НАУЧНОЕ ПРИБОРОСТРОЕНИЕ, 2026, том 36, № 4, с. 118–130 *ISSN 0868-5886*

ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ

УДК 004.932:612.17
EDN: IOXPAE

© С. Н. Глебов, А. Д. Панкратова, Д. М. Виноградов, Г. С. Парфёнов, А. К. Кожахметова, А. Д. Крель, А. Ю. Зайцева, 2026

СИСТЕМА ДИСТАНЦИОННОГО МОНИТОРИНГА С СЕРВЕРНОЙ ПОДСИСТЕМОЙ АНАЛИЗА ДАННЫХ ПРИСУТСТВИЯ, ВОВЛЕЧЕННОСТИ И УСТАЛОСТИ ПО ВИДЕОПОТОКУ

Представлена программно-аппаратная система компьютерного зрения для неинвазивного мониторинга присутствия, вовлеченности и признаков усталости участников на основе видеоизображения. Актуальность работы определяется потребностью в автоматизированной поддержке процессов, требующих контроля функционального состояния человека: оператору или лицу, принимающему решение, необходим инструмент, фиксирующий состояние членов группы в реальном времени без контактных датчиков. Предложенный подход объединяет в единый конвейер несколько методов: извлечение опорных точек и прямоугольника лица моделью MediaPipe Face Mesh, идентификацию личности по эмбеддингам InceptionResNetV1 с предварительным каскадом MTCNN, расчет коэффициентов открытости глаз (EAR) и рта (MAR), классификацию состояний "бодрствование/усталость" методом опорных векторов с RBF-ядром, оценку позы по ключевым точкам YOLOv8-Pose и бесконтактную видеоплетизмографию для регистрации пульсовой волны методом CHROM. Серверная часть построена по микросервисному принципу с событийно-ориентированной архитектурой на основе FastAPI, брокера сообщений Apache Kafka и СУБД PostgreSQL с расширением TimescaleDB. Это обеспечивает потоковый прием, долговременное хранение временны́х рядов в виде гипертаблиц и аналитику в реальном времени через непрерывные агрегаты. Аналитический модуль сопоставляет отфильтрованную телеметрию с индивидуальным базовым физиологическим профилем и формализованными правилами, генерирующими оповещения о критических событиях. Экспериментально получены: точность распознавания лиц 69%, средняя нормированная ошибка локализации ключевых точек 4.5% при времени обработки кадра 15–25 мс и средняя точность классификации признаков усталости 83% по результатам k-блочной кросс-валидации. Средняя абсолютная ошибка оценки частоты сердечных сокращений методом видеоплетизмографии составила 2.80 уд/мин. Система применима для автоматизированной поддержки процессов дистанционного наблюдения, контроля функционального состояния операторов, телемедицины на месте оказания помощи (point-of-care), а в особенности — в образовательной сфере для мониторинга состояния обучающихся.

Ключевые слова: компьютерное зрение, видеоплетизмография, мониторинг функционального состояния, EAR, MAR, MTCNN, YOLO, MediaPipe, метод опорных векторов, TimescaleDB, микросервисная архитектура, дистанционный мониторинг

ВВЕДЕНИЕ

Здоровье и работоспособность граждан остаются одной из основ экономической стабильности общества. Частота сердечных сокращений, микроциркуляция и транспорт крови относятся к базовым физиологическим процессам, определяющим жизнедеятельность организма. Их анализ служит обязательным условием достоверной оценки текущего функционального состояния человека и прогноза развития патологических состояний.

Снижение рождаемости и старение трудоспособного населения ведут к дефициту кадров во многих отраслях народного хозяйства. Не является исключением и здравоохранение: по оценкам Министерства здравоохранения Российской Федерации, в 2023 г. потребность отрасли составляла около 29 тыс. врачей, а по итогам 2024 г. дефицит оценивался примерно в 23.3 тыс. врачей и порядка 63 тыс. работников среднего медицинского звена [1]. Сохраняющаяся нехватка специалистов делает востребованной автоматизацию, способную облегчить труд персонала и одновременно повысить качество медицинской помощи, в том числе за счет рекомендательных систем неинвазивного мониторинга физиологических показателей.

В работе используется тот факт, что нарушениям внимания и вовлеченности, а также состояниям усталости соответствует определенная реакция мимики и положения тела в пространстве, сопро-

вождаемая характерной последовательностью изменений цвета тканей лица. Такая совокупность признаков может быть зафиксирована видеокамерой и обработана методами машинного обучения, что открывает возможность комплексного анализа разнородных сигналов средствами искусственных нейроподобных систем и методов обработки многомерной информации.

Цель настоящей работы — разработка и апробация программно-аппаратного комплекса, объединяющего методы компьютерного зрения, алгоритмы машинного обучения и серверную инфраструктуру потоковой обработки данных для дистанционного мониторинга присутствия, вовлеченности и признаков усталости по видеопотоку членов группы в режиме реального времени.

Научная новизна работы состоит в следующем. Впервые предложен единый конвейер мультимодальной видеоаналитики, объединяющий верификацию присутствия, оценку вовлеченности и детектирование усталости без контактных датчиков при групповом наблюдении. В качестве модельной конкретизации выбрана студенческая аудитория. Разработана методология интегрального показателя вовлеченности E_{pose}, объединяющего биомеханические параметры с когнитивным состоянием студента в одну скалярную метрику с весовыми коэффициентами. Для хранения и анализа физиологических временны́х рядов впервые применена событийно-ориентированная микросервисная архитектура, обеспечивающая горизонтальное масштабирование системы при групповом мониторинге. Использование механизма гипертаблиц и непрерывных агрегатов обеспечивает линейное масштабирование при росте числа одновременно наблюдаемых субъектов, что принципиально отличает предложенное решение от монолитных систем с локальной обработкой видео.

ОБЗОР СУЩЕСТВУЮЩИХ ПОДХОДОВ

Существующие методы оценки функционального состояния человека опираются либо на инструментальные исследования, либо на субъективную интерпретацию наблюдений. Классические медицинские подходы (анализы крови, полимеразная цепная реакция, магнитно-резонансная томография) обладают высокой точностью, однако требуют значительных временны́х и материальных затрат и непригодны для непрерывного дистанционного мониторинга в реальном времени.

С развитием информационно-коммуникационных технологий и методов машинного обучения активно формируются подходы к неинвазивной оценке состояния человека. Ключевым направлением выступает компьютерное зрение для анализа видеоизображений: такие системы обрабатывают визуальную информацию без специализированных контактных датчиков, что делает их перспективными для образовательных и удаленных сред [2]. Систематические обзоры показывают, что подавляющая часть работ по оценке вовлеченности использует внешние наблюдения или самоотчеты и лишь часть — автоматический анализ на основе машинного обучения [3].

Для определения присутствия применяются детекторы лиц на основе глубокого обучения. Модели семейства YOLO обеспечивают детектирование в реальном времени при высокой пропускной способности [4]; RetinaFace ориентирован на более точную локализацию в сложных условиях; каскад MTCNN реализует совместное детектирование и локализацию ключевых точек [5]. Идентификация выполняется по эмбеддингам лиц, получаемым сетями типа FaceNet [6], предобученными на больших наборах данных, например VGGFace2 [7]. Точность таких систем существенно зависит от освещения, положения головы и наличия перекрытий.

Для оценки вовлеченности используют анализ мимики, направления взгляда и положения головы. Модели, выделяющие ключевые точки лица (MediaPipe Face Mesh [8], регрессоры на базе MobileNet v3), позволяют отслеживать динамику мимики; на их основе рассчитываются коэффициенты EAR и MAR, интерпретируемые для детекции моргания, зевков и длительного смыкания век. При ограниченных обучающих выборках высокую устойчивость показывают классические алгоритмы, прежде всего метод опорных векторов с RBF-ядром. Ряд работ объединяет мимику, направление взгляда и оценку позы в мультимодальные модели вовлеченности [9], а также применяет детекторы YOLO для распознавания учебного поведения в реальной аудитории [10].

Неинвазивная оценка физиологических параметров обеспечивается дистанционной фотоплетизмографией (дФПГ), регистрирующей изменения кровенаполнения тканей по вариациям цвета кожи. Для выделения полезного сигнала применяются методы компонентного и проекционного анализов — ICA, PCA, GREEN, CHROM, POS [11–13], — а для спектральной оценки частоты сердечных сокращений используются быстрое преобразование Фурье и фильтры Баттерворта. Сравнительные исследования показывают, что выбор области интереса (ROI) и метода декомпозиции существенно влияет на точность, а проекционные методы CHROM и POS устойчивее к артефактам движения и освещения, чем простое усреднение зеленого канала [12]. Сопоставление дистанционной и контактной фотоплетизмографии подтверждает применимость камерных методов для оценки сердечного ритма [14]. Для хранения

физиологических временны́х рядов, накапливающихся при групповом мониторинге, применяются специализированные СУБД, ориентированные на time-series, в частности расширение TimescaleDB к PostgreSQL [15].

ПОСТАНОВКА ЗАДАЧИ

Анализ литературы показывает, что существующие методы решают задачи присутствия, вовлеченности и усталости преимущественно изолированно и не обеспечивают комплексной мультимодальной оценки с масштабируемой серверной частью [16, 17]. Это обосновывает актуальность настоящей работы и формулировку задачи: разработать систему, объединяющую методы компьютерного зрения и потоковую обработку данных для мониторинга присутствия, вовлеченности и функционального состояния человека в реальном времени. Перспективным применением предлагаемого решения является, например, дистанционный мониторинг обучающихся: в условиях онлайн-обучения система позволяет преподавателю своевременно выявлять снижение внимания и нарастающее утомление, а в очных аудиториях — оперативно корректировать темп и структуру занятия, поддерживая высокий уровень вовлеченности группы без привлечения дополнительного персонала или контактных датчиков.

АРХИТЕКТУРА СИСТЕМЫ И МЕТОДЫ КОМПЬЮТЕРНОГО ЗРЕНИЯ

Общая архитектура

Предлагаемая система (рис. 1) представляет собой распределенный программно-аппаратный комплекс, объединяющий клиентскую часть (цифровая видеокамера) и серверную часть, которая осуществляет прием, хранение и интеллектуальный анализ потоковой телеметрии.

