

А.А. Александров, М.А. Бутакова**ПОВЫШЕНИЕ КАЧЕСТВА РАЗНОМОДАЛЬНЫХ ДАННЫХ
НА ОСНОВЕ ГИБРИДИЗАЦИИ МЕТОДОВ ПРЕДВАРИТЕЛЬНОЙ ОБРАБОТКИ
И МУЛЬТИСЕНСОРНОГО СЛИЯНИЯ**

Исследуются методы предварительной обработки данных и слияния разномодальных данных. Собираемые с мультисенсорных устройств данные содержат шумы, аномалии и включают отказы датчиков. Использование неочищенных данных неизбежно приводит к получению ложных закономерностей при использовании моделей машинного обучения, в которых эти данные используются. Для улучшения качества данных необходимо применение специальных методов предварительной их обработки и слияния. Целью работы является создание гибридного конвейера предварительной обработки и мультисенсорного слияния данных, функционирующего на уровне периферийных вычислений. Для её достижения решены задачи удаления из данных последствий технических сбоях и шумов, временной синхронизации, масштабирования признаков и объединения дублирующих источников с учетом их динамического веса. Применены методы: межквартильного размаха для выявления выбросов, линейная интерполяция для их замены, фильтр высоких частот для подавления шумов и взвешенное мультисенсорное слияние с динамическим расчётом весов. Кроме того, представлены методы обработки визуальных, сенсорных и акустических данных в рамках гибридного конвейера, допускающие адаптацию для реализации на уровне периферийных вычислений. Экспериментальная проверка разработанного конвейера проведена на смоделированном наборе климатических данных с искусственно внесёнными зашумлениями и аномалиями. Предложенный подход продемонстрировал высокую эффективность восстановления данных. Для температуры воздуха коэффициент детерминации (R^2) увеличился с 0,9807 до 0,9964 при незначительном изменении средней абсолютной процентной ошибки (MAPE) с 10,11% до 10,48%. Для относительной влажности метрика R^2 выросла с 0,9248 до 0,9675, а ошибка MAPE снизилась с 2,87% до 2,57%. Наиболее значительное улучшение достигнуто для атмосферного давления. Метрика R^2 с отрицательного значения -0,124 выросла до 0,974, а ошибка MAPE сократилась с 0,12% до 0,06 %. Полученные результаты подтверждают, что предложенный конвейер снижает погрешность и повышает качество мультисенсорных данных, минимизирует объём передаваемых данных. Практическая ценность заключается в создании непрерывного цикла локальной фильтрации и слияния данных без накопления погрешностей.

Предварительная обработка данных; слияние данных; мультисенсорные системы; интернет вещей; периферийные вычисления; машинное обучение; качество данных; конвейер обработки данных.

A.A. Alexandrov, M.A. Butakova**IMPROVING THE QUALITY OF MULTI-MODAL DATA BASED
ON THE HYBRIDIZATION OF PREPROCESSING AND MULTISENSORY
FUSION METHODS**

The article explores the methods of data preprocessing and merging of multi-modal data. The data collected from multi-sensor devices contains noise, anomalies, and includes sensor failures. The use of raw data inevitably leads to false patterns when using machine learning models in which this data is used. To improve the quality of data, it is necessary to use special methods for preliminary processing and merging. The aim of the work is to create a hybrid preprocessing and multisensory data fusion controller operating at the peripheral computing level. To achieve this, the tasks of removing the consequences of technical failures and noise from the data, time synchronization, scaling of features and combining duplicate sources, considering their dynamic weight, have been solved. The following methods were applied: inter-quartile scale to detect outliers, linear interpolation to replace them, high-pass filter to suppress noise, and weighed multisensory fusion with dynamic weight calculation. In addition, methods for processing visual, sensory, and acoustic data within a hybrid pipeline are presented that can be adapted for implementation at the peripheral computing level. The experimental verification of the developed pipeline was carried out on a simulated set of climatic data with artificially introduced noises and anomalies. The proposed approach has demonstrated high efficiency of data recovery. For air temperature, the coefficient of determination (R^2) increased from 0.9807 to 0.9964 with a slight change in the mean absolute percentage error (MAPE) from 10.11% to 10.48%. For relative humidity, the R^2 metric increased from

0.9248 to 0.9675, and the MAPE error decreased from 2.87% to 2.57%. The most significant improvement has been achieved for atmospheric pressure. The R2 metric increased from a negative value of -0.124 to 0.974, and the MAPE error decreased from 0.12% to 0.06%. The results confirm that the proposed pipeline reduces the error and improves the quality of multisensory data, minimizes the amount of data being transferred. The practical value lies in creating a continuous cycle of local filtering and merging of data without accumulation of errors.

Data preprocessing; data fusion; multisensory systems; internet of things; peripheral computing; machine learning; data quality; data processing pipeline.

Введение. Достоверность любого научного исследования зависит от правильно подобранных и подготовленных данных. Полученные результаты должны быть повторяемыми и воспроизводимыми в рамках конкретного класса объектов при одинаковых условиях эксперимента. Обоснованное научное исследование предполагает проведение экспериментов строго в рамках научных стандартов и правил, действующих в соответствующей области знаний [1]. Ключевым требованием при этом является использование актуальных данных из различных разномодальных источников, которые позволяют всесторонне рассмотреть изучаемый объект. В эпоху больших данных исследователи часто сталкиваются с заблуждением доступности и высокого качества данных. Большой объем данных не гарантирует их качество.

Особую остроту эта проблема приобретает в системах мониторинга [2], промышленного интернета вещей (IoT) [3], в таких областях как медицина и биология [4], робототехника [5], экология [6], где мультисенсорные устройства собирают гетерогенные данные (температура, влажность, давление, акустические и визуальные сигналы) непосредственно в полевых условиях. Низкое качество первичных измерений ведёт к ошибочным прогнозам и неверным управляющим решениям. Избежать этого помогает обработка информации прямо на самих устройствах.

Также, скрытые системные ошибки часто искажают результаты, а аналитик об этом даже не подозревает. Если обучить классические модели машинного обучения (МО) на таких «грязных» массивах, они найдут ложные закономерности. Проявляется принцип, известный в зарубежной литературе как Garbage In, Garbage Out (GIGO, «мусор на входе – мусор на выходе») [7, 8]. Чтобы этого избежать, данные необходимо тщательно проверять и обрабатывать еще до запуска моделей.

Предварительная обработка состоит из нескольких обязательных шагов, таких как восстановление отсутствующих значений, удаление аномалий, нормировка и кодирование признаков. Комплексная реализация сложных научных исследований зачастую не основывается только на одном источнике или методе проведения исследований. Нужны методы применения гибридного подхода, в том числе подхода слияния данных, известного в научной литературе как Data Fusion – это методы, позволяющие объединить различными образом формируемые массивы данных в едином информационном пространстве [9–11]. Слияние может быть проведено на уровнях данных, признаков и решений. Профессиональное объединение данных позволяет получать более качественные данные и повысить точность исследования. Поэтому предварительная очистка обязательна. Если её пропустить, ошибки из разных источников наложатся друг на друга. В итоге погрешность вырастет лавинообразно.

Текущие методы очистки имеют два подхода: они либо собирают и обрабатывают весь массив данных на центральном сервере, либо разбивают фильтрацию на отдельные несвязанные шаги. В работах [12–14] показана эффективность отдельных методов, но не предложено целостного конвейера, способного одновременно удалять аномалии, компенсировать отказы датчиков и выполнять мультисенсорное слияние с пересчётом весов в реальном времени на периферийном оборудовании. Разработчики все еще предпочитают передавать «сырые» данные на единый сервер и проводить предварительную обработку на нем, не используя адаптивный конвейер, объединяющий фильтрацию, восстановление пропусков и взвешенное объединение резервирующих датчиков без накопления ошибок на самих периферийных устройствах.

Микроконтроллеры и технологии интернета вещей автоматизируют научные эксперименты, снижают трудозатраты и позволяют собирать большие объемы данных в реальном времени при проведении экспериментов *in situ* [15, 16]. Однако, каждый датчик подвержен шумам и погрешностям, а нарушения в работе электрических схем могут приводить к появлению аномальных значений. Поэтому непрерывный поток данных необходимо фильтровать мгновенно. Первичную обработку сигналов переносят на уровень периферийных вычислений (Edge Computing), чтобы отсекал искажения до отправки на сервер [17].

