 © Авторы статьи / Authors of the article

ИНФОРМАТИКА И ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ/INFORMATICS AND INFORMATION PROCESSES

DOI: https://doi.org/10.60797/IRJ.2026.173.67 EDN: IDTIBB

МЕТОД ФОРМИРОВАНИЯ ПРИЗНАКОВОГО ПРОСТРАНСТВА ПОЗ И ТЕМПОРАЛЬНЫЙ BILSTM-КЛАССИФИКАТОР С МЕХАНИЗМОМ АДДИТИВНОГО ВНИМАНИЯ ДЛЯ РАСПОЗНАВАНИЯ ПОВЕДЕНИЯ ПЛОВЦОВ

Научная статья

Козлов Э.Ю.[1], Минязев Р.Ш.[2], Гибадуллин Р.Ф.[3,] *, Габдрахманов Э.М.[4]

[3] ORCID : 0000-0001-9359-911X;

[1, 2, 3, 4] Казанский национальный исследовательский технический университет им. А.Н. Туполева – КАИ, Казань, Российская Федерация

* Корреспондирующий автор (landwatersun[at]mail.ru)

Предложена: 27.08.2026; Принята: 03.11.2026; Опубликована: 17.11.2026

Аннотация

Распознавание поведения пловцов на видеопотоке плавательного бассейна – центральная задача систем автоматического обнаружения утопающих. В работе предложено 89-мерное признаковое описание позы, в котором относительные координаты 17 ключевых точек тела (конфигурация COCO Keypoints), оценки уверенности их извлечения, нормированные параметры ограничивающего прямоугольника и поэлементные межкадровые скорости описывают наблюдаемого человека независимо от его положения и масштаба в кадре, но без потери информации о реальной интенсивности движений. Над этим описанием построен темпоральный классификатор – двунаправленная сеть с долгой краткосрочной памятью (BiLSTM) и слоем аддитивного внимания, разделяющая поведение на три класса: активное плавание, пассивное пребывание в воде и утопление. Промежуточный класс введён сознательно: без него отдых на воде неотличим от утопления, что и порождает основную долю ложных тревог в системах с бинарным разбиением. На собранном корпусе из 19 247 последовательностей поз модель достигает общей точности 0,920 и F1-меры 0,924 для класса «утопление». Сравнение с десятью альтернативами (от логистической регрессии до Transformer) и серия ablation-экспериментов показывают, что выбранный набор компонентов оптимален по соотношению «качество – вычислительные затраты».

Ключевые слова: обнаружение утопающих, распознавание поведения, оценка поз человека, ключевые точки COCO, двунаправленная LSTM, механизм аддитивного внимания, темпоральный анализ, признаковое пространство, глубокое обучение, компьютерное зрение.

THE METHOD OF FORMING A FEATURE SPACE OF POSES AND A TEMPORAL BILSTM CLASSIFIER WITH AN ADDITIVE ATTENTION MECHANISM FOR RECOGNIZING SWIMMERS' BEHAVIOR

Research article

Kozlov E.Y.[1], Minyazev R.S.[2], Gibadullin R.F.[3,] *, Gabdrahmanov E.M.[4]

[3] ORCID : 0000-0001-9359-911X;

[1, 2, 3, 4] Kazan National Research Technical University named after A.N. Tupolev – KAI, Kazan, Russian Federation

* Corresponding author (landwatersun[at]mail.ru)

Suggested: 27.08.2026; Accepted: 03.11.2026; Published: 17.11.2026

Abstract

Recognizing swimmers' behavior in a swimming pool video stream is the central task of automatic drowning detection systems. The paper proposes an 89-dimensional feature description of posture, in which the relative coordinates of 17 key body points (COCO Keypoints configuration), confidence estimates for their extraction, normalized bounding rectangle parameters, and element-wise inter-frame velocities describe the observed person regardless of their position and scale in the frame, but without losing information about the actual intensity of the movements. A temporal classifier has been built on top of this description – a bidirectional long short-term memory network (BiLSTM) with an additive attention layer, which divides behavior into three classes: active swimming, passive immersion in water, and drowning. The intermediate class was introduced deliberately: without it, rest on water is indistinguishable from drowning, which is what generates the majority of false alarms in systems with binary classification. On the assembled corpus of 19,247 pose sequences, the model achieves an overall accuracy of 0.920 and an F1 score of 0.924 for the "drowning" class. A comparison with ten alternatives (from logistic regression to Transformer) and a series of ablation experiments show that the selected set of components is optimal in terms of the "quality – computational cost" ratio.

Keywords: Drowning detection, behavior recognition, human pose estimation, COCO keypoints, bidirectional LSTM, additive attention mechanism, temporal analysis, feature space, deep learning, computer vision.

Введение

По оценкам Всемирной организации здравоохранения, ежегодно от утопления гибнет около 236 тысяч человек – это одна из ведущих причин непреднамеренной смерти от внешних воздействий. Особая сложность для спасательной службы состоит в том, что утопающий, в отличие от пострадавшего при большинстве других травм, как правило, не

 © Авторы статьи / Authors of the article

подаёт ни жестовых, ни голосовых сигналов о помощи: он погружается медленно, в вертикальном положении, и утрачивает способность дышать над водой в течение 20–60 с. В условиях загруженного бассейна – при бликах на поверхности воды, волнении, неудачно расположенном освещении и накапливающейся усталости оператора – такое состояние может остаться незамеченным даже при непосредственном визуальном контроле [1].

