# МАШИННОЕ ОБУЧЕНИЕ В ЛАБОРАТОРНОЙ МЕДИЦИНЕ

**В.П. Мудров[1,2], С.С. Иванов[3], М. Йовичич[4]**

*[1]ФГБОУ ДПО РМАНПО МЗ РФ, г. Москва, Россия*

*[2]ГБУЗ ДКЦ №1 ДЗМ, г. Москва, Россия*

*[3]РУДН, г. Москва, Россия*

*[4]Белградский университет, г.Белград, Сербия*

**Резюме**

Внедрение научных исследований геномики, протеомики, биоинформатики, биостатистики в клиническую практику открыло инновационные подходы в диагностике, терапии и прогнозе заболеваний. Машинное обучение способно использовать большие объемы данных для вывода сложных взаимосвязей и закономерностей, которые в противном случае могут быть за пределами возможностей системы, основанной на правилах, или эксперта-человека. Использование машинного обучения в лабораторной медицине становится все более важной областью. Аналитическая стратегия специалистов по биостатистике зависит от их патофизиологических знаний и опыта, когда прогнозный анализ на основе данных бросает вызов этой парадигме мышления, а растущая вычислительная мощность может выявить связи, не осознаваемые человеческим разумом. Числовой и структурированный формат данных в лабораторной медицине хорошо подходит для вычислительных методов, таких как машинное обучение. Базисом для внедрения машинного обучения стали лабораторные информационные технологии многопрофильного лечебного учреждения, содержащие обширный массив данных. Сравнение традиционной статистики и методов машинного обучения показывает, что традиционная статистика является фундаментальной основой машинного обучения, где алгоритмы черного ящика выводятся из базовой математики, но продвинуты с точки зрения автоматизированного анализа, обработки больших данных и предоставления интерактивных визуализаций. Традиционную статистику и машинное обучение лучше всего интегрировать для разработки инструментов автоматизированного анализа данных. Различные алгоритмы машинного обучения с учителем получили широкое применение в лабораторной медицине: дерево решений, случайный лес, экстремальное повышение градиента, логистическая регрессия, метод опорных векторов, искусственные нейронные сети и др. Широкое применение искусственный интеллект получил в онкогематологии, кардиологии, эндокринологии, нефрологии, урологии, нейрохирургии, фармакологии, в диагностике инфекционных заболеваний. Подход машинного обучения может быть преобразован в руководство для обучения студентов-медиков, преподавателей и научных сотрудников исследователей для проведения исследований, в том числе групповых в любой области здравоохранения. Внедрение инструментов машинного обучения в медицинскую практику сталкивается со многими проблемами. Технические препятствия включают сбор данных и потребность в «больших данных» для создания репрезентативных алгоритмов. Важным вопросом, требующим решения, является этика обмена данными и использования информации о пациентах. Тем не менее, машинное обучение будет приобретать все большее значение для лабораторной медицины. Машинное обучение обещает захватывающие достижения в медицине, но его применение в лабораторной медицине все еще находится на стадии становления. Поскольку область является молодой, существует дополнительная потребность в стандартизации того, как разрабатываются и представляются эти алгоритмы.

**Ключевые слова**: статистика, машинное обучение, дерево решений, случайный лес, экстремальное повышение градиента, логистическая регрессия, метод опорных векторов, искусственные нейронные сети.

DOI: 10.58953/15621790_2022_13_23

# MACHINE LEARNING IN LABORATORY MEDICINE

**V.P. Mudrov[1,2], S.S. Ivanov[3], M. Yovichich[4]**

*[1]Russian Medical Academy of Continuous Professional Education of the Ministry of Health of the Russian Federation, Moscow, Russia*

*[2]Diagnostic clinical center №1 of the Moscow City Department of Health, Moscow, Russia*

*[3]RUDN University, Moscow, Russia*

*[4]Belgrade University, Belgrade, Serbia*

**Abstract**

The introduction of scientific research of genomics, proteomics, bioinformatics, biostatistics into clinical practice has opened up innovative approaches in the diagnosis, therapy and prognosis of diseases. Machine learning is capable of using

large amounts of data to infer complex relationships and patterns that might otherwise be beyond the capabilities of a rule-based system or a human expert. The use of machine learning in laboratory medicine is becoming an increasingly important field. The analytical strategy of biostatistics specialists depends on their pathophysiological knowledge and experience, when predictive analysis based on data challenges this paradigm of thinking, and growing computing power can reveal connections that are not realized by the human mind. Numerical and structured data format in laboratory medicine is well suited for computational methods such as machine learning. The basis for the introduction of machine learning was the laboratory information technologies of a multidisciplinary medical institution containing an extensive array of data. A comparison of traditional statistics and machine learning methods shows that traditional statistics is the fundamental basis of machine learning, where black box algorithms are derived from basic mathematics, but are advanced in terms of automated analysis, big data processing and providing interactive visualizations. Traditional statistics and machine learning are best integrated to develop automated data analysis tools. Various machine learning algorithms with a teacher have been widely used in laboratory medicine: decision tree, random forest, extreme gradient enhancement, logistic regression, support vector method, artificial neural networks, etc. Artificial intelligence has been widely used in oncohematology, cardiology, endocrinology, nephrology, urology, neurosurgery, pharmacology, and in the diagnosis of infectious diseases. The machine learning approach can be transformed into a guide for teaching medical students, teachers and researchers to conduct research and group research in any field of healthcare. The introduction of machine learning tools into medical practice faces many challenges. Technical obstacles include data collection and the need for "big data" to create representative algorithms. An important issue that needs to be addressed is the ethics of data exchange and the use of patient information. Nevertheless, machine learning will become increasingly important for laboratory medicine. Machine learning promises exciting advances in medicine, but its application in laboratory medicine is still in its infancy. Since the field is young, there is an additional need to standardize how these algorithms are developed and presented.

**Keywords:** statistics, machine learning, decision tree, random forest, extreme gradient enhancement, logistic regression, support vector machine, artificial neural networks.

## Введение.

Обучение компьютерных программ в обработке «больших данных» стало острием прогресса медицины, в том числе клинической лабораторной диагностики. Внедрение научных исследований геномики, протеомики, биоинформатики, биостатистики в клиническую практику открыло инновационные подходы в диагностике, терапии и прогнозе заболеваний.

