

МАЦІЄВСЬКИЙ Вадим

Державного університету «Житомирська політехніка»

<https://orcid.org/0000-0003-3947-1363>

matciyevskiyv@gmail.com

НІКІТЧУК Тетяна

Державного університету «Житомирська політехніка»

<https://orcid.org/0000-0002-9068-931X>

ntn@ztu.edu.ua

АЛГОРИТМІЧНА КОРЕКЦІЯ «ЕФЕКТУ БАТЧУ» В ETL-КОНВЕЄРІ ІНТЕЛЕКТУАЛЬНОЇ ПЛАТФОРМИ ОПРАЦЮВАННЯ ГЕТЕРОГЕННИХ БІОМЕДИЧНИХ ДАНИХ

У статті розглянуто етап первинної обробки даних у складі ETL-конвеєра інтелектуальної платформи для опрацювання біомедичних показників, що агрегує дані з розподілених сенсорних мереж (IoT-пристроїв, медичних моніторів) та публічних репозиторіїв (MIMIC-III, PhysioNet). Обґрунтовано, що ключовим архітектурним викликом при об'єднанні гетерогенних джерел є «ефект батчу» – системний зсув розподілу даних, зумовлений технічними факторами (інструментальна похибка, умови середовища, міжлабораторна варіативність), а не фізіологічним станом пацієнта. Запропоновано математичну модель корекції на основі алгоритму ComBat, що враховує адитивні та мультиплікативні ефекти батчу за допомогою емпіричного методу Байєса, з окремим виділенням матриці клінічних коваріат (вік, стать, ІМТ, хронічні діагнози) для збереження біологічно значущої варіативності. Наведено порівняльний аналіз методів усунення «ефекту батчу» (Z-score, квантильна нормалізація, ComBat, денойзинг, Batch Normalization) та методів імпутації пропущених значень (Mean/Median, KNN, MICE, LOCF/NOCB, MissForest, GAIN) із зазначенням програмних бібліотек їх реалізації. Запропоновано архітектурне рішення – виділення модуля ComBat-корекції в окремий мікросервіс етапу первинної обробки даних, що дозволяє динамічно адаптувати платформу до нових типів сенсорів без перенавчання основної моделі.

Ключові слова: ефект батчу; ComBat; ETL-конвеєр; IoMT; гетерогенні біомедичні дані; емпіричний метод Байєса; матриця клінічних коваріат; імпутація даних; нормалізація сигналів; інтелектуальна платформа.

MATSHEVSKYI Vadym, NIKITCHUK Tetiana

Zhytomyr Polytechnic State University

ALGORITHMIC CORRECTION OF THE “BATCH EFFECT” IN THE ETL PIPELINE OF AN INTELLIGENT PLATFORM FOR PROCESSING HETEROGENEOUS BIOMEDICAL DATA

The article examines the data preprocessing stage within the ETL pipeline of an intelligent platform for processing biomedical indicators that aggregates data from distributed sensor networks (IoT devices and medical monitors) and public repositories (MIMIC-III, PhysioNet). It is substantiated that a key architectural challenge in integrating heterogeneous data sources is the “batch effect” – a systematic shift in data distribution caused by technical factors (instrumental error, environmental conditions, and interlaboratory variability) rather than by the patient’s physiological condition. A mathematical correction model based on the ComBat algorithm is proposed. The model accounts for additive and multiplicative batch effects using an empirical Bayes approach, while a separate clinical covariate matrix (age, sex, BMI, and chronic diagnoses) is incorporated to preserve biologically meaningful variability. A comparative analysis of methods for eliminating the “batch effect” (Z-score normalization, quantile normalization, ComBat, denoising, and Batch Normalization) and methods for imputing missing values (Mean/Median, KNN, MICE, LOCF/NOCB, MissForest, and GAIN) is presented, along with the corresponding software libraries used for their implementation. An architectural solution is proposed that involves separating the ComBat correction module into an independent microservice at the data preprocessing stage, enabling the platform to dynamically adapt to new sensor types without retraining the core model.

Keywords: batch effect; ComBat; ETL pipeline; IoMT; heterogeneous biomedical data; empirical Bayes method; clinical covariate matrix; data imputation; signal normalization; intelligent platform.