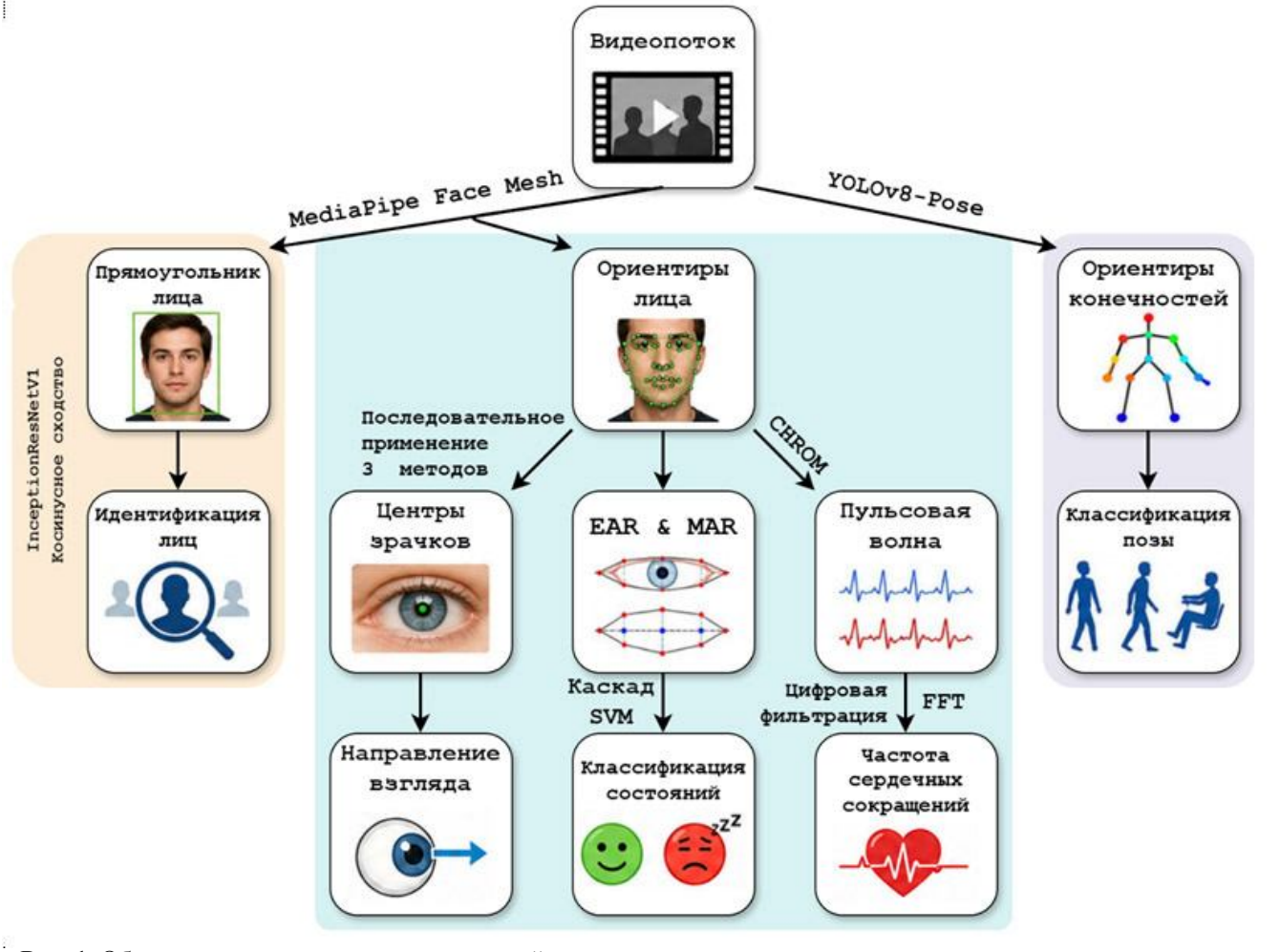

Рис. 1. Общая схема архитектуры предлагаемой системы

Биологические ткани являются гетерогенными многокомпонентными сильнорассеивающими средами, обладающими значительным диагностическим потенциалом. Для получения комплексной информации о физиологических процессах применен мультимодальный подход, интегрирующий разнородные биометрические данные, регистрируемые параллельно.

Многоуровневая обработка многомерных биомедицинских массивов методами многомерной статистики направлена на формирование интегрального показателя, характеризующего состояние биосистемы на разных уровнях ее структурно-функциональной организации. Этот показатель представляет собой формализованную модель, отражающую ключевые характеристики биологической среды через заданное параметрическое пространство, и служит основой для сравнительного анализа.

Выделение области лица и лицевых ориентиров

Общим для нескольких модулей обработки решением была выбрана модель компьютерного зрения MediaPipe Face Mesh (MPFM) — высокоточная модель от Google для отслеживания ключевых точек в реальном времени. Она способна выделять 468 трехмерных ориентиров (landmarks), включая контуры губ, бровей, глаз, лица и положение зрачков, что позволяет детально анализировать мимику и повороты головы. С архитектурной точки зрения подход сочетает легкую нейронную сеть для грубой локализации ключевых областей и последующую регрессию координат, что обеспечивает баланс между скоростью и точностью даже на мобильных устройствах. Благодаря этому MPFM широко применяется в фильтрах дополненной реальности, мониторинге усталости водителей и системах распознавания эмоций.

Скорость работы MPFM — одно из ее ключевых преимуществ. Благодаря легкой архитектуре и эффективным операциям, модель способна обрабатывать видеопоток в реальном времени со скоростью от 20 до более чем 60 кадров в секунду даже на среднестатистических смартфонах и вебкамерах ноутбуков. На выделенных GPU или десктопных системах производительность может достигать сотен кадров в секунду. Основные аналоги MPFM, такие как Dlib или OpenFace, заметно уступают ей в скорости работы из-за большей вычислительной сложности применяемых алгоритмов и моделей. Кроме того, многие альтернативы выдают меньшее число ориентиров.

В данной работе MPFM используется в качестве первого этапа обработки видеопотока для выделения основных используемых лицевых признаков. Результатом ее работы становится набор из 468 ориентиров. Из этого набора выделяются точки, необходимые для работы модулей обработки, а также по минимальным и максимальным координатам строится прямоугольник вокруг лица, что позволяет использовать единую модель как для детекции лиц, так и для выделения необходимых признаков.

Детектирование и идентификация лиц

Как уже было отмечено выше, результатом первого шага обработки видеопотока является, в том числе, прямоугольник, построенный вокруг лица. Выделенная область лица приводится к размеру 160 × 160 пикселей, и сигнал от них нормализуется вычитанием среднего и делением на стандартное отклонение. Для решения задачи идентификации применена сверточная сеть InceptionResNetV1, предобученная на VGGFace2. Перед выделением эмбеддингов выполняется каскадное детектирование MTCNN, состоящее из трех последовательных сетей: Proposal Network (P-Net) для быстрого формирования кандидатов, Refine Network (R-Net) для уточнения координат и Output Network (O-Net) для финальной локализации и выдачи пяти ключевых точек (глаза, нос, углы рта).

Следом вычисляется эмбеддинг размерностью 512, нормализуемый к единичной длине. База эталонных эмбеддингов формируется усреднением векторов нескольких изображений каждого наблюдаемого, что повышает устойчивость к вариативности условий съемки. Идентификация выполняется по косинусному сходству:

$$s(\mathbf{e}, \mathbf{e}_i) = \frac{\mathbf{e} \cdot \mathbf{e}_i}{\|\mathbf{e}\| \times \|\mathbf{e}_i\|}.$$

При превышении порога 0.6 лицу присваивается идентификатор; иначе лицо классифицируется как "неизвестное". Это позволяет вести непрерывный учет, например, присутствия обучающихся и отслеживать динамику вовлеченности каждого идентифицированного лица.

Анализ направления взгляда

Для локализации центра зрачка (после выделения контуров глаз MPFM) предложены три метода: градиентный (поиск точки с максимальным градиентом яркости после оператора Собеля); пороговый (бинаризация адаптивным порогом с поиском контура наибольшей площади); метод кругового преобразования Хафа. В комбинированном режиме методы применяются последовательно.

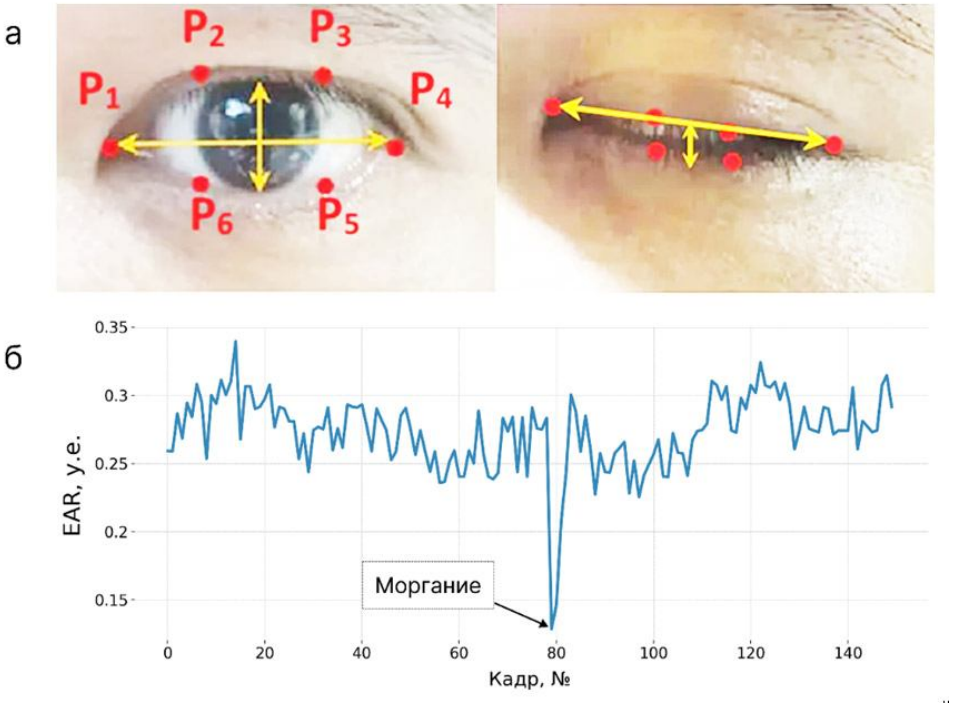

Рис. 2. Оценка состояния глаз наблюдаемого.
а — схема расчета степени открытости глаз EAR, б — отражение эффекта моргания на динамике EAR

После нахождения центра зрачка (x_p, y_p) вычисляется нормализованный вектор взгляда относительно центра области глаза (x_c, y_c):

$$g_x = 2\frac{(x_p - x_c)}{w}, \quad g_y = 2\frac{(y_p - y_c)}{h},$$

где w и h — ширина и высота области глаза. Полученные значения ограничиваются диапазоном [–1, 1] и сглаживаются скользящим средним. Калибровка выполняется автоматически в течение первых 30 кадров: фиксируется максимальное значение, относительно которого нормируются последующие измерения.

