Дата публикации: 01.09.2022
DOI: 10.51871/2588-0500_2022_06_03_14
УДК 004.85; 612.01

Publication date: 01.09.2022
DOI: 10.51871/2588-0500_2022_06_03_14
UDC 004.85; 612.01

ПРИМЕНЕНИЕ МАШИННОГО ОБУЧЕНИЯ И ОМИКСНЫХ ТЕХНОЛОГИЙ ДЛЯ ОЦЕНКИ ФУНКЦИОНАЛЬНОГО СОСТОЯНИЯ ЧЕЛОВЕКА

Н.Ф. Максютов, А.А. Муртазин, Е.И. Балакин, В.И. Пустовойт
Федеральное Государственное бюджетное учреждение «Государственный научный Центр Российской Федерации – Федеральный медицинский биофизический центр имени А.И. Бурназяна, г. Москва, Россия

Аннотация. До недавнего времени отсутствовала возможность получения полной информации о состоянии сложных биологических систем, к которым относится функциональное состояние человека: как правило, для анализа применялись только данные одного вида, либо ограниченное их количество. Благодаря развитию омиксных технологий, алгоритмов машинного обучения для анализа данных, а также увеличению производительности компьютеров стало возможным применение интеграции омиксных данных с данными другого вида и использование комплексного подхода для анализа функционального состояния человека. Однако эти процессы сопряжены с рядом трудностей. В связи с этим был проведен обзор актуальных на данный момент способов интеграции омиксных данных с неомиксными, а также алгоритмов машинного обучения, применяющихся для анализа омиксных данных. Особое внимание уделено глубокому обучению, в основе которого лежат нейронные сети, как наиболее эффективному подходу для решения этих задач. Для введения в эту область даны определения основных терминов, использующихся в анализе данных и машинном обучении. Также рассмотрены перспективы данной области, актуальные проблемы и способы их решения.
Ключевые слова: омиксные технологии, интеграция данных, машинное обучение в омиксах, глубокое обучение.

USING MACHINE LEARNING APPROACHES AND OMICS TECHNOLOGIES FOR ASSESSMENT OF HUMAN FUNCTIONAL STATE

N.F. Maksjutov, A.A. Murtazin, E.I. Balakin, V.I. Pustovoit
State Research Center – Burnasyan Federal Medical Biophysical Center of Federal Medical Biological Agency, Moscow, Russia

Annotation. It was not possible to get full information about states of complex biological systems, including human functional state, for a long time, because data of one type or limited number of data types was used for analysis. Development of omics technologies, machine learning algorithms, and increased computer performance allowed to apply integration of omics data with other data types and using a complex approach for analyzing the human functional state. However, these procedures involve a few of challenges. This aspect was the reason for making this review of state-of-the-art methods of integration of omics with non-omics data and machine learning algorithms for analysis of omics data. Special attention was paid to deep learning as the most effective group of algorithms for completing these tasks. The introduction to this field includes definitions of the basic terms used in data analysis and machine learning. Future prospects, actual problems of the field and ways to solve them are also discussed.
Keywords: omics, data integration, machine learning in omics, deep learning.

Введение. *Омиксные технологии*. Омиксы – группа дисциплин в биологии, которая изучает различные биологические пулы молекул, их состав и взаимодействие между собой и другими компонентами клетки. В зависимости от объекта исследования различают следующие виды омиксов: геномика (ДНК), транскриптомика (РНК),

протеомика (белки) и метаболомика (метаболиты). Основные виды омиксов представлены на рисунке 1. Завершение проекта по секвенированию генома человека в 2003 году [1] стало важным этапом, после которого началось интенсивное развитие омиксных технологии. Их роль крайне важна для получения полной информации о состоянии клетки или организма, так как биологические процессы регулируются не только на уровне ДНК, но и на уровне других биологических молекул [2]. Развитие омиксных технологий и удешевление процессов получения омиксных данных позволило применять системные подходы для анализа биологических систем [2].

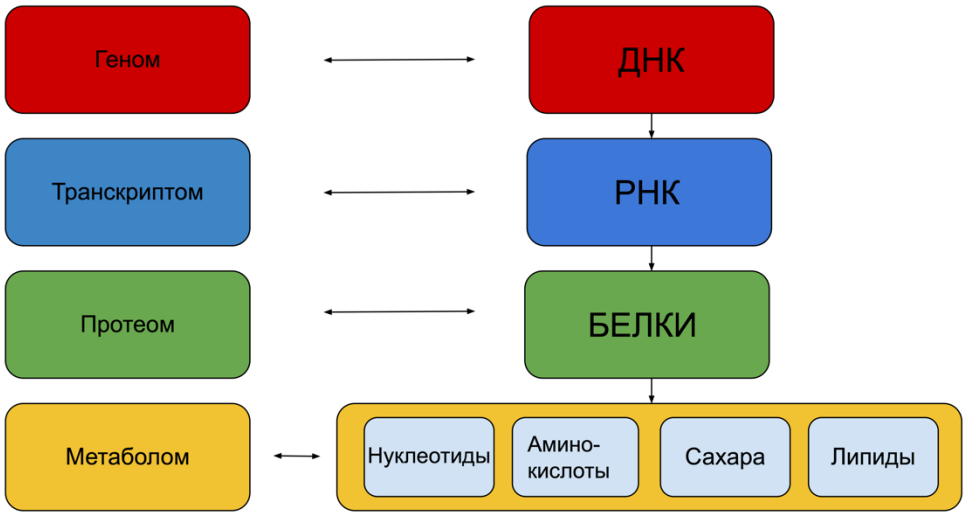


Рис. 1. Основные виды омиксов и биологических молекул, входящих в их состав.

Геномика. Геномика представляет собой междисциплинарную науку о картировании генов, а также структуре и функции генома, включающего всю последовательность ДНК, как кодирующие регионы, так и некодирующие, которые необходимы для развития и функционирования организма. Впервые этот термин был введен Томом Родериком в 1986 году при работе с генами человека [3]. Формирование этой науки произошло после завершения международного научного проекта по секвенированию генома человека в 2003 году [1]. Наиболее активно развивающиеся области геномики: функциональная геномика, метагеномика и эпигеномика [4]. Более 30 лет единственным способом получения информации о последовательности ДНК являлось секвенирование по Сенгеру, но в дальнейшем стали широко использоваться методы высокопроизводительного секвенирования: пиросеквенирование, секвенирование путем синтеза, секвенирование с помощью лигирования и др. [5].

Транскриптомика. Транскриптомика – наука о совокупности всех молекул РНК (транскриптом) в организме или клетке, а также изменении их состава при различных

условиях. Транскриптом включает в себя кодирующие и некодирующие участки РНК. К первой группе относятся матричные РНК, а ко второй – рибосомальные РНК, транспортные РНК, длинные некодирующие РНК, малые интерферирующие РНК и др. [6]. Транскриптом, в отличии от генома, который на протяжении всей жизни организма остается неизменным в соматических клетках, постоянно меняется в зависимости от внешних и внутренних факторов, поэтому изучение транскриптома позволяет получить информацию о состоянии клетки или организма. Также он позволяет получить не только информацию о геноме на уровне транскрипции, но и понимание о структуре и функции гена и регуляции его экспрессии. К современным методам секвенирования РНК относят ONT nanopore sequencing, Geo-seq, Slide seq [7].

Протеомика. Протеомика включает в себя анализ всех белков в клетке, органе или организме. Протеоформа – модификация канонической аминокислотной последовательности белка, причем разные протеоформы одного и того же белка могут выполнять различные функции [8]. Изучение протеома является более сложной задачей, чем изучение генома, так как геном человека насчитывает около 20 000 белок-кодирующих генов, а количество протеоформ по разным оценкам составляет от нескольких сот тысяч до миллионов [9]. Такое значительное количество обусловлено как постранскрипционными процессами, так и посттрансляционными, включающими в себя гидроксилирование, убиквитинилирование, метилирование и др. [8]. Состав протеома непостоянен и зависит от многих факторов, поэтому он является одним из основных индикаторов состояния клетки или организма, однако ввиду своей сложности его анализ сопряжен с большими трудностями. Основным методом исследования протеома является масс-спектроскопия, который основан на определении отношения массы к заряду ионов [2].

