либо незначительно превышает его, тогда как для более узких интервалов наблюдается умеренное недопокрытие. Вместе с тем частота нарушений монотонности квантильных оценок остается низкой и не превышает 2,62 % даже для наиболее распространенных контрольных точек. Это подтверждает практическую согласованность формируемых вероятностных прогнозов и эффективность используемой монотонной параметризации квантильных выходов.

Абляционный анализ показал, что наибольший вклад в итоговое качество модели вносят механизм СРСА, обеспечивающий взаимодействие между контрольными точками и наблюдаемым префиксом процесса, и межточечный трансформерный модуль ИСРА, моделирующий зависимости между прогнозируемыми контрольными точками. Исключение каждого из этих компонентов приводит к наиболее заметному росту значений CRPS, что подтверждает их ключевую роль в формировании согласованных вероятностных прогнозов. Напротив, использование более сложного временного кодирования не обеспечило улучшения качества и сопровождалось увеличением среднего значения CRPS, вследствие чего данный компонент не был включен в финальную версию архитектуры.

Сравнение с базовыми моделями демонстрирует преимущество APS2V на различных стадиях наблюдения процесса. При использовании префикса длиной 10 % от полной трассы APS2V достигает среднего значения CRPS 344,0 ч против 377,9 ч для GBM и 414,0 ч для LSTM. Наиболее выраженное превосходство наблюдается при длине префикса 40 %, где средний CRPS составляет 209,5 ч по сравнению с 260,8 и 303,3 ч соответственно. Для поздних стадий процесса (90 % длины трассы) преимущество APS2V сохраняется: среднее значение CRPS для оставшихся активных контрольных точек составляет 676,1 ч против 685,1 ч у GBM и 1041,3 ч у LSTM. Следует отметить, что при данной длине префикса события начала рецензирования, завершения рецензирования и принятия редакционного решения для большинства рукописей уже относятся к наблюдаемой части трассы. Поэтому сравнение моделей на этапе 90 % фактически характеризует качество прогнозирования двух поздних контрольных точек. Полученные результаты свидетельствуют о том, что предложенная архитектура эффективно использует информацию о динамике процесса как при ограниченной, так и при практически полной наблюдаемости трассы.

Для оценки возможного влияния временной утечки данных дополнительно проведены эксперименты с хронологическим разделением выборки, при котором первые 80 % рукописей по времени поступления использовались для обучения, а последние 20 % – для валидации. В данном сценарии абсолютные значения CRPS ожидаемо возросли для всех моделей (табл. 3), однако APS2V сохраняет наилучшее качество прогнозирования (285,4 ч против 301,5 ч для XGBoost и 352,7 ч для LSTM). Это позволяет заключить, что наблюдаемое преимущество APS2V не является следствием временного перекрытия экземпляров между обучающей и валидационной выборками.

Таблица 3
Сравнение моделей на валидационной выборке

Table 3
Comparison of models on the validation sample

Модель <i>Model</i>	Начало рецензирования <i>Start of review</i>	Завершение рецензирования <i>Completion of review</i>	Решение редакции <i>Editorial decision</i>	План публикации <i>Publication plan</i>	Публикация <i>Publication</i>	Среднее <i>Average</i>
GBM	325,0	157,6	210,5	409,0	405,2	301,5
LSTM	298,7	243,9	261,9	481,3	477,8	352,7
APS2V	319,9	149,5	171,0	393,3	393,3	285,4

Ключевым преимуществом APS2V по сравнению с LSTM и GBM является использование единой вероятностной архитектуры, формирующей согласованные прогнозы одновременно для всех контрольных точек процесса. Механизмы CPSCA и ICSPA позволяют учитывать как связь контрольных точек с наблюдаемым префиксом, так и зависимости между прогнозируемыми событиями. Архитектура поддерживает последовательное уточнение прогнозов по мере поступления новых данных и допускает расширение множества контрольных точек без принципиального изменения структуры модели. Эти свойства делают APS2V перспективным инструментом для задач предиктивного мониторинга и имитационного моделирования редакционных процессов.