Технические предпосылки для автоматизации такого контроля в настоящее время имеются: общественные бассейны, как правило, оснащены системами видеонаблюдения, а современные вычислительные средства позволяют обрабатывать видеопоток ресурсоёмкими нейросетевыми моделями в режиме реального времени. Основная трудность связана не с аппаратными возможностями, а со свойствами наблюдаемого сигнала. Один и тот же кадр – человек неподвижно лежит на спине – может означать и отдых на воде, и нырок, и начавшееся утопление [2]. Различие проявляется не в отдельном кадре, а в короткой временной окрестности – в изменении позы и движений конечностей на интервале 1–2 с. Поэтому обнаружение утопающего следует рассматривать не как задачу классификации изображений, а как задачу анализа поведения и формулировать её соответствующим образом.

Существуют коммерческие системы такого назначения – Poseidon (Франция), AngelEye (Италия), MyLifeguard (Израиль), – однако их алгоритмы закрыты, а адаптация к конкретному бассейну возможна только при участии производителя. В открытых публикациях последних лет также достигнут заметный прогресс:

– для выделения фигур пловцов применяются методы инстанс-сегментации, восходящие к Mask R-CNN [3];

– систематизация методов детекции по типу сенсоров и подходу к классификации приведена в обзорных работах [1], [2];

– в [4] покадровое обнаружение утопления в бассейне реализовано улучшенным детектором семейства YOLO.

Теоретическую базу настоящей работы составляют классические результаты по свёрточным сетям и глубокому обучению [5], рекуррентным архитектурам с долгой краткосрочной памятью [6] и механизмам внимания [7].

Общим ограничением большинства известных подходов является выбор уровня анализа. Покадровые методы дают высокую долю ложных тревог, поскольку статичная поза отдыха на воде визуально не отличается от позы утопающего. Бинарные постановки «норма / утопление» сталкиваются с той же проблемой с другой стороны: класс «норма» оказывается крайне неоднородным, и обученная на нём модель наследует эту неоднородность в виде систематической ошибки. В настоящей работе эти трудности преодолеваются по двум направлениям.

Во-первых, задача переформулирована как трёхклассовая: к классам «активное плавание» и «утопление» добавлен явный класс «пассивное пребывание в воде», на котором модель учится отличать отдых от опасности.

Во-вторых, входное пространство построено как описание позы, инвариантное к сдвигу и масштабу объекта в кадре, но сохраняющее физическую интенсивность движений; над ним работает двунаправленная LSTM с аддитивным вниманием.

Цель работы – построить признаковое описание и темпоральную модель, дающие в условиях плавательного бассейна более точное распознавание поведения пловца, чем известные аналоги. Соответственно, в работе формализуется трёхклассовая постановка задачи, обосновывается состав признакового вектора, описывается архитектура темпорального классификатора, приводятся результаты обучения и сравнение с альтернативными моделями.

Постановка задачи распознавания поведения пловца

Классификатор работает не с пикселями, а с последовательностями поз: для каждого человека, попавшего в кадр, отдельная модель оценки поз выдаёт набор ключевых точек тела, и уже над этим компактным описанием строится поведенческий анализ. Поза одного человека в одном кадре естественно записывается тензором:

$$\boldsymbol{K}_t^{(i)} \in \mathbb{R}^{17\times 3}, \tag{1}$$

где i – индекс человека в кадре, первая размерность – это 17 ключевых точек тела в конфигурации COCO Keypoints [8] (нос, глаза, уши, плечи, локти, запястья, бёдра, колени, лодыжки), а вторая – координаты x, y и оценка уверенности их извлечения. Координаты приводятся к ширине и высоте кадра, так что x, y ∈ [0, 1].

Сами по себе эти точки ещё не образуют пригодного для обучения сигнала: их нужно перевести в инвариантное представление и дополнить динамикой. После такой обработки получается 89-мерный вектор признаков на каждый кадр – 34 относительные координаты (17 × 2), 17 оценок уверенности, 4 параметра ограничивающего прямоугольника и 34 компоненты межкадровых скоростей. Подробное обоснование состава вектора и формулы для его компонент даны в Разделе 2; здесь существенна лишь его итоговая размерность:

$$x_t^{(i)} = \left[k_t^{\mathrm{rel}}\,; k_t^{\mathrm{conf}}\,; b_t; v_t\right] \in \mathbb{R}^{89}. \tag{2}$$

Темпоральный классификатор обрабатывает последовательности длиной T = 30 кадров – около секунды при частоте 30 Гц – и реализует отображение:

$$g_\theta : \mathbb{R}^{30\times 89} \to \Delta^3, \tag{3}$$

где Δ^3– вероятностный симплекс в $\mathbb{R}^3$ (множество распределений вероятностей по трём классам: «активное плавание», «пассивное пребывание в воде», «утопление»), а θ – обучаемые параметры. Длина окна около 1 с обусловлена двумя ограничениями: на меньшем интервале динамика движений рук и ног не успевает проявиться, а при большем задержка реакции системы становится критичной для своевременного вмешательства.

