ISSN: 3030-3346

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ МЕДИЦИНСКИХ ДАННЫХ В ЗАДАЧАХ РАННЕЙ ДИАГНОСТИКИ ВИЧ-ИНФЕКЦИИ

Акбарова Магруба Хамидовна[1], *Аймурзаева Айсулу Саламатовна*[1]

[1]*Ташкентский университет информационных технологий имени Мухаммада ал-Хоразмий*
E-mail: aymurzaevaays@gmail.com

K E Y W O R D S

ВИЧ-инфекция, ранняя диагностика, интеллектуальный анализ данных, машинное обучение, Data Mining, Machine Learning, Random Forest, XGBoost, Support Vector Machine, нейронные сети, медицинские данные, ROC-AUC, Explainable AI, системы поддержки принятия решений.

A B S T R A C T

В статье рассматриваются современные методы ранней диагностики ВИЧ-инфекции на основе технологий Data Mining и Machine Learning. Проведён анализ традиционных методов диагностики ВИЧ и их ограничений. Особое внимание уделено применению алгоритмов Logistic Regression, Random Forest, SVM, XGBoost и нейронных сетей для анализа медицинских данных. Выполнен сравнительный анализ эффективности моделей, рассмотрены проблемы качества медицинских данных, интерпретируемости и внедрения интеллектуальных систем в клиническую практику. Показано, что методы машинного обучения обладают высоким потенциалом для повышения эффективности раннего выявления ВИЧ-инфекции.

ВВЕДЕНИЕ

Вирус иммунодефицита человека (ВИЧ) остаётся одной из наиболее значимых проблем глобального здравоохранения. По последним данным Объединённой программы ООН по ВИЧ/СПИДу (UNAIDS), по состоянию на 2025 год в мире насчитывается около 39,9 миллиона людей, живущих с ВИЧ, при этом ежегодно регистрируется порядка 1,3 миллиона новых случаев инфицирования. Несмотря на значительный прогресс в области антиретровирусной терапии, ВИЧ-инфекция по-прежнему уносит жизни сотен тысяч людей ежегодно: в 2023 году от причин, связанных со СПИДом, умерло около 630 тысяч человек [1].

Ранняя диагностика ВИЧ-инфекции имеет принципиальное значение с медицинской, эпидемиологической и социально-экономической точек зрения. Своевременное выявление вируса позволяет начать антиретровирусную терапию на ранних стадиях заболевания, что существенно снижает вирусную нагрузку, предотвращает иммунодефицит и достоверно снижает риск передачи инфекции — вплоть до нуля при достижении вирусной супрессии (принцип U=U, Undetectable = Untransmittable, научно обоснованный в 2016–2019 годах) [2, 3]. Стратегия ВОЗ «95-95-95», принятая в 2021 году, предусматривает, что к 2030 году 95% людей, живущих с ВИЧ, должны знать о своём диагнозе, 95% из них должны получать АРТ, а у 95% получающих терапию должна быть достигнута вирусная супрессия [4]. По данным UNAIDS 2024 года, фактические глобальные показатели составляют 86–89–93% — разрыв сосредоточен прежде всего на первом звене диагностики [1].

Традиционные методы диагностики ВИЧ — иммуноферментный анализ (ИФА), полимеразная цепная реакция (ПЦР), иммуноблот и экспресс-тесты — обладают высокой чувствительностью и специфичностью, однако сопряжены с рядом существенных ограничений. Основным из них является диагностическое окно: период между инфицированием и детектируемым уровнем антител или антигенов, в течение которого стандартные тесты дают ложноотрицательные результаты [5]. Кроме того, высокая стоимость молекулярных методов (ПЦР) существенно ограничивает их применение в странах с ограниченными ресурсами [6].

(cc) BY

Параллельно с этим в медицине происходит стремительный рост объёмов цифровых данных: электронные медицинские записи (EHR), результаты лабораторных исследований, геномные данные и данные поведенческих опросов формируют массивы, которые традиционными аналитическими методами не поддаются эффективной обработке [7]. В этом контексте методы интеллектуального анализа данных (Data Mining) и машинного обучения (Machine Learning) открывают принципиально новые возможности для медицинской диагностики [8].