Метаболомика. Метаболом включает в себя молекулы с различными физико-химическими свойствами: нуклеотиды, аминокислоты, сахара и липиды, что вызывает большие трудности при его анализе [10]. Поэтому основной стратегией анализа метаболитов заключается в разбиении их по группам и последующим анализе каждой из них. Принадлежность к той или иной группе определяется полярностью молекул, общими функциональными группами, а также структурным сходством [11]. Состав метаболома обладает высокой динамичностью и поэтому, так же, как и протеом, наиболее точно характеризует текущее состояние клетки или организма. Существуют два подхода анализа метаболома – нетаргетный и таргетный [2]. В первом случае происходит анализ максимального возможного числа метаболитов, а во втором – только выбранных. Таким образом, нетаргетный подход позволяет сформулировать гипотезу: например, определенные метаболиты ответственны за тот или иной фенотип, а таргетный подход позволяет подтвердить или опровергнуть эту гипотезу. Для анализа метаболома используются спектроскопия ядерного магнитного резонанса, газовая хромато-масс-спектрометрия, жидкостная хроматография с тандемной масс-спектрометрией и другие методы [12].

Искусственный интеллект. Искусственный интеллект (ИИ) – это способность системы корректно интерпретировать данные, обучаться на этих данных и за счет этого обладать способностью выполнять различные задачи. Развитие этой науки началось в 50-х годах прошлого века. Также под термином "искусственный интеллект" имеют в виду способность компьютеров симулировать интеллект человека и решать задачи, которые невозможно решить классическими алгоритмами. Например, к таким задачам относятся предсказание будущих результатов на основе имеющихся данных, распознавание речи, естественная обработка языка и компьютерное зрение.

В ИИ встречаются такие термины, как «машинное обучение (МО)», «нейронные сети», «глубокое обучение». Из-за новизны этой области науки часто встречаются путанности в этих терминах. Определения этих терминов и их иерархия представлены на рисунке 2. В ИИ различают множество алгоритмов и МО – одна из таких групп, в основе которых лежат статистические методы. Эти алгоритмы получили стремительное развитие в последние годы и нашли практическое применение в самых различных областях, благодаря увеличению производительности компьютеров и накоплению большого количества данных.

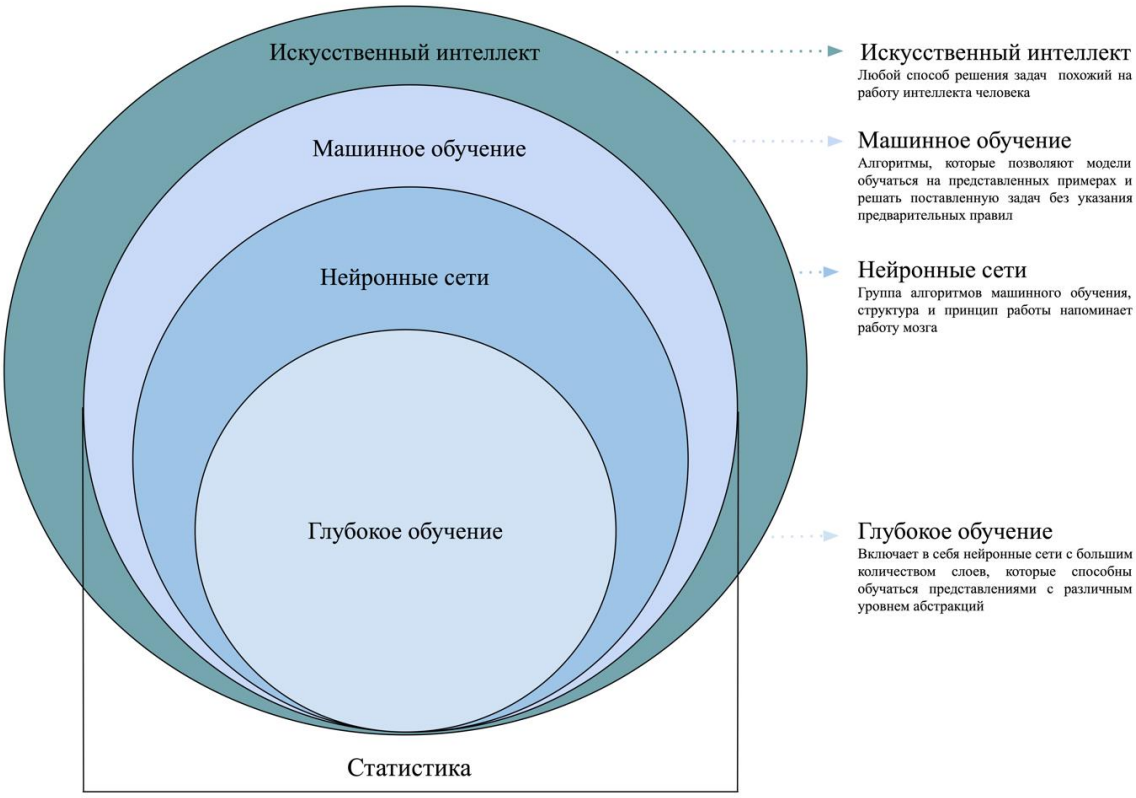


Рис. 2. Основные понятия ИИ

Этапы развития ИИ. Различают три этапа развития (или мощности ИИ): слабый ИИ (weak AI), сильный ИИ (strong AI) и супер-ИИ. Их основные параметры представлены в таблице 1. Слабый ИИ – единственный вид ИИ, который существует на данный момент. Его способности ограничиваются только решениями заранее определенного вида задач, однако даже на таком уровне модель способна показывать результаты лучше, чем человек. Примерами таких моделей является голосовые ассистенты, автопилоты и чат-боты. Для своей работы эти системы используют данные для обучения и не способны выполнять другие задачи [13].

Машинное обучение. Машинное обучение – это область ИИ, в которой используются статистические алгоритмы для создания систем, которые обладают способностью самостоятельно обучаться на данных и улучшать результаты, это отличает машинное обучение от классических алгоритмов в информатике. Главное отличие классических алгоритмов от алгоритмов МО заключается в том, что в первом случае вводятся правила, составленные человеком, которым должна следовать программа и

решить поставленную задачу, а программа, в основе которой лежат алгоритмы МО, должна найти способ самостоятельно находить эти правила (рис. 3). Также часто МО называют статистическим обучением.

Классификация алгоритмов машинного обучения. Существует большое количество способов классификации алгоритмов МО, один из них представлен на рисунке 4.

Классическое обучение. Алгоритмы классического обучения разделяются на две группы: с учителем (supervised learning) и без учителя (unsupervised learning).

Таблица 1

Уровни искусственного интеллекта

	Слабый ИИ	Сильный ИИ	Супер-ИИ
Области применения	Только определенные	Включает в себя несколько сфер	Любая область применения
Автономность	Не способен автономно решать проблемы в других областях	Способен автономно решать задачи в других областях	Способен решать проблемы в разных областях одновременно
Сравнение с человеческим интеллектом	Превосходит, либо равен человеческому только при решении определенных задач	Превосходит, либо равен человеческому в нескольких областях	Превосходит человеческий интеллект в любой сфере