Качество модели оценивается стандартным набором метрик: общая точность (accuracy), точность и полнота по целевому классу «утопление» (precision и recall), F1 – мера как их гармоническое среднее. Целевые значения, исходя из требований к практической работе в типовом бассейне, заданы единым порогом: accuracy, precision, recall, F1 ≥ 0,90.

 © Авторы статьи / Authors of the article

Этот порог не претендует на универсальность – он отражает уровень, ниже которого, по экспертной оценке, система перестаёт быть полезным дополнением к работе оператора.

Отдельного обоснования требует выбор числа классов. В подавляющем большинстве известных работ задача ставится бинарно: «утопление» против всего остального. Такая постановка удобна технически, но содержательно уязвима: в одном классе оказываются и быстрое плавание, и неподвижный отдых на воде, и стояние у бортика, и нырок. Модель, обученная отделять утопление от столь неоднородного класса, вынуждена компенсировать его разнородность за счёт ложных корреляций, что в условиях реальной эксплуатации приводит к росту числа ложных срабатываний именно на пассивных позах отдыха. Введение в постановку явного третьего класса – «пассивное пребывание в воде» – снимает это противоречие на уровне формулировки задачи, а не на уровне постобработки.

Метод формирования признакового пространства

Качество темпорального классификатора в значительной мере определяется его входным описанием. К признаковому описанию предъявляются два требования. Во-первых, оно должно быть инвариантно к факторам, изменяющимся при движении камеры или перемещении пловца по бассейну: сдвигу, масштабу, разрешению кадра. Во-вторых, оно должно сохранять информацию, существенную для распознавания поведения: взаимное расположение конечностей, амплитуду и направление их движений. Эти требования частично противоречат друг другу, поэтому вектор признаков формируется так, чтобы каждая его компонента отвечала одному из них.

Описание раскладывается на четыре блока. Первый – относительные координаты ключевых точек: их получают делением координат, отсчитанных от центра ограничивающего прямоугольника, на его размеры:

$$k_t^{rel}[i] = \frac{k_t^{abs}[i] - c_t}{s_t},$$

где $k_t^{abs}[i] = (x_i, y_i)$ – абсолютные координаты i-й точки, ct = (xc, yc) – центр ограничивающего прямоугольника, st = (w, h) – его размеры; деление поэлементное, к знаменателю прибавляется 10−6 для устойчивости к вырожденным случаям. После такого нормирования значения координат концентрируются вблизи нуля и не зависят ни от положения пловца в кадре, ни от его видимого размера – двух основных источников вариативности, несущественной для распознавания поведения.

Второй блок – четыре нормированных параметра самого ограничивающего прямоугольника:

$$b_t = \left(\frac{x_c}{W}, \frac{y_c}{H}, \frac{w}{w}, \frac{h}{H} \right), \tag{5}$$

где W и H – ширина и высота кадра в пикселях. Если первый блок исключает из описания положение и масштаб, то второй возвращает их в нормированном виде. Это необходимо в ситуациях, когда значение имеет сама геометрия объекта в кадре: небольшой объект у края кадра, вероятно, удалён от камеры, и его движения регистрируются с иными перспективными искажениями, чем движения объекта на переднем плане, занимающего половину кадра.

Третий блок – межкадровые скорости. Их можно было бы вычислять как разность относительных координат, однако это привело бы к нежелательному эффекту: при изменении размеров ограничивающего прямоугольника (например, когда пловец удаляется от камеры) разности нормированных координат начинают зависеть скорее от изменения масштаба, чем от реальных движений конечностей. Поэтому скорости вычисляются по абсолютным координатам:

$$v_t[i] = k_t^{abs}[i] - k_{t-1}^{abs}[i]. \tag{6}$$

Если в предыдущем кадре пловец отсутствовал, vt = 0. Размерность блока – 34 (17 точек, по две координаты). Именно эта компонента описания является решающей для разделения активного плавания и пассивного пребывания в воде: при плавании движения рук и ног быстрые и регулярные, а при пассивном дрейфе и тем более при утоплении – мелкие, нерегулярные, нередко хаотичные. По статическим позам эти состояния различить трудно: на отдельном кадре поза человека, отдыхающего на спине, может практически не отличаться от позы утопающего.

Четвёртый блок – оценки уверенности извлечения ключевых точек, которые модель оценки поз выдаёт вместе с координатами. Их 17 – по одной на точку. Их включение во входное описание обусловлено тем, что часть точек неизбежно извлекается ненадёжно или теряется: рука погружается под воду, плечо перекрывается другим пловцом, голову закрывает всплеск. Оценки уверенности позволяют модели снижать вклад недостоверно извлечённых компонент и опираться на более надёжные.