Детектирование состояния усталости

В основе модели детектирования состояния усталости лежит два физиологических индекса — уровень открытости глаз (EAR) и уровень открытости рта (MAR). В процессе анализа видеоизображения эти показатели позволяют количественно оценить степень сонливости и вероятность зевка. Таким образом, отслеживая динамику EAR и MAR во времени, система может своевременно распознавать критические фазы утомления.

Коэффициент открытости глаз EAR рассчитывается по шести характерным точкам контура глаза $P_1,\ldots, P_6$ (рис. 2):

$$\mathrm{EAR} = \frac{\|\mathbf{P}_2 - \mathbf{P}_6\| + \|\mathbf{P}_3 - \mathbf{P}_5\|}{2\|\mathbf{P}_1 - \mathbf{P}_4\|}.$$

При открытом глазе EAR колеблется около характерной для пользователя константы; при смыкании век числитель стремится к нулю. Моргание регистрируется при снижении EAR ниже порога 0.25 в течение трех последовательных кадров. Для сглаживания флуктуаций применяются скользящее окно из десяти последних значений и конечный автомат с состояниями "открыто", "закрывается", "закрыто", "открывается".

Коэффициент открытости рта MAR рассчитывается по характерным точкам контура губ $M_1,\ldots, M_6$ аналогичным образом — как отношение вертикального раскрытия рта к его ширине:

$$\mathrm{MAR} = \frac{\|\mathbf{M}_2 - \mathbf{M}_6\| + \|\mathbf{M}_3 - \mathbf{M}_5\|}{2\|\mathbf{M}_1 - \mathbf{M}_4\|}.$$

Резкое увеличение MAR свидетельствует о зевке.

Совместное использование EAR и MAR позволяет отличать обычную мимику от признаков утомления. Для автоматического распознавания

состояний "бодрствование" и "усталость" применен классификатор на базе метода опорных векторов (SVM) [19]. Входным вектором являются EAR, MAR и производные признаки (частота моргания, длительность закрытия глаз, частота зевков). SVM строит оптимальную разделяющую гиперплоскость, максимизирующую зазор между классами "бодрствование" (высокий EAR, низкий MAR) и "усталость" (низкий EAR либо высокий MAR). Поскольку зависимости между признаками носят нелинейный характер, используется нестандартное ядро радиальной базисной функции (RBF):

$$\mathrm{K}(\mathbf{x}_i, \mathbf{x}_j) = \mathrm{e}^{-\gamma\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2},$$

где γ определяет радиус влияния отдельного обучающего примера.

Бесконтактная видеоплетизмография

Для обеспечения возможности дистанционного мониторинга использован метод дистанционной фотоплетизмографии (дФПГ) — формирование сигнала ФПГ путем обработки видео, полученного цифровой камерой. Бесконтактный характер метода исключает необходимость ношения нательных датчиков, что делает метод оптимальным для длительной регистрации данных, не вызывая дискомфорта у обучающегося; для регистрации сигнала достаточно стандартной веб-камеры персонального компьютера. Для выделения полезного сигнала с видеопотока в данной работе используется область изображения, ограниченная контуром овала лица, полученного при помощи MPFM.

Существенной проблемой в дФПГ представляют значительные артефакты, вызванные микросмещениями лица и флуктуациями освещенности, поэтому простое усреднение зеленого канала (GREEN) не дает оценку пульса, т.к. подавляется артефактами движения. Поэтому необходимо использовать специальные методы, чтобы получить оценку пульса. К таким методам относятся ICA, PCA, PBV, CHROM, POS, SSR и LGI [11–14]. В данной работе, ввиду высокой надежности, и более высокой детерминированности в сравнении с ICA был использован CHROM.

Математическая модель взаимодействия света с кожей человека опирается на дихроматическую модель отражения. После ряда упрощений сигнал в RGB-каналах $C(t)$ представим в виде

$$C(t) = u_\mathrm{c} I_0 c_0 + \mathbf{u}_\mathrm{c} I_0 c_0 i(t) + \mathbf{u}_\mathrm{s} I_0 s(t) + \mathbf{u}_\mathrm{p} I_0 p(t),$$

где $i(t)$ — интенсивность освещения; $s(t)$ — зеркальное отражение, $p(t)$ — диффузное отражение, содержащее информацию о пульсе; $u_\mathrm{c} I_0 c_0$ — постоянная составляющая (шум); $\mathbf{u}_\mathrm{c}$ — вектор цвета кожи; I_0 — средняя интенсивность освещения; c_0 — коэффициент отражения; $\mathbf{u}_\mathrm{s}$ — вектор направления зеркального отражения; $\mathbf{u}_\mathrm{p}$ — вектор изменения цвета, обусловленного пульсом.

Выражение выше показывает, что наблюдение $C(t)$ является линейной смесью трех источников — $i(t)$, $s(t)$ и $p(t)$, — которые можно разделить линейной проекцией. Таким образом, задача извлечения пульсового сигнала из RGB-наблюдений сводится к определению системы проекции для декомпозиции $C(t)$.

Метод CHROM, один из методов декомпозиции, основан на проекции нормализованных RGB-сигналов на две ортогональные оси в плоскости, перпендикулярной вектору цвета кожи. Вычисляются промежуточные сигналы, подавляющие зеркальную компоненту:

$$X_\mathrm{s}(t) = 3R_\mathrm{n}(t) - 2G_\mathrm{n}(t),$$

$$Y_\mathrm{s}(t) = 1.5R_\mathrm{n}(t) + G_\mathrm{n}(t) - 1.5B_\mathrm{n}(t).$$

Для финальной очистки сигнала от остаточного движения применяется адаптивная альфа-настройка. Итоговый пульсовой сигнал $S(t)$ формируется как

$$S(t) = X_\mathrm{s}(t) - \alpha Y_\mathrm{s}(t),$$

где коэффициент α рассчитывается динамически, как отношение стандартных отклонений сигналов:

$$\alpha = \frac{\sigma(X_\mathrm{s})}{\sigma(Y_\mathrm{s})}.$$

Такая комбинация позволяет синфазным в данных проекциях компонентам движения взаимно уничтожаться, тогда как пульсовая волна, находящаяся в противофазе или имеющая иное направление в цветовом пространстве, сохраняется.

Полученный после CHROM временной ряд $S(t)$ еще содержит низкочастотные тренды и высокочастотный шум. Поскольку физиологически обусловленная частота сердечных сокращений (ЧСС) здорового человека в покое и при умеренной нагрузке лежит в диапазоне 30–180 уд/мин, сигнал подвергается полосовой фильтрации в интервале 0.5–4 Гц. На финальном этапе для перехода из временно́й области в частотную применяется быстрое преобразование Фурье (БПФ); в полученном спектре отыскивается доминирующий пик. Значение f_HR, выраженное в герцах, пересчитывается в удары в минуту по соотношению ЧСС = $f_\mathrm{HR}\cdot 60$.

Оценка позы и двигательной активности

Для оценки позы применена модель YOLOv8-Pose [18], детектирующая 17 ключевых точек тела в формате COCO. Архитектура на основе CSPNet и пирамиды признаков обеспечивает высокую скорость при сохранении точности. Вычисляются биомеханические параметры: углы наклона и поворота головы, плеч и корпуса, углы в локтевых и коленных суставах (по теореме косинусов). Положение классифицируется, как "стоя/сидя/наклон", с учетом относительного расположения ключевых точек, а наклон корпуса, — как "вперед" (более 15°), "назад" (менее 10°) или "нейтральный".

Двигательная активность A оценивается по смещению центра бедер за последние 30 кадров, нормированному относительно порога 10 пикселей. Стабильность позы S_{stab} вычисляется, как обратная нормированная дисперсия угла наклона головы за последние 10 кадров. Интегральный показатель вовлеченности по позе:

$$E_{\text{pose}} = w_1 S_{\text{lean}} + w_2 A + w_3 S_{\text{stab}},$$

где S_{lean} — нормированное отклонение корпуса от нейтрального положения; наклон вперед интерпретируется как признак вовлеченности, назад — как ее снижение.

СЕРВЕРНАЯ ЧАСТЬ АНАЛИЗА ДАННЫХ

Обоснование архитектурного подхода

Методы бесконтактной видеоплетизмографии и алгоритмы выделения областей интереса формируют непрерывный поток физиологических данных. Обработка этого потока непосредственно на периферийном устройстве затруднена из-за ограниченных вычислительных ресурсов, поэтому система включает выделенный вычислительный модуль (серверную часть), осуществляющий централизованную обработку, хранение и анализ данных.

Серверная часть построена по микросервисному принципу с применением событийно-ориентированной архитектуры (Event-Driven Architecture), которая разделяет процессы приема сырых данных и их обработки — критичное свойство для систем реального времени. Поток от клиентских устройств помещается в очередь сообщений, что сглаживает пиковые нагрузки, повышает отказоустойчивость и обеспечивает горизонтальное масштабирование.

В качестве брокера сообщений используется Apache Kafka. Поступающая телеметрия публикуется в тематические разделы (топики) с партиционированием по идентификатору устройства, что гарантирует строгий порядок сообщений в пределах одного источника и позволяет независимо масштабировать потребителей. Распределенный журнал Kafka выступает буфером между приемом и обработкой: при кратковременной недоступности аналитического сервиса данные не теряются, а накапливаются в журнале и обрабатываются после восстановления. Технологический стек дополняют язык Python с асинхронным веб-фреймворком FastAPI, обеспечивающим высокую пропускную способность при операциях ввода-вывода, и реляционная СУБД PostgreSQL с расширением TimescaleDB для работы с временны́ми рядами [15].

Модуль приема телеметрии

Данные, полученные при работе алгоритмов компьютерного зрения, а также со специализированных неинвазивных сенсоров (частота сердечных сокращений, температура кожи, сатурация, частота дыхательных движений, уровень активности), передаются на сервер по защищенному каналу. Модуль приема выполняет первичную валидацию структуры пакета, проверку временны́х меток и аутентификацию. Прошедшие проверку пакеты сериализуются и публикуются в брокер сообщений согласно идентификатору устройства, что обеспечивает строгое соблюдение временно́й последовательности измерений для каждого пользователя.

Хранение временны́х рядов

Одной из проблем систем дистанционного биомониторинга является быстрый рост объема сохраняемых данных. Классические реляционные структуры демонстрируют снижение производительности при большом числе записей. Для ее решения применяется расширение TimescaleDB к PostgreSQL: входящие измерения сохраняются в виде гипертаблиц (hypertables), автоматически разбиваемых по временны́м интервалам. Логическая схема базы данных приведена на рис. 3.