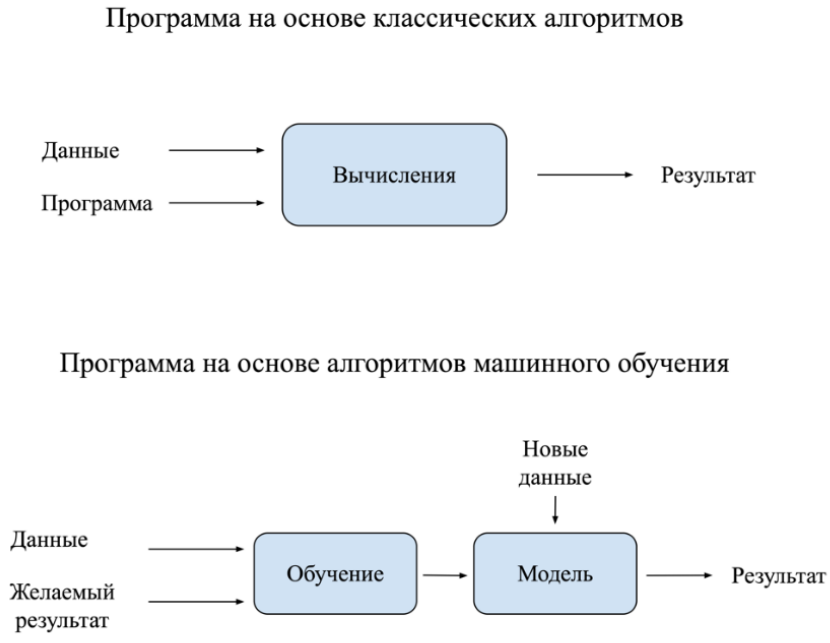


Рис. 3. Отличие алгоритмов МО от классических алгоритмов

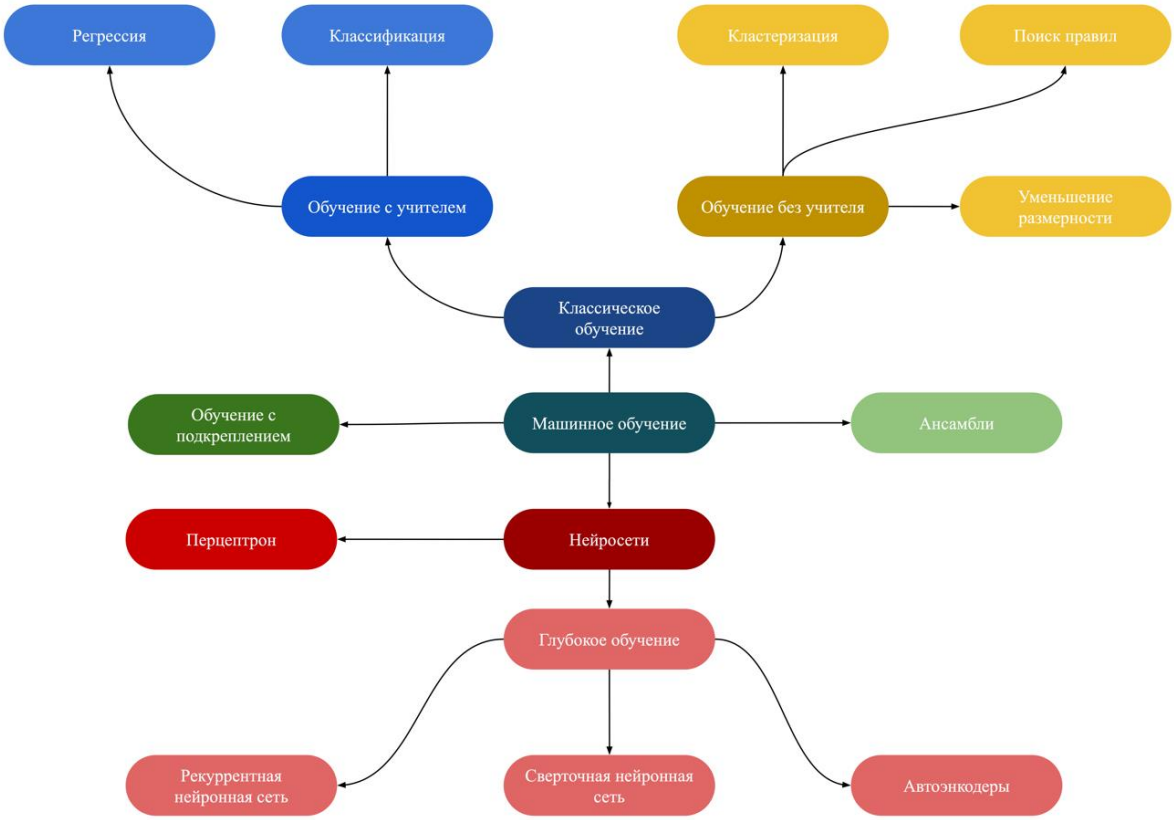


Рис. 4. Основные алгоритмы машинного обучения

Таблица 2

Сравнение алгоритмов обучения с учителем и без учителя

	Обучение с учителем	Обучение без учителя
Определение	Вид машинного обучения, работа которого осуществляется с помощью учителя, который предоставляет заранее размеченные данные	Вид машинного обучения, работа которого осуществляется без участия человека, программа самостоятельно пытается находить паттерны в неразмеченных данных
Входные данные	Размеченные	Неразмеченные
Условия применения	Заранее известно, что должно быть результатом работы программы	Заранее результат неизвестен
Цели	Если необходимо предсказать числовое значение, то используются регрессия, а если переменная категориальная – классификационные методы	Кластеризация и поиск ассоциаций

Продолжение таблицы 2

Примеры алгоритмов	• Наивный Байес • Метод опорных векторов • Линейная регрессия • Логистическая регрессия • Деревья решений	• Метод k-средних • Метод главных компонент • Алгоритм Apriori
Области применения	• Спам-фильтры • Прогнозирование цены или спроса • Распознавание изображений	• Рекомендательные системы • Поиск аномалий • Препроцессинг данных для дальнейшего анализа с помощью обучения с учителем

Обучение с учителем. В этой группе алгоритмов есть некий учитель, заранее разметивший все данные, на которых будет обучаться модель. Задачи, где используются алгоритмы классического обучения, делятся на два типа: предсказание категории объекта или предсказание значения параметра. Соответственно, в первом случае могут использоваться линейная или полиномиальная регрессия, а во втором – наивный Байес, деревья решений, логистическая регрессия, машины опорных векторов и др. [14]. К прикладным задачам классификации относятся спам-фильтры, определение языка, поиск похожих документов, анализ тональности текста, распознавание рукописных букв и цифр, определение подозрительных транзакций в банковской сфере.

Обучение без учителя. Алгоритмы обучения без учителя были изобретены значительно позже – в 90-е годы прошлого века. Они используются для кластеризации, уменьшения размерности и поиска ассоциативных правил.

Кластеризация. Кластеризация – это классификация без заранее известных классов. При этом алгоритм сам ищет похожие объекты, а затем объединяет их в кластеры. Количество кластеров можно задать заранее или позволить модели определить их самостоятельно. Основные алгоритмы кластеризации: метод K-средних, Mean-Shift, DBSCAN. Применяется в сжатии изображений, объединении точек на карте и разметке новых данных.

Уменьшение размерности. Алгоритм этой группы собирает признаки в абстракции. Используется в рекомендательных системах, визуализации, определении тематики, поиске похожих объектов, риск-менеджменте, при оверфиттинге модели, мультиколлинеарности, факторном анализе, сжатии изображений, уменьшении времени обучения модели. Алгоритмы: Метод главных компонент (PCA), Сингулярное разложение (SVD), Латентное размещение Дирихле (LDA), Латентно-семантический анализ (LSA, pLSA, GLSA), t-SNE.

Обучение на ассоциативных правилах. Используется для поиска взаимосвязей между переменными в датасете. Применяется для анализа товаров, покупаемых вместе, в медицине для установления диагноза и секвенировании белков и интеграции омиксных данных [15]. Популярные алгоритмы: Apriori, Euclat, FP-growth.

Ансамблевые методы. Ансамблевые методы представляют собой представляет собой подход, при котором несколько моделей обучаются на одних и тех же данных, а затем происходит их объединение, в результате итоговая точность модели превосходит точность каждой отдельной модели. Популярные алгоритмы: Random Forest, Gradient Boosting. Существуют три способа составления ансаблей: стекинг, беггинг, бустинг:

• Стеккинг. Происходит обучение различных алгоритмов, затем передача результатов их работы на вход последнему, который принимает финальное решение.

• Беггинг. В этом случае осуществляется обучение одного и того же алгоритма, но на разных выборках из исходного датасета. Затем происходит усреднение этих результатов. Популярные алгоритмы: Random Forest.

• Бустинг. Бустинг – принцип последовательного построения алгоритмов, при котором каждый следующий этап обучения происходит на тех данных, которые вызвали проблемы на предыдущем этапе обучения. Таким образом, комбинация нескольких алгоритмов позволяет значительно увеличить точность. Популярные методы бустинга: XGBoost, CatBoost и LightGBM.

Обучение с подкреплением. Задача – минимизировать ошибки, а не рассчитать все ходы. Применение в автопилотах, торговле на биржах. В отличие от других алгоритмов, основной задачей машинного обучения является не анализ данных, а существование интеллектуального агента в среде с минимализацией ошибок. Осуществляются повторяемые взаимодействия с окружающей средой, а ситуации запоминаются как простой Марковский процесс. Некоторые области применения: автопилоты, торговли на биржах. Популярные алгоритмы: генетический алгоритм, Q-Learning, SARSA, DQN, A3C. [16].

Искусственные нейронные сети (ИНС) и глубокое обучение. Искусственные нейронные сети – группа алгоритмов машинного обучения, чья структура и принцип работы напоминают работу мозга. ИНС состоит из нейронов или узлов и связей между ними, которые являются аналогами аксонов или дендритов у биологических нейронов и способны иметь различный вес. Нейрон ИНС можно представить в виде функции, на вход которой подаются n числа от w1 до wn, и которая производит некую операцию над ними, а затем выдает на выход результат со значением от 0 до 1 (рис 5) [17].