Полное описание одного кадра – конкатенация всех четырёх блоков:

$$x_t = \left[k_t^{\mathrm{rel}} ; k_t^{\mathrm{conf}} ; b_t ; v_t \right] \in \mathbb{R}^{17 \cdot 3 + 4 + 17 \cdot 2}. \tag{7}$$

Перед подачей в нейронную сеть вектор стандартизуется покомпонентно – каждая координата приводится к нулевому среднему и единичной дисперсии по обучающей выборке:

$$\tilde{x}_t = \frac{x_t - \mu}{\sigma}, \tag{8}$$

где μ, σ ∈ ℝ89 оцениваются по всем кадрам обучающего набора. Эта стандартная процедура исключает доминирование отдельных блоков (координат, скоростей, оценок уверенности) из-за различия масштабов их значений и обеспечивает устойчивую сходимость обучения. Размерность 89 – компромисс: она достаточно велика, чтобы вместить всё перечисленное без потерь, и достаточно невелика, чтобы рекуррентная модель оставалась лёгкой и быстрой.

 © Авторы статьи / Authors of the article

Архитектура темпорального классификатора

Выбор архитектуры здесь – компромисс между двумя соображениями. С одной стороны, темпоральный сигнал, который нужно извлечь, – это движение конечностей в окне порядка секунды. Такие зависимости имеют малую протяжённость, и для их моделирования архитектура Transformer с миллионами параметров избыточна. С другой – простая однонаправленная рекуррентная сеть видит контекст только из прошлого, тогда как окно длиной в секунду имеет смысл обрабатывать целиком, опираясь и на начало, и на конец. Двунаправленная LSTM удовлетворяет обоим требованиям, а её дополнение слоем аддитивного внимания [9] позволяет модели акцентировать те моменты последовательности, в которых поведенческое содержание максимально (например, момент потери равновесия в воде), и игнорировать остальные; результативность такого сочетания двунаправленной рекуррентной сети с механизмом внимания при извлечении темпоральных представлений подтверждена в смежных задачах анализа последовательностей [10], [11], в том числе при классификации поз по последовательностям ключевых точек тела [12].

Сеть собрана в PyTorch [13] и состоит из четырёх блоков. На вход подаётся тензор формы (B, 30, 89), где B – размер мини-батча; за ним следует двунаправленный двухслойный LSTM с размером скрытого состояния 128 единиц на каждое направление (256 в сумме), с Dropout 0,5 между слоями. Над выходом LSTM работает слой аддитивного внимания: линейное преобразование $256 \to 1$, softmax по временной оси и взвешенное усреднение скрытых представлений с полученными весами. Результат – один контекстный вектор размерности 256 на каждую последовательность. Далее следует полносвязный классификатор $256 \to 64 \to 3$ с функцией активации ReLU и слоем Dropout 0,5 между полносвязными слоями; на выходе формируются логиты по трём классам, преобразуемые функцией softmax.

Полное число обучаемых параметров – около 636 тысяч. Это в 40–160 раз (на 1,5–2 порядка) меньше, чем у современных свёрточных классификаторов с 25–100 млн параметров; для архитектуры, работающей не с пикселями, а с уже выделенным позным описанием, такая компактность естественна. На практике это обеспечивает быстрое обучение, умеренные требования к памяти GPU и низкую задержку при работе модели в реальном времени: на NVIDIA RTX 4080 одно окно последовательности обрабатывается менее чем за две миллисекунды, что оставляет запас на одновременный анализ нескольких десятков пловцов.

Обучение ведётся методом мини-батчей с оптимизатором AdamW [14]. Из-за заметного дисбаланса классов в обучающем наборе обычная перекрёстная энтропия даёт смещённое решение, поэтому функция потерь взвешена по обратной частоте классов:

$$L = -\sum_{c=1}^{C} w_c \cdot y_c \cdot \log\left(p_c\right) \tag{9}$$

где C = 3, yc – компоненты one-hot-кодировки истинной метки, pc – соответствующие предсказанные вероятности, а вес класса

$$w_c = \frac{N}{C \cdot n_c}, \tag{10}$$

где N – общее число обучающих примеров, nc – число примеров класса c. Фактическое соотношение классов в собранной выборке составляет примерно 2 : 1,4 : 1 (плавание : пассивное пребывание : утопление); без взвешивания решения модели смещаются в пользу наиболее представленного класса, и полнота по целевому классу «утопление» снижается.

Схема регуляризации включает четыре стандартных взаимодополняющих приёма. Аддитивный гауссов шум с амплитудой 0,02 добавляется к признакам на каждом проходе и снижает чувствительность модели к точным значениям координат. С вероятностью 0,3 случайно выбранный кадр в последовательности заменяется его предшественником – это имитирует пропуски кадров, возникающие в реальных видеопотоках. С вероятностью 0,2 амплитуда признаков масштабируется случайным коэффициентом из U [0,9;1,1] – небольшое сжатие или растяжение, не нарушающее семантики последовательности. Норма градиента ограничивается порогом 1,0, что существенно главным образом на первых эпохах, когда обучение LSTM наименее устойчиво. Обучение останавливается, если функция потерь на валидации не улучшается в течение 15 эпох подряд.

Экспериментальное исследование

Готовых наборов данных для трёхклассового разделения «плавание / пассивное пребывание / утопление» в открытом доступе на момент работы не было, поэтому корпус собран авторами вручную. Из 320 предварительно отсмотренных видеороликов на YouTube для дальнейшей работы оставлены 152 общей длительностью 7 ч 14 мин. После разметки ролики распределились по классам следующим образом: 47 роликов с активным плаванием, 42 – с пассивным пребыванием в воде, 38 – с эпизодами утопления, ещё 25 оставлены в резерв на случай досборки. Далее этот корпус обозначается как DDP (Drowning Detection Pose).