В данной статье предлагается гибридный конвейер предварительной обработки и мультисенсорного слияния, ориентированный на выполнение в среде периферийных вычислений для использования в научных исследованиях. Его новизна состоит в совместном применении методов фильтрации, обнаружения и удаления аномалий, и взвешенного мультисенсорного слияния с динамическим расчётом весовых коэффициентов. В отличие от существующих решений, конвейер не требует передачи сырых данных на сервер, а реализует полный цикл локальной очистки, временной синхронизации, масштабирования признаков и объединения дублирующих источников в реальном времени. Кроме того, в его архитектуре предусмотрена адаптация к разнотипным модальностям – визуальным, сенсорным и акустическим данным. Далее в статье описаны этапы разработанного конвейера, приведены используемые методы обработки, представлены результаты экспериментов на смоделированном климатическом наборе данных и выполнено сравнение с исходными показателями.

Методика исследования. Для повышения качества мультисенсорных данных, собираемых в системах интернета вещей, используют три подхода: централизованная облачная обработка, гибридная архитектура и полностью локальная на периферии. Облачный сервис даёт очень большие вычислительные ресурсы. Однако объёмы передачи данных нагружают сеть, вырастают задержки и появляются уязвимости для доступа к исходным данным, что критично для ресурсоограниченных автономных устройств. Альтернативой служат методы локальной обработки данных.

В работе выбран экспериментальный подход с моделированием на синтетическом наборе данных. Это позволило строго контролировать характер и интенсивность вносимых искажений, что невозможно сделать на реальных «полевых» данных без эталонных значений. Стратегия заключается в создании гибридного конвейера предварительной обработки и мультисенсорного слияния прямо на Edge-устройствах, и его верификации на специально подготовленных данных с известным истинным сигналом. Преимуществом подхода является возможность количественно оценить вклад каждого этапа конвейера в итоговое повышение точности и сравнить полученные показатели с исходными «сырыми» данными.

Процесс разработки и валидации конвейера включает следующие шаги:

1. Подбор методов и алгоритмов гибридного конвейера, включающего стадии фильтрации, синхронизации, удаления аномалий, масштабирования признаков и слияния модальностей. Система должна обрабатывать разные типы данных – визуальные, сенсорные и акустические.
2. Генерация синтетического климатического набора данных (температура воздуха, относительная влажность, атмосферное давление). Для имитации условий реальных сенсорных сетей искусственно добавлены в данные типичные дефекты: шум, выбросы и отказы датчиков.
3. Проверка работы конвейера в условиях смоделированных сбоев и помех.
4. Оценка эффективности конвейера с помощью регрессионных метрик путём сравнения эталонных (истинных) значений и исходными «сырыми» показателями до обработки.

Постановка задачи. Обычно исследователи работают с готовыми архивными файлами, которые уже прошли очистку. Но если обучать или тестировать модели МО на «сырых» данных, система примет шум и случайные всплески за реальные закономерности.

Справиться с этой задачей помогает локальная обработка прямо на Edge-устройствах. Такой конвейер убирает искажения «на месте», разгружает сеть и отправляет на сервер чистый, компактный массив данных.

Конвейер обработки разномодальных данных построен строго последовательно. Сначала отсекаются явные технические сбои, затем убираются шумы на каждом датчике, и синхронизируются данные по времени. После этого масштабируются признаки, и устраняется избыточность с помощью объединения дублирующих источников. Такая последовательность действий не позволяет ошибкам накапливаться, так как каждый этап жестко зависит от предыдущего. Конвейер сразу же исключает аномалии и объединяет дублирующие потоки на уровне «сырых» данных. Такая интеграция работает только для однородных датчиков. Вместо нескольких параллельных сигналов резервных датчиков поступает единственный эталонный сигнал, прошедший полную очистку, нормализацию и подготовку к обработке с использованием МО.

Ключевой особенностью конвейера является метод взвешенного мультисенсорного слияния с динамическим расчётом весов. Обычное усреднение сохраняет ошибки всех каналов, поэтому показаниям присваивают весовые коэффициенты. Если датчик выдает аномальное отклонение, конвейер мгновенно обнуляет его вес. Опираясь на избыточность, алгоритм вычисляет наиболее вероятное истинное значение параметра. Это превращает несколько сомнительных сигналов в одно надежное измерение, что резко повышает точность и отказоустойчивость системы.

Таким образом, разработанный инструментариум закладывает основу для создания унифицированных систем интеллектуального анализа гетерогенных данных. Конвейер универсален, но логика фильтрации, нормализации и комплексирования зависит от физической природы сигналов. Поэтому следующей задачей является разработка методов, применяемых для разномодальных данных – для обработки визуальных, сенсорных и акустических сигналов. Также необходима экспериментальная проверка предлагаемых подходов, которая выполнена и представлена в работе на основе на примере обработки климатических данных.

Методы обработки разномодальных данных. Для количественной оценки степени размытия и зашумленности видеоданных или последовательностей изображений широко применяется фильтр Лапласа. Фильтр использует вторую производную для обнаружения краев на изображении:

$$F(s) = \int_0^{\infty} f(t) e^{-st} dt, \quad (1)$$

где s – комплексное число, t – вещественное число, t' – первая производная функции $f(t)$.

В задачах дискретной обработки визуальных данных преобразование Лапласа работает с двумерными массивами значений яркости пикселей и позволяет количественно оценить градиент яркости в локальной окрестности каждой точки. Технически этот расчет выполняется путем свертки исходного кадра с маской (ядром 3×3) фильтр, коэффициенты которого аппроксимируют вторую производную функции яркости:

$$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}.$$

Вычисление дисперсии, полученной после применения фильтра Лапласа, позволяет оценить уровень шума и степень размытости изображения. Для вычисления дисперсии $D[I_f]$ отфильтрованного изображения I_f размером $n \times m$ сначала вычислим математическое ожидание $E[I_f]$ по формуле:

$$E[I_f] = \frac{1}{n \cdot m} \sum_{i=1}^n \sum_{j=1}^m I_f(i, j). \quad (2)$$

Итоговое значение дисперсии вычисляется согласно выражению:

$$D[I_f] = E\left[\left(I_f - E[I_f]\right)^2\right] = \frac{1}{n \cdot m} \sum_{i=1}^n \sum_{j=1}^m \left(I_f(i, j) - E[I_f]\right)^2. \quad (3)$$

Расчёты по формулам (2) и (3) дают оценку дисперсии. Чем выше дисперсия, тем выше уровень шума в отфильтрованном изображении.

Для удаления шума и сохранения структуры изображения часто применяется медианный фильтр. Медианная фильтрация хорошо удаляет импульсный шум и работает с однородными областями. Цифровое изображение представляет собой двумерный массив чисел $\{x_{i,j}\}$, где (i, j) изменяются по некоторому подмножеству Z^2 . Двумерный медианный фильтр с окном A и размером $n \times m$ для изображения $\{x_{i,j}, (i, j) \in Z^2\}$ определяется как

$$y_{i,j} = \underset{A}{med} x_{i,j} \equiv med[x_{i+r,j+s}; (r,s) \in A], (i, j) \in Z^2,$$

где y – отфильтрованное изображение, med – функция, находящая медиану, (r,s) – индексы смещения координаты внутри скользящего окна.

Данный метод обработки изображений демонстрирует наибольшую эффективность в условиях, когда шум проявляется в виде импульсных помех и представлен ограниченным набором пиковых значений на фоне нулевых значений. В результате применения медианного фильтра показатели яркости в наклонных участках и резких переходах изображения остаются неизменными. Метод показывает ограниченную эффективность при фильтрации флуктуационного шума. Для минимизации ухудшения четкости визуального отображения возможно применение адаптивной медианной фильтрации, которая позволяет лучше справляться с различными типами шумов, сохраняя при этом четкость контуров. Размер скользящего окна (маски) изменяется динамически и зависит от яркости соседних пикселей. Примем пороговый коэффициент отклонения яркости $S_{порог} \in [0,1]$, а величину отклонения яркости соседних пикселей – $A(r, n, m)$, попавших в окно размером $n \times m$. Тогда, относительно яркости центрального отсчета $A(r)$ запишется в виде:

$$S_{nm}(r) = \left| \frac{A(r, n, m) - A(r)}{A(r)} \right|,$$

где r – интенсивность центрального пикселя. Размер маски увеличиваем до тех пор, пока выполняется условие: $\max[S_{nm}(r)] < S_{порог}$.

Если медианный фильтр не дает достаточных результатов, переходят к фильтрации изображений на основе вейвлет-преобразования $\psi(t)$. Его используют для разложения изображения на составные части с разными частотами:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}} \psi\left(\frac{t-b}{a}\right), \quad (4)$$

где t – время (или пространственная координата), a – масштабный параметр, b – параметр смещения.