Искусственный интеллект — это современный подход, основанный на компьютерных науках, разрабатывающий программы и алгоритмы для выполнения задач, которые обычно требуют квалифицированного человеческого интеллекта. Искусственный интеллект — общий термин, описывающий использование технологий для выполнения задач, которые обычно требуют человеческого интеллекта, например, распознавания голоса или изображений. Машинное обучение (ML) является частью искусственного интеллекта. Технология ML позволяет машинам (компьютерам) учиться на полученных данных, используя статистические подходы и алгоритмы. Искусственный интеллект (ИИ) относится к моделированию человеческого разума в компьютерных системах, запрограммированных на то, чтобы думать, как люди, имитировать их действия в обучении и решении проблем. ИИ включает в себя различные подмножества, включая машинное обучение, глубокое обучение, обычные нейронные сети, нечеткую логику и распознавание речи. Такие интеллектуальные системы упрощают вмешательство человека в клиническую диагностику, медицинскую визуализацию и способность принимать решения. ИИ способен выполнять задачи визуального восприятия, принятия решений и общения [7].

Машинное обучение способно использовать большие объемы данных для вывода сложных взаимосвязей и закономерностей, которые в противном случае могут быть за пределами возможностей системы, основанной на правилах, или эксперта-человека. Кроме того, в то время как статические алгоритмы, основанные на правилах, основаны на ранее установленных знаниях, машинное обучение может выявлять новые шаблоны и приложения и постоянно использовать новые данные для улучшения своей производительности. Машинное обучение — область искусственного интеллекта, которая предоставляет компьютерным программам возможность обучения новым навыкам на основе опыта без дальнейшего программирования человеком. Алгоритмы могут быстро и точно анализировать большие наборы данных с помощью контролируемых и неконтролируемых методов обучения для получения результатов классификации и прогнозирования. Применение в лабораторной медицине может потенциально повысить эффективность на всех этапах всего лабораторного процесса тестирования [19].

Применение машинного обучения в медицине

привлекло огромное внимание за последнее десятилетие. В отличие от традиционных программ, которые определяются предварительно закодированными правилами, машинное обучение относится к компьютерным алгоритмам, которые учатся на предыдущих примерах. Целью большинства моделей машинного обучения с контролем является получение входных данных и вывод прогнозируемого результата. Алгоритм, который выполняет это предсказание, обучается на больших наборах данных предыдущих наблюдений [14]. Уже существует примеры методов машинного обучения, одобренные для использования Управлением по контролю за продуктами и лекарствами США (FDA) в области радиологии, кардиологии и патологии.

Использование машинного обучения в лабораторной медицине становится все более важной областью. Аналитическая стратегия специалистов по биостатистике зависит от их патофизиологических знаний и опыта, когда прогнозный анализ на основе данных бросает вызов этой парадигме мышления, а растущая вычислительная мощность может выявить связи, не осознаваемые человеческим разумом. Числовой и структурированный формат данных в лабораторной медицине хорошо подходит для вычислительных методов, таких как машинное обучение. Базисом для внедрения машинного обучения стали лабораторные информационные технологии многопрофильного лечебного учреждения, содержащие обширный массив данных [5]. Такие достижения открывают перспективы для будущего медицины, где лабораторное тестирование обеспечивает большую часть основы для принятия клинических решений.

Практика принятия медицинских решений стремительно меняется с развитием инновационных вычислительных технологий. Растущий интерес к анализу данных с усовершенствованием методов компьютерной обработки больших данных поднимает вопрос о том, можно ли интегрировать машинное обучение с традиционной статистикой в исследованиях в области здравоохранения [11].

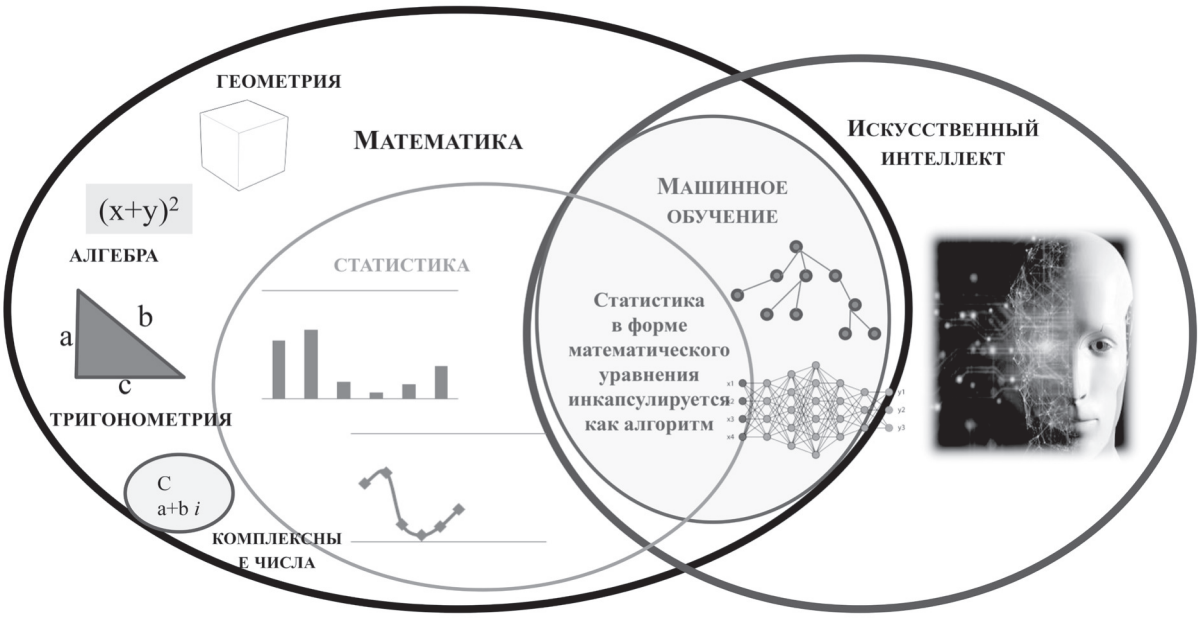
## Интеграция традиционной статистики с машинным обучением.

Искусственный интеллект и машинное обучение основаны на законах статистики. Статистика — это раздел математики, состоящий из комбинации математических методов для анализа и визуализации данных. С другой стороны, машинное обучение — это раздел искусственного интеллекта, состоящий из алгоритмов, выполняющих контролируемое и неконтролируемое обучение (рис. 1).