Аналитика в реальном времени обеспечивается непрерывными агрегатами (continuous aggregates): база данных в фоновом режиме рассчитывает средние, минимальные и максимальные значения показателей за фиксированные интервалы, что минимизирует задержки при визуализации. Применяются встроенная политика компрессии исторических данных и автоматическое удаление сырой телеметрии по истечении заданного срока хранения.

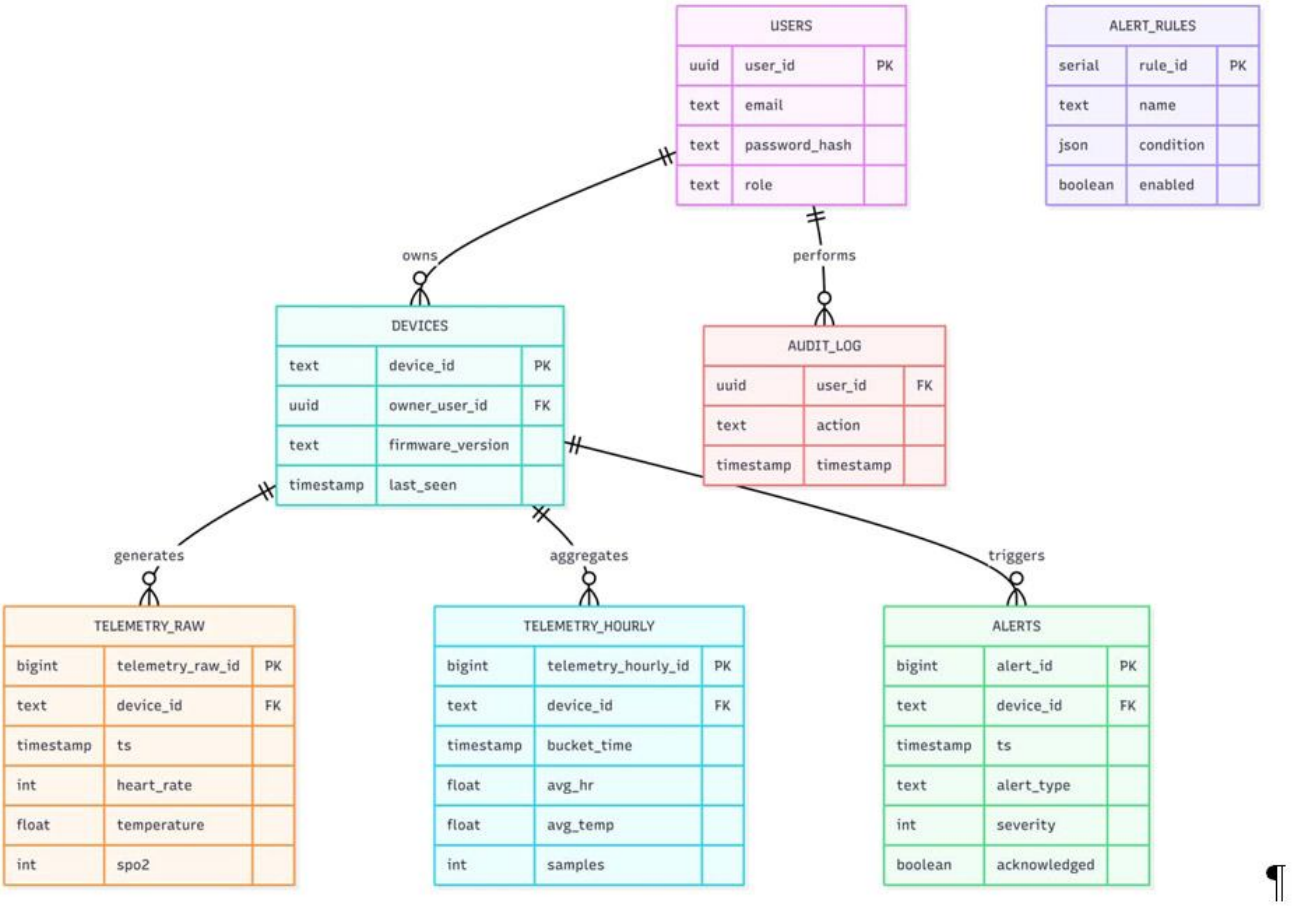

Рис. 3. Логическая схема базы данных (ER-диаграмма) системы телеметрии

Аналитический модуль мониторинга состояния

Центральным элементом серверной части является интеллектуальный модуль анализа функционального состояния, выявляющий паттерны усталости, когнитивного дефицита или стресса по динамике биометрических параметров. Модуль работает в режиме непрерывной обработки, сопоставляя данные с набором формализованных правил в рабочей базе данных; правила описывают пороговые значения и логические зависимости (например, сочетание снижения двигательной активности со специфическими изменениями сердечного ритма). Обработка включает нормализацию и фильтрацию (удаление шумов и артефактов медианными фильтрами), оценку текущего состояния (сопоставление с индивидуальным базовым физиологическим профилем), генерацию оповещений о критических событиях (запись в таблицу alerts) и уведомление наблюдателя и оператора системы.

ПРОГРАММНАЯ РЕАЛИЗАЦИЯ И ЭКСПЕРИМЕНТАЛЬНЫЕ РЕЗУЛЬТАТЫ

Конвейер обработки данных

Программная реализация выполнена на языке Python с использованием библиотек OpenCV (захват и предобработка видеопотока), MediaPipe (извлечение ключевых лицевых точек и метрик), scikit-learn (детектирование усталости, классификатор SVM), ultralytics (выделение позы при помощи YOLOv8-Pose), а также FastAPI и TimescaleDB на стороне сервера. Конвейер состоит из пяти этапов: захват и предобработка кадра (преобразование BGR→RGB); извлечение ключевых точек; формирование вектора признаков (евклидовы расстояния между парами точек, коэффициенты EAR и MAR); классификация методом опорных векторов; визуализация и агрегирование метки "устал / не устал" по всей видеозаписи.

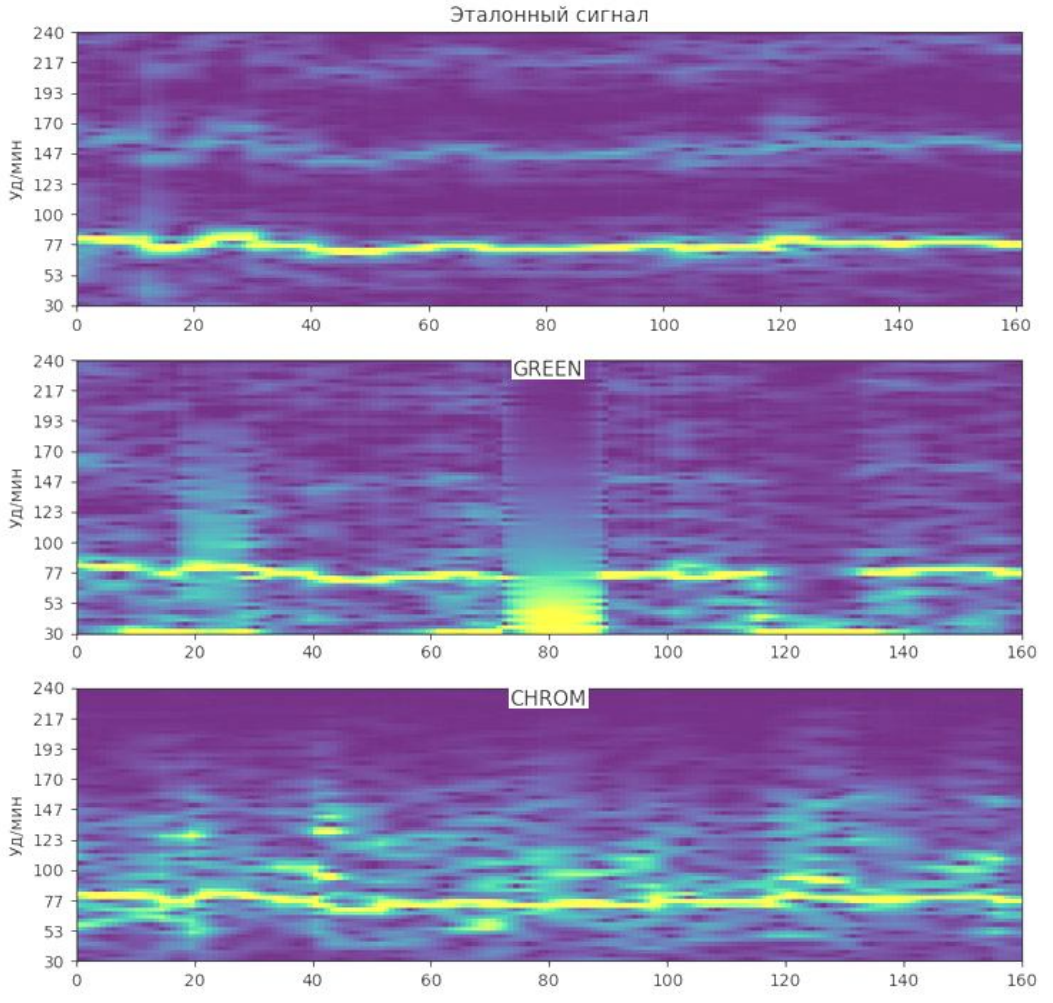

Рис. 4. Спектрограммы эталонного дФПГ-сигнала, сигнала зеленого канала (GREEN) и сигнала CHROM

Результаты экспериментальных исследований

Теоретическая модель, рассмотренная выше, обеспечивает аналитическое описание задачи. Для проверки устойчивости алгоритма в реальных условиях регистрации и оценки его предсказательной способности использовался открытый набор данных MCD-rPPG [20]. Из общего массива для тестирования устойчивости были отобраны записи, сделанные после легкой физической нагрузки.

Контрольное измерение проводилось фотоплетизмографом "Eldar". В качестве ROI выбрана область овала лица: она максимизирует объем выборки пикселей, снижая влияние шума, и остается видимой на протяжении всей записи [13]. Для декомпозиции сигнала применен метод CHROM, поскольку он обеспечивает инвариантность к зеркальным бликам и движениям субъекта и не требует предварительного обучения. Для удаления низкочастотного тренда и высокочастотного шума использован фильтр Баттерворта 5-го порядка. Спектрограммы эталонного дФПГ-сигнала, сигнала зеленого канала (GREEN) и сигнала CHROM приведены на рис. 4.

Видно, что GREEN-сигнал по сравнению с эталоном сильно искажен вследствие световых и двигательных артефактов и демонстрирует недостаточную устойчивость в реальных условиях регистрации, тогда как метод CHROM позволяет справиться с возникающими артефактами движения.

Для количественной оценки эффективности коррекции артефактов методом CHROM проведен статистический анализ (рис. 5).