В отличие от преобразования Фурье, преобразование на основе (4) обеспечивает возможность локального анализа (временного или пространственного) сигнала. С помощью преобразования (4) также очень эффективно удаляются мелкие детали, не содержащие в себе никакой полезной информации, которые проявляются в изображении в виде шума. Вейвлет-преобразование изображения $f(x, y)$ производится следующим образом:

$$W_f(a, b) = \iint f(x, y) \psi_{a,b}(x, y) dx dy, \quad (5)$$

где $W_f(a, b)$ – коэффициенты вейвлет-преобразования при масштабе a и смещении b .

Процесс фильтрации изображений осуществляется посредством дискретного вейвлет-преобразования. Алгоритм раскладывает изображение на низкочастотную и высокочастотную составляющие (аппроксимирующую и детализирующую матрицы). Анализ полученных коэффициентов вейвлет-преобразования (5) позволяет провести очистку изображения от шумов. Процесс уменьшения избыточных данных реализуется через пороговую фильтрацию, где решающим правилом выступает величина T :

$$W_f(a, b) = \begin{cases} W_f(a, b), & \text{если } |W_f(a, b)| > T; \\ 0, & \text{если } |W_f(a, b)| \leq T. \end{cases}$$

После фильтрации изображение восстанавливается из отфильтрованных коэффициентов с помощью обратного вейвлет-преобразования:

$$f(x, y) = \iint W_f(a, b) \psi_{a,b}(x, y) da db.$$

Альтернативный подход базируется на методах нелинейной фильтрации. Алгоритмы Калмана и Винера используют априорную информацию о сигнале. Это помогает минимизировать артефакты при обработке. В основе фильтра Калмана лежит рекурсивный подход. Также метод опирается на модель состояний. Благодаря этому удастся эффективно обрабатывать динамические последовательности. За счет экстраполяции оценок фильтр Калмана проявляет устойчивость к аддитивным шумам даже в условиях неопределенности. Процессы прогнозирования и коррекции ковариационной матрицы описываются классическими выражениями (6) и (7) соответственно:

$$\hat{x}_k^- = \mathbf{F}\hat{x}_{k-1} + \mathbf{B}u_{k-1}, \quad (6)$$

$$\mathbf{P}_k^- = \mathbf{F}\mathbf{P}_{k-1}\mathbf{F}^T + \mathbf{Q}. \quad (7)$$

Переменная в (6) $\hat{x}_k^-$ отвечает за прогноз состояния. Априорный характер этой оценки традиционно отмечают индексом «минус». Остальные параметры определяются следующим образом: $\mathbf{F}$ – матрица перехода между состояниями; $\hat{x}_{k-1}$ – состояние системы в предыдущий момент времени; $\mathbf{B}$ – матрица управляющего воздействия; u_{k-1} – управляющее воздействие в предыдущий момент времени; $\mathbf{P}$ – ковариационная матрица ошибки модели; $\mathbf{Q}$ – ковариационная матрица шума процесса.

Стандартный алгоритм фильтрации Калмана работает по двухэтапной схеме. Она последовательно совмещает шаги экстраполяции и апостериорной коррекции. Для точной проверки прогноза с учетом «свежих» данных рассчитывается оптимальный матричный коэффициент усиления Калмана $\mathbf{K}_k$. Его вычисляют в соответствии с выражением

$$\mathbf{K}_k = \mathbf{P}_k^- \mathbf{H}^T (\mathbf{H} \mathbf{P}_k^- \mathbf{H}^T + \mathbf{R})^{-1},$$

где $\mathbf{H}$ – матрица измерений, отображающая соотношение измерений и состояний, $\mathbf{R}$ – ковариационная матрица шума измерений.

Последующая оптимизация текущей оценки на основе векторов реальных измерений z_k осуществляется путем математического развертывания расчетной формулы

$$\hat{x}_k = \hat{x}_k^- + \mathbf{K}_k (z_k - \mathbf{H}\hat{x}_k^-),$$

где $\hat{x}_k$ – оценка состояния системы с учетом измерения, z_k – измерение в текущий момент времени.

Завершающий пересчет и обновление матрицы плотности ошибок модели выполняется в строгом соответствии со структурной формулой

$$\mathbf{P}_k = (\mathbf{I} - \mathbf{K}_k \mathbf{H}) \mathbf{P}_k^-,$$

где $\mathbf{I}$ – единичная матрица соответствующей размерности.

В случае работы с RGB-изображениями данные операции проводятся локально для каждой матрицы интенсивностей. Математический аппарат такой коррекции, направленный на устранение артефактов освещения и расфокусировки подробно описан в научной литературе [18].

Контрастность монохромного изображения можно определить согласно формуле

$$C = z_{\max} - z_{\min},$$

где $z_{\max}$ – максимальная яркость пикселей, $z_{\min}$ – минимальная яркость пикселей в изображении. Относительная контрастность монохромного изображения вычисляется по формуле

$$C_i = \frac{N}{z_m},$$

где N – яркость цвета, $z_m = 255$ – максимально возможная яркость.

Если динамический диапазон яркости исходного изображения не достигает максимально допустимого значения z_m , то усиление контрастности может быть достигнуто за счет линейного растяжения гистограммы яркостей по формуле

$$z' = \frac{z - z_{\min}}{C_i},$$

где z , $z_{\min}$ – яркость пикселя до и после преобразования.

Это преобразование заключается в перераспределении значений интенсивности пикселей таким образом, чтобы минимальное значение яркости отображалось в нулевое, а максимальное – в z_m , что приводит к расширению диапазона яркости и, как следствие, к улучшению визуального восприятия изображения. Если диапазон яркостных характеристик изображения приближен к максимальным значениям, то повышение различимости смежных фрагментов может быть достигнуто посредством применения нелинейного преобразования шкалы яркости с использованием методов гамма-коррекции или соляризации. Метод гамма-коррекции яркости и контраста реализуется в соответствии с формулой

$$z' = z_m \left(\frac{z}{z_m} \right)^n,$$

где $\frac{z}{z_m}$ – относительная яркость.

Для обработки изображения с высокой яркостью рекомендуется задавать n в пределах $\{2 \dots 10\}$. Если же яркость фрагментов недостаточна, параметр n следует выбирать в интервале $\{0,1 \dots 0,5\}$.

Применение метода соляризации позволяет выделить границы смежных участков за счёт повышения контрастности в соответствии с формулой:

$$z' = 4 \frac{z}{z_m} (z_m - z).$$

Изображения нормализуют, чтобы модели учились быстрее и работали лучше. Чаще всего используют минимаксную нормализацию. Этот метод переводит яркость пикселей в диапазон от 0 до 1. Границы данных фиксируются жестко: минимум становится нулем, а максимум единицей. Нормализация реализуется в соответствии формулой

$$z' = \frac{z - z_{\min}}{z_{\max} - z_{\min}}.$$

Выходные данные измерительных приборов строятся по принципу «параметр-значение». По сути, они представляют собой временные ряды. Данные с приборов часто подвержены воздействию шумов. Их вызывают как внешние факторы, так и внутренние

свойства самих датчиков [19, 20]. Под шумом понимаются случайные отклонения или искажения сигнала, которые маскируют полезный сигнал. Фильтрация – один из методов борьбы с шумом. Аналоговые или цифровые фильтры позволяют избавиться от шумов в определенных частотных диапазонах. Фильтр низких частот (ФНЧ) пропускает низкие частоты и подавляет высокие. Одним из применяемых ФНЧ является Фильтр Баттерворта, который имеет максимально плоскую амплитудную характеристику в полосе пропускания. Фильтр Баттерворта определяется своим порядком n и частотой среза f_c . Передаточная функция $H(s)$ в аналоговой области задается согласно формуле:

$$H(s) = \frac{1}{\sqrt{1 + \left(\frac{s}{\omega_c}\right)^{2N}}},$$

где s – комплексная частота, $\omega_c = 2\pi f_c$ – частота среза, N – порядок фильтра.

Для перехода в цифровую область используется билинейное преобразование по формуле $s = \frac{2}{T} \cdot \frac{1 - z^{-1}}{1 + z^{-1}}$, где T – период дискретизации, z – оператор сдвига.

Фильтр высоких частот (ФВЧ) пропускает высокие частоты и подавляет низкие. Фильтр Баттерворта может быть реализован как ФНЧ, так как и ФВЧ. Передаточной функцией $H(s)$ в аналоговой области является величина, описываемая формулой:

$$H(s) = \frac{\left(\frac{s}{\omega_c}\right)^N}{\sqrt{1 + \left(\frac{s}{\omega_c}\right)^{2N}}}.$$

Полосовой фильтр пропускает сигналы только в определенном диапазоне частот и подавляет сигналы за пределами этого диапазона. Фильтр может быть реализован каскадным соединением на основе ФНЧ и ФВЧ согласно формуле:

$$H(s) = \frac{1}{\sqrt{1 + \left(\frac{s}{\omega_0}\right)^2}} \cdot \frac{\frac{s}{\omega_1}}{\sqrt{1 + \left(\frac{s}{\omega_1}\right)^2}},$$

где ω_0 и ω_1 – частоты среза ФНЧ и ФВЧ соответственно.