Сравнение традиционной статистики и методов машинного обучения показывает, что традиционная статистика является фундаментальной основой машинного обучения, где алгоритмы черного ящика выводятся из базовой математики, но продвинуты с точки зрения автоматизированного анализа, обработки боль-

**Рисунок 1.**

***Интеграция между статистикой, искусственным интеллектом, машинным обучением. (адаптировано по [9].***



ших данных и предоставления интерактивных визуализаций. Хотя природа обоих этих методов различна, они концептуально схожи. Традиционную статистику и машинное обучение лучше всего интегрировать для разработки инструментов автоматизированного анализа данных [9].

Традиционная статистика считается более эффективной в вычислительном отношении и более приемлемой в медицинской сфере. Машинное обучение требует высокой вычислительной мощности с точки зрения вычислительной мощности и памяти. В обычном статистическом анализе ученые используют базовые программные инструменты, которым не хватает возможностей для обработки больших данных и визуализации результатов. Алгоритмы «черного ящика» машинного обучения способны обнаруживать тонкие скрытые закономерности в многомодельных данных.

Классическая статистика — это проверка гипотез и статистический вывод для классификации признака. Проверка гипотез — интерпретация результатов путем выдвижения предположений (гипотез) на основе экспериментальных данных. Статистические тесты (например, t -критерий Стьюдента, ANOVA) используются для интерпретации результатов на основе таких показателей, как p-значение (значимая разница). Биостатистики и ученые-медики выполняют статистический анализ с использованием обычных программных инструментов, потому что основная цель состоит в том, чтобы сосредоточиться на анализе, основанном на проверке гипотез насколько является значимым биомаркёр для прогноза заболевания, дает ли лечение положительные результаты, как контролировать определенные факторы риска для конкретной патологии. Машинное обучение часто бывает очень сложным и трудным для интерпретации клиницистами, потому что оно использует вычислительное программирование, а обычные статистические данные легко интерпретируются и имеют меньшую емкость, поэтому представляют меньший риск неспособности обобщить непричинные эффекты.

Причинно-следственный вывод играет жизненно важную роль в понимании механизмов переменных, чтобы найти порождающую модель и предсказать результаты, которым подвергаются переменные. Характеристика причинного вывода состоит в том, чтобы найти ответы на вопросы о механизмах, посредством которых переменные принимают значения. Например, эпидемиологи собирают данные, связанные с питанием, и находят факторы, влияющие на ожидаемую продолжительность жизни, чтобы предсказать последствия изменения диеты людьми. Большое количество переменных, небольшой размер выборки и пропущенные значения считаются серьезными препятствиями для надлежащего анализа данных и принятия точных решений в области медицины.

Обычная статистика больше подходит для более простых наборов данных, тогда как машинное обучение может управлять сложными наборами данных. Как правило, обычный статистический анализ проводится, если для исследования имеется предварительная литература по интересующей теме, количество переменных, участвующих в исследовании, относительно невелико, а количество наблюдений (выборок) больше, чем количество переменных. Это может помочь ученому лучше понять тему, выбирая важные переменные из предшествующих знаний, а также применяя соответствующие аналитические модели для проверки связи между переменными (независимыми) и результатами (зависимыми). Традиционный статистический подход отдает больший приоритет типу набора данных, например, когортное исследование, следующее определенной гипотезе. Для вывода о значении важную роль играет количество наблюдений в ассоциативном исследовании. Алгоритмы машинного обучения могут обрабатывать данные с пропущенными значениями и служат альтернативой обычной статистике для обычных анализов, таких как определение размера эффекта, значимых факторов, анализа выживаемости. алгоритмы машинного обучения

Наиболее часто используемыми моделями в традиционной статистике являются модели логистической регрессии или регрессии Кокса для бинарных результатов, линейная регрессия для непрерывных результатов, обобщенные линейные модели, основанные на распределении данных. Это популярно в исследованиях значимости для общественного здравоохранения, особенно когда анализ включает популяционное исследование.

Регрессионный анализ — это набор статистических методов для оценки взаимосвязи между зависимой переменной и набором независимых переменных. Регрессия широко используется в исследованиях в области здравоохранения для анализа и прогнозирования различных заболеваний. Выбор конкретного типа регрессии зависит от типа зависимой переменной, такой как непрерывная и категориальная. Линейная регрессия используется для определения взаимосвязи между непрерывной зависимой переменной и набором независимых переменных.

Модель прогнозирования с использованием алгоритмов машинного обучения требует надежной связи между наблюдениями (пациентами) и переменными (независимыми переменными). Модели прогнозирования генерируют меры точности для

определения качества данных и прогнозирования конечного результата, используя наблюдения (пациенты), входные данные (независимые переменные) и выходные данные (зависимая переменная). Обучение представлению — это процесс обучения алгоритмов машинного обучения обнаружению интерпретируемых представлений. Различные представления могут запутать различные объясняющие факторы в конкретном наборе данных. Результат обучения представлению должен облегчить последующую задачу в процессе принятия решений. Например, репрезентативное обучение обрабатывает и группирует очень большие объемы неразмеченных обучающих данных в неконтролируемом или частично контролируемом обучении. Группировка немаркированных данных используется для соответствующей задачи, такой как выбор функций и дерево решений, для прогнозирования результатов. Сложный фактор обучения представлению заключается в том, что оно должно сохранять столько информации, сколько содержится во входных данных, чтобы получить точные прогнозы.

Медицинские исследования используют репрезентативное обучение в основном при распознавании изображений. Обучение с подкреплением обучает модели машинного обучения принимать последовательность решений, в отличие от обучения с учителем, которое зависит от одноразового или единственного зависимого фактора. Его основная цель — наделить человека навыками делать прогнозы на основе опыта работы с окружающей средой и развивать оценочную обратную связь. Эта уникальная особенность обучения с подкреплением помогает найти преобладающие решения в различных схемах диагностики и лечения в здравоохранении, которые обычно характеризуются длительной и последовательной процедурой. Обучение с подкреплением следует нескольким методам последовательного принятия решений, а именно, эффективным методам, таким как уровень опыта, уровень модели и уровень задачи, и репрезентативным методам, таким как представление для функции ценности, функции вознаграждения и задачи или моделей). Несоответствие в точности между обучающими и проверочными наборами указывает на «переобучение». Бонгард-Полонский М.М. [1] в 1967 году сформулировал проблему переобучения, заключающуюся в хорошей интерпретации построенной моделью примеров из обучающей выборки, но плохо работающей на данных, не участвовавших в обучении. Построенная модель машинного обучения запоминает огромное количество всевозможных данных вместо того, чтобы научиться выявлять особенности, в анализе биомаркёров такая переобученная система может выдать результат, желаемый исследователю.