В качестве примера можно представить нейрон, на вход которому подаются числа 2, 3 и 4, затем он считает их среднее значение Z = (4+3+2)/3=3, и если это среднее больше определенного числа (например 2), то нейрон на выход подает единицу, если меньше – то ноль (рис. 6).

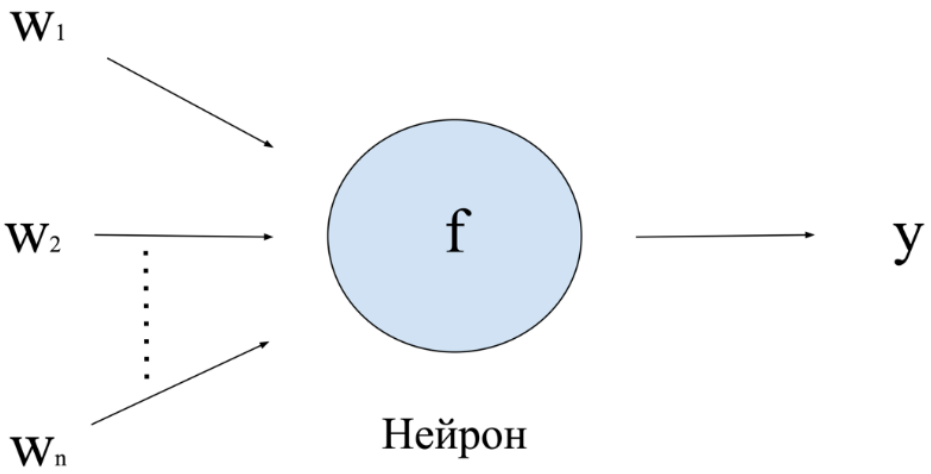


Рис. 5. Схема искусственного нейрона

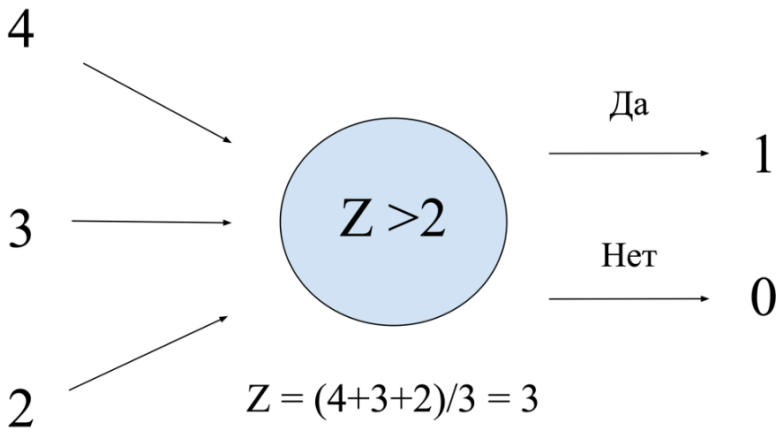


Рис 6. Принцип работы искусственного нейрона

С точки зрения математического представления все выходные значения из одного слоя являются вектором, а веса связей между нейронами – матрицами, у которых каждая колонка – это совокупность связей одного слоя к определенному нейрону следующего слоя. При обучении нейросети, которое может быть как с учителем, так и без него, подбираются значения весов связей таким образом, чтобы обеспечить наиболее точный результат. Наиболее широко используемым методом обучения нейронной сети является метод обратного распространения ошибки [18]. На текущий момент, благодаря достаточной производительности современных компьютеров, нейросети способны заменить многие алгоритмы машинного обучения и обладают большей точностью.

Топология нейронных сетей. В ИНС различают три вида слоя: входящий, скрытый и выходящий. Создание нейронной сети включает в себя определение количества слоев и нейронов в каждом из них, а также конфигурацию связей между нейронами. Определение значений этих параметров происходит, как правило, в процессе обучения нейросети.

Взаимосвязь внутри нейронной сети может быть построена простым способом, когда нейроны одного слоя могут связаться только с нейронами другого слоя, либо нейроны одно слоя могут связываться с друг другом и нейронами других слоев. Также нейросети могут быть прямого распространения, либо обратного распространения. Некоторые виды ИНС представлены на рисунке 7.

Глубокое обучение. Не существует строгого разграничения между обычными нейронными сетями и нейронными сетями глубокого обучения. Однако в последнем случае подразумевают ИНС с большим количеством скрытых слоев, которые также обладают способностью к обучению представлениями, т.е. нейросеть сама способна определять признаки, необходимые для ее работы [19]. К преимуществам глубокого обучения относят высокую точность по сравнению с другими алгоритмами машинного обучения, но для этого необходимы очень большие объемы данных и высокая вычислительная мощность компьютера (рис. 8). Глубокие нейронные сети нашли применение в таких сложных задачах, как распознавание объектов на изображении, машинный перевод и голосовые помощники.

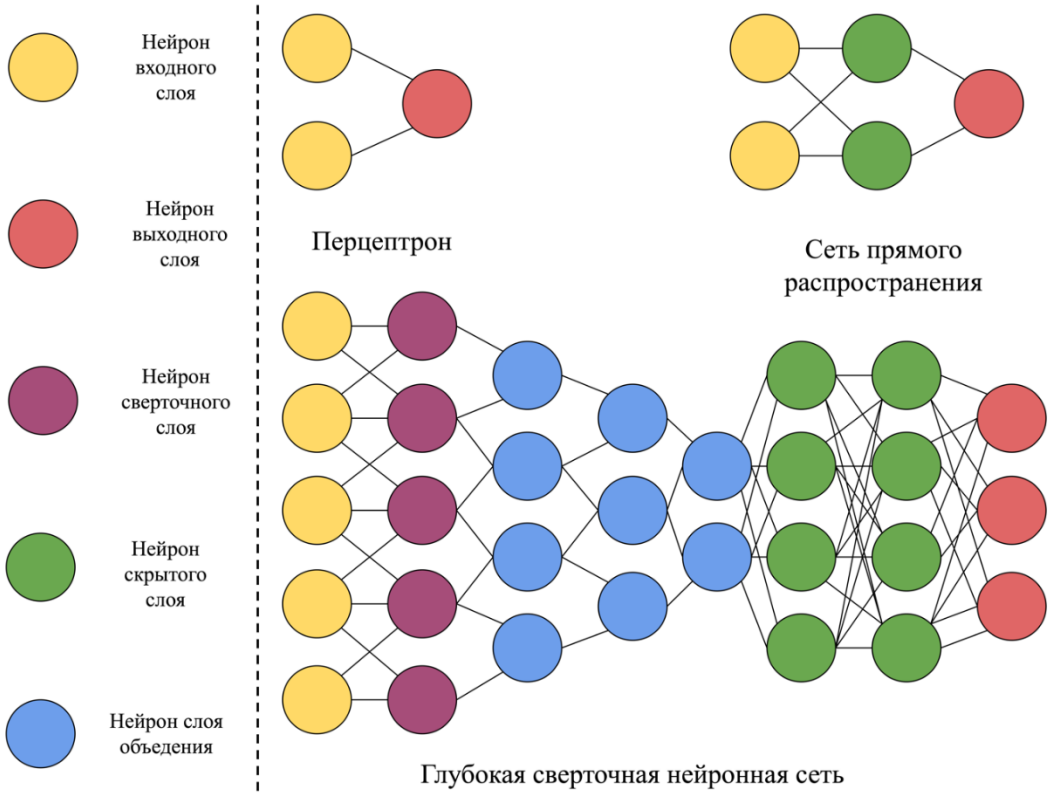


Рис. 7. Архитектура нейронных сетей

Примечание: представлены простейшая нейронная сеть – перцептрон, сеть прямого распространения и глубокая сверточная нейронная сеть