Данные временных рядов могут содержать не только шумы, но также и искажения или выбросы, которые представляют собой аномалии, отклоняющиеся от ожидаемого паттерна данных. Аномалии могут быть вызваны ошибками измерений, сбоем в работе оборудования или внешними воздействиями, не имеющими ничего общего с исследуемым процессом. Аномалии во временных рядах могут искажать результаты анализа и прогнозирования, поэтому их обнаружение и устранение являются критически важным этапом.

Применение метода межквартильного размаха (Interquartile Range или IQR) оправдано в условиях, когда специфика формы распределения не является определяющим фактором. Сам показатель IQR охватывает центральную часть выборки. В этот диапазон попадают 50% всех значений. Расчет проводится по формуле $IQR = Q3 - Q1$, где $Q1$ – первый квартиль, ниже которого находится 25% данных, $Q3$ – третий квартиль, ниже которого находится 75% данных. Определение нижней B_l и верхней B_h границ для выбросов осуществляется на основе формул: $B_h = Q1 - 1.5 \times IQR$; $B_l = Q3 + 1.5 \times IQR$.

Найденные аномалии необходимо классифицировать. Это может быть сбой оборудования, случайный выброс или новый тренд. Далее выбираем метод очистки: сглаживаем значения, заменяем их логичными аналогами либо подключаем прогнозные модели.

Чтобы уменьшить влияние случайных выбросов, данные сглаживают. Обычно для этого используют скользящее среднее, медианный фильтр или фильтр Калмана. При работе со временными рядами $y = \{y_1, y_2, \dots, y_n\}$ значение $\tilde{y}_k$ после медианной фильтрации с заданным окном $2m + 1$ вычисляется по формуле

$$\tilde{y}_k = \text{median}(y_{k-m}, y_{k-m+1}, \dots, y_k, \dots, y_{k+m-1}, y_{k+m}),$$

где y_n – элементы вариационного ряда ранжированные в порядке возрастания значений.

Восстановить пропуски можно с помощью линейной или сплайновой интерполяции. Также подойдут прогностические модели, например, ARIMA или LSTM (Long Short-Term Memory). Они рассчитывают нужные значения по соседним точкам или прошлым данным.

Использование нескольких датчиков компенсирует недостатки одиночных приборов [21–24]. Сенсоры дублируют друг друга и при поломке одного из них мониторинг продолжится без потери данных. Комбинация разных устройств решает две задачи – повышает точность измерений и отсеивает информационный шум.

Для количественной оценки согласованности дублирующих потоков информации применяется коэффициент корреляции Пирсона. Он позволяет установить степень их линейной зависимости [25]. Пусть X и Y будут случайными величинами – наблюдаемыми двумерными точками данных. Тогда коэффициент корреляции Пирсона r вычисляется по формуле:

$$r = \frac{\sum_{i=1}^k (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^k (x_i - \bar{x})^2 \sum_{i=1}^k (y_i - \bar{y})^2}},$$

где $\bar{x} = \frac{\sum_{i=1}^k x_i}{k}$, $\bar{y} = \frac{\sum_{i=1}^k y_i}{k}$ – выборочные средние значения.

Поскольку для переменных X и Y стандартные отклонения являются константными. В подобной нотации данная математическая структура традиционно идентифицируется в научной литературе как классический коэффициент корреляции смешанных моментов Пирсона, либо фигурирует под более лаконичным термином «коэффициент корреляции»:

$$r = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y},$$

где σ_x , σ_y – среднеквадратические (стандартные) отклонения соответствующих выборочных совокупностей X и Y .

При сильной корреляции, когда значение коэффициента стремится к единице, исследуемые сигналы становятся практически идентичными. Так мы можем напрямую объединять временные ряды, сохраняя при этом практически весь объем полезной информации. В промежуточном диапазоне (от 0,3 до 0,7), то есть при наличии заметной, но не исчерпывающей связи, решение о слиянии данных принимается уже на основе детального сравнительного анализа измерительных характеристик. Исследователи смотрят на точность и разрешение приборов, а также на физическую специфику самих сигналов. В таком случае используют прецизионному сенсору для максимального сохранения детализации, либо прибору с наименьшим уровнем шума для четкого выделения общего тренда. Выбор зависит от конкретной цели текущего исследования. Естественно, любые собираемые данные требуют обязательной нормализации и глубокого изучения статистических свойств еще до их слияния. Только после приведения всех параметров к единому масштабу можно безопасно приступить к непосредственному слиянию временных рядов.

Выбор метода объединения временных рядов напрямую зависит от качества исходных данных. При разной степени их надежности применяется взвешенное среднее, где веса отражают уровень доверия к источнику. Для метрологически идентичных каналов – с одинаковой погрешностью и частотой записи – достаточно вычислить простое среднее арифметическое:

$$y_t = \frac{\sum_{i=1}^n \omega_i \cdot x_{i,t}}{\sum_{i=1}^n \omega_i},$$

где y_t – объединенное значение в момент времени t , ω_i – вес i -го источника, $x_{i,t}$ – значение из i -го источника в момент времени t .

Анализ акустических сигналов традиционно требует декомпозиции исходных данных и перехода от временной развертки к частотному спектру. Такая трансформация позволяет извлекать ключевые параметры процесса. К ним могут относиться спектральный состав, амплитудное распределение и фазовые соотношения. С математической точки зрения сигналы могут быть описаны в двух плоскостях: временной $x = f(t)$ и частотной $X = F(\omega)$. В первом случае физический процесс интерпретируется как функция $x(t)$, определяющая мгновенное значение амплитуды в произвольный момент времени t , выступающий в качестве независимой переменной.

В частотной области сигнал описывается как функция частоты $X(\omega)$. В данном контексте $X(\omega)$ выступает спектральной характеристикой сигнала, $\omega = 2\pi f$ – независимая переменная, представляющая угловую частоту, а f – частота в герцах. Одним из инструментов для получения спектральных характеристик является преобразование Фурье. Преобразование Фурье позволяет представлять временную функцию, например, звуковую волну, в частотной области. Непрерывное преобразование Фурье (НПФ) определяется по формуле

$$X(\omega) = \int_{-\infty}^{\infty} x(t) \cdot e^{-i\omega t} dt,$$

где $X(\omega)$ – спектральная характеристика сигнала, $x(t)$ – значение исходного сигнала в момент времени t , t – время, j – мнимая единица, ω – угловая частота.

На практике часто используется дискретное преобразование Фурье (ДПФ), которое применяется к дискретным сигналам, поскольку компьютеры работают с конечными выборками данных. ДПФ вычисляется по аналогичной формуле, но с использованием конечных сумм:

$$X(k) = \frac{1}{N} \sum_{n=0}^{N-1} x(nT) \cdot e^{-i2\pi nk/N} = \frac{1}{N} \sum_{n=0}^{N-1} x(nT) \cdot W^{nk},$$

где T – период дискретизации, $n = \{0, 1, \dots, N-1\}$ – номер отсчета дискретизированного сигнала, $k = \{0, 1, \dots, N-1\}$ – номер гармоники сигнала, частота гармоник равна $\frac{k}{T}$, T – период измерения, W – вспомогательная функция.

Если сигнал нестационарный, Фурье-преобразование не сможет показать, в какой момент времени произошли эти изменения. Вейвлет-преобразование, наоборот, позволяет анализировать сигнал одновременно в частотной и временной областях. Оно использует базисные функции (вейвлеты), которые локализованы как по частоте, так и по времени. Это позволяет увидеть, когда и на каких частотах происходят изменения в сигнале. Вейвлет-преобразования делятся на два основных типа – непрерывное Вейвлет-преобразование (НВП) и дискретное Вейвлет-преобразование (ДВП). Непрерывное Вейвлет-преобразование является основным инструментом вейвлет-анализа. Вычисление прямого и обратного НВП осуществляется в соответствии с выражениями (8) и (9):

$$W_s(a, b) = (S(t), \Psi_{ab}(t)) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} S(t) \Psi\left(\frac{t-b}{a}\right) dt, \quad (8)$$

$$S(t) = \frac{1}{C_\Psi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W_s(a, b) \Psi_{ab}(t) \frac{dad b}{a^2}, \quad (9)$$

где $W_s(a, b)$ – результат вейвлет-преобразования, $S(t)$ – исходный сигнал, a – параметр масштаба (определяет частоту), b – параметр сдвига (определяет временное положение), $\Psi(t)$ – вейвлет-функция, масштабированная и сдвинутая, C_Ψ – нормирующий коэффициент, для случая ортонормированных вейвлетов его значение принимается равным единице.