Переобучение — это явление, при котором алгоритм очень хорошо моделируется на тренировочном наборе до такой степени, что его невозможно обобщить на другие данные.

С другой стороны, прогнозный анализ, основанный на алгоритмах машинного обучения, учится на данных, не полагаясь на программирование на основе правил, не делающих никаких предварительных предположений, а основанных на предоставленных исходных данных. Алгоритмы машинного обучения могут работать с многомерными большими данными, но обычная статистика может работать только с одним конкретным форматом данных. Кроме того, алгоритмы машинного обучения могут обрабатывать данные из различных источников данных, таких как внешние базы данных или онлайн-хранилища данных.

Оценка модели в машинном обучении аналогична анализу мощности в обычной статистике для оценки качества данных. Это ключевой шаг в машинном обучении, поскольку способность модели делать прогнозы для невидимых или будущих выборок повышает доверие к модели, которая будет использоваться в конкретном наборе данных. Используя специфичность и чувствительность алгоритмов, можно построить ROC-кривую с использованием коэффициентов истинно-положительных и ложноположительных результатов и рассчитать площадь под кривой (AUC), применяемую для оценки алгоритмов классификации с бинарными результатами.

Шесть различных алгоритмов машинного обучения с учителем нашли широкое применение в медицинской информатике: дерево решений, случайный лес, экстремальное повышение градиента, логистическая регрессия, метод опорных векторов, искусственные нейронные сети (таб.1).

## Модели (алгоритмы) машинного обучения

Дерево решений широко используется в медицинской информатике, поскольку оно является базовой концепцией, используемой другими алгоритмами, такими как случайный лес и повышение градиента, но с некоторыми различиями в процессах для прогнозирования конечного результата. Алгоритм дерева решений следует модели древовидной структуры, где есть корневой узел, узел решения и конечный узел. Корневой узел начинается с наиболее важной независимой переменной, за которой следуют узлы принятия решений (другие независимые переменные). Конечный узел указывает зависимую переменную, которая является окончательным предсказанным выходом. Древовидная структура строится на основе наблюдения,

**Таблица 1.**
***Термины, используемые в машинном обучении.***

| | |
|---|---|
| **Искусственный интеллект (ИИ)** | **Отрасль компьютерных наук, занимающаяся задачами, которые обычно требуют человеческого интеллекта.** |
| **Машинное обучение** | **Отрасль ИИ, в которой статистические алгоритмы устанавливают свои собственные шаблоны, подвергаясь воздействию репрезентативных данных для интерпретации и обработки новых данных.** |
| **Машина опорных векторов** | **Статистический метод для различения классов данных с максимально возможным запасом путем обучения. Количество обучаемых параметров: 10-100.** |
| **Случайный лес** | **Статистический метод, использующий сеть деревьев принятия решений для классификации данных.** |
| **k-ближайший сосед** | **Статистический метод классификации и регрессии данных, основанный на количестве k соседей.** |
| **Глубокие нейронные сети** | **Также называется глубоким обучением (DL), которое представляет собой подмножество машинного обучения с использованием сложных многоуровневых архитектур, включающих несколько скрытых уровней и большое количество узловых соединений.** |
| **Искусственные нейронные сети** | **Набор многоуровневых взаимосвязанных искусственных нейронов на основе глубоких нейронных сетей для изучения функций более высокого уровня, имитирующих биологический мозг.<br>Количество обучаемых параметров 100 000.** |
| **Рекуррентные нейронные сети** | **Тип глубоких нейронных сетей, но соединения между узлами в скрытом слое циклические.** |
| **Сверточные нейронные сети** | **Тип нейронных сетей, специально разработанный для машинного зрения. Чаще всего применяются для анализа изображений, распознавания и классификации** |
| **Компьютерное зрение** | **Быстрый и точный анализ тенденций и закономерностей из цифровых изображений, имитирующий биологическое зрение.** |
| **Машинное зрение** | **Аналог компьютерного зрения, ориентированный на эффективность, автоматический контроль, процесс роботизированного наведения.** |

попадающего в каждую область, и среднего значения прогноза.

Случайный лес — это алгоритм обучения ансамбля, полученный из дерева решений и позволяющий управлять многомерными данными. Он следует правилу дерева решений, но строит множество деревьев решений во время обучения и выводит класс с максимальным количеством факторов. Случайный лес известен как улучшенная версия дерева решений, поскольку он строит более одного дерева для выбора наилучшего результата, тогда как дерево решений строит только одно дерево. Количество деревьев, построенных в процессе обучения, не задано по умолчанию, так как пользователи могут указать его на основе количества выборок. Количество деревьев прямо пропорционально количеству образцов.

Экстремальное повышение градиента следует принципу случайного леса, но с дополнительной интерпретацией для прогнозирования конечного результата. Этот алгоритм также строит несколько деревьев, называемых усиленными деревьями. Оценка прогноза присваивается каждому листу в усиленных деревьях (градиентах), тогда как случайный лес содержит только окончательное значение решения для одного дерева Оценки всех листьев на деревьях суммируются для получения значений градиента, а окончательный прогноз делается на основе среднего значения, называемого усилением градиента.

Логистическая регрессия используется в медицине для прогнозной аналитики. Логистическая регрессия

предсказывает категориальный вывод, например, статус выживания (живой или мертвый). Прогнозы делаются на основе вероятностей, показанных кривой. Этот процесс повторяется для всех образцов. Кривая сдвигается для расчета новых вероятностей выборок, попадающих на эту линию. Наконец, вероятность данных рассчитывается путем умножения всех вероятностей вместе, и в качестве окончательного результата выбирается максимальная вероятность.

Машина опорных векторов разделяет данные на разные классы, но требует обнаружения гиперплоскостей. Гиперплоскость делит данные на две группы (классы). Точки, расположенные ближе к границе решения или гиперплоскости, называются опорными векторами. Окончательный прогноз делается на основе значений независимых переменных и опорных векторов, соответствующих гиперплоскости. Количество гиперплоскостей зависит от количества независимых переменных. Структура алгоритма сложна, имеет более трех функций, но ее способность одновременно обрабатывать несколько переменных с несколькими гиперплоскостями для прогнозирования конечного результата является одним из преимуществ этого алгоритма.