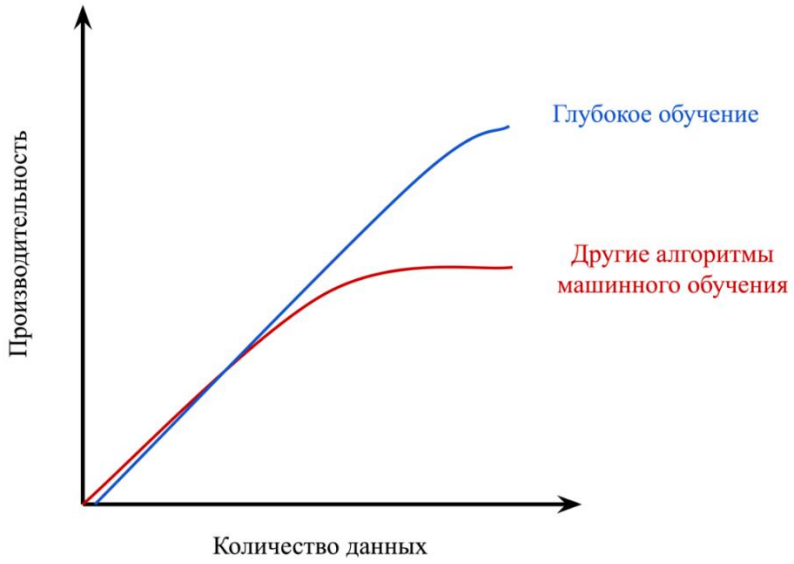


Рис. 8. Производительность алгоритмов глубокого обучения по сравнению с другими алгоритмами машинного обучения

Проблемы анализа мультиомиксных данных с помощью машинного обучения. Гетерогенность, недостаточность данных и выбросы. Омиксные данные, как правило, очень гетерогенные [20]. Например, транскриптомика и протеомика используют разные способы нормализации, это приводит к отличающимся друг от друга динамическими диапазонам и распределениям [21]. Кроме того, некоторые омиксы имеют недостаточное количество данных. Это особенно актуально для метаболомики: некоторые метаболиты присутствуют в количестве, недостаточном для их обнаружения [22]. Таким образом, вменение и обнаружение выбросов должны осуществляться для каждого омикса по отдельности перед проведением дальнейшего анализа [23-24].

Дисбаланс классов и переобучение. В задачах классификации с применением мультиомиксных технологий некоторые классы встречаются значительно реже, чем другие, это приводит к дисбалансу классов в датасете [25]. Модель, обученная на таких данных, может быть переобученной, то есть обладать высокой точностью на тренировочном датасете, но низкой – на анализируемых данных. Существуют несколько способов решения этой проблемы:

1) сбор наибольшего возможного количества;

2) использование нормализованных метрик для оценки модели, например, F1-Score или Kappa;

3) применение oversampling или undersampling;

4) использование искусственной генерации выборки, например, SMOTE [26] или ADASYN для недопредставленного класса [27].

Также для балансировки компромисса «отклонение-дисперсия» могут быть использованы регуризация, бэггинг, оптимизация и кросс-валидация [28].

Количество признаков больше, чем данных. Одним из главных недостатков омиксных датасетов является "проклятие размерности", то есть количество семплов меньше, чем количество омиксных признаков [29]. Итоговое мультиразмерное пространство часто содержит коррелирующие признаки, которые могут отрицательно сказаться на обучении модели [30]. Размерность пространства можно уменьшить с помощью выделения признаков. Среди часто используемых способов уменьшения размерности выделяют метод главных компонент [31], линейный дискриминантный анализ [32], многомерное шкалирование [33]. Также выделение признаков позволяет найти те признаки, которые повышают точность модели.

Методы и организация исследования. Проведен обзор научных публикаций в электронных базах данных PubMed, Scopus и ScienceDirect за 2016-2022 гг., используя различные комбинации ключевых слов: Multi-omics, Integration, Machine Learning, Deep Learning, Integrative Analysis, Data Integration, Sportomics, Non-Omics Data.

Результаты исследования и их обсуждение. *Методы интеграции омиксных данных.* Анализ биологических систем является трудной задачей ввиду их высокой сложности, и лишь в последние годы благодаря развитию статистических методов, омиксных технологий и увеличению производительности компьютеров появилась возможность применения интегративного подхода к анализу группы омиксов различного вида. Разделяют следующие подходы интегративного анализа: метод слияния, метод модели и метод трансформации, каждый из них которых обладает своими преимуществами и недостатками (табл. 3) [21, 34, 35, 36] (рис. 9). Другие авторы также предлагают интеграцию, ориентированную вверх и ориентированную вниз. [37-38].

Метод слияния. При этом способе интеграции происходит объединение матриц с омиксными данными в одну общую матрицу до момента создания модели. После этого этапа выбор статистического метода анализа обычно не вызывает проблем. К основным недостаткам можно отнести то, что этот метод не учитывает различные распределения у разных омиксов,

также происходит увеличение размерности данных, что осложняет дальнейший анализ. Для решения этих проблем можно воспользоваться нормализацией признаков и их отбором. Также при большом количестве признаков может быть использован их отбор с помощью таких алгоритмов, как метод главных компонент, линейный дискриминантный анализ и многомерное шкалирование [39].

Метод модели. При использовании этого подхода для каждой матрицы с данными происходит создание промежуточных моделей, которые затем интегрируются в итоговую модель. Основным преимуществом является то, что эти методы могут быть использованы для слияния моделей, основанных на разных видах омиксов [27, 36].

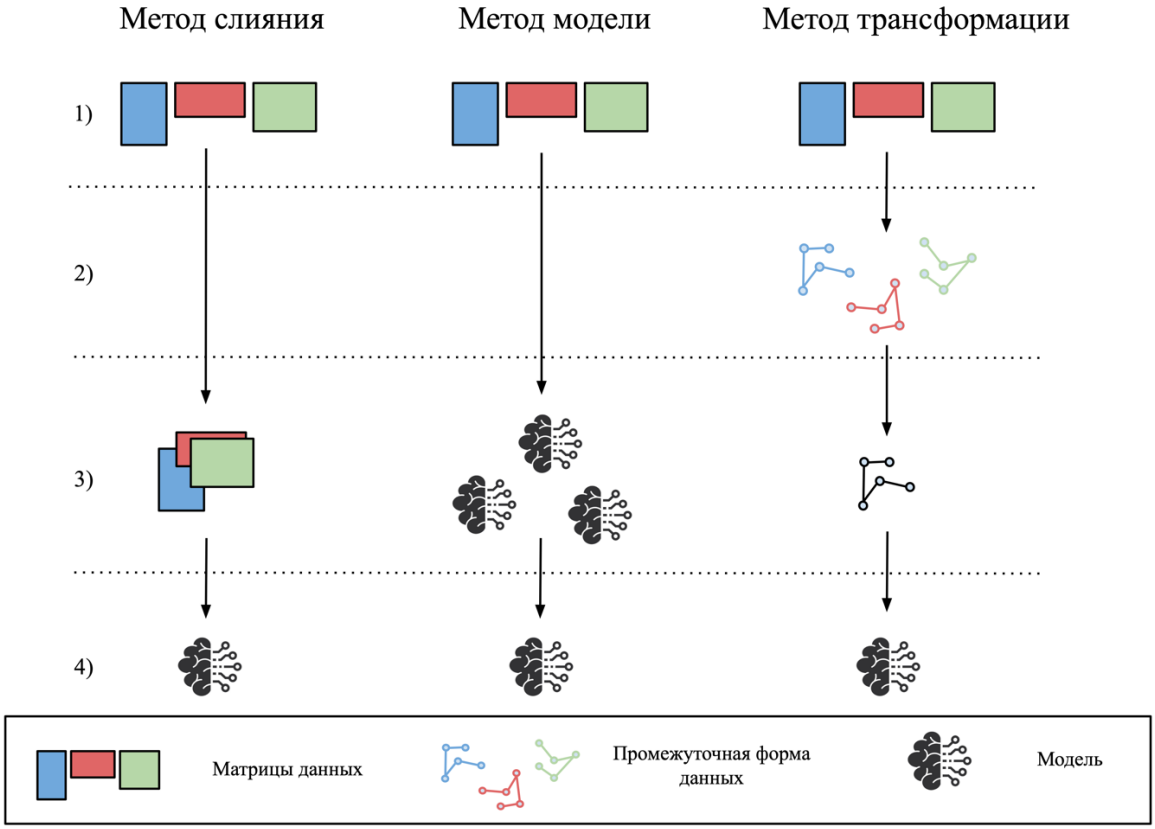


Рис. 9. Основные методы интеграции омиксных данных

Таблица 3

Алгоритмы машинного обучения для каждого из подходов интеграции омиксных данных

Интегративный подход	Примеры алгоритмов		Преимущества	Недостатки
	С учителем	Без учителя		
Основанные на слиянии	• Дерево решений • Наивный Байес • Нейронные сети • Метод опорных векторов • Метод k-ближайших соседей • Случайный лес	Множество методов классификации и анализа ассоциаций (например, основанные на неотрицательном матричном разложении)	• Легкость в реализации • Возможность использования классических методов обучения	• Для реализации желательно иметь в наличии все омиксные данные • Необходимость нормализации перед объединением