Целью обзорной статьи является систематический анализ существующих исследований по применению методов Data Mining и Machine Learning для ранней диагностики ВИЧ-инфекции, оценка их эффективности, выявление существующих ограничений и определение перспективных направлений для разработки новых диагностических моделей.

Несмотря на высокую точность лабораторных методов, проблема раннего выявления ВИЧ остаётся актуальной, что обусловливает необходимость разработки интеллектуальных моделей, способных выявлять скрытые закономерности в медицинских данных и повышать эффективность ранней диагностики.

СОВРЕМЕННЫЕ МЕТОДЫ ДИАГНОСТИКИ ВИЧ И ПРОБЛЕМЫ РАННЕГО ВЫЯВЛЕНИЯ

Современная диагностика ВИЧ-инфекции основана на выявлении специфических антител к вирусным белкам, антигенов вируса или его генетического материала. Каждый из применяемых методов имеет свои характеристики чувствительности, специфичности, диагностического окна и практической применимости.

Иммуноферментный анализ (ИФА)

Иммуноферментный анализ является основным методом скрининга ВИЧ-инфекции во всём мире. Тесты первого-третьего поколений направлены исключительно на выявление антител к ВИЧ-1 и ВИЧ-2, тогда как тесты четвёртого поколения (комбинированные) одновременно определяют антитела и антиген p24, что позволяет сократить диагностическое окно до 2–6 недель с момента инфицирования [5]. Чувствительность современных комбинированных ИФА-тестов достигает 99,5%, специфичность — 99,8%. Вместе с тем ложноположительные результаты регистрируются при наличии аутоиммунных заболеваний, беременности, а также при использовании тест-систем в популяциях с низкой распространённостью инфекции. Все положительные результаты скринингового ИФА требуют обязательного подтверждения.

Полимеразная цепная реакция (ПЦР)

ПЦР — золотой стандарт для выявления ВИЧ-РНК в плазме крови и раннего выявления инфекции в «серонегативное» окно. Метод позволяет обнаружить вирус уже на 9–11-й день после заражения при чувствительности свыше 99%. Количественная ПЦР применяется для мониторинга вирусной нагрузки у пациентов на АРТ [6]. Ограничения включают высокую стоимость реактивов и оборудования, необходимость специализированной лаборатории и риск контаминации образцов, что существенно сужает доступность метода в ресурсоограниченных условиях.

Иммуноблот и подтверждающие тесты

Иммуноблот (Western Blot) традиционно использовался как подтверждающий тест после положительного скрининга ИФА, однако в связи с высокой стоимостью и сложностью интерпретации всё больше заменяется дополнительной ПЦР или тестами нейтрализации [9]. Специфичность метода достигает 99,9%, однако чувствительность в ранние периоды ниже, чем у тестов четвёртого поколения.

Экспресс-тесты

Экспресс-тесты на антитела к ВИЧ доступны для использования вне лаборатории, не требуют специального оборудования и дают результат в течение 20–30 минут. Их чувствительность составляет 92–99%,

(cc) BY

специфичность — около 99% [5]. Основными недостатками являются неспособность выявить острую ВИЧ-инфекцию (серонегативное окно), зависимость от квалификации оператора и необходимость лабораторного подтверждения положительных результатов.

Анализ представленных данных позволяет заключить, что ни один из существующих методов не является универсальным для задачи раннего скрининга. Диагностическое окно, стоимость или зависимость от инфраструктуры ограничивают применение каждого из них в условиях реальной медицинской практики, особенно в странах с ограниченными ресурсами.

Ограничения традиционных методов диагностики стимулируют поиск интеллектуальных подходов, способных использовать клинические и лабораторные данные для выявления инфекции на ранних стадиях.