Искусственные нейронные сети — искусственное представление нервной системы человека. Дендриты собирают информацию от других нейронов в виде электрических импульсов (вход). Тело ячейки генерирует выводы на основе входных данных и решает, какие действия следует предпринять. Выходы передаются через экзонные терминали в виде электрических импульсов к другим нейронам. Та же концепция используется в искусственных нейронных сетях. Входные данные относятся к независимым переменным и образцам, предоставленным алгоритму. Входные данные умножаются на веса для вычисления функции суммирования. Чем выше вес входных данных, тем более значимыми являются входные данные для прогнозирования конечного результата. Функция активации предсказывает вероятности на основе обучающих данных и генерирует окончательный результат. Это известно как однослойный персептрон. В искусственной нейронной сети есть три типа слоев: входной слой, скрытый слой и выходной слой.

За оценкой модели следует переменная важность в машинном обучении. Переменная важность (оценка важности) является альтернативой идентификации значимых факторов (значение p) в традиционной статистике с использованием меры доверительного интервала и проверки гипотез. После выполнения оценки модели необходимо дополнительно изучить элементы (переменные) входных данных в отношении того, как они влияют на показатель точности.

Глубокое обучение — это подмножество машинного обучения, которое подходит к проблемам способом, который больше похож на человеческий подход. В глубоком обучении алгоритмы могут выводить характеристики/паттерны данных посредством нескольких уровней обработки (например, нейронных сетей).

## Диагностика и распознавание на основе изображений

Распознавание изображений и диагностика имеют важное значение при многих заболеваниях, включая злокачественные и доброкачественные гематологические заболевания [8,10]. Основной функцией машинного обучения, используемой в диагностике на основе изображений, является классификация гистопатологических препаратов. Подход к использованию машинного обучения для распознавания и классификации изображений обычно начинается с предварительной обработки изображений, которая включает цифровую маркировку слайдов. Маркировка слайдов особенно важна при контролируемом машинном обучении, которое обычно требует маркировки как входных, так и выходных данных процесса. После этого изображения обычно сегментируют на разные части (например, цитоплазму и ядро), после чего идентифицируют признаки (извлечение признаков). Алгоритм машинного обучения затем можно применить к выборке (обычно с известным результатом), создав модель, которая может классифицировать изображения на основе их характеристик. В большинстве моделей диагностики на основе слайдов используется бинарный подход (диагноз/отсутствие диагноза), который значительно упрощает реальную сложность гистопатологической диагностики. Впоследствии все модели должны пройти внутреннюю или внешнюю проверку для обеспечения применимости.

Сверточные нейронные сети для диагностики острого лимфолейкоза достигают AUC близкую к 95%.

## Машинное обучение в гематологии

Широкое применение искусственный интеллект получил в онкогематологии [13]. Интеграция моделей ИИ в лабораторную гематологию проводится давно. Использование инструментов машинного и глубокого обучения было исследовано в различных областях гематологической диагностики, включая лабораторные исследования, гистопатологию, проточную цитометрию и молекулярные данные. Были разработаны различные методы цифровой визуализации для помощи в простой лабораторной диагностике, например, при железодефицитной анемии или анализе периферических мазков.

В многочисленных исследованиях показано применение машинного обучения и глубокого обучения в проточной цитометрической диагностике доброкачественных и злокачественных гематологических заболеваний. Это распознавание пациентов с множественной миеломой с помощью искусственных нейронных сетей в сочетании с масс-спектрометрией плазмы периферической крови, обнаружение железодефицитной анемии по гематологическим параметрам с использованием деревьев решений, выявление остаточной болезни с помощью машинного обучения и многоцветной проточной цитометрии при миелодиспластическом синдроме, применение байесовской кластеризации данных проточной цитометрии для диагностики В-хронического лимфолейкоза. Для диагностики острого миелоидного и лимфобластного лейкоза применяются масштабируемое прогнозирование с использованием многомерного машинного обучения и транскриптомики крови, предварительно глубоко обученные сверточные нейронные сети, кластеризация и метод опорных векторов, гибридные иерархические классификаторы.

Имеются данные о полезности алгоритмов машинного обучения для получения диагностической и прогностической информации для минимальной остаточной болезни при остром миелоидном лейкозе и миелодиспластическом синдроме, используя данные многопараметрической проточной цитометрии. В модели использовались результаты 5000 образцов костного мозга от 1700 пациентов, и была достигнута AUC выше 0,90. Другие исследования применяли масс-спектрометрию с использованием искусственных нейронных сетей для достижения 100% чувствительности и 95% специфичности в диагностике множественной миеломы.

Учитывая достижения в области геномики и объем данных, генерируемых передовыми технологиями, машинное обучение и глубокое обучение предлагают аналитические инструменты для подхода и использования этих данных в клинической практике гематологов и онкологов. Так машинное и глубокое обучение использовались в классификации лимфомы с использованием профиля экспрессии генов и ДНК-микрочипов. Однако для построения моделей, которые будут иметь клиническое значение, требуются высококачественные исследования и хорошо зарекомендовавшие себя базы данных, рассмотренных в обзоре литературы [13]. Gal O. et al. (2019) использовали алгоритм k -ближайших соседей для прогнозирования полной ремиссии у пациентов с острым миелоидным лейкозом, используя около 75 генов. Модель смогла достичь AUC 0,81. Другие прогностические модели были применены при использовании данных однонуклеотидного полиморфизма для прогноза множественной миеломы, экспрессии генов для прогнозирования прогноза лимфомы Ходжкина, ансамблевого алгоритма для разработки риска инфекционных осложнений у пациентов с хроническим лимфоцитарным лейкозом.