Стаття надійшла до редакції / Received 27.07.2026

Прийнята до друку / Accepted 31.08.2026

Опубліковано / Published 10.09.2026



This is an Open Access article distributed under the terms of the [Creative Commons CC-BY 4.0](https://creativecommons.org/licenses/by/4.0/)

© МАЦІЄВСЬКИЙ Вадим, НІКІТЧУК Тетяна

ПОСТАНОВКА ПРОБЛЕМИ У ЗАГАЛЬНОМУ ВИГЛЯДІ ТА ЇЇ ЗВ'ЯЗОК ІЗ ВАЖЛИВИМИ НАУКОВИМИ ЧИ ПРАКТИЧНИМИ ЗАВДАННЯМИ

Розвиток інтелектуальних платформ моніторингу стану здоров'я супроводжується стрімким зростанням гетерогенності джерел біомедичних даних: до аналізу одночасно залучаються показники з носимих IoT-пристроїв та медичних моніторів, а також дані з публічних клінічних репозиторіїв (MIMIC-III, PhysioNet) [1, 2]. Кожне з цих джерел має власні технічні характеристики – частоту дискретизації, розрядність аналого-цифрового перетворення, конструктивні особливості сенсорів, умови калібрування, – що породжує так званий «ефект батчу»: системний зсув розподілу значень показника, не пов'язаний із фізіологічним станом

пацієнта. Якщо цей ефект не усунути на етапі первинної обробки, він призводить до хибної ідентифікації патернів моделями машинного навчання та знижує достовірність подальшого прогнозного моделювання.

Наявні підходи до нормалізації гетерогенних даних (Z-score, квантильна нормалізація) здебільшого не враховують клінічно значущі коваріати пацієнта (вік, стать, індекс маси тіла, супутні діагнози), що створює ризик «статистичного нівелювання» – усунення разом із технічним шумом і реальної, діагностично важливої варіативності. Це визначає потребу в методах корекції, здатних розрізняти технічну та біологічну складові варіативності даних ще на етапі проєктування архітектури програмного забезпечення [3-7].

ФОРМУЛЮВАННЯ ЦІЛЕЙ СТАТТІ

Метою статті є розробка та обґрунтування алгоритмічного рішення для етапу первинної обробки даних ETL-конвеєра інтелектуальної платформи опрацювання біомедичних показників, що забезпечує усунення «ефекту батчу» при об'єднанні даних розподілених сенсорних мереж і зовнішніх датасетів із збереженням клінічно значущої варіативності пацієнтів, а також порівняльний аналіз наявних методів нормалізації та імпутації даних для обґрунтування вибору оптимального інструментарію в архітектурі платформи.

Для досягнення поставленої мети в роботі здійснено математичне обґрунтування методу ComBat на основі емпіричного методу Байєса, визначено структуру та роль матриці клінічних коваріат, а також проведено порівняльний аналіз програмних методів усунення «ефекту батчу» та імпутації пропущених значень із зазначенням відповідних бібліотек реалізації (Python/R).

ВИКЛАД ОСНОВНОГО МАТЕРІАЛУ

Первинна обробка медичних даних у межах інтелектуальної платформи, що розробляється, розглядається як критичний етап, що визначає достовірність подальшого прогнозного моделювання. Основним архітектурним викликом при агрегації даних із розподілених сенсорних мереж (IoT-девайси, медичні монітори) та зовнішніх публічних репозиторіїв (MIMIC-III, PhysioNet) є виникнення «ефекту батчу» (системного зсуву доменів), що порушує вимоги до синтаксичної та семантичної інтероперабельності програмного забезпечення.

Ефект батчу проявляється як системна варіативність, що не корелює з біологічним станом пацієнта, а є наслідком технічних факторів, а саме:

інструментальна похибка, що проявляється як різниця в частоті дискретизації та розрядності АЦП при знятті біоелектричних сигналів (ЕКГ, ЕЕГ);

умови середовища, наприклад, вплив освітленості та щільності прилягання фотоплетизмографічних сенсорів при вимірюванні сатурації (SpO2) та HRV;

міжлабораторна варіативність, адже різниця в реагентах та калібруванні обладнання при аналізі лабораторних маркерів (глюкоза, холестерин) впливає на результат.