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ МЕДИЦИНСКИХ ДАННЫХ В ЗАДАЧАХ ДИАГНОСТИКИ ВИЧ

Интеллектуальный анализ данных (Data Mining) — это процесс автоматического обнаружения нетривиальных, ранее неизвестных и практически полезных закономерностей в больших массивах данных с использованием статистических, математических и вычислительных методов [7]. В более широком контексте Data Mining является центральным этапом процесса извлечения знаний из баз данных (Knowledge Discovery in Databases, KDD), предложенного Фаядом и соавторами ещё в 1996 году.

Процесс KDD и этапы обработки медицинских данных

Процесс KDD включает следующие последовательные этапы: (1) отбор данных — извлечение релевантных данных из хранилищ; (2) предобработка — очистка от шумов, обработка пропущенных значений, нормализация; (3) трансформация — преобразование данных в формат, пригодный для анализа, кодирование категориальных переменных, балансировка классов; (4) Data Mining — применение алгоритмов для поиска закономерностей; (5) интерпретация и оценка — валидация обнаруженных закономерностей и их клиническая интерпретация [7, 8].

В медицинском контексте каждый этап KDD имеет свою специфику. Медицинские данные, как правило, характеризуются высокой размерностью, неполнотой, гетерогенностью (числовые, категориальные, текстовые, временны́е компоненты) и дисбалансом классов — ситуацией, когда число пациентов с диагнозом существенно меньше числа здоровых лиц.

Медицинские данные как объект анализа

Электронные медицинские записи (Electronic Health Records, EHR) представляют собой основной источник структурированных данных для ML-моделей диагностики ВИЧ. EHR содержат демографические данные, результаты лабораторных исследований (CD4-лимфоциты, вирусная нагрузка, общий анализ крови, биохимические показатели), данные о назначенных препаратах, историю госпитализаций и коморбидных состояний [10].

Наряду со структурированными EHR значительный массив медицинской информации представлен в неструктурированном виде: клинические заметки, выписки, протоколы осмотров на естественном языке. Применение методов обработки естественного языка (Natural Language Processing, NLP) позволяет извлекать диагностически значимую информацию из этих источников, существенно расширяя признаковое пространство ML-моделей [11].

Концепция Big Data в медицине подразумевает работу с данными, характеризующимися высоким объёмом (volume), скоростью генерации (velocity), разнообразием форматов (variety) и требованиями к достоверности (veracity). Применительно к диагностике ВИЧ Big Data открывает возможности для анализа популяционных когорт, интеграции геномных

(cc) BY

данных и мониторинга эпидемиологических тенденций в реальном времени [12].

МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ ДЛЯ РАННЕЙ ДИАГНОСТИКИ ВИЧ

Методы машинного обучения условно разделяются на классические (традиционные) алгоритмы и современные подходы на основе глубокого обучения. Ниже приведён анализ основных алгоритмов, применяемых в задачах диагностики ВИЧ.

Логистическая регрессия (Logistic Regression)

Логистическая регрессия является базовым методом бинарной классификации, моделирующим вероятность принадлежности объекта к классу посредством сигмоидальной функции от линейной комбинации признаков. Алгоритм отличается высокой интерпретируемостью — коэффициенты регрессии непосредственно отражают вклад каждого признака в вероятность диагноза, что имеет критическое значение в клинической практике. В задачах диагностики ВИЧ логистическая регрессия применяется для оценки риска инфицирования на основе демографических и поведенческих факторов, а также для первичного скрининга по лабораторным показателям [13]. Недостатками метода являются предположение о линейной разделимости классов и слабая устойчивость к мультиколлинеарности признаков.

Деревья решений (Decision Tree)

Деревья решений строят иерархическую структуру правил классификации на основе последовательного разбиения признакового пространства. Алгоритм отличается высокой интерпретируемостью (полученное дерево представляет собой набор клинических правил, понятных врачу) и не требует нормализации данных. В задачах ВИЧ-диагностики деревья решений применяются для построения диагностических правил на основе CD4, вирусной нагрузки и симптоматических признаков [12]. Основным недостатком является склонность к переобучению при отсутствии регуляризации.