В трансплантации гемопоэтических клеток применение искусственного интеллекта предназначено для прогнозирования, отбора пациентов до трансплантации и др. Реакция «трансплантат против хозяина» является серьезной проблемой после аллогенной трансплантации стволовых клеток. В исследовании Lee C. et al. (2018) [13] для создания модели для прогнозирования острой болезни «трансплантат против хозяина» был использован суперобучаемый алгоритм (сочетающий несколько алгоритмов), но достиг скромной AUC около 0,60, что указывает на важность выбора исходных данных при построении моделей. На сегодняшний день существует несколько баз данных по трансплантации, используемых Европейским обществом по трансплантации крови и костного мозга и Центром международных исследований по трансплантации крови и костного мозга. В работе Gunčar G. et al. (2018) [13] метод случайного леса (контролируемый алгоритм машинного обучения, объединяющий несколько деревьев решений) использовался для создания модели, способной анализировать закономерности между различными показателями крови, чтобы направлять клиницистов к пяти наиболее возможным гематологическим диагнозам (как доброкачественным, так и злокачественным) на ранней стадии исследования. Точность модели превзошла оценку врачей-терапевтов и была сравнима с точностью специалистов-гематологов. Область машинного обучения, особенно в гематологии, имеет потенциал для будущего воздействия. Этому следует способствовать путем разработки баз данных/узлов, более эффективного управления сбором данных и совершенствования существующих методов проведения исследований.

## Искусственный интеллект и машинное обучение в кардиологии, эндокринологии, нефрологии, урологии, нейрохирургии, диагностике инфекционных заболеваний

Улучшение раннего скрининга, диагностики и прогнозирования заболевания являются важными шагами в оказании медицинской помощи [18]. Пандемия Sars-CoV-2 и глобальная неготовность к ее решению стали важным стимулом для определения нового способа решения связанных с этим проблем; сложность проведения конкретных тестов побудила медицинское сообщество искать новые подходы. С тех пор как ВОЗ объявила вспышку COVID-19 пандемией, было прове-

дено несколько исследований с использованием методов искусственного интеллекта для оптимизации этих шагов в клинических условиях с точки зрения качества, точности, времени. Было определено 14 моделей для скрининга, 38 диагностических моделей для выявления COVID-19 и 50 прогностических моделей для отделения интенсивной терапии, потребности в ИВЛ, риска смертности, оценки тяжести, продолжительности пребывания в больнице. Исследования были основаны на медицинской визуализации, на использовании клинических параметров, результатов лабораторных исследований, демографических характеристиках. Было идентифицировано несколько гетерогенных предикторов, полученных из мультимодальных данных. Был проведен анализ этих мультимодальных данных, полученных из различных источников, с точки зрения значимости каждой категории включенных исследований. Был проведен анализ риска систематической ошибки (RoB) для изучения применимости включенных исследований в клинических условиях [6,17].

Модели машинного обучения при анализе биомаркеров хронического пародонтита показали вовлеченность в процесс остеодеструкции как инфекционной пародонтопатогенной составляющей, так и компонентов мукозальной иммунной системы [4]. Модель «случайный лес» показала связь иммунного ответа на хроническую инфекцию при остеопорозе [3].

Широту возможных приложений искусственного интеллекта показывает исследование с применением алгоритмов регрессии и алгоритмов классификации для прогнозирования уровня ферритина и повышения его точности, поскольку на него обычно влияют многие другие биологические процессы. Полученная модель достигла точности AUC 0,97. Bigorra L. et al. (2019) [13] применили машинное обучение с использованием случайного леса, наивного байесовского классификатора, k -ближайших соседей, нейронных сетей для построения модели, использующей данные клеточной популяции для улучшения обнаружения заболеваний печени и анемии в образцах с аномальными диаграммами рассеяния.

В кардиологии машинное обучение может предсказать выживаемость пациентов с сердечной недостаточностью только на основе сывороточного креатинина и фракции выброса. В эндокринологии клиническое исследование на основе модели глубокого обучения показало, что развитие диабетической полинейропатии связано со снижением относительного количества нейтрофилов (AUC=0,8988). Корреляционный анализ показал, что возраст пациента и средний объем тромбоцитов имеют положительную корреляцию с полинейропатией.

В области нефрологии у многих пациентов, нуждавшихся в гемодиализе и получавших стимуляторы эритропоэза в связи с терминальной стадией почечной недостаточности, наблюдается феномен циклического гемоглобина. Lobo B. et al. (2020), используя данные по 1972 пациентам, создали систему, которая может предсказать тенденцию гемоглобина в зависимости от терапии, чтобы позволить клиницистам предсказать, что произойдет, если они проведут или не назначат запланированную терапию [13].

Модели логистической регрессии используются в урологии для динамического прогнозирования реадмиссии с использованием рутинных послеоперационных лабораторных результатов после радикальной цистэктомии. Kirk P. S. et al. (2020) использовали данные 996 пациентов, чтобы оценить, может ли интеграция обычных послеоперационных данных в прогностической модели 30-дневной повторной госпитализации после цистэктомии улучшить ее прогностическую эффективность [13]. Множественные модели логистической регрессии были обучены с использованием различных комбинаций переменных и пороговых значений, а клинические данные использовались для изучения влияния на риск повторной госпитализации. Наиболее значимыми были лейкоциты, бикарбонат, мочевина, креатинин у повторно госпитализированных пациентов, чем у негоспитализированных повторно.

В нейрохирургии модель машинного обучения смогла предсказать возможность послеоперационной гипонатриемии после резекции поражений гипофиза (т.е. сывороточный натрий <130 ммоль/л в течение 30 дней после операции). Были определены наиболее важные характеристики: предоперационные концентрации пролактина, инсулиноподобного фактора роста 1 (IGF-1), натрия в сыворотке крови и индекс массы тела. Чувствительность составила 81,4%, специфичность — 77,5% [16].

## Машинное обучение в фармакологии

В области фармакологии модель машинного обучения, сочетающая функции алгоритмов с различными аналитическими методологиями для обнаружения сигналов побочных реакций на лекарства, связанных с лабораторными событиями. Jeong et al. (2018) решали задачу выявления и оценки нежелательных реакций на лекарственные препараты с помощью модели машинного обучения, объединяющей уже существующие алгоритмы на основе электронной медицинской карты и справочного набора данных из 1674 пар лекарство-событие (778 с известными ассоциациями и 896 с неизвестными ассоциациями). AUROC составляли от 0,629-0,709 до 0,737-0,816 [16].

## Машинное обучение в лабораторной медицине — метод сокращения числа лабораторных анализов

Неправильный выбор лабораторных тестов в виде чрезмерного или недостаточного использования часто имеет место, несмотря на имеющиеся руководства. Среди лабораторных специалистов, а также клиницистов существует широкое одобрение того, что стратегии управления спросом являются полезными инструментами, позволяющими избежать этой проблемы. Большинство этих инструментов основаны на автоматизированных алгоритмах или других типах машинного обучения.