Продолжение таблицы 3

				• Не учитывается уникальное распределение для каждого омикса • Крайне ресурсозатратны
Основанные на модели	• Алгоритм большинства голосов Бойера-Мура • Иерархические классификаторы • Ансамблевые методы • Метод k-ближайших соседей	• Непараметрические Байевсовские модели • Анализ формальных понятий • Интеграция множественных датасетов • Консенсусная кластеризация	• Позволяет получить представление о взаимодействиих между омиксами • Омиксы могут быть из разных образцов при условии одинаковых фенотипов • Не происходит увеличение размерности данных	• Неэффективны при гетерогенности даннных • Высокая вероятность оверфиттинга • Слабые сигналы могут быть утеряны
Основанные на трасформации	• Полуопределенное программирования опорных векторов • Метод релевантных векторов	• Методы, основанные на графах и нулевых пространств матриц • Мета-аналитический метод опорных векторов	• Представление в виде графов наглядно и не так требовательно к наличию больших вычислительных ресурсов • Kernel methods обеспечивают самую высокую производительность	В некоторых случаях трансформация может вызывать трудности

Метод трансформации. При этом подходе каждый датасет сначала трансформируется в графы или матрицы нулевого пространства, затем происходит их объединение перед созданием модели. Основным достоинством данных методов является возможность комбинирования большого количества омиксов разных видов при наличии уникального идентификатора каждого из них. Также этот подход позволяет

сохранить информацию о свойствах каждого из типа данных [36].

Интеграция омиксных данных с другими видами данных. В работе Evangelina López de Maturana было предложено три подхода к интеграции омиксных и неомиксных данных: независимое моделирование, моделирование с условием и совместное моделирование (рис. 10) [40]. В исследованиях, в которых была произведена интеграция данных, была показана большая точность, чем при использовании отдельно только омиксных или неомиксных данных. Однако в некоторых случаях это увеличение точности было незначительным [41-42]. Оно зависело от таких факторов, как выбор стратегии объединения данных, видов омиксных и неомиксных данных и результата (например, количество фенотипов) [43].

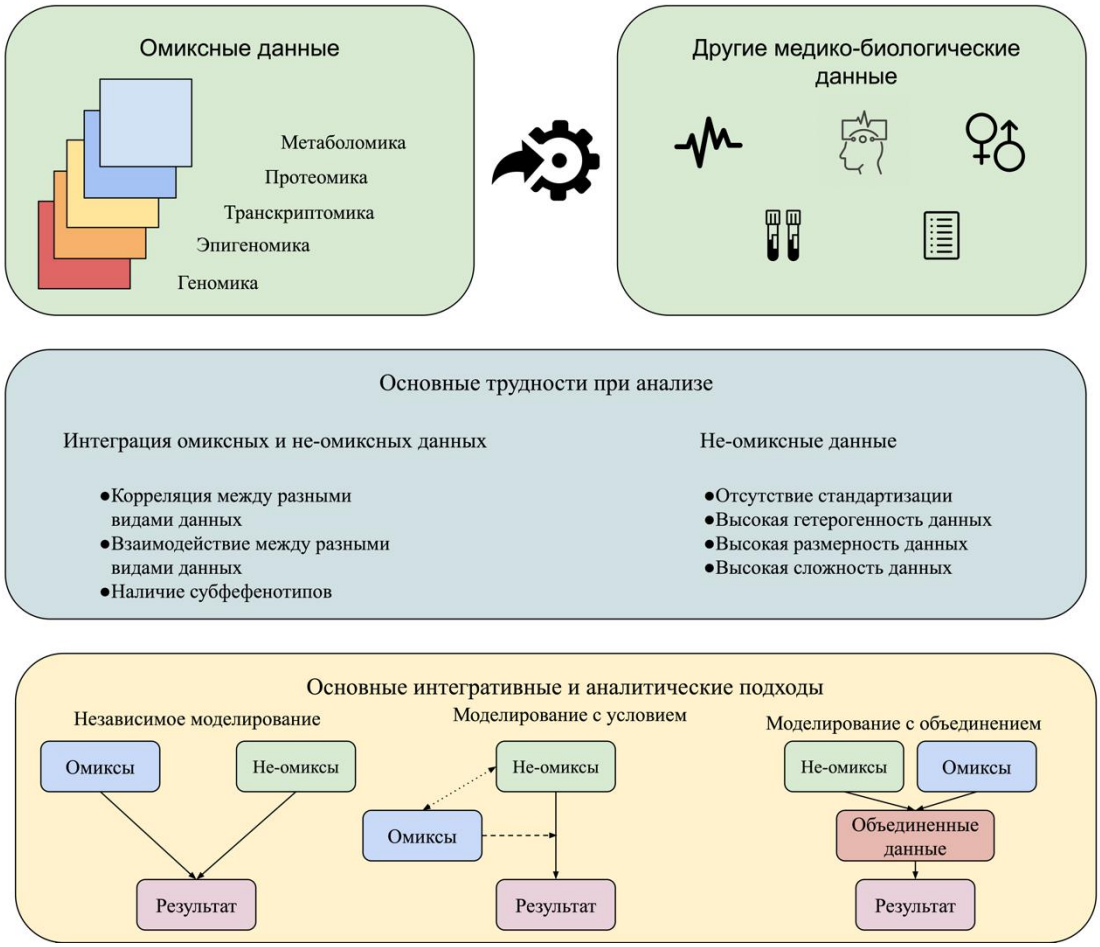


Рис. 10. Стратегии для интеграции омиксных данных с данными другого вида

Независимое моделирование. При независимом подходе, который также имеет название «поздняя интеграция», омиксные и неомиксные модели создаются независимо друг от друга: омиксная модель включает в себя только омиксные данные, неомиксная модель создается независимо от омиксных переменных, и фиттинг модели происходит с участием предварительно отобранных неомиксных параметров [44]. В обоих случаях обязательным этапом будет являться отбор признаков и уменьшение размерности. Независимо отобранные омиксные и неомиксные переменные используются в финальной модели. Благодаря своей простоте именно эта стратегия является часто используемой, но к недостаткам относится то, что в этом случае не всегда удается определить корреляцию и взаимодействия омиксных и неомиксных данных. Для решения этой проблемы были предложены подходы, основанные на деревьях решений,

однако их применение было реализовано на клинических данных и данных с уровнем экспрессией генов, поэтому возможность применения данного способа при анализе с другими видами омиксов остается под вопросом [45-46].

Моделирование с условием. Первым этапом при моделировании с условием является определение модели, основанной только на неомиксных данных, к которой затем добавляются омиксные параметры. То есть, основным моментом будет являться решение, какие именно омиксные параметры будут добавлены в модель. Это может быть осуществлено в различных вариантах: универиантный отбор и частичное уменьшение размерности.

При универиантом отборе параметр омиксных данных будет добавлен в модель, если его включение увеличит точность модели. Довольно часто при выборе этого подхода итоговая модель имеет точность ниже, чем при использовании только неомиксных данных, поэтому его использование оправдано в редких случаях [47].

Другой подход, показывающий лучшие результаты, включает в себя уменьшение размерности в объединенной модели только омиксных данных. Одним из методов его реализаций является метод наименьших квадратов – частично наименьших квадратов (least squares-partial least squares – LS-PLS) [43]. Стоит иметь в виду, что при использовании этого метода часто приходится сталкиваться с проблемой сходимости, а точность модели зависит от степени коллинеарности между различными типами данных. Помимо LS-PLS для уменьшения размерности омиксных данных могут использоваться ридж-регрессия и регрессия наименьшего абсолютного значения и оператора выбора [43, 48, 49]. Основным недостатком перечисленных методов является их высокая ресурсозатратность.

Совместное моделирование. Различают два подхода при совместном моделировании: поэтапный и параллельный. В первом случае происходит отдельный анализ ассоциаций между различными типами данных и затем с результатом, а во втором случае – одновременный анализ различных типов данных. Ritchie с соавт. предложил классифицировать параллельный подход так же, как и методы анализа омиксных данных, на три группы: интеграция, основанная на слиянии, основанная на трансформации и основанная на модели (рис. 11) [36]. Одна из первых работ, где использовался параллельный подход, была выполнена Sun с соавт., в ней была выполнена интеграция со слиянием и совместный отбор признаков омиксных и неомиксных данных с помощью алгоритма i-relief [50].