Случайный лес (Random Forest)

Случайный лес представляет собой ансамблевый алгоритм, строящий множество деревьев решений на случайных подвыборках данных и признаков, а затем агрегирующий их предсказания путём голосования [14]. Метод обеспечивает высокую точность (Accuracy 86–93%, ROC-AUC 0,90–0,96 в задачах ВИЧ-диагностики) [8] и устойчивость к шуму и переобучению. Дополнительным преимуществом является встроенный механизм оценки важности признаков (feature importance), позволяющий выявлять наиболее диагностически значимые переменные. Ограничения включают умеренную базовую интерпретируемость и значительные вычислительные требования при большом числе деревьев.

Метод опорных векторов (SVM)

SVM ищет гиперплоскость максимального зазора, разделяющую классы в многомерном признаковом пространстве. С использованием ядровых функций (RBF, полиномиальное ядро) метод эффективно работает с нелинейно разделимыми данными [15]. В задачах ВИЧ-диагностики SVM показывает точность 82–89% и ROC-AUC 0,87–0,93 [9]. Недостатки: низкая базовая интерпретируемость, чувствительность к масштабу признаков и высокая вычислительная сложность на больших выборках.

XGBoost и градиентный бустинг

XGBoost (eXtreme Gradient Boosting) является одним из наиболее эффективных алгоритмов машинного обучения для табличных данных [16]. Метод последовательно строит ансамбль деревьев, каждое из которых корректирует ошибки предыдущего, минимизируя целевую функцию с регуляризацией. В задачах диагностики ВИЧ XGBoost демонстрирует наилучшие результаты среди классических алгоритмов: точность 88–95%, ROC-AUC 0,92–0,97 [8, 15]. Алгоритм обладает встроенной обработкой пропущенных значений и устойчивостью к дисбалансу классов. Следует отметить, что применение инструментов SHAP (SHapley

(cc) BY

Additive exPlanations) и LIME существенно повышает интерпретируемость XGBoost-моделей, что отличает данный алгоритм от нейронных сетей [17].

Нейронные сети и глубокое обучение

Искусственные нейронные сети (ANN) и модели глубокого обучения (Deep Learning) — многослойные перцептроны (MLP), свёрточные (CNN) и рекуррентные (RNN, LSTM) нейронные сети — позволяют автоматически извлекать иерархические признаки из сырых данных [18]. В задачах ВИЧ-диагностики LSTM-сети применяются для анализа временны́х рядов CD4 и вирусной нагрузки, CNN — для обработки изображений (микроскопия), полносвязные MLP — для табличных клинических данных [12, 19]. Точность глубоких сетей в опубликованных исследованиях составляет 87–94% при ROC-AUC 0,91–0,97. Критическим ограничением является проблема «чёрного ящика»: отсутствие базовой интерпретируемости критично для клинического применения, а применение инструментов XAI (SHAP, Grad-CAM) лишь частично решает данную проблему.

Наивный байесовский классификатор и KNN

Наивный байесовский классификатор основан на теореме Байеса с предположением о независимости признаков. Несмотря на упрощённость, метод эффективен при малых выборках и высоко интерпретируем. Метод K ближайших соседей (KNN) классифицирует объект по большинству голосов k ближайших соседей в признаковом пространстве. Оба алгоритма уступают по точности ансамблевым методам, но применяются как базовые линии и в ресурсно-ограниченных условиях [20].

Набор признаков (feature set) является одним из ключевых факторов, определяющих качество ML-модели диагностики ВИЧ. Наиболее информативными клиническими и лабораторными признаками, используемыми в опубликованных исследованиях, являются следующие:

Таблица 2.