Из видео последовательности поз извлекались скользящим окном длиной 30 кадров с шагом 15 (то есть с перекрытием в половину окна) – это даёт более плотное покрытие при той же длительности и снижает чувствительность к выбору точки старта. Использовалась модель YOLOv8m-Pose — вариант средней ёмкости из семейства YOLOv8-Pose [15]; кадры с уверенностью детекции ниже 0,3 или менее чем 5 валидными ключевыми точками отбрасывались. В итоге получено 19 247 последовательностей: 8 731 класса swimming (45,4 %), 6 102 класса idle (31,7 %) и 4 414 класса drowning (22,9 %). Примерно одна пятая (3 850 последовательностей) отделена в тестовую подвыборку при стратифицированном разделении.

Гиперпараметры подобраны сеточным поиском по основным осям: размерность скрытого состояния LSTM (значения 64, 96, 128, 192, 256), число слоёв LSTM (1, 2, 3), Dropout (0,2 – 0,6 с шагом 0,1), длина окна (15, 24, 30, 45, 60 кадров), размер мини-батча (16, 32, 64) и начальная скорость обучения (0,0001; 0,0005; 0,001; 0,005). Лучшая

 © Авторы статьи / Authors of the article

комбинация – hidden_size = 128, num_layers = 2, dropout = 0,5, seq_length = 30, batch_size = 32, lr = 0,0005. Полное обучение на этой конфигурации занимает 1 мин 24 с на NVIDIA RTX 4080 при 100 эпохах с ранней остановкой по терпению 15 эпох; как правило, ранняя остановка срабатывает на 60–70-й эпохе.

На тестовой подвыборке получены следующие значения метрик: общая точность 0,920; макросредняя F1 0,915; F1 по классам swimming, idle и drowning – 0,933, 0,889 и 0,924 соответственно; precision и recall по классу drowning – 0,920 и 0,929. Все цели, сформулированные в Разделе 1 (порог 0,90 по основным метрикам), достигнуты.

Чтобы оценить статистическую погрешность полученных значений, обусловленную конечным объёмом тестовой подвыборки, для метрик построены 95-процентные доверительные интервалы. Процедура построения интервала выбиралась в зависимости от того, какой статистической величиной является метрика.

Общая точность, а также точность и полнота по отдельному классу являются биномиальными долями: это доли верных решений соответственно среди всех n = 3 850 примеров тестовой подвыборки, среди примеров, отнесённых моделью к данному классу, и среди примеров этого класса (столбец «Поддержка» в таблице 1). Для них использован интервал Уилсона [16]. Он предпочтительнее интервала Вальда, основанного на нормальной аппроксимации: фактическая вероятность накрытия интервала Вальда, как правило, ниже номинальной и немонотонно колеблется при изменении объёма выборки и значения доли, особенно при долях, близких к 0 или 1, тогда как у интервала Уилсона она остаётся близкой к номинальной во всём диапазоне значений [17].

F1-мера биномиальной долей не является, поэтому формула Уилсона к ней непосредственно не применима. Для F1 по отдельному классу интервал построен с использованием её взаимно однозначной связи с индексом Жаккара J:

$$\mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} = \frac{2J}{1+J}, \quad J = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}, \tag{11}$$

где TP – число примеров класса, верно отнесённых к нему моделью; FP – число примеров других классов, ошибочно отнесённых к данному классу; FN – число примеров класса, ошибочно отнесённых к другим классам. При фиксированном числе m = TP + FP + FN примеров, принадлежащих классу по истинной метке или по решению модели, величина TP имеет биномиальное распределение с параметрами m и J. Поэтому для J строится интервал Уилсона $[J_L; J_U]$, границы которого переводятся в границы интервала для F1 монотонным преобразованием из (11):

$$\left[\frac{2J_L}{1+J_L}; \frac{2J_U}{1+J_U}\right]. \tag{12}$$

Поклассовые значения F1 вычисляются по одной и той же тестовой подвыборке и потому статистически зависимы. Для макро- и взвешенно-усреднённых F1, которые представляют собой линейные комбинации поклассовых значений с фиксированными коэффициентами (1/3 и долями классов в тестовой подвыборке соответственно), приведены консервативные интервалы: поклассовые интервалы строятся по формуле (12) с уровнем доверия 1 − 0,05/3 (поправка Бонферрони), а их границы усредняются с теми же коэффициентами. Такой интервал накрывает истинное значение с вероятностью не ниже 0,95 при любой зависимости между поклассовыми оценками. Результаты сведены в таблицу 1.