Ограничения исследования

Проведенное исследование имеет ряд ограничений, которые следует учитывать при интерпретации полученных результатов. Во-первых, модель обучалась и оценивалась на данных журналов событий одного домена, что ограничивает переносимость выводов на другие редакционные платформы и предметные области. Во-вторых, объем выборки (2568 завершенных рукописей) остается относительно небольшим для трансформерных архитектур, вследствие чего дальнейшее расширение набора данных может способствовать повышению качества прогнозирования. В-третьих, основная серия экспериментов проводилась на стратифицированном разбиении данных по экземплярам рукописей. Дополнительная проверка на хронологическом разделении выборки подтвердила устойчивость полученных результатов, однако дальнейшая валидация на независимых временных интервалах представляет практический интерес.

Анализ вероятностных характеристик показал, что модель обеспечивает удовлетворительную согласованность квантильных прогнозов и низкую частоту нарушений их монотонности. Вместе с тем для узких прогнозных интервалов сохраняется тенденция к недопокрытию, что указывает на перспективность дальнейшего развития методов калибровки вероятностных распределений и оценки неопределенности.

Заключение

В настоящей работе формализована и решена задача совместного вероятностного мониторинга многоэтапных бизнес-процессов на примере прогнозирования длительностей стадий редакционного документооборота. Предложенная архитектура APS2V формирует согласованные вероятностные оценки времени достижения нескольких контрольных точек процесса. Это позволяет учитывать взаимосвязи между этапами редакционного цикла и получать более информативное представление о неопределенности будущего развития процесса.

Проведенное исследование позволяет сделать следующие выводы:

1. Использование механизма перекрестного внимания совместно с межточечным трансформерным модулем оказывается более эффективным по сравнению с независимым прогнозированием отдельных контрольных точек. Совместное моделирование зависимостей между прогнозируемыми событиями позволяет учитывать структуру развития экземпляра процесса и формировать более точные вероятностные прогнозы на различных стадиях наблюдения.

2. Применение адаптивного пропускного механизма для интеграции гетерогенных признаков позволяет модели динамически регулировать вклад событийного контекста, информации о рецензировании и агрегированных характеристик процесса. Такой подход повышает устойчивость модели к вариативности и неполноте данных, характерным для журналов событий редакционных систем.

3. Экспериментальные результаты подтверждают перспективность использования единой трансформерной архитектуры для задач предиктивного мониторинга. В отличие от ансамблей независимых моделей и классических рекуррентных сетей APS2V обеспечивает согласованное вероятностное прогнозирование нескольких контрольных точек в рамках одной модели и допускает последовательное обновление прогнозов по мере поступления новых событий. Это делает предложенный подход перспективным для использования в задачах имитационного моделирования и анализа редакционных бизнес-процессов.

В перспективе предложенный подход отходит от парадигмы реактивного анализа логов. Дальнейшее развитие данного направления связано с разработкой методов постобработки и калибровки квантилей для узких интервалов, а также адаптацией предложенных механизмов внимания для задач оперативной оптимизации распределения ресурсов в реальном времени.

References

1. Dumas M., La Rosa M., Mendling J., Reijers H. A. *Fundamentals of Business Process Management*, 2nd ed. Berlin, Springer, 2018, 527 p. <https://doi.org/10.1007/978-3-662-56509-4>.
2. Aalst W. M. P. van der. *Process Mining: Data Science in Action*, 2nd ed. Berlin, Springer, 2016, 467 p. <https://doi.org/10.1007/978-3-662-49851-4>.
3. Verenich I., Dumas M., La Rosa M., Maggi F. M., Teinemaa I. Survey and cross-benchmark comparison of remaining time prediction methods in business process monitoring. *ACM Transactions on Intelligent Systems and Technology*, 2019, vol. 10, iss. 4, pp. 1–34. <https://doi.org/10.1145/3331449>.
4. Teinemaa I., Dumas M., La Rosa M., Maggi F. M. Outcome-oriented predictive process monitoring: review and benchmark. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2019, vol. 13, iss. 2, pp. 1–57. <https://doi.org/10.1145/3301300>.
5. Maggi F. M., Di Francescomarino C., Dumas M., Ghidini C. Predictive monitoring of business processes. *Advanced Information Systems Engineering: 26th International Conference, CAiSE 2014, Thessaloniki, Greece, 16–20 June 2014*, pp. 457–472 (Lecture Notes in Computer Science; vol. 8484). https://doi.org/10.1007/978-3-319-07881-6_31.
6. Teinemaa I., Dumas M., Maggi F. M., Di Francescomarino C. Predictive business process monitoring with structured and unstructured data. *Business Process Management: 14th International Conference, BPM 2016, Rio de Janeiro, Brazil, 18–22 September 2016*, pp. 401–417 (Lecture Notes in Computer Science; vol. 9850). https://doi.org/10.1007/978-3-319-45348-4_23.
7. Kratsch W., Manderscheid J., Röglinger M., Seyfried J. Machine learning in business process monitoring: A comparison of deep learning and classical approaches used for outcome prediction. *Business & Information Systems Engineering*, 2021, vol. 63, iss. 3, pp. 261–276. <https://doi.org/10.1007/s12599-020-00645-0>.
8. Evermann J., Rehse J.-R., Fettke P. Predicting process behaviour using deep learning. *Decision Support Systems*, 2017, vol. 100, pp. 129–140. <https://doi.org/10.1016/j.dss.2017.04.003>.