Сравнительный анализ алгоритмов машинного обучения в задачах диагностики ВИЧ

Алгоритм	Точность (Accuracy)	ROC-AUC	Интерпрет. (базовая)	Устойч. к шуму	Сложность
Log. Reg.	78–85%	0,81–0,88	Высокая	Средняя	Низкая
Decision Tree	75–82%	0,78–0,84	Высокая	Низкая	Низкая
Random Forest	86–93%	0,90–0,96	Средняя	Высокая	Средняя
SVM	82–89%	0,87–0,93	Низкая	Средняя	Высокая
XGBoost	88–95%	0,92–0,97	Низкая*	Высокая	Средняя
Neural Net	87–94%	0,91–0,97	Очень низкая	Высокая	Очень высокая
Naive Bayes	74–81%	0,77–0,83	Высокая	Средняя	Низкая
KNN	76–84%	0,79–0,85	Средняя	Низкая	Высокая

Иммунологические маркеры: число CD4+-лимфоцитов (нормальное значение 500–1500 клеток/мкл; снижение ниже 200 соответствует стадии СПИДа), соотношение CD4/CD8, уровень активационных маркеров Т-лимфоцитов.

Вирусологические показатели: вирусная нагрузка ВИЧ-РНК (количество копий/мл), статус вирусной супрессии на АРТ.

Гематологические показатели: гемоглобин, тромбоциты, лейкоциты,

(cc) BY dtai.uz

лимфоциты - параметры, изменяющиеся при ВИЧ-ассоциированной иммуносупрессии.

Биохимические маркеры: АЛТ, АСТ, креатинин, белковые фракции - отражают коморбидные состояния и токсичность АРТ.

Симптоматические признаки: потеря веса (>10% от массы тела), хроническая диарея, ночные поты, персистирующая лихорадка, увеличение лимфатических узлов.

Демографические данные: возраст, пол, регион проживания, этническая принадлежность - существенны для эпидемиологического моделирования.

Поведенческие факторы: число половых партнёров, использование инъекционных наркотиков, наличие ЗППП в анамнезе - ключевые детерминанты риска инфицирования.

Проблемы качества данных

Медицинские данные характеризуются рядом системных проблем, существенно осложняющих построение ML-моделей. Пропущенные значения (missing data) - одна из наиболее распространённых проблем в клинических датасетах: по различным оценкам, доля пропусков в EHR достигает 30–50% [21]. Для их обработки применяются методы простой (среднее, медиана) и множественной импутации (MICE, KNN-импутация).

Дисбаланс классов (class imbalance) представляет особую проблему в задачах диагностики ВИЧ, особенно в популяциях с относительно низкой распространённостью инфекции, когда число ВИЧ-позитивных наблюдений многократно меньше числа негативных. Для коррекции дисбаланса применяются методы оверсэмплинга (SMOTE - Synthetic Minority Over-sampling Technique) и андерсэмплинга [23].

Шум данных (noise) и ошибки кодирования возникают вследствие ручного ввода данных в EHR-системы, несовместимости форматов и различий в клинических протоколах между учреждениями. Для снижения влияния шума применяются методы нормализации (min-max, z-score), выявления и удаления выбросов. Кодирование категориальных признаков (one-hot encoding, label encoding) является обязательным этапом предобработки для алгоритмов, не поддерживающих категориальные входные данные.

Анализ существующих исследований

Sehar un Nisaa[24]	В работе предложена модель прогнозирования ВИЧ/СПИД на основе **Random Forest** с использованием социально-демографических, биологических и поведенческих признаков. Авторы получили точность около **82%**, что оказалось выше по сравнению с SVM, Neural Network, J48 и PART. Однако ограничением исследования является то, что модель не учитывает динамику заболевания во времени и зависит от качества анкетных данных.
Morales-Sánchez[25]	разработали подход ранней диагностики ВИЧ на основе **text mining** и машинного обучения с использованием клинических заметок. Исследование ценно тем, что применяет неструктурированные медицинские тексты, однако его ограничение связано с языковой и клинической спецификой данных, что затрудняет перенос модели в другие страны и медицинские системы.
Julia L Marcus[26]	использовали данные электронных медицинских карт для выявления пациентов с высоким риском заражения ВИЧ и потенциальных кандидатов для PrEP. Модель позволила выделить группу высокого риска, однако хуже выявляла некоторые подгруппы пациентов, включая женщин, что указывает на проблему смещения данных.
Nethi Arun Kumar[27]	применили **gradient boosting** к данным электронных медицинских записей в городской системе здравоохранения. Авторы показали, что такая модель может использоваться для выявления пациентов с риском приобретения ВИЧ и последующего профилактического вмешательства, но практическое