Таблица 1 - Метрики качества темпорального классификатора на тестовой подвыборке с 95 % доверительными интервалами для F1

DOI: https://doi.org/10.60797/IRJ.2026.173.67.1

Класс	Precision	Recall	F1	95 % CI для F1	Поддержка
swimming	0,936	0,930	0,933	[0,924; 0,941]	1 743
idle	0,877	0,902	0,889	[0,876; 0,902]	1 218
drowning	0,920	0,929	0,924	[0,911; 0,936]	889
Macro avg	0,911	0,920	0,915	[0,901; 0,929]	3 850
Weighted avg	0,914	0,920	0,917	[0,903; 0,930]	3 850

Примечание: для отдельных классов интервал построен по формуле (12), для усреднённых значений приведены консервативные интервалы с поправкой Бонферрони (уровень доверия не ниже 95 %)

Для общей точности интервал Уилсона составляет [0,911; 0,928], для точности и полноты по классу «утопление» – [0,900; 0,936] и [0,910; 0,944] соответственно. Полуширина интервалов для F1 не превышает 0,013 для отдельных классов и 0,014 – для усреднённых значений. Целевой порог 0,90, заданный в Разделе 1 для общей точности и метрик по классу «утопление», выполняется не только для точечных оценок, но и для нижних границ доверительных интервалов; для точности по этому классу нижняя граница совпадает с порогом. Приведённые интервалы характеризуют погрешность, обусловленную конечным объёмом тестовой подвыборки при фиксированной обученной модели. Разброс, связанный с выбором конкретного разбиения данных и случайной инициализацией весов, ими не учитывается; его оценка требует повторного обучения на нескольких разбиениях.

Сравнительный анализ архитектур

Само по себе достижение целевых метрик ещё не доказывает, что выбранная архитектура – лучший выбор для задачи. Чтобы оценить вклад архитектурных решений, на тех же данных и в тех же условиях обучены десять альтернативных моделей – от логистической регрессии до Transformer с 1,2 млн параметров. У всех архитектур

 © Авторы статьи / Authors of the article

одинаковыми оставались признаковое описание, протокол аугментации, валидационная схема и стратегия подбора гиперпараметров. Результаты сведены в таблицу 2.

Таблица 2 - Сравнение различных архитектур темпорального классификатора

DOI: https://doi.org/10.60797/IRJ.2026.173.67.2

Архитектура	Параметров, тыс.	Время обучения, с	Accuracy	F1 (drowning)
Логистическая регрессия	0,3	5	0,671	0,639
MLP (2 слоя × 128)	44,1	15	0,760	0,736
1D-CNN (3 блока)	147,3	35	0,825	0,812
GRU (1 слой, 128)	85,4	55	0,866	0,861
LSTM (1 слой, 128)	112,8	62	0,877	0,872
LSTM (2 слоя, 128)	244,7	85	0,896	0,891
BiLSTM (1 слой, 128)	225,5	78	0,903	0,900
BiLSTM (2 слоя, 128)	489,3	112	0,913	0,910
BiLSTM + Att. (предлаг.)	636,4	84	0,920	0,924
Transformer (4 слоя)	1 217,8	188	0,917	0,919
BiLSTM + Self-Att.	928,2	151	0,918	0,921

Результаты сравнения позволяют сделать ряд выводов. Простые модели – логистическая регрессия и MLP – существенно уступают остальным: они не учитывают временную структуру последовательности явным образом, и их общая точность в трёхклассовой постановке ограничена значениями 0,67–0,76. 1D-CNN [18] за счёт локальных свёрток в значительной мере сокращает это отставание, однако уступает однослойным рекуррентным моделям 4–5 процентных пунктов. Однонаправленные GRU и LSTM обеспечивают общую точность 0,866–0,896; переход к двунаправленному варианту прибавляет 1,7–2,6 процентного пункта общей точности в зависимости от глубины. Углубление с одного слоя до двух даёт ещё 1–2 пункта при примерно вдвое большем числе параметров. Слой аддитивного внимания добавляет 0,7 пункта общей точности (0,920 против 0,913) и 1,4 пункта F1 по классу «утопление» поверх BiLSTM (2 слоя). Существенно, что Transformer и BiLSTM с самовниманием, имея в 1,5–1,9 раза больше параметров и в 1,8–2,2 раза большее время обучения, не превосходят модель с аддитивным вниманием ни по общей точности, ни по F1. Это подтверждает, что для коротких поведенческих окон механизмы самовнимания избыточны.

Сравнение архитектур позволяет выбрать модель, но не показывает, какие элементы предлагаемого метода действительно влияют на качество классификации. Для ответа на этот вопрос проведена серия ablation-экспериментов: каждый элемент поочерёдно отключался или заменялся упрощённым вариантом, после чего измерялось изменение метрик. Результаты приведены в таблице 3.

Таблица 3 - Результаты ablation study по влиянию отдельных компонент метода на качество классификации

DOI: https://doi.org/10.60797/IRJ.2026.173.67.3

Конфигурация	Accuracy	F1 (drowning)	Δ от baseline
Полный набор + все техники (baseline)	0,920	0,924	–
Без аугментации шумом	0,905	0,908	−0,016
Без замены кадров	0,912	0,915	−0,009
Без масштабирования	0,916	0,919	−0,005
Без всех видов аугментации	0,889	0,886	−0,038
Без класса idle (2 класса)	0,931	0,940	+0,016
Без признаков скорости (55-мерн.)	0,879	0,872	−0,052
Без относит. координат (абс.)	0,896	0,890	−0,034