Математично це описується як зміщення розподілу ймовірностей, що призводить до хибної ідентифікації патернів моделями машинного навчання.

Для усунення систематичних відхилень між різними джерелами даних («батчами») у межах інтелектуальної платформи використовується лінійна регресійна модель, що враховує адитивні та мультиплікативні ефекти. Для нівелювання вказаних відхилень у роботі використано алгоритм ComBat, заснований на емпіричному методі Байєса. На відміну від стандартних методів масштабування (Min-Max або Z-score), ComBat припускає, що ефект батчу впливає як на середнє значення, так і на дисперсію показника.

Припустимо, що Y_{ijg} – це значення g -го показника (наприклад, амплітуда сегмента ЕКГ або рівень глюкози) для j -го вимірювання у i -му батчі (конкретному сенсорі чи датасеті). Згідно з фундаментальним дослідженням Johnson et al. [8, 9], модель має вигляд:

$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\epsilon_{ijg} \quad (1)$$

де:

α_g – середнє базове значення показника g ;

X – матриця клінічних коваріат (наприклад, вік, стать пацієнта), які необхідно зберегти;

β_g – вектор коефіцієнтів для відповідних коваріат;

γ_{ig} – адитивний ефект батчу i на показник g (зміщення середнього);

δ_{ig} – мультиплікативний ефект батчу i на показник g (зміна дисперсії/масштабу);

ϵ_{ijg} – випадкова помилка з нульовим середнім та дисперсією σ_g^2 .

Варто зазначити, що γ_{ig} та δ_{ig} – адитивний та мультиплікативний ефекти батчу для показника g . Це дозволяє «вирівняти» дані з різних клінік або пристроїв, зберігаючи коваріати (вік, стать, клінічний статус), що підтверджує доцільність його використання в архітектурі платформи.

Опишемо детально етапи реалізації в інтелектуальній платформі.

На основі методу найменших квадратів обчислюється оцінки параметрів $\hat{\alpha}_g, \hat{\beta}_g, \hat{\gamma}_{ig}, \hat{\delta}_{ig}$. Дані

стандартизуються для усунення впливу загального середнього коваріат:

$$Z_{ijg} = \frac{Y_{ijg} - \hat{\alpha}_g - X\hat{\beta}_g}{\hat{\sigma}_g} \quad (2)$$

Байєсівська оцінка параметрів батчу для уточнення значень γ_{ig} та δ_{ig} .

Корекція значень, де кінцеві скориговані дані $\hat{Y}_{ijg}$, які позбавлені ефекту батчу, але зберігають клінічну варіативність, обчислюються як:

$$\hat{Y}_{ijg} = \frac{\hat{\sigma}_g}{\delta_{ig}^*} (Z_{ijg} - \hat{\gamma}_{ig}^*) + \hat{\alpha}_g + X\hat{\beta}_g \quad (3)$$

де $\hat{\gamma}_{ig}^*$ та δ_{ig}^* уточнені байєсівські оцінки.

У межах інтелектуальної платформи матриця X відіграє роль «захисного фільтра» для біологічно значущих сигналів. При проведенні корекції «ефекту батчу» алгоритм повинен чітко розрізняти, які коливання показників спричинені типом сенсора, а які індивідуальними характеристиками пацієнта.

1. Структура матриці коваріат.

Матриця X включає вектори параметрів, що мають стабільний кореляційний зв'язок із цільовими біосигналами. Для нашої платформи основними коваріатами є:

вік X_1 та стать X_2 . Наприклад, відомо, що базова частота серцевих скорочень (ЧСС) та її варіабельність природно змінюються з віком. Без врахування цієї коваріати алгоритм міг би помилково «нормалізувати» пульс літньої людини до показників молоді;

індекс маси тіла (ІМТ, X_3). Це критично для аналізу амплітуди ЕКГ-сигналу та точності вимірювання сатурації (SpO2) через товщину тканин;

наявність підтверджених хронічних діагнозів X_4 , наприклад, гіпертензія). Це дозволяє системі зрозуміти, що високий артеріальний тиск у конкретного батчу даних є ознакою патології пацієнта, а не несправністю тонометра.