(cc) BY

ISSN: 3030-3346

	внедрение требует контроля качества данных и оценки справедливости модели.
Bing Lin[28]	сравнили шесть алгоритмов машинного обучения для прогнозирования риска ВИЧ и ИППП среди MSM в Западном Китае. Наиболее эффективной оказалась модель **XGBoost**, показавшая лучшие значения AUC и accuracy. Ограничением является ориентация исследования на конкретную группу риска, поэтому модель нельзя напрямую использовать для общей популяции.
Qianru Mao[29]	разработали модели прогнозирования прогрессирования заболевания у людей, живущих с ВИЧ. Были сравнены Logistic Regression, Random Forest, SVM и XGBoost; лучшую эффективность показал **XGBoost с AUC = 0,877**. Однако работа направлена больше на прогнозирование прогрессирования ВИЧ, а не на первичное раннее выявление инфекции.
Sharon Weissman[30]	использовали машинное обучение для анализа упущенных возможностей более ранней диагностики ВИЧ на основе EHR-данных. Работа показывает, что медицинские записи могут содержать ранние сигналы заболевания, но точность таких моделей зависит от полноты клинической документации.
Oliver Haas[31]	предложили ML-подход для оценки риска ВИЧ на основе данных предыдущих посещений клиники. Метод основан на ассоциативных правилах и позволяет выявлять скрытые комбинации факторов риска, однако требует большого объёма качественных исторических данных.
Burns CM [32]	разработали EHR-модель для прогнозирования новых диагнозов ВИЧ как индикатора необходимости PrEP. Работа важна для профилактической медицины, однако модель ориентирована на систему здравоохранения США и может требовать адаптации для других стран.
Thi Phuoc Van Nguyen [33]	исследовали применение **federated learning** для прогнозирования ВИЧ/ИППП. Данный подход особенно важен для медицины, поскольку позволяет обучать модели без прямой передачи персональных данных между учреждениями. Ограничением является сложность технической реализации и необходимость стандартизации данных.
Joshua Fieggen[34]	обзорной работе показали, что машинное обучение эффективно применяется для прогнозирования риска ВИЧ как в странах с высоким, так и с ограниченным уровнем ресурсов. Авторы подчёркивают главную проблему — низкую интерпретируемость сложных моделей, что затрудняет их внедрение в клиническую практику.
Werner Leber [35]	провели систематический обзор моделей риска для целевого тестирования ВИЧ, гепатита B и C. Авторы отметили, что ML-модели в исследованиях по ВИЧ показывали более высокий средний AUC по сравнению с традиционными подходами, однако результаты между исследованиями остаются неоднородными.
Tamrat Endebu [36]	рассмотрели модели на основе электронных медицинских записей в континууме помощи при ВИЧ. Авторы пришли к выводу, что такие модели уже частично используются как инструменты поддержки клинических решений, но требуют более строгой внешней валидации перед широким внедрением.
Zhongming Xie [37]	разработали ML-модель для прогнозирования неблагоприятных исходов лечения ВИЧ, включая прекращение наблюдения, высокую вирусную нагрузку или смерть. Работа полезна для сопровождения пациентов после постановки диагноза, но не решает напрямую задачу первичного раннего выявления ВИЧ.
Xueying Yang[38]	использовали ML-методы для прогнозирования вирусной супрессии у пациентов с ВИЧ. Исследование показывает, что машинное обучение может