Серийное лабораторное тестирование является обычным явлением, особенно в отделениях интенсивной терапии. Такое повторное тестирование стоит дорого и может даже навредить пациентам. Однако определение конкретных тестов, которые можно пропустить, является сложной задачей. Пространство поиска различных лабораторных тестов велико, и оптимальное сокращение трудно определить без моделирования временной траектории решений, что является нетривиальной задачей оптимизации. Данные в таких ситуациях могут быть получены из лабораторной информационной системы, интегрированной с медицинской информационной системой [2]. Стимулирующая область исследований представлена способностью моделей машинного обучения предсказывать значения лабораторных испытаний без их выполнения. Yu L. et al. (2020) [13] разработали модель нейронной сети с целью сократить количество выполняемых тестов, потеряв лишь небольшой процент точности. Хотя приложение предназначено для повторных лабораторных тестов, появилась возможность пропустить 15% лабораторных тестов с потерей точности прогноза <5%.

## Значение машинного обучения для здравоохранения, образования и общества

Интеграция между традиционной статистикой и машинным обучением является ключевым фактором, убеждающим клиницистов и исследователей в том, что алгоритмы машинного обучения основаны на основных традиционных статистических идеях; таким образом, их можно использовать для дополнения анализа данных с использованием традиционной статистики. Алгоритмы машинного обучения широко используются для разработки инструментов поддержки принятия клинических решений. Эти алгоритмы объединяют четыре шага (сбор данных, генерация гипотез, интерпретация данных и оценка гипотез) традиционного принятия решений в один. Преимущества алгоритмов машинного обучения в медицинской информатике зависят от целей исследования и типов используемых данных. Алгоритмы машинного обучения, такие как дерево решений, случайный лес, повышение градиента, регрессия, машина опорных векторов и искусственные нейронные сети, подходят для медицинской информатики, поскольку они способны обрабатывать большие данные, комбинацию числовых и категориальных данных и пропущенных значений. Кроме того, эти алгоритмы генерируют визуализации, которые могут быть автоматически преобразованы (интегрированы в инструменты) для использования клиницистами в качестве рекомендаций для пациентов [15].

В любом анализе машинного обучения по-прежнему требуются эксперты в предметной области, чтобы повысить надежность машины и понять результаты. В частности, в медицинской информатике решение клиницистов о состоянии здоровья конкретного пациента играет важную роль в предоставлении пациенту рекомендаций. В соответствии с принципом доказательной медицины принятие решений на основе данных и проверки должно быть более гибким и гибким, чтобы лучше преобразовывать базовые знания о сложностях в растущие достижения.

Подход машинного обучения может быть преобразован в руководство для обучения студентов-медиков, преподавателей и исследователей для проведения исследований в том числе и групповых исследований в любой области здравоохранения. Специалисты по биостатистике могут рассмотреть возможность использования методов машинного обучения и автоматизированных инструментов вместе с традиционной статистикой, чтобы повысить эффективность аналитики и надежность результатов. Интеграция между статистикой и машинным обучением может помочь специалистам по биостатистике получить новые результаты исследований [12].

Интеграция статистики и машинного обучения способствует не только дополнению данных, но и медицинской диагностике с использованием данных нескольких моделей. В будущем этот подход вместе с методами глубокого обучения возможно применить в биоинформатическом анализе с использованием геномных данных или комбинации геномных и клинических данных для улучшения автоматизированного процесса принятия решений. Глубокое обучение, являющееся одним из беспрецедентных технических достижений в исследованиях в области здравоохранения, помогает клиницистам понять роль искусственного интеллекта в принятии клинических решений.

Машинное обучение имеет дополнительное преимущество автоматизированного анализа, который можно

преобразовать в инструменты поддержки принятия решений, предоставляя удобные интерфейсы на основе интерактивных визуализаций и настройки значений данных. Такие инструменты могут помочь клиницистам взглянуть на данные с разных точек зрения, что поможет им принимать более обоснованные решения. Несмотря на споры между традиционной статистикой и машинным обучением, их интеграция ускоряет время принятия решений, обеспечивает автоматизированное принятие решений и повышает объяснимость. Этот обзор предполагает, что клиницисты могут рассмотреть возможность интеграции машинного обучения с традиционной статистикой для получения дополнительных преимуществ. И машинное обучение, и традиционная статистика лучше всего интегрируются для создания мощных автоматизированных инструментов принятия решений, не ограничивающихся клиническими данными, но также и для анализа биоинформатики.

## Заключение

Машинное обучение на основе искусственного интеллекта как развивающаяся дисциплина продемонстрировала большие перспективы для развития клинической лабораторной диагностики. Хотя машинное обучение все еще находится на ранних стадиях, оно используется для автоматизации лабораторных задач, оптимизации использования и предоставления персонализированных эталонных диапазонов и интерпретации тестов.

Искусственный интеллект и машинное обучение — многообещающее направление лабораторной диагностики в области гематологии, но использование ИИ пока ограничено с точки зрения качества и количества исходных данных. «Большие данные» трудно получить без достаточных баз данных с внутренне однородными данными, особенно в диагностике гистопатологии. Новые базы данных следует разрабатывать с такими процессами управления и сбора данных, которые позволяют оптимально использовать алгоритмы машинного обучения. Так Американское общество гематологов в 2017 г. начало создание центра данных для использования больших данных.

Внедрение инструментов машинного обучения в медицинскую практику сталкивается со многими проблемами. Технические препятствия включают сбор данных и потребность в «больших данных» для создания репрезентативных алгоритмов.

Основными ограничениями в этом направлении являются:

- нехватка специалистов;
- ограниченность ресурсов здравоохранения;
- постоянно растущий объем доступных медицинских данных, включающий цифровые изображения, омиксные данные, клинические записи и демографическую информацию о пациентах;
- повышенная сложность, возникающая при управлении и интеграции данных из разных источников;
- алгоритмы, основанные на машинном обучении, должны быть эффективно использованы для обработки больших данных.
- Отдельные лица не должны подвергаться окончательному решению, основанному исключительно на автоматизированной обработке или машинном обучении с использованием алгоритмов, но в равной степени важна интеграция статистики и принятия решений человеком.