2. Математична роль у моделі

Включення матриці X у рівняння $Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\epsilon_{ijg}$ дозволяє оцінити вектор коефіцієнтів β_g .

Під час ітераційного обчислення алгоритм спочатку визначає частку варіації, що пояснюється параметрами X .

Результат ефекту батчу γ_{ig} обчислюється лише на основі залишкової варіації, яка не пояснюється клінічними коваріатами.

3. Приклад реалізації для показника ЧСС

Припустимо, ми об'єднуємо дані з домашнього смарт-годинника (Batch 1) та госпітального монітора (Batch 2).

Якщо в Batch 1 середній вік пацієнтів 20 років, а в Batch 2 – 70 років, звичайна нормалізація (Z-score) стерла б різницю в їхніх фізіологічних нормах.

ComBat із матрицею X «бачить» вік як коваріату, тому він скоригує лише технічне зміщення (наприклад, +5 ударів через похибку годинника), але залишить різницю, зумовлену віковими особливостями пацієнтів.

Застосування матриці клінічних коваріат у процесі трансформації даних дозволяє реалізувати принцип персоніфікованої нормалізації. Це мінімізує ризик «статистичного нівелювання» патологічних станів, забезпечуючи високу чутливість платформи до індивідуальних відхилень у межах різних демографічних груп пацієнтів.

При впровадженні алгоритму в архітектуру інтелектуальної платформи було враховано наступні теоретичні припущення моделі.

Припущення про лінійність ефектів де модель передбачає, що вплив батчу на медичний показник є лінійним комбінованим ефектом (адитивне зміщення середнього та мультиплікативна зміна дисперсії).

Для більшості сенсорних похибок (калібрування, чутливість фотодіодів) лінійна апроксимація є достатньою і статистично стійкою.

Алгоритм припускає, що випадкові помилки ϵ_{ijg} розподілені за нормальним законом з нульовим середнім, тоді як γ_{ig} є адитивним ефектом батчу з апіорним нормальним розподілом, а δ_{ig}^2 – мультиплікативним ефектом з апіорним оберненим гамма-розподілом; ці припущення стосуються різних рівнів моделі та не є взаємозамінними. Але, якщо дані (наприклад, викиди в ЕКГ) мають занадто «важкі хвости» розподілу, перед ComBat обов'язково необхідно провести попередній денойзинг або логарифмічне перетворення.

ComBat найкраще працює, коли в різних батчах (різних наборах даних або групах пристроїв) присутні пацієнти з подібними коваріатами (наприклад, є і чоловіки, і жінки в обох групах). Необхідно мати на увазі, якщо в одному батчі лише молоді пацієнти, а в іншому – лише літні, модель може «переплутати» ефект віку з ефектом пристрою (явище колінеарності). На етапі вибірки платформа проводить перевірку на

збалансованість вибірок.

Класичний ComBat обробляє кожен показник g окремо. Оскільки медичні параметри (пульс і тиск) корелюють, ми використовуємо модифікацію ComBat-M, яка враховує коваріати X для збереження цих міжпараметричних зв'язків.

Припускається, що похибка пристрою є стабільною протягом сесії вимірювання. Для динамічних умов (наприклад, зміна положення сенсора під час руху) передбачено повторну рекалібрування або сегментацію даних на коротші «стаціонарні» батчі.

Підсумовуючи, у нашому контексті «ефект батчу» – це викривлення даних, що виникає через технічні відмінності між джерелами (різні сенсори, моделі пристроїв або умови збору даних), а не через реальний стан пацієнта.

У межах ETL-процесу він виявляється через:

об'єднуємо дані з різних медичних пристроїв (наприклад, пульсометр Apple Watch та професійний медичний монітор) або завантажуюмо частину даних із відкритих наборів даних. Кожен пристрій має власну похибку, частоту дискретизації або рівень чутливості.

система може помилково прийняти технічну різницю між пристроями за медичний показник. Наприклад, якщо пристрій А завжди дає на 2% вищу температуру, ніж пристрій Б, алгоритм може вирішити, що група пацієнтів з пристроями А «більш хвора», хоча це просто системна помилка обладнання.