а б

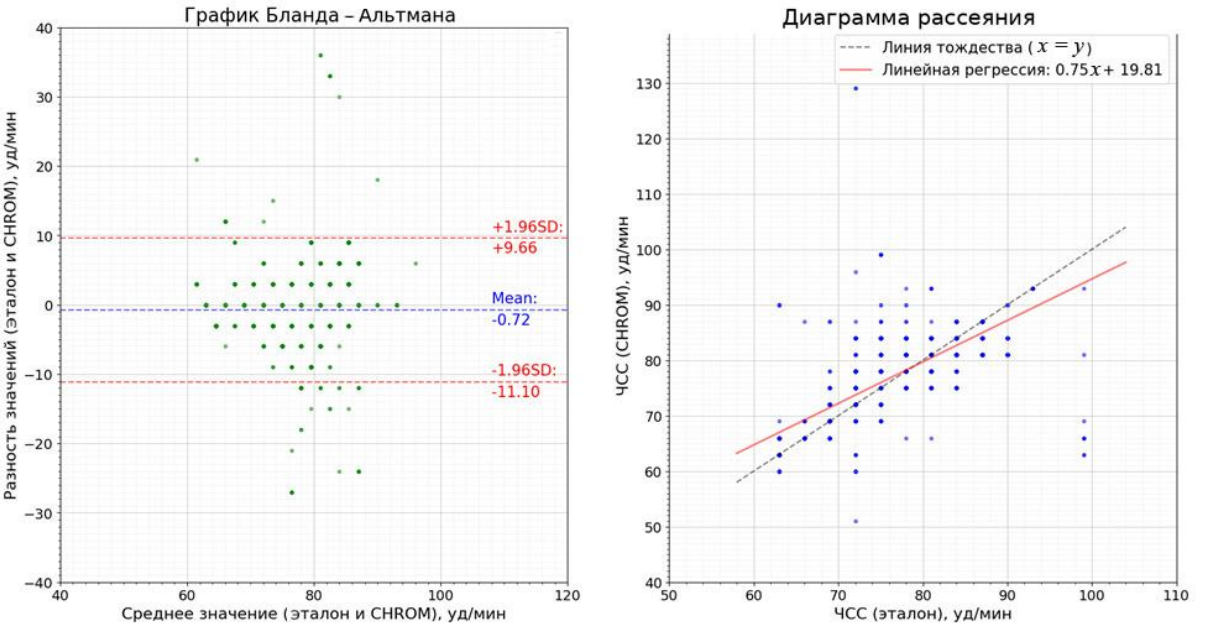

Рис. 5. Статистический анализ точности алгоритма CHROM. График Бланда – Альтмана (а) и диаграмма рассеяния (б)

На диаграмме Бланда – Альтмана среднее отклонение составляет −0.72 уд/мин, что указывает на отсутствие значимого систематического смещения оценок CHROM относительно эталона. Подавляющее большинство измерений сосредоточено внутри доверительного интервала от −11.10 до +9.66 уд/мин; единичные выбросы носят случайный характер и соответствуют моментам интенсивной мимической активности, не оказывая критического влияния на статистическую достоверность метода.

Для оценки согласованности данных построена диаграмма рассеяния, отражающая зависимость между измерениями ЧСС методом CHROM и эталонными значениями. Уравнение линейной регрессии $y = 0.75x + 19.81$ свидетельствует о сильной положительной корреляции; большинство экспериментальных точек сосредоточено вдоль линии тождества $x = y$, что подтверждает достоверность метода. Средняя абсолютная ошибка (MAE) по всей выборке составила 2.8 уд/мин, или 3.68 % от среднего эталонного значения (76 уд/мин). Несмотря на ограниченное временно́е разрешение видео (30 кадров/с), заметно более низкое, чем у контрольного пульсометра, дистанционная оценка устойчиво сходится с эталоном, что подтверждает достаточную точность метода CHROM для применения в подобных системах.

Таблица. Характеристики обученной модели распознавания лиц

Метрика	Значение
Accuracy	69%
Precision	71%
Recall	66%

Итоговые значения метрик для модели распознавания лиц, приведены в таблице.

Для классификатора SVM применялась k-блочная кросс-валидация ($k = 5$) на выборке видеозаписей обучающихся в состояниях бодрствования и усталости. Средняя точность классификации составила 83%. Разработанный программный комплекс обладает низкой вычислительной сложностью и высокой скоростью отклика и пригоден для интеграции в образовательные и телемедицинские информационные системы.

Сочетание бесконтактной видеоплетизмографии и физиологических показателей с неинвазивных сенсоров с высокопроизводительной системой обработки на базе специализированных СУБД и брокеров сообщений обеспечивает надежную регистрацию физиологических параметров.

Предложенный подход реализует автоматический непрерывный мониторинг уровня вовлеченности и утомления группы пользователей в реальном времени при сохранении масштабируемости системы.

ЗАКЛЮЧЕНИЕ

В работе предложена и программно реализована комплексная система компьютерного зрения для дистанционного мониторинга присутствия, вовлеченности и признаков усталости по видеопотоку с серверной подсистемой анализа данных. Система объединяет бесконтактную видеоплетизмографию, современные архитектуры сверточных нейронных сетей (MediaPipe Face Mesh, MTCNN, InceptionResNetV1, YOLOv8-Pose) и классический метод опорных векторов с RBF-ядром для формирования интегральной оценки функционального состояния обучающегося.

Проведено обучение моделей детектирования лица (Accuracy = 69%) и ключевых точек лица. Реализован конвейер извлечения признаков EAR и MAR на базе MediaPipe с классификацией методом опорных векторов, обеспечившей по результатам *k*-блочной кросс-валидации среднюю точность 83%. Сопоставление видеоплетизмографии с эталонным фотоплетизмографом дало среднюю абсолютную ошибку оценки ЧСС 2.8 уд/мин. Серверная часть на основе FastAPI, Apache Kafka и PostgreSQL/TimescaleDB обеспечивает масштабируемую потоковую обработку временны́х рядов, автоматическую агрегацию и генерацию оповещений о критических состояниях.

Дальнейшее развитие системы предполагает расширение обучающей выборки, интеграцию дополнительных физиологических сенсоров и внедрение моделей глубокого обучения для уточненной оценки когнитивной нагрузки. Выполнение данных целей позволит подходу быть применимым в системах дистанционного мониторинга комплексного функционального состояния наблюдаемого, экспресс-диагностики в географически удаленных и труднодоступных районах, а также для автоматизации наблюдения за обучающимися.

СПИСОК ИСТОЧНИКОВ

1. *Мурашко М.* Дефицит врачей в России превысил 23 тысячи // Парламентская газета. 2025. [Электронный ресурс] Режим доступа: https://www.pnp.ru/politics/murashko-deficit-vrachey-v-rossii-prevysil-23-tysyachi.html (дата обращения: 31.05.2026).
2. *Pillai A.S.* Student Engagement Detection in Classrooms through Computer Vision and Deep Learning: A Novel Approach Using YOLOv4 // SSRN Electronic Journal. 2025. DOI: 10.5281/zenodo.15061626
3. *Karimah S.N., Hasegawa S.* Automatic engagement estimation in smart education/learning settings: a systematic review of engagement definitions, datasets, and methods // Smart Learning Environments. 2022. Vol. 9. Id. 31. DOI: 10.1186/s40561-022-00212-y
4. *Wang C.-Y., Bochkovskiy A., Liao H.-Y.M.* YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors // IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023. P. 7464–7475. DOI: 10.1109/CVPR52729.2023.00721
5. *Zhang K. et al.* Joint face detection and alignment using multitask cascaded convolutional networks // IEEE Signal Processing Letters. 2016. Vol. 23, no. 10. P. 1499–1503. DOI: 10.48550/arXiv.1604.02878
6. *Schroff F., Kalenichenko D., Philbin J.* FaceNet: A unified embedding for face recognition and clustering // Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015. P. 815–823. DOI: 10.1109/CVPR.2015.7298682
7. *Cao Q. et al.* VGGFace2: A Dataset for Recognising Faces across Pose and Age // Proc. 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018). 2018. P. 67–74. DOI: 10.48550/arXiv.1710.08092
8. *Lugaresi C. et al.* MediaPipe: A framework for building perception Pipelines // arXiv (Cornell University). 2019. DOI: 10.48550/arXiv.1906.08172
9. *Sukumaran A., Manoharan A.* Multimodal engagement recognition from image traits using deep learning techniques // IEEE Access. 2024. Vol. 12. P. 25228–25244. DOI: 10.1109/ACCESS.2024.3353053
10. *Trabelsi Z. et al*. Real-Time Attention Monitoring System for Classroom: A deep learning approach for student's behavior recognition // Big Data and Cognitive Computing. 2023. Vol. 7, no. 1. Id. 48. DOI: 10.3390/bdcc7010048
11. *De Haan G., Jeanne V.* Robust pulse rate from Chrominance-Based RPPG // IEEE Transactions on Biomedical Engineering. 2013. Vol. 60, no. 10. P. 2878–2886. DOI: 10.1109/TBME.2013.2266196
12. *Wang W. et al*. Algorithmic principles of remote PPG // IEEE Transactions on Biomedical Engineering. 2016. Vol. 64, no. 7. P. 1479–1491. DOI: 10.1109/TBME.2016.2609282
13. *Kim D.-Y., Lee K., Sohn C.-B.* Assessment of ROI selection for Facial Video-Based RPPG // Sensors. 2021. Vol. 21, no. 23. Id. 7923. DOI: 10.3390/s21237923
14. *Van Es. V.A.A. et al*. Contactless Cardiovascular Assessment by Imaging Photoplethysmography: A Comparison with Wearable Monitoring // Sensors. 2023. Vol. 23, no. 3. Id. 1505. DOI: 10.3390/s23031505
15. Tiger data documettation. TimescaleDB: a time-series database on PostgreSQL [Электронный ресурс]. 2024. URL: https://docs.timescale.com/ (дата обращения: 31.05.2026).

16. *Singh A. et al.* VisioPhysioENET: Visual Physiological Engagement Detection Network // arXiv (Cornell University). 2024. DOI: 10.48550/arXiv.2409.16126
17. *Georgiadis P. et al.* Head engagement during visuomotor tracking is determined by postural challenges and aging // Journal of Neurophysiology. 2025. Vol. 134, no. 6. P. 2086–2098. DOI: 10.1152/jn.00319.2025
18. *Cai S. et al.* A human pose estimation network based on YOLOv8 framework with efficient multi-scale receptive field and expanded feature pyramid network // Scientific Reports. 2025. Vol. 15. Id. 15284. DOI: 10.1038/s41598-025-00259-0
19. *Сапрыкин Я.Д., Рязанцев В.И., Смирнов А.А.* Обзор подходов к распознаванию усталости водителя и существующих технических решений // Известия МГТУ "МАМИ". 2020. Т. 14, № 3. С. 48–58. DOI: 10.31992/2074-0530-2020-45-3-48-58
20. MCD-rPPG: Multi-CameraDataset for Remote Photoplethysmography [Электронный ресурс]. URL: https://huggingface.co/datasets/wengziheng/mcd_rppg (дата обращения: 16.07.2026).