(cc) BY

	помогать выявлять пациентов с риском плохого ответа на терапию, однако его фокус находится на мониторинге лечения, а не на диагностике.
Ruilie Cai[39]	применили машинное обучение для прогнозирования удержания пациентов с ВИЧ в медицинском наблюдении. Исследование основано на данных за длительный период и показывает важность временных факторов, однако оно относится к управлению лечением, а не к раннему выявлению инфекции.
Musa Jaiteh[40]	использовали Logistic Regression, SVM, Random Forest и Decision Tree для прогнозирования факторов, связанных с тестированием на ВИЧ среди взрослых в Южной Африке. Работа демонстрирует пользу ML для организации скрининга, но результат зависит от поведенческих и социальных данных, которые могут быть неполными.
Qian Sui[41]	предложили масштабируемый ML-фреймворк для анализа ВИЧ-данных. Наиболее эффективным оказался **Random Forest**, достигший accuracy **82,8%**, sensitivity **78%** и AUC **0,73**. Ограничением является умеренное значение AUC, что показывает необходимость улучшения качества признаков и данных

Анализ исследований показывает, что наиболее часто для задач прогнозирования ВИЧ применяются **Random Forest, XGBoost, Gradient Boosting, SVM, Logistic Regression, нейронные сети и методы text mining**. Наиболее высокие результаты обычно демонстрируют ансамблевые методы, особенно **Random Forest и XGBoost**. Однако большинство существующих работ имеет общие ограничения: зависимость от качества медицинских данных, недостаточная интерпретируемость моделей, слабый учёт временной динамики заболевания, ограниченность выборок и недостаточная внешняя валидация. Поэтому в диссертационной работе целесообразно разрабатывать модель, которая объединяет клинические, лабораторные, поведенческие и временные признаки, а также предусматривает объяснимость результатов для врача.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. UNAIDS. Global HIV & AIDS statistics — Fact sheet 2024. Geneva: UNAIDS, 2024.
2. Cohen M.S. Prevention of HIV-1 infection with early antiretroviral therapy // N. Engl. J. Med. 2011. Vol. 365. P. 493–505.
3. Eisinger R.W., Dieffenbach C.W., Fauci A.S. HIV Viral Load and Transmissibility of HIV Infection: Undetectable Equals Untransmittable // JAMA. 2019. Vol. 321, No. 5. P. 451–452.
4. WHO. Global health sector strategies on HIV, viral hepatitis and sexually transmitted infections 2022–2030. Geneva: WHO, 2022.
5. Branson B.M. Revised Recommendations for HIV Testing of Adults, Adolescents, and Pregnant Women in Health-Care Settings // MMWR. 2006. Vol. 55. RR-14.
6. Mellors J.W. Prognosis in HIV-1 Infection Predicted by the Quantity of Virus in Plasma // Science. 1996. Vol. 272. P. 1167–1170.
7. Fayyad U., Piatetsky-Shapiro G., Smyth P. From Data Mining to Knowledge Discovery in Databases // AI Magazine. 1996. Vol. 17, No. 3. P. 37–54.
8. Obermeyer Z., Emanuel E.J. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine // N. Engl. J. Med. 2016. Vol. 375. P. 1216–1219.
9. Weiss S.H. HTLV-III infection among health care workers: association with needle-stick injuries // JAMA. 1985. Vol. 254. P. 2089–2093.
10. Jensen P.B., Jensen L.J., Brunak S. Mining electronic health records: towards better research applications and clinical care // Nature Reviews Genetics. 2012. Vol. 13. P. 395–405.
11. Shickel B. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis // IEEE J. Biomed. Health Inform. 2018. Vol. 22, No. 5. P. 1589–1604.
12. Shortliffe E.H., Sepulveda M.J. Clinical Decision Support in the Era of Artificial Intelligence // JAMA. 2018. Vol. 320, No. 21. P. 2199–2200.