В отличие от НВП, которое за счет анализа на всех масштабах и сдвигах требует значительных вычислительных мощностей, ДВП оперирует фиксированными значениями этих параметров. В основе ДВП лежит концепция многомасштабного анализа с использованием ортогональных вейвлетов. Это позволяет рекурсивно разлагать исследуемый сигнал $S(t)$ на детализирующие $y_{high}[n]$ и аппроксимирующие $y_{low}[n]$ компоненты, последовательно уточняя его представление на разных уровнях разрешения. Разложение сигнала $S(t)$ реализуется через систему фильтров, вычисляющих коэффициенты по формулам:

$$y_{low}[n] = \sum_{k=-\infty}^{\infty} S[k] g[2n-k] \quad (35), \text{ и } y_{high}[n] = \sum_{k=-\infty}^{\infty} S[k] h[2n-k] \quad (36), \text{ где } g[k] \text{ – низкочастотный (аппроксимирующий) фильтр, } h[k] \text{ – высокочастотный (детализирующий) фильтр.}$$

В данном разделе рассмотрены существующие методы очистки и объединения данных для создания на их основе единого гибкого конвейера. По отдельности каждый метод хорошо работает только в «своей» узкой задаче. Ни один из них не позволяет одновременно убирать шумы, подстраиваться под изменения и выдавать точный результат при работе с разными источниками. Поэтому рассмотренные методы объединяются в общую структуру для повышения качества данных. Такая гибридная система позволяет подстроиться под логику обработки разнородных данных и минимизировать вычислительные затраты. Оптимизация конвейера позволяет использовать рассмотренные методы так же на устройствах с ограниченными вычислительными ресурсами.

Результаты. Для численной оценки эффективности и валидации предложенных подходов был проведен имитационный эксперимент на основе верифицированных эталонных климатических данных (температура, относительная влажность, атмосферное давление) за месячный период. С целью моделирования реальных условий эксплуатации измерительных систем в исходные эталонные временные ряды были искусственно внедрены случайные высокочастотные шумы и единичные аномальные пики (выбросы). Параметры генерируемых искажений рассчитывались индивидуально для имитации выходных сигналов конкретных типов датчиков различного класса точности: температуры (DHT22, DS18B20, BMP280), влажности (DHT22, DHT11) и давления (BMP280, BMP180).

Далее к сформированным зашумленным массивам данных был применен конвейер предварительной обработки. Качество фильтрации оценивалось путем прямого сравнения обработанных временных рядов с исходными эталонными значениями. На первом этапе (рис. 1) было выполнено обнаружение аномальных значений с помощью метода IQR. Замена аномалий выполнялась с помощью линейной интерполяции.

Далее выполнена фильтрация высокочастотных компонентов сигнала с помощью ФВЧ. На основе метода взвешенного среднего реализовано слияние данных. Сопоставление обработанных данных с зашумленными наглядно представлено на рис. 2.

При генерации выборки для датчика влажности DHT11 и барометрического сенсора BMP180 была смоделирована повышенная дисперсия случайной составляющей погрешности. Это обусловлено их более низким технологическим классом по сравнению с другими сенсорами. В итоговой таблице весов (табл. 1) более стабильные и прецизионные сенсоры (DHT22, DS18B20, BMP280) получили больший математический приоритет. Веса сенсоров температуры распределились почти поровну.

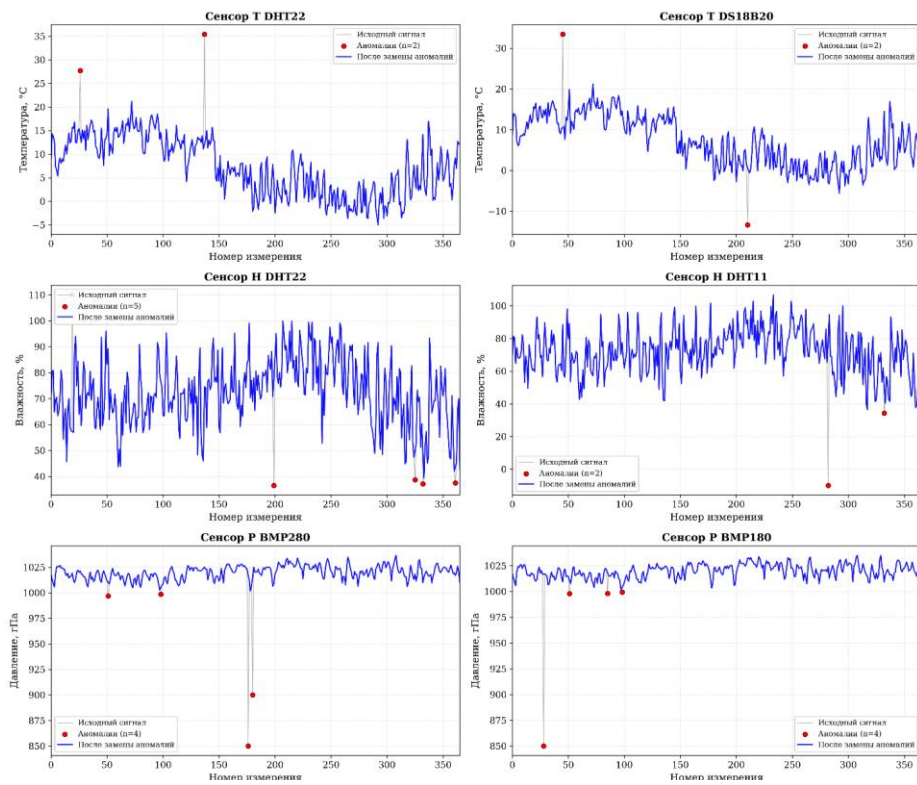


Рис. 1. Обнаружение и замена аномальных значений в показаниях сенсоров

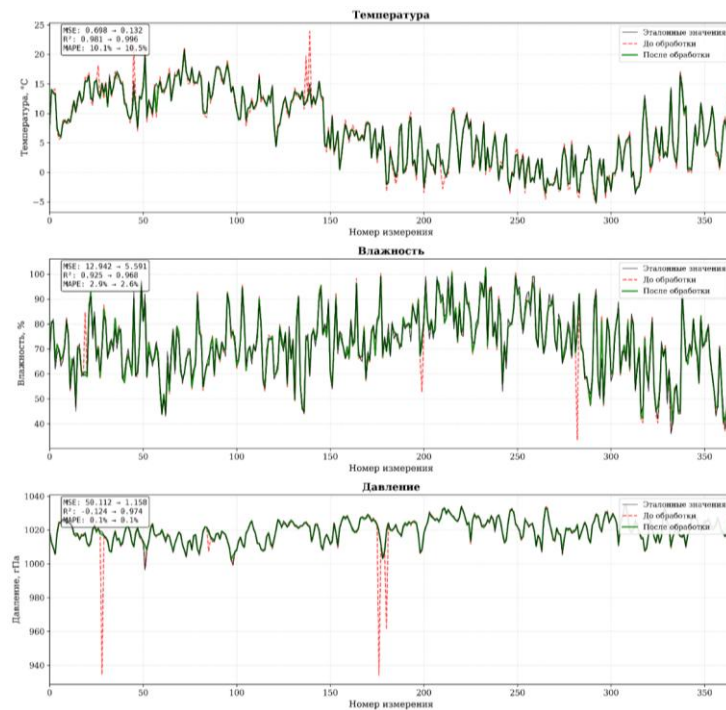


Рис. 2. Сравнение результатов слияния данных до и после обработки

Таблица 1

Весовые коэффициенты сенсоров

	Температура			Влажность		Давление	
Сенсор	T_DHT22	T_DS18B20	T_BMP280	H_DHT22	H_DHT11	P_BMP280	P_BMP180
Значение весов	0.336	0.336	0.329	0.522	0.478	0.504	0.496

Для глубокого анализа точности фильтрации было исследовано эмпирическое распределение плотности вероятности ошибок между эталоном и обработанным результатом. Из графиков (рис. 3) отчетливо видно, что распределения погрешностей для всех трех исследуемых параметров (температуры, влажности и давления) обладают свойствами симметричности и высокой концентрации значений вблизи нуля.