Выбор системы машинного обучения не является кратчайшим путем для получения более простых и лучших результатов по сравнению с традиционными методами статистики. Хорошая система машинного обучения требует адекватного объема данных, надлежащего качества данных и надежного управления недостающими значениями (т.е. вменения данных), обоснованного предварительного выбора переменных для ввода в систему и правильного использования набора для обучения, набора для проверки (или настройки) и набора для тестирования. Имеющиеся в настоящее время данные показывают ограниченное количество проспективных исследований в области ИИ без каких-либо доказательств улучшения клинических результатов. Модели машинного обучения, представленные в различных исследованиях, не воспроизводимы, что ограничивает их применимость. Точный алгоритм машинного обучения может быть невоспроизводим на другой популяции/наборе данных. Затраты на воспроизведение/воспроизведение алгоритмов огромны, что представляет собой сложный аспект их реализации в реальной практике. Однако проблемы, стоящие перед интеграцией машинного обучения, многогранны. Когда данные доступны, создание модели может быть технически осуществимо, но необходимо учитывать многие другие вопросы. Например, важным вопросом, требующим решения, является этика обмена данными и использования информации о пациентах. Эти вопросы требуют регулирования для избежания неправильного использования данных, увеличения систематической ошибки и др.

Тем не менее, машинное обучение будет приобретать все большее значение для лабораторной медицины. Лаборатории будущего будут использовать эти методы для значительного повышения эффективности и точности диагностики. Машинное обучение обещает захватывающие достижения в медицине, но его применение в лабораторной медицине все еще находится

на стадии становления. Поскольку область является молодой, существует дополнительная потребность в стандартизации того, как разрабатываются и представляются эти алгоритмы. Также ведется постоянная работа по использованию машинного обучения для оптимизации использования лабораторий.

## Список литературы


1. Бонгард М. М. Проблема узнавания.— М.: Физматгиз, 1967.- 321 с.

2. Казаков С.П., Скворцов С.В., Тарасов А.К., Изгородин А.С., Тихонов Ю.Г., Карпов В.О. Информационные технологии в лабораторном отделении многопрофильного лечебного учреждения // Клиническая лабораторная диагностика. -2000. - № 9. - С. 21-22.

3. Мудров В.П., Йовичич М. Машинное обучение в поиске прогностически значимых тестов для лабораторной диагностики остеопороза // Справочник заведующего КДЛ. -2020. - №1. – С. 44-56

4. Мудров В.П. Модели машинного обучения при анализе биомаркеров хронического пародонтита // Медицинский алфавит. Современная лаборатория (2) № 19 / 2022 С.43-47 DOI: 10.33667/2078-5631-2022-19-55-59

5. Тарасов А.К., Казаков С.П. Основные направления и перспективы развития информационных лабораторных технологий в многопрофильном лечебном учреждении // Клиническая лабораторная диагностика. - 2003. - № 9. - С. 24.

6. Adamidi E.S., Mitsis K., Nikita K.S. Artificial intelligence in clinical care amidst COVID-19 pandemic: A systematic review // ComputStructBiotechnol J.- 2021. – Vol. 19. – P.2833-2850. doi: 10.1016/j.csbj.2021.05.010.

7. Cui M., Zhang D. Artificial intelligence and computational pathology // Lab Invest.- 2021. – Vol.101. – P.:412-422. doi: 10.1038/s41374-020-00514-0.

8. Deo R.C. Machine learning in medicine// Circulation. - 2015/ - Vol. 132. - P. 1920–1930. doi: 10.1161/CIRCULATIONAHA.115.001593

9. Dhillon S., Ganggayah M., Sinnadurai S., Lio P., Taib N. Theory and practice of integrating machine learning and conventional statistics in medical data analysis // Diagnostics (Basel). 2022 Oct 18;12(10):2526. doi: 10.3390/diagnostics12102526

10. Douglas M. Machine intelligence in cardiovascular medicine // Cardiology in review. - 2020. – Vol. 28 , Iss. 2. – P. 53-64. doi: 10.1097/CRD.0000000000000294

11. Handelman G., Kok H., Chandra R., Razavi A., Lee M., Asadi H. eDoctor: machine learning and the future of medicine // Journal of Internal Medicine. – 2018. - Vol.284. P:603-619 https://doi.org/10.1111/joim.12822

12. Mrazek C., Haschke-Becher E., Felder T.K., Keppel, Oberkofler H., Cadamuro J. Laboratory demand management strategies-an overview // Diagnostics (Basel). 2021 Jun 23;11(7). P:1141. doi: 10.3390/diagnostics11071141.

13. Muhsen I.N., Shyr D., Sung A.D., Hashmi S.K. Machine learning applications in the diagnosis of benign and malignant hematological diseases // Clin Hematol Int. 2021 Mar; 3(1): 13–20.doi: 10.2991/chi.k.201130.001

14. Rabbani N., Kim G., Suarez C., Chen J. Applications of machine learning in routine laboratory medicine: Current state and future directions // Clin Biochem. - 2022. – Vol.103. – P.1-7. doi: 10.1016/j.clinbiochem.2022.02.011.

15. Rashidi H., Tran N., Albahra S., Dang L. Machine learning in health care and laboratory medicine: General overview of supervised learning and Auto-ML // The International Journal of Laboratory Hematology. - 2022. - Vol. 43, Iss. S1. - P. 15–22. doi.org/10.1111/ijlh.13537

16. Ronzio L., Cabitza F., Barbaro A., Banfi G. Has the flood entered the basement? A systematic literature review about machine learning in laboratory medicine // Diagnostics (Basel). 2021 Feb; 11(2): 372. doi: 10.3390/diagnostics11020372

17. Smith K., Kirby J. Image analysis and artificial intelligence in infectious disease diagnostics // Clin Microbiol Infect. – 2020/ - Vol. 26. - P.1318-1323. doi: 10.1016/j.cmi.2020.03.012.

18. Tran N., Albahra S., May L., Waldman S., Crabtree S., Bainbridge S., Rashidi H. Evolving applications of artificial intelligence and machine learning in infectious diseases testing // Clin Chem. – 2022. – Vol. 68. – P.125–133. doi: 10.1093/clinchem/hvab239

19. Wen X., Leng P., Wang J., Yang G., Zu R. et al. Clinlabomics: leveraging clinical laboratory data by data mining strategies // BMC Bioinformatics. 2022 Sep 24;23(1):387. doi: 10.1186/s12859-022-04926-1