При цьому саме денойзинг (очищення від шумів конкретного сенсора), імпуація (заповнення пропусків, які в різних батчах виникають по-різному) та нормалізація (приведення всіх значень до єдиної шкали) нівелюють ці розбіжності (роль етапу первинної обробки даних). Це дозволяє платформі аналізувати сутність даних, а не особливості заліза, на якому вони були зібрані.

Отже, математичне обґрунтування методу ComBat дозволяє продемонструвати глибину пропрацювання етапу первинної обробки даних. Цей метод базується на емпіричному методі Байєса, який дозволяє «позичати» інформацію між різними показниками (генами або сенсорами) для стабільної оцінки похибок навіть у невеликих вибірках.

Для усунення «ефекту батчу» при роботі з гетерогенними медичними даними (сенсори + набори даних) зазвичай використовують комбінацію статистичних та алгоритмічних методів нормалізації.

Зазначимо основні підходи, які варто закласти в етап первинної обробки даних.

1. Статистична стандартизація (Z -score), де кожне значення перераховується відносно середнього значення (μ) та стандартного відхилення (σ) конкретного батчу (пристрою).

Отож, дані кожного сенсора центруються навколо нуля. В результаті у калібруванні пристроїв нівелюється, і система бачить лише відхилення від індивідуальної норми.

2. Метод квантильної нормалізації прийшов із генетики і є надзвичайно потужним для медичних даних.

При застосуванні саме цього методу розподіл значень у всіх батчах робиться ідентичним. Найвище значення в батчі А стає рівним найвищому в батчі Б, друге – другому і так далі. В результаті повністю прибираємо технічну варіативність між різними типами сенсорів, зберігаючи при цьому внутрішню структуру сигналу.

3. Якщо існує необхідність об'єднати дані з різних лікарень або публічних наборів даних, метод ComBat вважається класичним «стандартом», адже він використовує моделі Байєса, щоб оцінити, яка частина варіації в даних зумовлена біологією, а яка – батчем (джерелом). Потім він «віднімає» ефект джерела. В результаті – дуже точне вирівнювання навіть при малих обсягах даних у батчах.

4. Якщо це можливо, один пристрій або один набір даних обирається як «еталонний». Всі інші потоки даних масштабуються так, щоб їхні базові показники (наприклад, пульс у спокої) відповідали еталонним значенням. Таким чином відбувається регуляризація на основі референсного зразка.

5. Якщо платформа використовує Deep Learning, цей шар додається безпосередньо в модель, мережа автоматично перераховує середнє та дисперсію для кожного «батчу» під час навчання. В результаті модель стає стійкою до того, що дані з різних джерел мають різний масштаб.

Отже, для потокових даних із сенсорів найкраще підходить Z -score або адаптивна нормалізація. Для об'єднання з офлайн наборами даних – ComBat або квантильна нормалізація.

Тому структуру етапу первинної обробки даних для інтелектуальної платформи можна представити у вигляді таблиці 1. Вона охоплює основні методи усунення «ефекту батчу» та підготовки гетерогенних медичних даних разом із програмними бібліотеками їх реалізації.

Оскільки ми розібрали технічну сторону «ефекту батчу» та методи його усунення (від простого Z -score до потужного ComBat), інтелектуальна платформа тепер може мати фундамент для етапу первинної обробки даних.

Для завершення цілісної картини ETL-процесу, розглянемо етап заповнення пропущених значень, адже сенсорні мережі часто дають збої, а пропуски в даних можуть виникати в різних батчах по-різному, що теж створює системні перекоси.