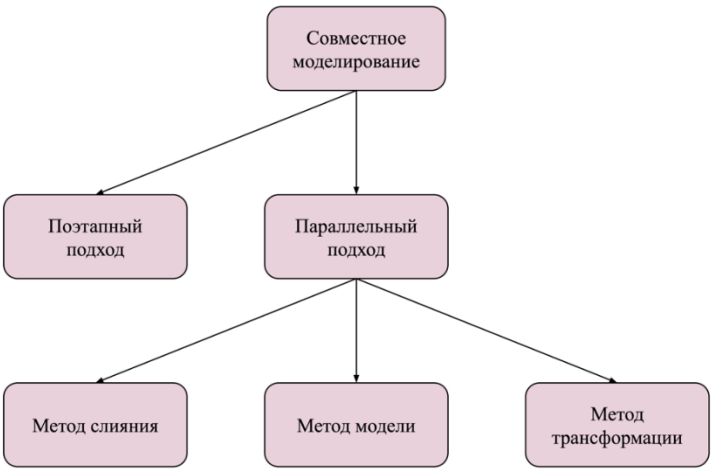


Рис. 11. Виды совместного моделирования омиксных и неомиксных данных

Заключение. Применение системных подходов для анализа омиксных и неомиксных данных стало возможным благодаря развитию омиксных технологий и статистических методов их анализа. Различают три подхода интеграции данных разного вида для создания гибридных моделей: независимый, с условием и совместный. На данный момент в литературе описываются только случаи интеграции многомерных омиксных с малоразмерными неомиксными данными, однако все три подхода интеграции также могут применяться и для многомерных неомиксных данных, причем эти случаи будут встречаться чаще ввиду увеличения количества информации из различных источников, например, из медицинских электронных карт и индивидуальных сенсоров, регистрирующих различные физиологические сигналы. Как правило, любой подход интеграции омиксных и неомиксных данных требует предварительного отбора признаков, уменьшения размерности и регуляризации.

Совместный подход создания модели является предпочтительным для интеграции многомерных омиксных и неомиксных данных, так как он позволяет определить корреляцию между двумя типами данных, даже несмотря на большую сложность системы. Дальнейший выбор поэтапного или параллельного подхода зависит от цели анализа: увеличение точности предсказания фенотипов или тест ассоциаций. Поэтапный подход, который моделирует отношения между различными размерностями данных, будет предпочтительным, когда целью является получение информации о механизмах, которые обуславливают тот или иной фенотип. Параллельный подход, в свою очередь, используется, когда основная цель – улучшить точность прогнозирования классификации, а моделирование и понимание механизмов, лежащих в основе этой классификации, отходят на второй план. Для понимания комплекса характеристик, ответственных за то или иное состояние, требуется не только понимание биологической системы, но и комбинаторного эффекта набора других факторов.

Интеграция омиксных и неомиксных данных позволяет улучшить точность классификации, определить новые параметры, оказывающие наибольшее влияние, и определенно внесет свой вклад в развитие прецизионных подходов в медицине. Исходя из вышеизложенного, дальнейшее развитие методов стандартизации и валидации является актуальной задачей для интегративного подхода к анализу биологических систем.

СПИСОК ЛИТЕРАТУРЫ

1. Hood, L. The human genome project: big science transforms biology and medicine / L. Hood, L. Rowen // Genome Medicine. – 2013. – Vol. 5. – № 9 – pp. 79.
2. Narad, P. Omics Approaches, Technologies And Applications. / P. Narad, P. Arivaradarajan G. Misra. – Singapore: Springer Singapore, 2018. – pp. 1-10.
3. McKusick, V. A. Toward a complete map of the human genome / V. A. McKusick, F. H. Ruddle. // Genomics. – 1987. – Vol. 1. – № 2. – pp. 103-106.
4. Feinberg, A. P. Epigenomics reveals a functional genome anatomy and a new approach to common disease / A. P. Feinberg // Nature Biotechnology. – 2010. – Vol. 28. – № 10. – pp. 1049-1052.
5. Heather, J. M. The sequence of sequencers: The history of sequencing DNA / J. M. Heather, B. Chain // Genomics. – 2016. – Vol. 107. – № 1 – pp. 1-8.
6. Reviving the Transcriptome Studies: An Insight Into the Emergence of Single-Molecule Transcriptome Sequencing / B. Wang, V. Kumar, A. Olson, D. Ware // Frontiers in Genetics. – 2019. – Vol. 10. – pp. 384.
7. RNA sequencing: new technologies and applications in cancer research / Hong, M., Tao S. S., Zhang L. [et al] // Journal of Hematology & Oncology. – 2020. – Vol. 13. – № 1. – pp. 166.
8. Carbonara, K. Proteomes Are of Proteoforms: Embracing the Complexity / K. Carbonara, M. Andonovski, J. R. Coorssen // Proteomes. – 2021. – Vol. 9. – № 3. – pp. 38.
9. The Size of the Human Proteome: The Width and Depth / Ponomarenko E.A., Poverennaya E. V.,

Ilgisonis E. V. [et al] // International Journal of Analytical Chemistry. – 2016. – Vol. 2016. – pp. 1-6.
10.Kuehnbaum, N. L. New Advances in Separation Science for Metabolomics: Resolving Chemical Diversity in a Post-Genomic Era / N. L. Kuehnbaum, P. Britz-McKibbin // Chemical Reviews. – 2013. – Vol. 113. – № 4. – pp. 2437-2468.
11.Clish, C. B. Metabolomics: an emerging but powerful tool for precision medicine / C. B. Clish // Cold Spring Harbor Molecular Case Studies. – 2015. – Vol. 1. – № 1. – pp. a000588.
12.Targeted metabolomics / L. Roberts, A. Souza, R. Gerszten, C. Clish // Current Protocols in Molecular Biology. – 2012. – Chapter 30. – Unit 30.2.1-24.
13.Kaplan, A. Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence / A. Kaplan, M. Haenlein // Business Horizons. – 2019. – Vol. 62. – № 1 – pp. 15-25.
14. Jo, T. Machine learning foundations: supervised, unsupervised, and advanced learning / T. Jo. – Switzerland: Springer, 2021. – 411 p.
15.OmicsARules: a R package for integration of multi-omics datasets via association rules mining / D. Chen, F. Zhang, Q. Zhao, J. Xu // BMC bioinformatics. – 2019. – Vol. 20. – № 1 – pp. 554.
16. Shin J. Reinforcement Learning – Overview of recent progress and implications for process control / J. Shin, T. Badgwell, K. Liu, J. Lee // Computers & Chemical Engineering. – 2019. – Vol. 127. – pp. 282-294.
17.Zou, J. Artificial Neural Networks. / J. Zou, Y. Han, S-S. So. – Totowa, NJ: Humana Press, 2008. – pp. 14–22.
18.Rumelhart, D. E. Learning representations by back-propagating errors / D. E. Rumelhart, G. E. Hinton, R. J. Williams // Nature. – 1986. – Vol. 323. – № 6088 – pp. 533-536.
19. LeCun, Y. Deep learning / Y. LeCun, Y. Bengio, G. Hinton. // Nature. – 2015. – Vol. 521. – № 7553 – pp. 436-444.
20.Methods for the integration of multi-omics data: mathematical aspects / Bersanelli M., Mosca E., Remondini D. [et al] // BMC Bioinformatics. – 2016. – Vol. 17. – № S2 – pp. S15.
21.Using machine learning approaches for multi-omics data analysis: A review / Reel P. S., Reel S., Pearson E. [et al] // Biotechnology Advances. – 2021. – Vol. 49. – pp. 107739.
22.Statistical Workflow for Feature Selection in Human Metabolomics Data / Antonelli, Claggett, Henglin [et al] // Metabolites. – 2019. – Vol. 9. – № 7 – pp. 143.
23.Liew, A. W.-C. Missing value imputation for gene expression data: computational techniques to recover missing data from available information / A.W.-C. Liew, N.-F. Law, H. Yan // Briefings in Bioinformatics. – 2011. – Vol. 12. – № 5 – pp. 498-513.
24. Bayesian Framework for Detecting Gene Expression Outliers in Individual Samples / Vivian J., Eizenga J. M., Beale H. C. [et al] // JCO Clinical Cancer Informatics. – 2020. – № 4 – pp. 160–170.
25. Designing and interpreting 'multi-omic' experiments that may change our understanding of biology / Haas R., Zelezniak A., Iacovacci J. [et al] // Current Opinion in Systems Biology. – 2017. – Vol. 6. – pp. 37-45.
26.SMOTE: Synthetic Minority Over-sampling Technique / N. Chawla, K. Bowyer, L. Hall, W. Kegelmeyer // Journal of Artificial Intelligence Research. – 2002. – Vol. 16. – pp. 321-357.
27.Application of Clinical Bioinformatics / Wang X., Baumgartner C., Shields D. C. [et al]. – Dordrecht: Springer Netherlands, 2016. – pp. 23-57.
28.Hoboken, N. J. Statistical bioinformatics: a guide for life and biomedical science researchers / N. J. Hoboken. – Wiley-Blackwell, 2010. – 386 p.
29.Integrated omics: tools, advances and future approaches / B. Misra, C. Langefeld, M. Olivier, L. Cox // Journal of Molecular Endocrinology. – 2019. – Vol. 62. – № 1. – pp. R21-R45.
30.An introduction to statistical learning: with applications in R / Gareth M. James G. M., Witten D. [et al.] – New York: Springer, 2013. – 622 p.
31.Jolliffe, I. T. Principal component analysis 2nd ed / I. T. Jolliffe. – New York: Springer, 2002. – 488 p.
32.Martinez, A. M. PCA versus LDA / A. M. Martinez, A. C. Kak // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2001. – Vol. 23. – № 2 – pp. 228-233.
33.Young, F. W. Multidimensional scaling: history, theory, and applications Hillsdale / F. W. Young. – Erlbaum Associates, 1987. – 336 p.
34.Rappoport, N. Multi-omic and multi-view clustering algorithms: review and cancer benchmark / N. Rappoport, R. Shamir // Nucleic Acids Research. – 2018. – Vol. 46. – № 20. – pp. 10546-10562.
35.Integration strategies of multi-omics data for machine learning analysis / Picard M., Scott-Boyer M.-P., Bodein A. [et al] // Computational and Structural Biotechnology Journal. – 2021. – Vol. 19. – pp. 3735-3746.
36.Methods of integrating data to uncover genotype-phenotype interactions / Ritchie M. D.,