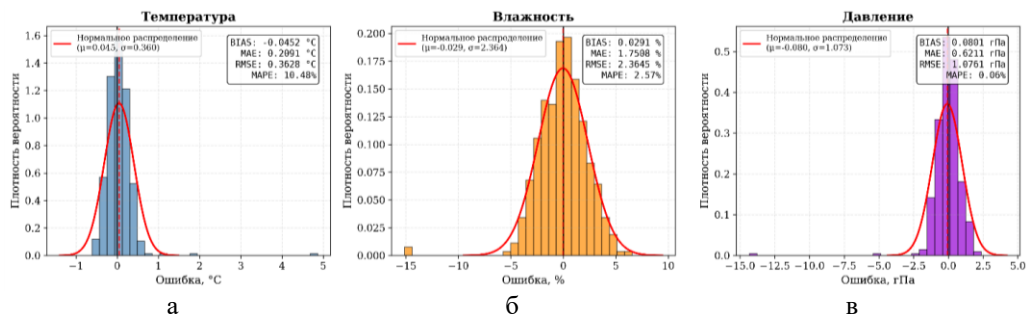


Рис. 3. Распределение ошибок обработанных данных

Эффективность предложенных решений была оценена количественно на основе расчета ключевых показателей качества. На рис. 4 представлен сравнительный анализ критериев точности до и после фильтрации.

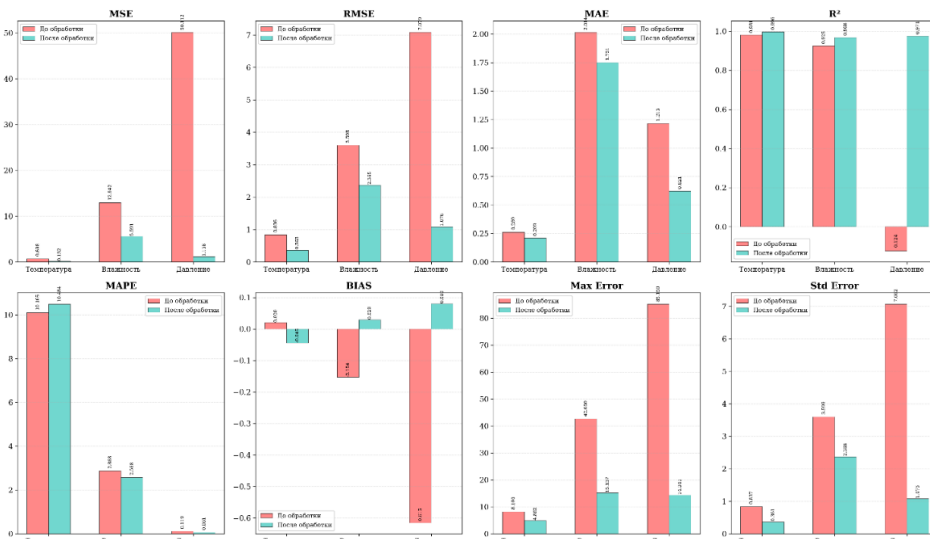


Рис. 4. Сравнительный анализ метрик качества до и после обработки

Гибридный конвейер позволил значительно повысить итоговое качество данных по нескольким ключевым метрикам. Средняя ошибка с учётом знака (BIAS) показывает систематическую погрешность данных. Среднеквадратическая ошибка (MSE) чувствительна к выбросам и аномалиям. Корень из среднеквадратической ошибки (RMSE) характеризует типичный разброс ошибок с усилением влияния крупных отклонений. Средняя абсо-

лютная ошибка (MAE) даёт линейную меру средней ошибки, менее чувствительна к выбросам, чем MSE/RMSE. Средняя абсолютная процентная ошибка (MAPE) позволяет сравнивать точность данных разного масштаба. Максимальная ошибка (Max Error) помогает обнаружить грубые промахи, выбросы, сбои датчиков или ошибки ввода. Стандартное отклонение ошибок (Std Error) отвечает за воспроизводимость данных. Коэффициент детерминации (R^2) показывает, какая доля разброса зависимой переменной объясняется построенной моделью зависимости.

Конвейер предварительной обработки и слияния данных демонстрирует универсальную высокую эффективность для всех рассматриваемых климатических параметров (табл. 2–4). Применение алгоритма привело к радикальному снижению величины ошибок (MSE, RMSE, MAE), минимизации отклонений Max Error и значительному росту коэффициента детерминации (R^2).

Таблица 2

Метрики качества обработки данных температуры

Метрика	До обработки	После обработки	Улучшение
MSE	0.6984	0.1317	+81.15%
RMSE	0.8357	0.3628	+56.58%
MAE	0.2604	0.2091	+19.71%
R^2	0.9807	0.9964	+0.0157
MAPE	10.1053	10.4844	-3.75%
BIAS	0.0201	-0.0452	—
Max Error	8.1000	4.8624	+39.97%
Std Error	0.8366	0.3605	+56.91%

Небольшой рост относительной ошибки (MAPE) для температуры на 3,75% обусловлен спецификой самой метрики. При этом абсолютные показатели (MSE, RMSE) резко улучшились. При приближении реальной температуры к 0 °С деление на близкие к нулю значения искусственно завышает MAPE. Этот скачок не указывает на реальное ухудшение качества обработки.

Таблица 3

Метрики качества обработки данных влажности

Метрика	До обработки	После обработки	Улучшение
MSE	12.9422	5.5911	+56.80%
RMSE	3.5975	2.3645	+34.27%
MAE	2.0145	1.7508	+13.09%
R^2	0.9248	0.9675	+0.0427
MAPE	2.8683	2.5681	+10.47%
BIAS	-0.1536	0.0291	—
Max Error	42.6500	15.1372	+64.51%
Std Error	3.5992	2.3676	+34.22%

Таблица 4

Метрики качества обработки данных давления

Метрика	До обработки	После обработки	Улучшение
MSE	50.1119	1.1580	+97.69%
RMSE	7.0790	1.0761	+84.80%
MAE	1.2148	0.6211	+48.87%
R^2	-0.1240	0.9740	+1.0980
MAPE	0.1191	0.0610	+48.77%
BIAS	-0.6153	0.0801	—
Max Error	85.1500	14.3415	+83.16%
Std Error	7.0618	1.0746	+84.78%

Регрессионный анализ подтверждает точность метода. На диаграммах рассеяния (рис. 5) видна сильная линейная связь между эталонными и обработанными значениями температуры, влажности и давления. Точки плотно ложатся на линию идеального соответствия, а коэффициент детерминации (R^2) более 0,96 доказывают, что конвейер сохраняет исходные физические закономерности.

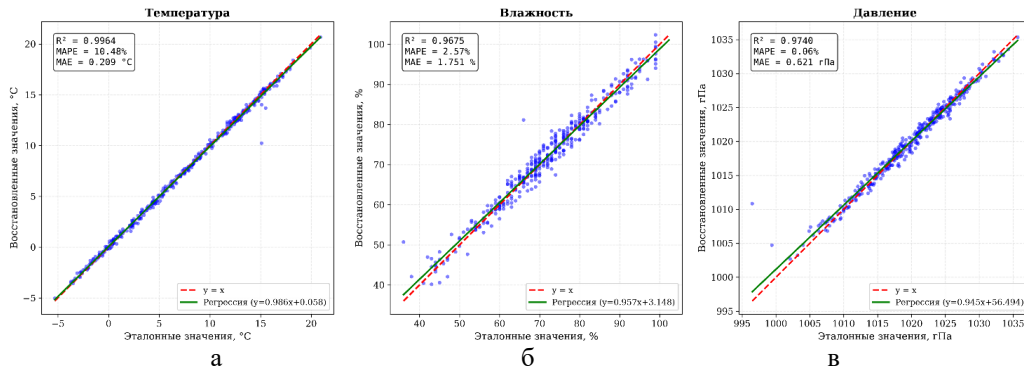


Рис. 5. Корреляция между эталонными и обработанными значениями

Эксперименты показали, что гибридный конвейер действительно справляется с шумами и аномалиями. Совместная работа IQR-фильтрации выбросов, высокочастотной очистки и динамического взвешенного слияния датчиков превращает сырые мультисенсорные данные в материал, пригодный для обучения моделей МО. Для всех трёх климатических параметров (табл. 2–4) R^2 вырос до 0,99, причём для сигнал атмосферного давления ($R^2 = -0,124$) удалось восстановить до 0,974. Конвейер не просто сглаживает шум, а возвращает реальную динамику измеряемой величины. Полученные цифры говорят о том, что подход снимает жёсткие требования к точности сенсоров: даже бюджетные датчики могут давать надёжный эталонный сигнал. В сравнении с обычным усреднением или фиксированными весами динамическое перераспределение приоритетов делает систему отказоустойчивой – временная деградация одного канала почти не влияет на итоговое значение. Локальная очистка и слияние данных без пересылки «сырых» потоков на сервер снижают нагрузку на сеть и энергопотребление, а в перспективе позволяют строить полностью децентрализованные самонастраивающиеся измерительные сети с высокой повторяемостью результатов.

Закключение. В ходе работы обоснован выбор методов предобработки разнородных данных и реализован конвейер, позволяющий повысить их качество. Для верификации конвейера сгенерирован синтетический набор данных.