 © Авторы статьи / Authors of the article

Конфигурация	Accuracy	F1 (drowning)	Δ от baseline
Без bbox в признаках	0,913	0,916	−0,008
Без взвешивания классов	0,915	0,897	−0,027
Без двунаправленности (LSTM)	0,896	0,891	−0,033
Без механизма внимания	0,909	0,903	−0,021
Без Dropout (dropout = 0)	0,902	0,906	−0,018
seq_length = 15 (вместо 30)	0,880	0,871	−0,053
seq_length = 60 (вместо 30)	0,918	0,921	−0,003

Здесь и далее приращения приводятся в процентных пунктах F1 по классу «утопление» – именно эта величина сведена в колонку Δ таблицы 3; соответствующие изменения общей точности указаны в скобках. Самый большой вклад в качество вносят компоненты, связанные с темпоральной природой задачи: признаки скорости – 5,2 пункта F1 (4,1 пункта accuracy), длина окна 30 кадров вместо 15 – 5,3 пункта F1 (4,0 пункта accuracy), двунаправленность LSTM – 3,3 пункта F1 (2,4 пункта accuracy). На втором месте – устройство признакового описания: переход к относительным координатам вместо абсолютных стоит 3,4 пункта F1 (2,4 пункта accuracy), отсутствие любых аугментаций – 3,8 пункта F1 (3,1 пункта accuracy). Меньший, но измеримый вклад вносят взвешивание классов – 2,7 пункта F1 (0,5 пункта accuracy) – и аддитивное внимание – 2,1 пункта F1 (1,1 пункта accuracy). Параметры ограничивающего прямоугольника, отдельные виды аугментации и Dropout вносят меньший, но ненулевой вклад – порядка 0,5–1,8 пункта.

Отдельного обсуждения заслуживает строка «без класса idle». На первый взгляд она противоречит исходному построению работы: двухклассовая постановка даёт прирост accuracy на 1,1 пункта и F1 на 1,6 пункта. Однако этот прирост является кажущимся. Он возникает потому, что при объединении классов «пассивное пребывание» и «утопление» все нейтральные сцены отдыха на воде автоматически получают метку «утопление» – и попадают в ложноположительные срабатывания. На уровне внутренних метрик это выглядит как улучшение, но на уровне практического использования означает, что система генерирует постоянный поток ложных тревог. Поэтому формальное численное преимущество двухклассовой постановки – это артефакт метрики, а не реальное улучшение качества; именно отказ от такой постановки в пользу трёхклассовой и есть один из главных содержательных вкладов работы.

Обобщая результаты ablation-экспериментов, можно выделить минимальный набор компонент, без каждой из которых F1 по классу «утопление» опускается ниже целевого порога 0,90: признаки скорости, относительные координаты, двунаправленная LSTM, взвешивание классов, окно длиной 30 кадров, а также аугментация данных (достаточно хотя бы одного из трёх её видов). Удаление любой из этих компонент снижает F1 по классу «утопление» на 2,7 пункта и более. Остальные элементы – аддитивное внимание, параметры ограничивающего прямоугольника, отдельные виды аугментации и Dropout – дают прирост от 0,5 до 2,1 пункта каждый и могут включаться или исключаться в зависимости от вычислительного бюджета.

Заключение

В работе предложен подход к автоматическому распознаванию поведения пловцов, опирающийся на две идеи: компактное описание позы, инвариантное к сдвигу и масштабу пловца в кадре, и трёхклассовое разделение поведения вместо привычной бинарной формулировки. Каждая из идей по отдельности уже встречалась в литературе, но именно их сочетание оказывается ключевым.

Признаковое описание собирается из 89 компонент: 34 относительные координаты 17 ключевых точек, 17 оценок уверенности их извлечения, 4 нормированных параметра ограничивающего прямоугольника и 34 поэлементные межкадровые скорости. Использование относительных координат для статической позы и абсолютных – для скоростей не случайно: первое позволяет модели не зависеть от того, где и как близко пловец оказался к камере, второе – сохранить физическую интенсивность движений конечностей, без которой активное плавание невозможно отличить от пассивного пребывания в воде.

Темпоральная модель – двунаправленная двухслойная LSTM (скрытое состояние 128) с аддитивным вниманием – обучена разделять три класса: «активное плавание», «пассивное пребывание в воде», «утопление». Введение явного промежуточного класса принципиально меняет картину: вместо объединения отдыха на воде и истинного утопления в один разнородный класс модель обучается явно различать эти состояния, что избавляет от необходимости компенсировать противоречие постановки на этапе постобработки.

Эксперименты на собранном корпусе из 19 247 последовательностей поз дают следующие итоги: общая точность 0,920, F1 по классу «утопление» 0,924, precision и recall по нему же 0,920 и 0,929. Полуширина 95-процентных доверительных интервалов для этих метрик и для общей точности (интервал Уилсона [16] для долей и построенный на его основе через индекс Жаккара интервал для F1) не превышает 0,018, а их нижние границы не опускаются ниже

 © Авторы статьи / Authors of the article

0,90. Таким образом, все целевые значения, заданные на этапе постановки задачи, достигнуты с учётом статистической погрешности, связанной с объёмом тестовой подвыборки.