9. Camargo M., Dumas M., González-Rojas O. Learning accurate LSTM models of business processes. *Business Process Management: 17th International Conference, BPM 2019, Vienna, Austria, 1–6 September 2019*, pp. 286–302 (Lecture Notes in Computer Science; vol. 11675). https://doi.org/10.1007/978-3-030-26619-6_19.
10. Tax N., Verenich I., La Rosa M., Dumas M. Predictive business process monitoring with LSTM neural networks. *Advanced Information Systems Engineering: 29th International Conference, CAiSE 2017, Essen, Germany, 12–16 June 2017*, pp. 477–492 (Lecture Notes in Computer Science; vol. 10253). https://doi.org/10.1007/978-3-319-59536-8_30.
11. Navarin N., Vincenzi B., Polato M., Sperduti A. LSTM networks for data-aware remaining time prediction of business process instances. *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, Honolulu, HI, USA, 27 November – 01 December 2017. Available at: <https://arxiv.org/pdf/1711.03822> (accessed 10.03.2026). <https://doi.org/10.1109/SSCI.2017.8285184>.
12. Polato M., Sperduti A., Burattin A., de Leoni M. Time and activity sequence prediction of business process instances. *Computing*, 2018, vol. 100, iss. 9, pp. 1005–1031. <https://doi.org/10.1007/s00607-018-0593-x>.
13. Senderovich A., Di Francescomarino C., Ghidini C., Jorbina K., Maggi F. M. Intra and inter-case features in predictive process monitoring: A tale of two dimensions. *Business Process Management: 15th International Conference, BPM 2017, Barcelona, Spain, 10–15 September 2017*, pp. 306–323 (Lecture Notes in Computer Science; vol. 10445). https://doi.org/10.1007/978-3-319-65000-5_18.
14. Kazemi S. M., Goel R., Eghbali S., Ramanan J., Sahota J., ..., Brubaker M. *Time2Vec: Learning a vector representation of time*, 2019. Available at: <https://arxiv.org/pdf/1907.05321> (accessed 10.03.2026). <https://doi.org/10.48550/arXiv.1907.05321>.
15. Lim B., Arik S. Ö., Loeff N., Pfister T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 2021, vol. 37, iss. 4, pp. 1748–1764. <https://doi.org/10.1016/j.ijforecast.2021.03.012>.
16. Shazeer N., Mirhoseini A., Maziarz K., Davis A., Le Q., Hinton G., Dean J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April 2017*. Available at: <https://arxiv.org/pdf/1701.06538> (accessed 10.03.2026). <https://doi.org/10.48550/arXiv.1701.06538>.
17. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., ..., Polosukhin I. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, vol. 30, pp. 6000–6010. <https://doi.org/10.48550/arXiv.1706.03762>.
18. Gneiting T., Raftery A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 2007, vol. 102, no. 477, pp. 359–378. <https://doi.org/10.1198/016214506000001437>.
19. Gorishniy Y., Rubachev I., Khrulkov V., Babenko A. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021, vol. 34, pp. 18932–18943. <https://doi.org/10.48550/arXiv.2106.11959>.

Информация об авторе

Сьянов Дмитрий Андреевич, старший преподаватель кафедры веб-технологий и компьютерного моделирования, Белорусский государственный университет.
E-mail: syanov@bsu.by
<https://orcid.org/0009-0004-5347-8806>

Information about the author

Dmitry A. Syanov, Senior Lecturer, Department of Web Technologies and Computer Modeling, Belarusian State University.
E-mail: syanov@bsu.by
<https://orcid.org/0009-0004-5347-8806>