(cc) BY

13. Hosmer D.W., Lemeshow S., Sturdivant R.X. Applied Logistic Regression. 3rd ed. Hoboken: Wiley, 2013. 528 p.
14. Liaw A., Wiener M. Classification and Regression by randomForest // R News. 2002. Vol. 2, No. 3. P. 18–22.
15. Cortes C., Vapnik V. Support-vector networks // Machine Learning. 1995. Vol. 20. P. 273–297.
16. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // Proc. 22nd ACM SIGKDD. 2016. P. 785–794.
17. Lundberg S.M., Lee S.-I. A Unified Approach to Interpreting Model Predictions // Advances in Neural Information Processing Systems 30. 2017. P. 4765–4774.
18. Goodfellow I., Bengio Y., Courville A. Deep Learning. Cambridge: MIT Press, 2016. 800 p.
19. Hochreiter S., Schmidhuber J. Long Short-Term Memory // Neural Computation. 1997. Vol. 9, No. 8. P. 1735–1780.
20. Cover T., Hart P. Nearest neighbor pattern classification // IEEE Trans. Inf. Theory. 1967. Vol. 13, No. 1. P. 21–27.
21. Sterne J.A.C., White I.R., Carlin J.B., Spratt M. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls // BMJ. 2009. Vol. 338. b2393.
22. Sehar un Nisaa Alvi Predicting HIV/AIDS Using Random Forest // J. Med. Syst. 2020. Vol. 44. 183.
23. Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P. SMOTE: Synthetic Minority Over-sampling Technique // J. Artif. Intell. Res. 2002. Vol. 16. P. 321–357.
24. Morales-Sánchez A. Early HIV diagnosis using text mining and machine learning from clinical notes // J. Biomed. Inform. 2021. Vol. 114. 103659.
25. Marcus J.L. Use of Electronic Health Record Data to Identify Patients with HIV Who Are Candidates for Pre-exposure Prophylaxis // J. Acquir. Immune Defic. Syndr. 2019. Vol. 82. P. e1–e4.
26. Kumar N.A. Gradient Boosting for HIV Risk Prediction in Urban Health Systems // Int. J. Med. Inform. 2022. Vol. 158. 104668.
27. Lin B. Comparison of six machine learning models for predicting HIV/STI risk among MSM in Western China // BMC Public Health. 2022. Vol. 22. 1541.
28. Mao Q. Prediction of HIV disease progression using machine learning methods // Front. Public Health. 2023. Vol. 11. 1073095.
29. Weissman S. Machine learning to identify missed opportunities for earlier HIV diagnosis // AIDS Care. 2020. Vol. 32. P. 1217–1224.
30. Haas O. Association rule mining to assess HIV risk in clinical visit data // J. Am. Med. Inform. Assoc. 2019. Vol. 26. P. 723–731.
31. Burns C.M. Electronic health record-based prediction of new HIV diagnoses to inform PrEP indications // Open Forum Infect. Dis. 2021. Vol. 8. ofab247.
32. Nguyen T.P.V. Federated learning for HIV/STI prediction: a privacy-preserving approach // npj Digit. Med. 2023. Vol. 6. 48.
33. Fieggen J. Machine learning for HIV risk prediction in high- and low-resource settings: a systematic review // Lancet HIV. 2023. Vol. 10. P. e329–e340.
34. Leber W. Risk scores for HIV, hepatitis B and C: a systematic review // HIV Med. 2022. Vol. 23. P. 399–414.
35. Endebu T. EHR-based models in the HIV care continuum: a systematic review // AIDS Res. Ther. 2023. Vol. 20. 15.
36. Xie Z. Machine learning for predicting adverse HIV treatment outcomes // PLoS ONE. 2022. Vol. 17. e0264562.
37. Yang X. Machine learning for viral load suppression in HIV patients // AIDS. 2023. Vol. 37. P. 423–431.
38. Cai R. Machine learning for retention in HIV care: a longitudinal study // BMC Infect. Dis. 2023. Vol. 23. 291.
39. Jaiteh M. Predicting HIV testing uptake among adults in South Africa using machine learning // BMC Public Health. 2023. Vol. 23. 710.
40. Chen Y. Machine learning for HIV/STI decision support: a review // Sex. Transm. Infect. 2023. Vol. 99. P. 1–7.
41. Sui Q. A scalable machine learning framework for HIV data analysis: Random Forest achieves accuracy 82.8% with AUC 0.73 // Comput. Biol. Med. 2023. Vol. 162. 107089.

(cc) BY