Институт аналитического приборостроения РАН, Санкт-Петербург (*Глебов С.Н., Зайцева А.Ю.*)

Санкт-Петербургский национальный исследовательский академический университет им. Ж.И. Алферова РАН, Санкт-Петербург (*Панкратова А.Д., Виноградов Д.М., Парфенов Г.С.*)

Академический лицей "Физико-техническая школа" им. Ж.И. Алферова, Санкт-Петербург (*Кожахметова А.К.*)

Санкт-Петербургский политехнический университет, Санкт-Петербург (*Крель А.Д.*)

Контакты: *Глебов Степан Николаевич*, stepangleboff@yandex.ru

Материал поступил в редакцию 26.06.2026

A VIDEOSTREAM BASED SYSTEM FOR REMOTE MONITORING OF PRESENCE, ENGAGEMENT AND FATIGUE WITH DATA-ANALYSIS SERVER BACKEND

S. N. Glebov[1], A. D. Pankratova[2], D. M. Vinogradov[2], G. S. Parfyonov[2], A. K. Kozhakhmetova[3], A. D. Krel[4], A. Yu. Zaytseva[1]

[1]*Institute for Analytical Instrumentation of RAS, Saint Petersburg, Russia*
[2]*Saint Petersburg Academic University of RAS, Saint Petersburg, Russia*
[3]*Academic Lyceum "Physical-Technical High School", Saint Petersburg, Russia*
[4]*SaintPetersburg Polytechnic University, Saint Petersburg, Russia*

A hardware-software computer vision system based on video stream is presented for non-invasive monitoring of presence, engagement, and signs of fatigue in both in-person and remote collectives. The relevance of this work is determined by the need for automated workplace process support: decision-making requires a tool that assesses the functional state in real time without contact sensors. The proposed approach combines several methods into a single pipeline: facial landmark and bounding box extraction using the MediaPipe Face Mesh model, person identification via InceptionResNetV1 embeddings with a preliminary MTCNN cascade, calculation of the Eye Aspect Ratio (EAR) and Mouth Aspect Ratio (MAR), classification of "alert/fatigue" states using a Support Vector Machine with an RBF kernel, pose estimation based on YOLOv8-Pose keypoints, and contact-free photoplethysmography for pulse wave recording using the CHROM method. The server-side component is built on a microservice principle with an event-driven architecture based on FastAPI, an Apache Kafka message broker, and a PostgreSQL database with the TimescaleDB extension; this enables streaming ingestion, long-term storage of time-series data as hypertables, and real-time analytics via continuous aggregates. The analytical module compares filtered telemetry with an individual baseline physiological profile and formalized rules to generate alerts about critical events. The following results were obtained experimentally: face recognition accuracy of 68.7%, a mean normalized landmark localization error of 4.5% with a frame processing time of 15–

25 ms, and a mean fatigue feature classification accuracy of 83% based on *k*-block cross-validation. The mean absolute error of heart rate estimation using videoplethysmography was 2.80 bpm. The system is applicable for automated support of the workplace processes, remote control of operator functional state, point-of-care telemedicine and, in particular, monitoring in educational environments.

Keywords: computer vision, videoplethysmography, functional state monitoring, EAR, MAR, MTCNN, YOLO, MediaPipe, support vector machine, TimescaleDB, microservice architecture, distance learning

REFERENCES

1. Murashko M. [The shortage of doctors in Russia has exceeded 23,000]. *Parlamentskaya gazeta* [Parliamentary Newspaper], 2025. URL: https://www.pnp.ru/politics/murashko-deficit-vrachey-v-rossii-prevysil-23-tysyachi.html (accessed: 31.05.2026). (In Russ.).
2. Pillai A.S. Student Engagement Detection in Classrooms through Computer Vision and Deep Learning: A Novel Approach Using YOLOv4. *SSRN Electronic Journal*, 2025, vol. 5, pp. 87–97. DOI: 10.5281/zenodo.15061626
3. Karimah S.N., Hasegawa S. Automatic engagement estimation in smart education/learning settings: a systematic review of engagement definitions, datasets, and methods. *Smart Learning Environments*, 2022, vol. 9, id. 31. DOI: 10.1186/s40561-022-00212-y
4. Wang C.-Y., Bochkovskiy A., Liao H.-Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 7464–7475. DOI: 10.1109/CVPR52729.2023.00721
5. Zhang K. et al. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 2016, vol. 23, no. 10, pp. 1499–1503. DOI: 10.48550/arXiv.1604.02878
6. Schroff F., Kalenichenko D., Philbin J. FaceNet: A unified embedding for face recognition and clustering. *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823. DOI: 10.1109/CVPR.2015.7298682
7. Cao Q. et al. VGGFace2: A Dataset for Recognising Faces across Pose and Age. *Proc. 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018)*, 2018, pp. 67–74. DOI: 10.48550/arXiv.1710.08092
8. Lugaresi C. et al. MediaPipe: A framework for building perception Pipelines. *arXiv (Cornell University)*, 2019. DOI: 10.48550/arXiv.1906.08172
9. Sukumaran A., Manoharan A. Multimodal engagement recognition from image traits using deep learning techniques. *IEEE Access*, 2024, vol. 12, pp. 25228–25244. DOI: 10.1109/ACCESS.2024.3353053
10. Trabelsi Z. et al. Real-Time Attention Monitoring System for Classroom: A deep learning approach for student's behavior recognition. *Big Data and Cognitive Computing*, 2023, vol. 7, no. 1, id. 48. DOI: 10.3390/bdcc7010048
11. De Haan G., Jeanne V. Robust pulse rate from Chrominance-Based RPPG. *IEEE Transactions on Biomedical Engineering*, 2013, vol. 60, no. 10, pp. 2878–2886. DOI: 10.1109/TBME.2013.2266196
12. Wang W. et al. Algorithmic principles of remote PPG. *IEEE Transactions on Biomedical Engineering*, 2016, vol. 64, no. 7, pp. 1479–1491. DOI: 10.1109/TBME.2016.2609282
13. Kim D.-Y., Lee K., Sohn C.-B. Assessment of ROI selection for Facial Video-Based RPPG. *Sensors*, 2021, vol. 21, no. 23, id. 7923. 7923. DOI: 10.3390/s21237923
14. Van Es. V.A.A. et al. Contactless Cardiovascular Assessment by Imaging Photoplethysmography: A Comparison with Wearable Monitoring. *Sensors*, 2023, vol. 23, no. 3, id. 1505. DOI: 10.3390/s23031505
15. Tiger Data. *Tiger data documettation. TimescaleDB: a time-series database on PostgreSQL*. 2024. URL: https://docs.timescale.com/ (accessed: 31.05.2026).
16. Singh A. et al. VisioPhysioENET: Visual Physiological Engagement Detection Network. *arXiv (Cornell University)*, 2024. 35 p. DOI: 10.48550/arXiv.2409.16126
17. Georgiadis P. et al. Head engagement during visuomotor tracking is determined by postural challenges and aging. *Journal of Neurophysiology*, 2025, vol. 134, no. 6, pp. 2086–2098. DOI: 10.1152/jn.00319.2025
18. Cai S. et al. A human pose estimation network based on YOLOv8 framework with efficient multi-scale receptive field and expanded feature pyramid network. *Scientific Reports*, 2025, vol. 15, id. 15284. DOI: 10.1038/s41598-025-00259-0
19. Saprykin Y.D., Ryazantsev V.I., Smirnov A.A. [Overview of approaches to driver fatigue recognition and existing technical solutions]. *Izvestiya MGTU MAMI* [Proceedings of MSTU "MAMI"], 2020, vol. 14, no. 3, pp. 48–58. DOI: 10.31992/2074-0530-2020-45-3-48-58 (In Russ.).
20. Hugging Face. *MCD-rPPG: Multi-CameraDataset for Remote Photoplethysmography*. URL: https://huggingface.co/datasets/wengziheng/mcd_rppg (accessed: 16.07.2026).

Contacts: *Glebov Stepan Nikolaevich,*
stepangleboff@yandex.ru

The article was received by the editorial office on 26.06.2026

INTRODUCTION

The health and performance of citizens remain one of the foundations of society's economic stability. Heart rate, microcirculation, and blood flow are among the basic physiological processes that determine the body's vital functions. Their analysis is essential for a reliable assessment of a person's current functional state and the prognosis of the development of pathological conditions.

Declining birth rates and the aging of the employable population are leading to a labor shortage in many sectors of the national economy. Healthcare is no exception: according to estimates by the Russian Ministry of Health, the industry's need in 2023 was approximately 29 000 doctors, and by the end of 2024, the shortage was estimated at approximately 23 300 doctors and approximately 63 000 mid-level medical workers [1]. The ongoing shortage of specialists creates a demand for automation, which can ease the workload of staff and simultaneously improve the quality of medical care, including through recommendation systems for non-invasive monitoring of physiological parameters.

This study exploits the fact that disturbances in attention and engagement, as well as states of fatigue, are associated with a specific facial expression and body posture, accompanied by a characteristic sequence of changes in facial tissue color. This set of features can be recorded by a video camera and processed using machine learning methods, opening the possibility of comprehensive analysis of heterogeneous signals using artificial neural systems and multidimensional information processing methods.

The goal of this study is to develop and test a hardware and software system that combines computer vision methods, machine learning algorithms, and a server infrastructure for streaming data processing for remotely monitoring the presence, engagement, and signs of fatigue from a video stream of group members in real time.

The scientific novelty of this work lies in the following. For the first time, a unified multimodal video analytics pipeline has been proposed, combining presence verification, engagement assessment, and fatigue detection without contact sensors during group monitoring. A student audience was chosen as the model specificity. A methodology for the integral engagement indicator E_{pose} has been developed, combining biomechanical parameters with the student's cognitive state into a single scalar metric with weighting coefficients. An event-driven microservice architecture has been applied for the first time to store and analyze physiological time series, enabling horizontal scaling of the system during group monitoring. The use of hypertables and continuous aggregates ensures linear scaling as the number of simultaneously monitored subjects increases, fundamentally distinguishing the proposed solution from monolithic systems with local video processing.