Таблиця 1

Методи усунення «ефекту батчу» та засоби їх програмної реалізації

Назва методу	Як це працює	Результат	Програмні бібліотеки (Python / R)
Z-score нормалізація	обчислює відхилення кожного значення від середнього арифметичного батчу	дані всіх пристроїв зводяться до єдиного масштабу (середнє 0, дисперсія 1)	Scikit-learn (StandardScaler), Pandas
Квантильна нормалізація	робить розподіл значень у всіх наборах даних ідентичним за формою	усуваються нелінійні технічні розбіжності між різними типами сенсорів	Scikit-learn (QuantileTransformer), preprocessCore (R)
Алгоритм ComBat	використовує моделі Байєса для вилучення варіативності, зумовленої джерелом (батчем)	найточніше розділення медичних показників від «технічного шуму» джерела	neuroCombat (Python), sva (R/Bioconductor)
Денойзинг (Wavelet / FFT)	видаляє високочастотні шуми, характерні для електроніки конкретного сенсора	отримання «чистого» сигналу, незалежного від електричних завад девайса	SciPy (signal), PyWavelets, NumPy
Імпутація (KNN / MICE)	заповнює пропуски на основі кореляцій з іншими показниками того самого батчу	відновлення цілісності при різній частоті опитування пристроїв	Scikit-learn (IterativeImputer), Fancyimpute, mice (R)
Batch Normalization	шар у нейромережі, що автоматично перераховує параметри даних під час навчання	стабілізація навчання та стійкість моделі до даних із нових (невідомих) джерел	PyTorch (BatchNorm1d), TensorFlow/Keras

Для медичних сенсорних мереж етап «заповнення пропусків» заповнення пропущених значень є критичним, оскільки пропуски часто виникають не випадково (наприклад, через збій датчика або розрядку акумулятора в конкретній моделі пристрою), що також підсилює «ефект батчу».

Наведемо порівняльну таблицю методів імпутації для інтелектуальної платформи, що відображена таблицею 2.

Таблиця 2

Порівняльна характеристика методів імпутації для платформи, що розробляється

Назва методу	Як це працює	Результат	Інструменти (Python / R)
Mean/Median Imputation	заповнює пропуски середнім або медіанним значенням по всьому батчу	найшвидший метод, але спотворює варіативність і кореляції	Pandas (fillna), Scikit-learn
K-Nearest Neighbors (KNN)	Шукає k подібних записів (пацієнтів) і заповнює пропуск середнім їхніх значень	добре зберігає локальну структуру даних, підходить для невеликих датасетів	Scikit-learn (KNNImputer), VIM (R)
MICE (Multivariate Imputation)	створює регресійну модель для кожного показника, враховуючи зв'язки між усіма змінними	є стандартом для клінічних даних, зберігає складні взаємозв'язки між симптомами	Scikit-learn (IterativeImpute), mice (R)
LOCF / NOCB	заповнює пропуск останнім відомим значенням	ідеально для швидких часових рядів (ЕКГ, пульс), де стан не змінюється миттєво	Pandas (ffill, bfill)
MissForest (Random Forest)	використовує ансамблі дерев рішень для передбачення пропущених значень	висока точність на нелінійних даних, не потребує нормалізації перед роботою	Missingpy (Python), missForest (R)
GAIN (Generative Adversarial Imputation)	нейромережа (GAN) намагається «вгадати» пропуски, щоб обманути дискримінатор	найкращий вибір для Deep Learning платформ та дуже складних, зашумлених сигналів	PyTorch, TensorFlow

Як показує практика, для медичних сенсорів (часових рядів) необхідно комбінувати LOCF (для коротких пауз у сигналі) та MICE або MissForest (для заповнення втрачених блоків даних при порівнянні з публічними наборами даних).

Вибір алгоритму ComBat як основного методу нормалізації обґрунтований його здатністю ефективно розділяти технічну варіативність джерел (різні моделі сенсорів, умови збору в публічних наборах даних) та реальні фізіологічні показники пацієнта. На відміну від стандартної Z-score корекції, ComBat зберігає внутрішню біологічну структуру даних навіть при малих обсягах вибірок, що є критичним для медичної діагностики.

Для вирішення проблеми втрачених пакетів даних у сенсорних мережах обрано метод MICE (Multivariate Imputation by Chained Equations). Його перевага полягає в ітераційному моделюванні кожного пропущеного показника на основі кореляцій з іншими доступними параметрами (наприклад, відновлення рівня сатурації на основі частоти дихання та пульсу). Це дозволяє уникнути зміщення статистичних оцінок, яке властиве методам простого заповнення середнім значенням, та забезпечує цілісність даних для подальшого аналізу моделями машинного навчання.