Сравнение с десятью альтернативными архитектурами – от логистической регрессии до Transformer и BiLSTM с самовниманием – и серия ablation-экспериментов показывают, что ни усложнение архитектуры, ни увеличение числа параметров в 1,5–1,9 раза не дают заметного прироста по сравнению с предлагаемой конфигурацией. Это делает её разумным компромиссом между качеством и стоимостью обучения и инференса.

Конфликт интересов

Не указан.

Рецензия

Гаврилов Н.К., Казанский национальный исследовательский технический университет им. А.Н. Туполева – КАИ, Казань Российская Федерация
DOI: https://doi.org/10.60797/IRJ.2026.173.67.4

Conflict of Interest

None declared.

Review

Gavrilov N.K., Kazan National Research Technical University named after A.N. Tupolev – KAI, Kazan Russian Federation
DOI: https://doi.org/10.60797/IRJ.2026.173.67.4

Список литературы на английском языке / References in English

1. Shatnawi M. Advances and challenges in automated drowning detection and prevention systems / M. Shatnawi, F. Albreiki, A. Alkhoori et al. // Information. — 2024. — Vol. 15, № 11. — Art. 721. — DOI: 10.3390/info15110721.

2. Jalalifar S. Enhancing water safety: Exploring recent technological approaches for drowning detection / S. Jalalifar, A. Belford, E. Erfani et al. // Sensors. — 2024. — Vol. 24, № 2. — Art. 331. — DOI: 10.3390/s24020331.

3. He K. Mask R-CNN / K. He, G. Gkioxari, P. Dollár et al. // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 2020. — Vol. 42, № 2. — P. 386–397. — DOI: 10.1109/TPAMI.2018.2844175.

4. Zhang W. A pool drowning detection model based on improved YOLO / W. Zhang, L. Chen, J. Shi // Sensors. — 2025. — Vol. 25, № 17. — Art. 5552. — DOI: 10.3390/s25175552.

5. LeCun Y. Deep learning / Y. LeCun, Y. Bengio, G. Hinton // Nature. — 2015. — Vol. 521, № 7553. — P. 436–444. — DOI: 10.1038/nature14539.

6. Hochreiter S. Long short-term memory / S. Hochreiter, J. Schmidhuber // Neural Computation. — 1997. — Vol. 9, № 8. — P. 1735–1780. — DOI: 10.1162/neco.1997.9.8.1735.

7. Vaswani A. Attention is all you need / A. Vaswani, N. Shazeer, N. Parmar et al. — DOI: 10.48550/arXiv.1706.03762.

8. Papandreou G. Towards accurate multi-person pose estimation in the wild / G. Papandreou, T. Zhu, N. Kanazawa et al. // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). — IEEE, 2017. — P. 3711–3719. — DOI: 10.1109/CVPR.2017.395.

9. Bahdanau D. Neural machine translation by jointly learning to align and translate / D. Bahdanau, K. Cho, Y. Bengio. — DOI: 10.48550/arXiv.1409.0473.

10. Li C. Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition / C. Li, Z. Bao, L. Li et al. // Information Processing & Management. — 2020. — Vol. 57, № 3. — Art. 102185. — DOI: 10.1016/j.ipm.2019.102185.

11. Liu G. Bidirectional LSTM with attention mechanism and convolutional layer for text classification / G. Liu, J. Guo // Neurocomputing. — 2019. — Vol. 337. — P. 325–338. — DOI: 10.1016/j.neucom.2019.01.078.

12. Srivastava R.P. Real-time yoga pose classification with 3-D pose estimation model with LSTM / R.P. Srivastava, L.S. Umrao, R.S. Yadav // Multimedia Tools and Applications. — 2024. — Vol. 83, № 11. — P. 33019–33030. — DOI: 10.1007/s11042-023-17036-8.

13. Imambi S. PyTorch / S. Imambi, K.B. Prakash, G.R. Kanagachidambaresan // Programming with TensorFlow. — Springer, 2021. — P. 87–104. — DOI: 10.1007/978-3-030-57077-4_10.

14. Zhou P. Towards understanding convergence and generalization of AdamW / P. Zhou, X. Xie, Z. Lin et al. // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 2024. — Vol. 46, № 9. — P. 6486–6493. — DOI: 10.1109/TPAMI.2024.3382294.

15. Lu T. KSL-POSE: A real-time 2D human pose estimation method based on modified YOLOv8-Pose framework / T. Lu, K. Cheng, X. Hua et al. // Sensors. — 2024. — Vol. 24, № 19. — Art. 6249. — DOI: 10.3390/s24196249.

16. Wilson E.B. Probable inference, the law of succession, and statistical inference / E.B. Wilson // Journal of the American Statistical Association. — 1927. — Vol. 22, № 158. — P. 209–212. — DOI: 10.1080/01621459.1927.10502953.

17. Brown L.D. Interval estimation for a binomial proportion / L.D. Brown, T.T. Cai et al. // Statistical Science. — 2001. — Vol. 16, № 2. — P. 101–133. — DOI: 10.1214/ss/1009213286.

18. Kiranyaz S. 1D convolutional neural networks and applications: A survey / S. Kiranyaz, O. Avci, O. Abdeljaber et al. // Mechanical Systems and Signal Processing. — 2021. — Vol. 151. — Art. 107398. — DOI: 10.1016/j.ymssp.2020.107398.