Тесты на смоделированных климатических данных доказали высокую точность предложенного метода. Для значения параметра температуры R^2 увеличился на 0,0157 и составил 0,9964, а MAPE в процентах составляет 10,48%. Для относительной влажности R^2 вырос на 0,0427 до 0,9675, MAPE составляет 2,57%. Для атмосферного давления R^2 повысился на 1,0980, став положительным и составив 0,9740, MAPE составляет 0,06%.

Эксперименты подтвердили эффективность работы гибридного конвейера предобработки и мультисенсорного слияния данных. Алгоритм полностью устраняет шумы и аномалии, сводя к минимуму погрешности измерений. Объединение дублирующих каналов позволяет формировать единые информативные признаки и избегать ошибок, характерных для одиночных датчиков. Предложенные алгоритмы могут служить основой для систем мониторинга в режиме реального времени, комплексов проведения экспериментов in situ, а также для построения энергоэффективных сетей IoT с возможностью непосредственного развертывания на периферийных вычислительных устройствах.

Дальнейшие исследования будут направлены на оптимизирование вычислительной сложности алгоритмов для ещё более жёстких аппаратных ограничений. Также планируется разработать и интегрировать предложенное решение с протоколами передачи данных в беспроводных сенсорных сетях для создания сквозной самонастраивающейся системы сбора и обработки информации без участия центрального сервера.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Грязнов С.А., Кузнецов М.И. Факторы, влияющие на воспроизводимость научных исследований // Международный журнал гуманитарных и естественных наук. – 2024. – № 8-1 (95). – С. 69-71.
2. Jesus G. Casimiro A., Oliveira A. A Survey on Data Quality for Dependable Monitoring in Wireless Sensor Networks // Sensors. – 2017. – Vol. 17, No. 9. – 2010 p.
3. Zhang L. Jeong D., Lee S. Data Quality Management in the Internet of Things // Sensors. – 2021. – Vol. 21, No. 17. – 5834 p.
4. Боровская В.Г., Гомон Ю.М. Оценка качества данных реальной клинической практики // Реальная клиническая практика: данные и доказательства. – 2022. – Т. 2, № 4. – С. 10-16.
5. Papavasileiou A., Michalos G., Makris S. Quality control in manufacturing – review and challenges on robotic applications // International Journal of Computer Integrated Manufacturing. – 2025. – Vol. 38, No. 1. – P. 79-115.
6. Alexakis T., Adamopoulou E., Peppes N. [et al.]. Evaluating Data Quality: Comparative Insights on Standards, Methodologies, and Modern Software Tools // Electronics. – 2025. – Vol. 14, No. 15. – 3038 p.
7. Bogani R., Theodorou A., Arnaboldi L., Wortham R.H. Garbage in, toxic data out: a proposal for ethical artificial intelligence sustainability impact statements // AI and Ethics. – 2023. – Vol. 3, No. 4. – P. 1135-1142.
8. Kilkenny M.F., Robinson K.M. Data quality: “Garbage in – garbage out” // Health Information Management Journal. – 2018. – Vol. 47, No. 3. – P. 103-105.
9. Liu C., Ye F. A review of multimodal medical data fusion techniques for personalized medicine // Association for Computing Machinery. – New York: ACM, 2025. – P. 338-347.
10. Duan J., Xiong J., Li Y., Ding W. Deep learning based multimodal biomedical data fusion: An overview and comparative review // Information Fusion. – 2024. – Vol. 112. – 102536 p.
11. Azcarate S.M., Ríos-Reina R., Amigo J.M., Goicoechea H.C. Data handling in data fusion: Methodologies and applications // TrAC Trends in Analytical Chemistry. – 2021. – Vol. 143. – 116355 p.
12. Mourad A., Bertrand-Krajewski J.L., El-Youssef A. Edge Computing for Data Anomaly Detection of Multi-Sensors in IoT Environmental Monitoring // Electronics. – 2021. – Vol. 10, No. 3. – 302 p.
13. Zraunig M., Susto G.A., Cenedese A. Detecting Faults at the Edge via Sensor Data Fusion Echo State Networks // Sensors. – 2022. – Vol. 22, No. 8. – 2858 p.
14. Li L., Wang J., Zheng Y. Hybrid Machine Learning-Based Fault-Tolerant Sensor Data Fusion and Anomaly Detection for Fire Risk Mitigation in IIoT Environment // Sensors. – 2025. – Vol. 25, No. 7. – 2146 p.
15. Chauhan D., Shukla K., Desai D. [et al.]. A Comprehensive Review of IoT-Enabled Microgrids: Architectures, Challenges, and Future Prospects to achieve Sustainability // Next Research. – 2026. – 102156 p.
16. Bikakis N., Maroulis S., Papastefanatos G., Vassiliadis P. In-situ visual exploration over big raw data // Information Systems. – 2021. – Vol. 95. – 101616 p.
17. Maia A., Boutouchent A., Kardjadjia Y., Gherari M., Soyak E.G., Saqib M., Boussekar K., Cilbir I., Habibi S., Ali S.O., Ajib W. A survey on integrated computing, caching, and communication in the cloud-to-edge continuum // Computer Communications. – 2024. – Vol. 219. – P. 128-152.
18. Сергеев М.Б., Соловьев Н.В., Стадник А.И. Методы повышения контрастности растровых изображений для систем цифровой обработки видеoinформации // Информационно-управляющие системы. – 2007. – № 1. – С. 2-7.
19. Li D., Wang Y., Wang J., Wang C., Duan Y. Recent advances in sensor fault diagnosis: A review // Sensors and Actuators A: Physical. – 2020. – Vol. 309. – 111990 p.
20. Gasparyan L., Gasparyan F., Simonyan V. Internal Electrical Noises of BioFET Sensors Based on Various Architectures // OJBIPHY. – 2021. – Vol. 11, No. 02. – P. 177-204.
21. Jadhav S.P., Srinivas A., Dipak R.P., Ramkumar P.M., Suryawanshi J., Haldorai A. Deep learning approaches for multi-modal sensor data analysis and abnormality detection // Measurement: Sensors. – 2024. – Vol. 33. – 101157 p.
22. Tang Q., Liang J., Zhu F. A comparative review on multi-modal sensors fusion based on deep learning // Signal Processing. – 2023. – Vol. 213. – 109165 p.
23. Mitsis K., Zarkogianni K., Kalafatis E., Dalakleidi K., Jaafar A., Mourkousis G., Nikita K.S. A Multi-modal Approach for Real Time Recognition of Engagement towards Adaptive Serious Games for Health // Sensors. – 2022. – Vol. 22, No. 7. – 2472 p.
24. Chernov A.V., Savvas I.K., Alexandrov A.A., Kartashov O.O., Polyanchenko D.S., Butakova M.A., Soldatov A.V. Integrated Video and Acoustic Emission Data Fusion for Intelligent Decision Making in Material Surface Inspection System // Sensors. – 2022. – Vol. 22, No. 21. – 8554 p.
25. Wang Q.-G., Li X., Qin Q. Feature Selection for Time Series Modeling // JILSA. – 2013. – Vol. 05, No. 03. – P. 152-164.