На рис. 1 наведено діаграму потоків даних процесу первинної обробки. Особливістю архітектури є виділення окремого модуля ComBat-корекції в межах етапу первинної обробки даних, що дозволяє динамічно

усувати технічну варіативність (батч-ефект) між даними сенсорних мереж та клінічними датасетами до моменту їх подачі на вхід класифікатора, що, в свою чергу, дозволяє динамічно адаптувати платформу до нових типів сенсорів без перенавчання основної моделі.

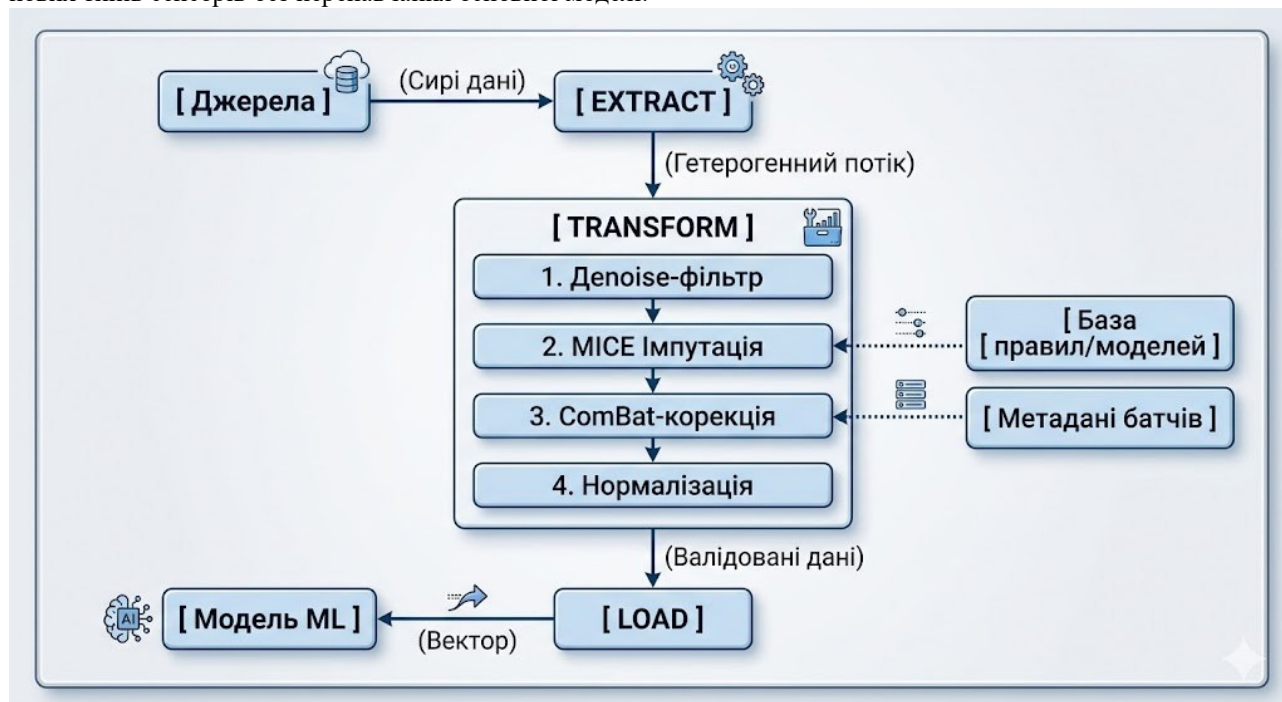


Рис. 1. Схема потоків даних (DFD) процесу первинної обробки та мінімізації «ефекту батчу» в інтелектуальній платформі

Новизна запропонованої моделі полягає не у складі шарів, а у властивості, що з нього випливає і піддається вимірюванню, – апаратній інваріантності конвеєра, тобто можливості підключення нового джерела даних без перенавчання предиктивної моделі.