OVERVIEW OF EXISTING APPROACHES

Existing methods for assessing human functional status rely either on instrumental studies or subjective interpretation of observations. Classical medical approaches (blood tests, polymerase chain reaction, magnetic resonance imaging) are highly accurate but require significant time and material costs and are unsuitable for continuous, real-time remote monitoring.

With the development of information and communication technologies and machine learning methods, approaches to non-invasive assessment of human status are actively developing. A key area is computer vision for video image analysis: such systems process visual information without specialized contact sensors, this makes them promising for educational and remote environments [2]. Systematic reviews show that the overwhelming majority of work on engagement assessment uses external observations or self-reports and only a portion uses automated analysis based on machine learning [3].

Deep learning-based face detectors are used to detect presence. Models from the YOLO family provide real-time detection at high throughput [4]; RetinaFace is focused on more accurate localization in complex environments; the MTCNN cascade implements joint detection and localization of keypoints [5]. Identification is performed using face embeddings obtained by networks such as FaceNet [6], pretrained on large datasets, such as VGGFace2 [7]. The accuracy of such systems significantly depends on lighting, head position, and the presence of occlusions.

Facial expression analysis, gaze direction, and head position are used to assess engagement. Models that extract facial keypoints (MediaPipe Face Mesh [8], regressors based on MobileNet v3) allow tracking the dynamics of facial expressions; based on these, the EAR and MAR coefficients are calculated, which are interpreted for detecting blinks, yawns, and prolonged eyelid closures. Classical algorithms, primarily support vector machines with an RBF kernel, demonstrate high stability with limited training sets. A number of studies combine facial expressions, gaze direction, and posture assessment into multimodal engagement models [9], and also apply YOLO detectors to recognize learning behavior in a real classroom [10].

Non-invasive assessment of physiological parameters is provided by remote photoplethysmography (rPPG), which records changes in tissue blood flow based on variations in skin color. Component and projection analysis methods — ICA, PCA, GREEN, CHROM, POS [11–13] — are used to extract the useful signal, while the fast Fourier transform and But-

terworth filters are used for spectral assessment of heart rate. Comparative studies show that the choice of region of interest (ROI) and decomposition method significantly impacts accuracy, while the CHROM and POS projection methods are more robust to motion and lighting artifacts than simple green channel averaging [12]. A comparison of remote and contact photoplethysmography confirms the applicability of chamber methods for assessing heart rate [14]. Specialized time-series-oriented DBMSs, in particular the TimescaleDB extension to PostgreSQL [15], are used to store physiological time series accumulated during group monitoring.

PROBLEM STATEMENT

A literature review reveals that existing methods address presence, engagement, and fatigue primarily in isolation and do not provide comprehensive multimodal assessment with a scalable backend [16, 17]. This underscores the relevance of this work and the formulation of the problem: to develop a system combining computer vision methods and streaming data processing to monitor human presence, engagement, and functional state in real time. A promising application of the proposed solution is, for example, remote student monitoring: in online learning, the system allows teachers to promptly detect waning attention and increasing fatigue, and in in-person classrooms, it allows them to quickly adjust the pace and structure of the lesson, maintaining high levels of group engagement without the need for additional personnel or contact sensors.

SYSTEM ARCHITECTURE AND COMPUTER VISION METHODS

General architecture

The proposed system (Fig. 1) is a distributed hardware and software system combining a client (digital video camera) and a server that receives, stores, and intelligently analyzes streaming telemetry.

Fig. 1. General architecture diagram of the proposed system

Biological tissues are heterogeneous, multicomponent, highly scattering media with significant diagnostic potential. To obtain comprehensive information on physiological processes, a multimodal approach is used, integrating heterogeneous biometric data recorded in parallel.

Multilevel processing of multidimensional biomedical data arrays using multivariate statistics methods is aimed at generating an integrated indicator characterizing the state of a biosystem at different levels of its structural and functional organization. This metric is a formalized model that reflects key characteristics of the biological environment through a given parametric space and serves as the basis for comparative analysis.

Extraction of face region and facial marks

The MediaPipe Face Mesh (MPFM) computer vision model, a high-precision model from Google for real-time keypoint tracking, was chosen as a common solution for several processing modules. It is capable of extracting 468 3D landmarks, including the contours of the lips, eyebrows, eyes, face, and pupil position, enabling detailed analysis of facial expressions and head movements. Architecturally, the approach combines a lightweight neural network for coarse localization of key regions and subsequent coordinate regression, which ensures a balance between speed and accuracy, even on mobile devices. This makes MPFM widely used in augmented reality filters, driver fatigue monitoring, and emotion recognition systems.

MPFM's speed is one of its key advantages. Thanks to its lightweight architecture and efficient operations, the model is capable of processing real-time video streams at frame rates from 20 to over 60 frames per second, even on average smartphones and laptop webcams. On dedicated GPUs or desktop systems, performance can reach hundreds of frames per second. Key MPFM analogs, such as Dlib or OpenFace, are significantly slower due to the greater computational complexity of the algorithms and models used. Furthermore, many alternatives produce fewer landmarks.

In this paper, MPFM is used as the first stage of video stream processing to extract the main facial features. The result is a set of 468 landmarks. From this set, the points required for the processing modules are extracted, and a rectangle around the face is constructed using the minimum and maximum coordinates, allowing a single model to be used for both face detection and feature extraction.

Face detection and identification

As noted above, the result of the first step of video stream processing is, among other things, a rectangle constructed around the face. The extracted facial region is resized to 160 × 160 pixels, and the signal from these pixels is normalized by subtracting the mean and dividing by the standard deviation. To solve the identification problem, the InceptionResNetV1 convolutional network, pretrained on VGGFace2, was used. Before embedding extraction, a cascaded MTCNN detection is performed, consisting of three sequential networks: a Proposal Network (P-Net) for quickly generating candidates, a Refine Network (R-

Net) for refining coordinates, and an Output Network (O-Net) for final localization and output of five keypoints (eyes, nose, corners of the mouth).

Next, 512 dimension embedding is calculated, normalized to unit length. The database of reference embeddings is formed by averaging the vectors of several images of each subject to increase robustness to variability in shooting conditions. Identification is performed using cosine similarity:

$$s(\mathbf{e}, \mathbf{e_i}) = \frac{\mathbf{e} \cdot \mathbf{e_i}}{\|\mathbf{e}\| \times \|\mathbf{e_i}\|}.$$

If the threshold exceeds 0.6, the face is assigned an identifier; otherwise, the face is classified as "unknown." This allows for continuous monitoring, for example, of student attendance and tracking the dynamics of engagement of each identified face.

Gaze direction analysis

Three methods are proposed for pupil center localization (after MPFM eye contour extraction): gradient-based (searching for the point with the maximum brightness gradient after the Sobel operator); threshold-based (binarization using an adaptive threshold to search for the contour with the largest area); and the circular Hough transform. In combined mode, the methods are applied sequentially.

After finding the pupil center (x_p, y_p), the normalized gaze vector relative to the center of the eye region (x_c, y_c) is calculated:

$$g_x = 2\frac{(x_p - x_c)}{w}, \quad g_y = 2\frac{(y_p - y_c)}{h},$$

where w and h are the width and height of the eye region. The resulting values are limited to the range [–1, 1] and smoothed by a moving average. Calibration is performed automatically during the first 30 frames: the maximum value is recorded, against which subsequent measurements are normalized.

Fatigue detection

The fatigue detection model is based on two physiological indices: the eye aspect ratio (EAR) and the mouth aspect ratio (MAR). Through video image analysis, these indices allow for a quantitative assessment of the degree of drowsiness and the likelihood of yawning. Thus, by tracking the dynamics of EAR and MAR over time, the system can promptly detect critical phases of fatigue.

The eye aspect ratio (EAR) is calculated using six characteristic eye contour points P_1,…, P_6 (Fig. 2):

$$EAR = \frac{\|\mathbf{P_2} - \mathbf{P_6}\| + \|\mathbf{P_3} - \mathbf{P_5}\|}{2\|\mathbf{P_1} - \mathbf{P_4}\|}.$$

Fig. 2. Assessing the eye condition of the subject. a — schematic diagram for calculating the eye aspect ratio (EAR); б — reflection of the effect of blinking on the EAR dynamics

When the eye is open, EAR fluctuates around a constant value characteristic of the user; when the eyelids are closed, the numerator approaches zero. A blink is registered when EAR drops below a threshold of 0.25 for three consecutive frames. To smooth out fluctuations, a sliding window of the last ten values and a finite state machine with the states "open," "closing," "closed," and "opening" are used.

The mouth aspect ratio (MAR) is calculated using characteristic lip contour points M_1,…, M_6 in a similar manner — as the ratio of the vertical mouth opening to its width:

$$MAR = \frac{\|\mathbf{M_2} - \mathbf{M_6}\| + \|\mathbf{M_3} - \mathbf{M_5}\|}{2\|\mathbf{M_1} - \mathbf{M_4}\|}.$$

A sharp increase in MAR indicates a yawn.

Using EAR and MAR together allows one to distinguish normal facial expressions from signs of fatigue. To automatically recognize the states of "awakeness" and "fatigue," a classifier based on the support vector machine (SVM) [19] was used. The input vectors are EAR, MAR, and derived features (blink rate, eye closure duration, yawn rate). The SVM constructs an optimal separating hyperplane that maximizes the gap between the classes "awakeness" (high EAR, low MAR) and "fatigue" (low EAR or high MAR). Since the dependencies between features are nonlinear, a nonstandard radial basis function (RBF) kernel is used:

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2},$$

where γ defines the radius of influence of an individual training example.

Contactless video plethysmography

For remote monitoring, remote photoplethysmography (rPPG) was used to obtain a PPG signal by processing video captured by a digital camera. The contactless nature of the method eliminates the need for body sensors, making it ideal for long-term data recording without causing discomfort to the student. A standard personal computer webcam is sufficient for signal recording. To extract the useful signal from the video stream, this study utilizes an image region bounded by the facial contour obtained using MPFM.

A significant problem with dPPG is the artifacts caused by micro-movements of the face and illumination fluctuations. Therefore, simple averaging of the green channel (GREEN) does not provide a heart rate estimate, as it is suppressed by motion artifacts. Therefore, specialized methods must be used to obtain a heart rate estimate. Such methods include ICA, PCA, PBV, CHROM, POS, SSR, and LGI [11–14].

In this study, CHROM was used due to its high reliability and higher determinism compared to ICA.