REFERENCES

1. Gryaznov S.A., Kuznetsov M.I. Faktory, vliyayushchie na vosproizvodimost' nauchnykh issledovaniy [Factors affecting the reproducibility of scientific research], *Mezhdunarodnyy zhurnal gumanitarnykh i estestvennykh nauk* [International Journal of Humanities and Natural Sciences], 2024, No. 8-1 (95), pp. 69-71.
2. Jesus G. Casimiro A., Oliveira A. A Survey on Data Quality for Dependable Monitoring in Wireless Sensor Networks, *Sensors*, 2017, Vol. 17, No. 9, 2010 p.
3. Zhang L. Jeong D., Lee S. Data Quality Management in the Internet of Things, *Sensors*, 2021, Vol. 21, No. 17, 5834 p.
4. Borovskaya V.G., Gomon Yu.M. Otsenka kachestva dannykh real'noy klinicheskoy praktiki [Quality assessment of real-world data], *Real'naya klinicheskaya praktika: dannye i dokazatel'stva* [Real-World Data & Evidence], 2022, Vol. 2, No. 4, pp. 10-16.
5. Papavasileiou A., Michalos G., Makris S. Quality control in manufacturing – review and challenges on robotic applications, *International Journal of Computer Integrated Manufacturing*, 2025, Vol. 38, No. 1, pp. 79-115.
6. Alexakis T., Adamopoulou E., Peppes N. [et al.]. Evaluating Data Quality: Comparative Insights on Standards, Methodologies, and Modern Software Tools, *Electronics*, 2025, Vol. 14, No. 15, 3038 p.
7. Bogani R., Theodorou A., Arnaboldi L., Wortham R.H. Garbage in, toxic data out: a proposal for ethical artificial intelligence sustainability impact statements, *AI and Ethics*, 2023, Vol. 3, No. 4, pp. 1135-1142.
8. Killenny M.F., Robinson K.M. Data quality: “Garbage in – garbage out”, *Health Information Management Journal*, 2018, Vol. 47, No. 3, pp. 103-105.
9. Liu C., Ye F. A review of multimodal medical data fusion techniques for personalized medicine, *Association for Computing Machinery*. New York: ACM, 2025, pp. 338-347.
10. Duan J., Xiong J., Li Y., Ding W. Deep learning based multimodal biomedical data fusion: An overview and comparative review, *Information Fusion*, 2024, Vol. 112, 102536 p.
11. Azcarate S.M., Ríos-Reina R., Amigo J.M., Goicoechea H.C. Data handling in data fusion: Methodologies and applications, *TrAC Trends in Analytical Chemistry*, 2021, Vol. 143, 116355 p.
12. Mourad A., Bertrand-Krajewski J.L., El-Youssef A. Edge Computing for Data Anomaly Detection of Multi-Sensors in IoT Environmental Monitoring, *Electronics*, 2021, Vol. 10, No. 3, 302 p.
13. Zraunig M., Susto G.A., Cenedese A. Detecting Faults at the Edge via Sensor Data Fusion Echo State Networks, *Sensors*, 2022, Vol. 22, No. 8, 2858 p.
14. Li L., Wang J., Zheng Y. Hybrid Machine Learning-Based Fault-Tolerant Sensor Data Fusion and Anomaly Detection for Fire Risk Mitigation in IIoT Environment, *Sensors*, 2025, Vol. 25, No. 7, 2146 p.
15. Chauhan D., Shukla K., Desai D. [et al.]. A Comprehensive Review of IoT-Enabled Microgrids: Architectures, Challenges, and Future Prospects to achieve Sustainability, *Next Research*, 2026, 102156 p.
16. Bikakis N., Maroulis S., Papastefanatos G., Vassiliadis P. In-situ visual exploration over big raw data, *Information Systems*, 2021, Vol. 95, 101616 p.
17. Maia A., Boutouchent A., Kardjadjia Y., Gherari M., Soyak E.G., Saqib M., Boussekar K., Cilbir I., Habibi S., Ali S.O., Ajib W. A survey on integrated computing, caching, and communication in the cloud-to-edge continuum, *Computer Communications*, 2024, Vol. 219, pp. 128-152.
18. Sergeev M.B., Solov'ev N.V., Stadnik A.I. Metody povysheniya kontrastnosti rastrovyykh izobrazheniy dlya sistem tsifrovoy obrabotki videoinformatsii [Methods for increasing the contrast of raster images for digital video processing systems], *Informatsionno-upravlyayushchie sistemy* [Information and Control Systems], 2007, No. 1, pp. 2-7.
19. Li D., Wang Y., Wang J., Wang C., Duan Y. Recent advances in sensor fault diagnosis: A review, *Sensors and Actuators A: Physical*, 2020, Vol. 309, 111990 p.
20. Gasparyan L., Gasparyan F., Simonyan V. Internal Electrical Noises of BioFET Sensors Based on Various Architectures, *OJBIPHY*, 2021, Vol. 11, No. 02, pp. 177-204.
21. Jadhav S.P., Srinivas A., Dipak R.P., Ramkumar P.M., Suryawanshi J., Haldorai A. Deep learning approaches for multi-modal sensor data analysis and abnormality detection, *Measurement: Sensors*, 2024, Vol. 33, 101157 p.
22. Tang Q., Liang J., Zhu F. A comparative review on multi-modal sensors fusion based on deep learning, *Signal Processing*, 2023, Vol. 213, 109165 p.
23. Mitsis K., Zarkogianni K., Kalafatis E., Dalakleidi K., Jaafar A., Mourkousis G., Nikita K.S. A Multi-modal Approach for Real Time Recognition of Engagement towards Adaptive Serious Games for Health, *Sensors*, 2022, Vol. 22, No. 7, 2472 p.
24. Chernov A.V., Savvas I.K., Alexandrov A.A., Kartashov O.O., Polyanichenko D.S., Butakova M.A., Soldatov A.V. Integrated Video and Acoustic Emission Data Fusion for Intelligent Decision Making in Material Surface Inspection System, *Sensors*, 2022, Vol. 22, No. 21, 8554 p.
25. Wang Q.-G., Li X., Qin Q. Feature Selection for Time Series Modeling, *JILSA*, 2013, Vol. 05, No. 03, pp. 152-164.

Александров Александр Алексеевич – Научно-исследовательский и проектно-конструкторский институт информатизации, автоматизации и связи на железнодорожном транспорте АО «НИИАС», Ростовский филиал; e-mail: aa.alexandrov@vniias.ru; г. Ростов-на-Дону, Россия; тел.: + 79064278038.

Бутакова Мария Александровна – Ростовский государственный университет путей сообщения; e-mail: butakova@rgups.ru; г. Ростов-на-Дону Россия; тел.: +791859766 71; д.т.н.; профессор.

Aleksandrov Aleksandr Alekseevich – JSC NIIAS; e-mail: a.alexandrov@vniias.ru; Rostov-on-Don, Russia; phone: + 79064278038; Master's Degree in Computer Science.

Butakova Maria Aleksandrovna – Rostov State Transport University; e-mail: butakova@rgups.ru; Rostov-on-Don, Russia; phone: +79185976671; dr. of eng. sc., professor.

УДК 621.372

DOI 10.18522/2311-3103-2026-4-38-50

Н.Н. Войт

МОДИФИЦИРОВАННЫЙ АЛГОРИТМ ИНТЕГРАЦИИ СИСТЕМЫ УПРАВЛЕНИЯ ЖИЗНЕННЫМ ЦИКЛОМ ИЗДЕЛИЯ С ВИРТУАЛЬНОЙ ПЛОЩАДКОЙ ПРОЕКТИРОВАНИЯ

Низкая успешность разработки аппаратно-программных средств сложных изделий является проблемой в САПР, которая влечет траты материальных, технических и людских ресурсов, поэтому с целью повышения успешности разработки изделий работа вносит вклад в решение широко известной и важной задачи твердотельного моделирования – это задача по интеграции виртуальных прототипов. Задача включает цифровое представление изделия, отражающее пространственное взаимодействие компонентов и позволяющее оценивать работоспособность конструкций, и относится к PSPACE-классу сложных задач. В работе описана интеграция виртуальных прототипов на пример системы управления жизненным циклом изделия ЛОЦМАН:PLM с виртуальной площадкой проектирования Unity 3D. Работа содержит разработку модифицированного алгоритма поиска и извлечения в репозиторий 3D-сборок. Алгоритм отличается блоком поиска окружения сборки api.getinfo, позволяющим найти и поместить в графический репозиторий конструкторскую, технологическую и эксплуатационную документацию (данные, метаданные и прочее). По помещенным данным в виртуальную площадку специалистом выполняется оценка эргономики 3D-сборок. В рамках крупного проектно-промышленного предприятия проведены вычислительные эксперименты с модифицированным алгоритмом, реализованном как интегратор 3D-сборок, и получены примеры применения, а также выполнено внедрение интегратора. В виртуальной реальности специалистом было оценено расположение рабочих органов и блоков по примерам. Модифицированный алгоритм вносит вклад в решение проблемы успешности, сокращая сроки построения, оценки и правок 3D-сборок на 10 % (на 1,25 рабочих дня) по результатам экспериментов. Интегратор используется в конструкторском проектировании 3D-сборок и в процессе обучения (повышении квалификации, переподготовке) специалистов без отрыва от производства. Благодаря интерфейсной реализации интегратор взаимодействует с продуктами САПР компаний России и других стран. В работе используется теоретико-множественное обоснование для математического описания выражений, параметров и значений. Дальнейшие исследования связаны с выявлением ошибок по поведенческой модели специалиста.

Виртуальная реальность; ЛОЦМАН; Unity 3D; проектирование; интеграция; изделие; прототип.

N.N. Voit

MODIFY INTEGRATION ALGORITHM BETWEEN CONTROL SYSTEM OF PART LIFE CYCLE AND VIRTUAL DESIGN PLATFORM

The low hardware and software development success is a problem in CAD. The problem entails large expenditures of material, technical and human resources. In order to increase the product development success need to solve the issue of integrating virtual prototypes in the design stage. The issue includes a digital representation that allows evaluate the design's operability. The paper describes the integration between LOTSMAN:PLM and Unity 3D, that includes the modified algorithm development for searching and extracting 3D assemblies in the repository. This algorithm has a name api.getinfo, that