Holzinger E. R., Li R. [et al] // Nature Reviews Genetics. – 2015. – Vol. 16. – № 2 – pp. 85-97.
37.Gligorijević, V. Methods for biological data integration: perspectives and challenges / V. Gligorijević, N. Pržulj. // Journal of The Royal Society Interface. – 2015. – Vol. 12. – № 112 – pp. 20150571.
38.Yu, X.-T. Computational Systems Biology / X.-T. Yu. – Springer New York, 2018. – pp. 109–135.
39.Sorzano, C. O. S. A survey of dimensionality reduction techniques / C. O. S. Sorzano, J. Vargas, A. P. Montano // Published Online First: 2014. DOI:10.48550/ARXIV.1403.2877
40.Challenges in the Integration of Omics and Non-Omics Data / López de Maturana, E., Alonso L., Alarcón P. [et al] // Genes. – 2019. – Vol. 10. – № 3 – pp. 238.
41.Prediction of non-muscle invasive bladder cancer outcomes assessed by innovative multimarker prognostic models / López de Maturana E., Picornell A., Masson-Lecomte A. [et al] // BMC Cancer. – 2016. – Vol. 16. – № 1 – pp. 351.
42.Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer / K. Chaudhary, O. Poirion, L. Lu, L. Garmire // Clinical Cancer Research: An Official Journal of the American Association for Cancer Research. – 2018. – Vol. 24. – № 6 – pp. 1248-1259.
43.Bazzoli, C. Classification based on extensions of LS-PLS using logistic regression: application to clinical and multiple genomic data / C. Bazzoli, S. Lambert-Lacroix // BMC Bioinformatics. – 2018. – Vol. 19. – № 1 – pp. 314.
44.Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks / Gevaer O., De Smet F., Timmerman D., [et al] // Bioinformatics (Oxford, England). – 2006. – Vol. 22. – № 14 – pp. e184-190.
45.Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction / Nevins J. R., Huang E. S., Dressman H. [et al] // Human Molecular Genetics. – 2003. – Vol. 12. – Spec. № 2. – pp. R153-157.
46.Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes / Pittman J., Huang E., Dressman H. [et al] // Proceedings of the National Academy of Sciences. – 2004. – Vol. 101. – № 22 – pp. 8431-8436.
47.Bøvelstad, H. M. Survival prediction from clinico-genomic models--a comparative study / H. M. Bøvelstad, S. Nygård, O. Borgan. // BMC bioinformatics. – 2009. – Vol. 10. – pp. 413.
48.Hoerl, A. E. Ridge Regression: Biased Estimation for Nonorthogonal Problems / A. E. Hoerl, R. W. Kennard. // Technometrics. – 1970. – Vol. 12. – № 1 – pp. 55-67.
49.Tibshirani, R. Regression Shrinkage and Selection Via the Lasso / R. Tibshirani // Journal of the Royal Statistical Society: Series B (Methodological). – 1996. – Vol. 58. – № 1 – pp. 267-288.
50.Improved breast cancer prognosis through the combination of clinical and genetic markers / Sun Y., Goodison S., Li J. [et al] // Bioinformatics. – 2007. – Vol. 23. – № 1. – pp. 30-37.

СВЕДЕНИЯ ОБ АВТОРАХ:

Наиль Фанисович Максютов – младший научный сотрудник лаборатории больших данных и прецизионной спортивной медицины центра спортивной медицины и реабилитации ГНЦ ФМБЦ им. А.И. Бурназяна ФМБА России, Москва, e-mail: Maksjytov.nail@gmail.com.

Артур Амирович Муртазин – младший научный сотрудник лаборатории больших данных и прецизионной спортивной медицины центра спортивной медицины и реабилитации ГНЦ ФМБЦ им. А.И. Бурназяна ФМБА России Москва, e-mail: aa.murtazin@gmail.com.

Евгений Игоревич Балакин – кандидат медицинских наук, старший научный сотрудник лаборатории больших данных и прецизионной спортивной медицины ГНЦ ФМБЦ им. А.И. Бурназяна ФМБА России, Москва, e-mail: evgbalakin@yandex.ru.

Василий Игоревич Пустовойт – кандидат медицинских наук, заведующий лабораторией больших данных и прецизионной спортивной медицины центра спортивной медицины и реабилитации ГНЦ ФМБЦ им. А.И. Бурназяна ФМБА России, Москва, e-mail: vipust@yandex.ru.

INFORMATION ABOUT THE AUTHORS:

Nail' Fanisovich Maksjutov – Junior Researcher in the Big Data and Precision Sports Medicine Laboratory, Sports Medicine and Rehabilitation Center, State Research Center – Burnasyan Federal Medical Biophysical Center of Federal Medical Biological Agency, Moscow, e-mail: Maksjytov.nail@gmail.com.

Artur Amirovich Murtazin – Junior Researcher in the Big Data and Precision Sports Medicine Laboratory, Sports Medicine and Rehabilitation Center, State Research Center – Burnasyan Federal Medical Biophysical Center of Federal Medical Biological Agency, Moscow, e-mail: aa.murtazin@gmail.com.
Evgenij Igorevich Balakin – Candidate of Medical Sciences, Senior in the Big Data and Precision Sports Medicine Laboratory, Sports Medicine and Rehabilitation Center, State Research Center – Burnasyan Federal Medical Biophysical Center of Federal Medical Biological Agency, Moscow, e-mail: evgbalakin@yandex.ru.
Vasilij Igorevich Pustovoit – Candidate of Medical Sciences, Head of the Big Data and Precision Sports Medicine Laboratory, Sports Medicine and Rehabilitation Center, State Research Center – Burnasyan Federal Medical Biophysical Center of Federal Medical Biological Agency Moscow, e-mail: vipust@yandex.ru.

Для цитирования: Применение машинного обучения и омиксных технологий для оценки функционального состояния / Н.Ф. Максютов, А.А. Муртазин, Е.И. Балакин, В.И. Пустовойт // Современные вопросы биомедицины – 2022 – Т.6 – № 3. DOI: 10.51871/2588-0500_2022_06_03_14

For citation: Maksjutov N.F., Murtazin A.A., Balakin E.I., Pustovoit V.I. Using machine learning approaches and omics technologies for assessment of human functional state. *Modern Issues of Biomedicine*, 2022, vol. 6, no. 3. DOI: 10.51871/2588-0500_2022_06_03_14