The mathematical model of light interaction with human skin is based on the dichromatic reflectance model. After several simplifications, the signal in the RGB channels $C(t)$ can be represented as follows:

$$C(t) = u_c I_0 c_0 + \mathbf{u}_c I_0 c_0 i(t) + \mathbf{u}_s I_0 s(t) + \mathbf{u}_p I_0 p(t),$$

where $i(t)$ is the illuminance intensity; $s(t)$ is the specular reflection, $p(t)$ is the diffuse reflection, containing information about the pulse; $u_c I_0 c_0$ is the constant component (noise); $\mathbf{u}_c$ is the skin color vector; I_0 is the average illuminance intensity; c_0 is the reflectance; $\mathbf{u}_s$ is the specular reflection direction vector; $\mathbf{u}_p$ is the color change vector due to the pulse.

The above expression shows that the observation $C(t)$ is a linear mixture of three sources — $i(t)$, $s(t)$ and $p(t)$ — which can be separated by a linear projection. Thus, the task of extracting a pulse signal from RGB observations is reduced to defining a projection system for decomposing $C(t)$.

The CHROM method, one of the decomposition methods, is based on projecting normalized RGB signals onto two orthogonal axes in a plane perpendicular to the skin color vector. Intermediate signals that suppress the specular component are calculated:

$$X_s(t) = 3R_n(t) - 2G_n(t),$$

$$Y_s(t) = 1.5R_n(t) + G_n(t) - 1.5B_n(t).$$

Adaptive alpha tuning is used to finally remove residual motion from the signal. The final pulse signal $S(t)$ is formed as

$$S(t) = X_s(t) - \alpha Y_s(t),$$

where the coefficient α is calculated dynamically as the ratio of the standard deviations of the signals:

$$\alpha = \frac{\sigma(X_s)}{\sigma(Y_s)}.$$

This combination allows in-phase components of motion in these projections to cancel each other out, while the pulse wave, which is out of phase or has a different direction in color space, is preserved.

The time series $S(t)$ obtained after CHROM processing still contains low-frequency trends and high-frequency noise. Since the physiologically determined heart rate (HR) of a healthy person at rest and under moderate exertion lies in the range of 30–180 bpm, the signal is bandpass-filtered in the range of 0.5–4 Hz. In the final step, a fast Fourier transform (FFT) is applied for the transition from the time domain to the frequency domain; the dominant peak is found in the resulting spectrum. The f_{HR} value, expressed in hertz, is converted into beats per minute using the ratio $HR = f_{HR} \cdot 60$.

Pose and motion activity estimation

For pose estimation, the YOLOv8-Pose model [18] was used. It detects 17 key body points in the COCO format. The architecture, based on CSPNet and a feature pyramid, ensures high speed while maintaining accuracy. Biomechanical parameters are calculated: tilt and rotation angles of the head, shoulders, and torso, as well as elbow and knee joint angles (using the cosine law). Posture is classified as "standing/sitting/tilted" based on the relative position of the key points, and torso tilt is classified as "forward" (more than 15°), "backward" (less than 10°), or "neutral."

Motion activity A is estimated based on the displacement of the hip center normalized to a threshold of 10 pixels over the last 30 frames, Pose stability S_{stab} is calculated as the inverse normalized variance of the head tilt angle over the last 10 frames. The integral postural engagement index is:

$$E_{pose} = w_1 S_{lean} + w_2 A + w_3 S_{stab},$$

where S_{lean} is the normalized torso deviation from the neutral position; forward tilt is interpreted as an indication of engagement, and backward tilt as a decrease in engagement.

SERVER COMPONENT OF DATA ANALYSIS

Justification of the architectural approach

Non-contact video plethysmography methods and region-of-interest detection algorithms generate a continuous stream of physiological data. Processing this stream directly on the peripheral device is difficult due to limited computing resources, so the system includes a dedicated computing module (server) that performs centralized data processing, storage, and analysis.

The server component is built on a microservices principle using an event-driven architecture, which separates raw data reception and processing — a critical feature for real-time systems. Streams from client devices are placed in a message queue to smooth out peak loads, increase fault tolerance, and enable horizontal scaling.

Apache Kafka is used as a message broker. Incoming telemetry is published into topics partitioned by device ID to ensure strict message ordering within a single source and enable independent scaling of consumers. A distributed Kafka log acts as a buffer between reception and processing: if the analytics service is briefly unavailable, data is not lost but accumulates in the log and is processed upon recovery. The technology stack is complemented by the Python language with the asynchronous web framework FastAPI, which provides high throughput for input-output operations, and the relational PostgreSQL DBMS with the TimescaleDB extension for working with time series [15].

Telemetry reception module

Data obtained from computer vision algorithms, as well as from specialized non-invasive sensors (heart rate, skin temperature, oxygen saturation, respiratory rate, activity level), is transmitted to the server via a secure channel. The reception module performs initial packet structure validation, timestamp verification, and authentication. Validated packets are serialized and published to the message broker according to the device identifier to ensure strict adherence to the time sequence of measurements for each user.

Time series storage

One of the challenges of remote biomonitoring systems is the rapid growth of the volume of stored data. Classic relational structures exhibit performance degradation with a large number of records. To address this, the TimescaleDB extension to PostgreSQL is used: incoming measurements are stored as hypertables, automatically partitioned by time intervals. The database logical diagram is shown in Fig. 3.

Fig. 3. Database logical diagram (ER diagram) of the telemetry system

Real-time analytics is provided by continuous aggregates: the database calculates average, minimum, and maximum values for indicators over fixed intervals in the background, minimizing visualization delays. A built-in historical data compression policy and automatic deletion of raw telemetry after a specified retention period are applied.

Analytical module for functional state monitoring

The central element of the server-side component is an intelligent functional state analysis module that identifies patterns of fatigue, cognitive deficits, or stress based on the dynamics of biometric parameters. The module operates in continuous processing mode, comparing data with a set of formalized rules in the working database; The rules describe threshold values and logical dependencies (for example, the combination of decreased motor activity with specific changes in heart rate). Processing includes normalization and filtering (removal of noise and artifacts using median filters), assessment of the current state (comparison with the individual baseline physiological profile), generation of alerts about critical events (writing to the alerts table), and notification of the observer and system operator.

SOFTWARE IMPLEMENTATION AND EXPERIMENTAL RESULTS

Data processing pipeline

The software implementation was written in Python using OpenCV (video stream capture and preprocessing), MediaPipe (extracting facial keypoints and metrics), scikit-learn (fatigue detection, SVM classifier), ultralytics (pose extraction using YOLOv8-Pose), as well as FastAPI and TimescaleDB on the server side. The pipeline consists of five stages: frame capture and preprocessing (BGR→RGB conversion); keypoint extraction; feature vector generation (Euclidean distances between point pairs, EAR and MAR coefficients); support vector machine classification; visualization and aggregation of the "tired/not tired" label across the entire video recording.

Experimental results

The theoretical model discussed above provides an analytical description of the problem. To test the algorithm's robustness under real-world recording conditions and evaluate its predictive power, the open MCD-rPPG dataset [20] was used. Recordings made after light physical exercise were selected from the overall dataset for robustness testing.

A control measurement was performed with an Eldar photoplethysmograph. The facial contour was selected as the ROI: it maximizes the pixel sampling volume, reducing the influence of noise, and remains visible throughout the recording [13]. The CHROM method was used for signal decomposition, as it ensures invariance to specular glare and subject movement and does not require prior training. A 5th-order Butterworth filter was used to remove low-frequency trends and high-frequency noise. Spectrograms of the reference dPPG signal, the green channel signal (GREEN), and the CHROM signal are shown in Fig. 4.

Fig. 4. Spectrograms of the reference dPPG signal, green channel signal (GREEN), and CHROM signal

It is evident that the GREEN signal is significantly distorted compared to the reference signal due to light and motion artifacts and exhibits insufficient stability under real-world recording conditions, whereas the CHROM method can cope with the resulting motion artifacts.

To quantitatively evaluate the effectiveness of artifact correction using the CHROM method, a statistical analysis was performed (Fig. 5).

Fig. 5. Statistical analysis of the CHROM algorithm's accuracy. Bland–Altman plot (a) and scatterplot (б)

In the Bland – Altman plot, the mean deviation is -0.72 bpm, indicating the absence of significant systematic bias in the CHROM estimates relative to the reference signal. The overwhelming majority of measurements are concentrated within the confidence interval from –11.10 to +9.66 bpm; isolated outliers are random and correspond to moments of intense facial expression, without critically affecting the statistical validity of the method.

To assess the consistency of the data, a scatterplot was constructed, reflecting the relationship between HR measurements by the CHROM method and reference values. The linear regression equation $y = 0.75x + 19.81$ indicates a strong positive correlation; most experimental points are concentrated along the identity line $x = y$, confirming the validity of the method. The mean absolute error (MAE) for the entire sample was 2.8 bpm, or 3.68% of the mean reference value (76 bpm). Despite the limited time resolution of the video (30 frames/s), which is significantly lower than that of the control heart rate monitor, the remote assessment consistently converges with the standard, which confirms the sufficient accuracy of the CHROM method for use in such systems.

The final metric values for the face recognition model are shown in the Table.

Table. Characteristics of the trained face recognition model

For the SVM classifier, k-block cross-validation ($k = 5$) was used on a sample of video recordings of students in states of wakefulness and fatigue. The average classification accuracy was 83%. The developed software package has low computational complexity and high response speed and is suitable for integration into educational and telemedicine information systems.

The combination of contactless video plethysmography and physiological parameters from non-invasive sensors with a high-performance processing system based on specialized DBMS and message brokers ensures reliable recording of physiological parameters. The proposed approach implements automatic, continuous monitoring of the level of engagement and fatigue of a group of users in real time while maintaining system scalability.

CONCLUSION

This paper proposes and implements a comprehensive computer vision system for remotely monitoring presence, engagement, and fatigue indicators via a video stream, complete with a server-side data analysis subsystem. The system combines contactless video plethysmography, modern architectures of convolutional neural networks (MediaPipe Face Mesh, MTCNN, InceptionResNetV1, YOLOv8-Pose), and a classic support vector machine (SVM) with an RBF kernel to generate an integrated assessment of the learner's functional state.

Face detection models (Accuracy = 69%) and facial keypoint detection models were trained. A MediaPipe-based EAR and MAR feature extraction pipeline was implemented with SVM classification, achieving an average accuracy of 83% based on k-block cross-validation. Comparison of video plethysmography with a reference photoplethysmograph yielded an average absolute error in heart rate estimation of 2.8 bpm. The server component, based on FastAPI, Apache Kafka and PostgreSQL/ TimescaleDB, provides scalable stream processing of time series, automatic aggregation, and the generation of alerts about critical conditions.

Further development of the system involves expanding the training set, integrating additional physiological sensors, and implementing deep learning models for more refined cognitive load assessment. Fulfilling these goals will allow the approach to be applicable in systems for remote monitoring of the complex functional state of subjects, express diagnostics in geographically remote and hard-to-reach areas, and monitor students automatically.