Нехай $B = \{b_1, \dots, b_n\}$ – множина джерел даних (типів сенсорів або датасетів), на яких навчено предиктивний компонент; $Q(\cdot)$ – метрика якості цього компонента (F_1 -міра для задачі класифікації критичних станів); T – модуль корекції джерела, що реалізує перетворення (1)–(3) з тим самим складом матриці клінічних коваріат X . Платформа задовольняє властивість апаратної інваріантності з допуском ε , якщо для будь-якого нового джерела $b_{n+1} \notin B$, обробленого модулем T , виконується умова:

$$|Q(T(b_{n+1})) - Q(b_1, \dots, b_n)| \leq \varepsilon \quad (4)$$

причому предиктивний компонент не піддається донавчанню або перенавчанню. Значення допуску ε задається як нефункціональна вимога до платформи; у цій роботі прийнято $\varepsilon = 0,03$ за метрикою F_1 .

Отже, платформа задовольняє властивість апаратної інваріантності з допуском ε , якщо для будь-якого нового джерела виконується $|Q(T(b_{n+1})) - Q(b_1, \dots, b_n)| \leq \varepsilon$, причому предиктивний компонент не піддається донавчанню. У роботі прийнято $\varepsilon = 0,03$ за метрикою F_1 .

ВИСНОВКИ З ДАНОГО ДОСЛІДЖЕННЯ І ПЕРСПЕКТИВИ ПОДАЛЬШИХ РОЗВІДОК У ДАНОМУ НАПРЯМІ

Запропоновано архітектурне рішення для етапу первинної обробки даних ETL-конвеєра інтелектуальної платформи опрацювання біомедичних показників, що полягає у виділенні байєсівської корекції «ефекту батчу» (алгоритм ComBat) в окремий мікросервіс із конкретизованим для класу IoMT-даних складом матриці клінічних коваріат (вік, стать, ІМТ, хронічні діагнози). Показано, що врахування коваріат дозволяє реалізувати принцип персоніфікованої нормалізації – розрізняти технічну варіативність джерел і реальну фізіологічну варіативність пацієнта, що критично для збереження діагностичної чутливості системи при об'єднанні даних із різнорідних сенсорів та публічних наборів даних. Проведений порівняльний аналіз методів усунення «ефекту батчу» та імпутації пропущених значень обґрунтовує вибір ComBat як основного методу нормалізації для малих і середніх вибірок та MICE – як методу відновлення пропусків, що зберігає міжпараметричні кореляції. Запропонована схема потоків даних підтверджує технічну реалізованість підходу та його придатність до динамічної адаптації платформи під нові типи сенсорів без перенавчання основної моделі прогнозування.

References

- 1 Johnson A. E. W., Pollard T. J., Shen L. et al. MIMIC-III, a freely accessible critical care database. *Scientific Data*. 2016. Vol. 3. 160035. DOI: <https://doi.org/10.1038/sdata.2016.35>.
- 2 Goldberger A. L., Amaral L. A. N., Glass L. et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*. 2000. Vol. 101, № 23. P. 215–220. DOI: <https://doi.org/10.1161/01.CIR.101.23.e215>.
- 3 Мацієвський, В. А., Нікітчук, Т. М., Богоявленська, Ю. В. (2026). Перспективи та виклики інженерії програмного забезпечення для інтелектуального опрацювання біомедичних показників. *Технічна інженерія*. 2026. №1 (97). С. 241–248. [https://doi.org/10.26642/ten-2026-1\(97\)-241-248](https://doi.org/10.26642/ten-2026-1(97)-241-248)
4. ISO/IEC 27001:2022. Information security, cybersecurity and privacy protection – Information security management systems – Requirements. ISO, 2022.
5. ISO/IEC 27799:2016. Health informatics – Information security management in health using ISO/IEC 27002. ISO, 2016.
6. IEC 62304:2006+AMD1:2015. Medical device software – Software life cycle processes. IEC, 2015.
7. ISO 13485:2016. Medical devices – Quality management systems – Requirements for regulatory purposes. ISO, 2016.
8. Zhang Y., Parmigiani G., Johnson W. E. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics and Bioinformatics*. 2020. Vol. 2, № 3. lqaa078. DOI: <https://doi.org/10.1093/nargab/lqaa078>.
9. Johnson W. E., Li C., Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007. Vol. 8, № 1. P. 118–127. DOI: <https://doi.org/10.1093/biostatistics/kxj037>