
Секция 2 | ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ

УДК 620.22–4:004.65:004.82

doi: 10.15622/rcai.2025.010

МАЛЫЕ ДАННЫЕ – ЭТО ВСЕ ЧТО У ВАС ЕСТЬ

А.В. Аментес (*Artem.amentes@yandex.ru*)

Федеральный исследовательский центр
«Информатика и управление» РАН, Москва

В работе описывается подход к решению задач связанных с расширяющимися малыми наборами данных (МНД). Приведены описания малых наборов данных, их размеры и ограничения, с которыми сталкиваются исследователи. Также приведены методы интеллектуального анализа данных (ИАД), глубокого обучение и некоторые другие методы работы с МНД. ДСМ-метод рассмотрен как эффективная методология построения интеллектуального анализа данных и построения прогнозных моделей даже при очень малом размере data set'ов.

Ключевые слова: малые наборы данных, искусственный интеллект (ИИ), нейронные сети, большие данные, ДСМ-метод.

Введение

Модели искусственных нейронных сетей часто зависимы от больших объемов данных, необходимых для обучения. Актуальность представляемой работы – в том, что она посвящена отдельно стоящему подклассу задач, связанных с МНД. Существуют предметные области, данные в которых накапливаются крайне медленно, или вообще находятся в очень ма-

лом количестве, а собрать больше данных здесь не представляется возможным [Faraway, 2017]. Избежать обработки МНД не получится. Вытянутые вправо таблицы данных ОБЪЕКТЫ x ПРИЗНАКИ всегда есть в отраслях с глубокой экспертизой. Чем выше экспертность специалиста тем структурно сложнее данные, являющиеся не только многопараметричными, но и, нередко, разнородными. При этом в реальных задачах модели ИИ часто сталкиваются с примерами, которые не похожи на обучающие, что позволяет говорить о проблеме нехватки репрезентативных данных и необходимости надёжного выявления таких случаев для повышения устойчивости и надёжности ИИ-систем и решений. Ошибки на данных, не похожих на обучающие, являются серьезным вызовом при обработке МНД. За годы исследований методов работы с МНД, накоплено значительное количество подходов. При этом, многие идеи, разработанные еще в прошлом веке, только сегодня стали активно применяться для работы с накапливаемыми данными. Вместе с тем, в 2020-е годы все чаще появляются работы, связанные исследованием причинно-следственных связей в постоянно накапливаемых эмпирических данных. Все дело в том, что «объяснимость» как одно из свойств надежного ИИ должна отвечать пользователю на вопрос «Почему?», что связано с причинами формирования моделью ИИ конкретного прогноза или ответа. Новизна нашего исследования заключается в демонстрации методологии, позволяющей не только давать уверенные ответы относительно наличия или отсутствия целевого свойства у изучаемого объекта посредством ИЭ модели ИАД, но и отвечать на вопрос «Почему получен именно такой результат?». А это решает проблему объяснимости работы такого класса программного обеспечения.

Практическая значимость предлагаемых методов, основанных на анализе причинно-следственных связей, заключается в том, что такие виды деятельности как медицина, военное дело, кибербезопасность, инженерное дело и др. предъявляют к интеллектуальным системам особые требования:

- 1) надежность – интерпретация работы модулей и функций;
- 2) устойчивость – сохранение интерполяционных связей каузального характера при пополнении базы прецедентов;
- 3) объяснимость – демонстрация причинных связей в найденном решении.

Эти три понятия создают основы для доверия к результату работы программного обеспечения. Лица принимающие решения должны обладать исчерпывающими обоснованиями. Современные модели искусственного интеллекта, в особенности обученные на ограниченном числе прецедентов, часто не обеспечивают качественной поддержки в этих вопросах.

1. Малые наборы данных

Современная гонка за большими данными сформировалась во многом благодаря особому классу ИЭ задач, когда на больших данных появляется возможность прогнозировать явления или свойства с достаточно большой статистической уверенностью. Сегодня приоритетом в ИИ принято считать интерполяционно-экстраполяционные модели (ИЭМ), которые в наибольшей мере представлены нейросетевыми архитектурами. Считается что, набор методов, связанных с нейросетевым обучением, позволяет существенно вычислительными средствами имитировать человеческие познавательные когнитивные функции, а именно индуктивное обучение на примерах и экстраполяцию интерполированных зависимостей на новые данные. Современные ИЭМ часто зависимы от больших объемов данных, необходимых для их обучения (см., например, искусственные нейронные сети – ИНС). Актуальность представляемой работы связана с необходимостью обрабатывать расширяющиеся data set’ы малого объема и заключается в том, что требуется работать с отдельно стоящим подклассом задач, где предметом анализа оказываются МНД. Существуют предметные области, данные в которых накапливаются крайне медленно, или вообще находятся в очень малом количестве, а собрать больше данных не представляется возможным. Грубо говоря, существование МНД обусловлено отсутствием возможности собрать хоть сколько бы то ни было значимую выборку [Ballantyne, 2023].

МНД – это самостоятельно направление в науке об искусственном интеллекте. Природа данных такова, что существуют ограничения по сбору обучающих данных для некоторых событий, которые могут случаться раз в столетие, и собрать даже 100 таких примеров не представляется возможным. Также, существует ресурсный дефицит у исследователей, который не дает возможность ученым разметить 200 000 фото деревьев, когда ресурсов достаточно только для 200 образцов. Специфические задачи связанные, например, с редкими (мертвыми) языками сильно зависят от найденных артефактов. Существуют и другие регуляторные, политические, военные и этические барьеры (согласие пациентов, конфиденциальность, гостайна). Кроме того, данные должны быть надлежащего качества, что в значительной степени усложняет их сбор. МНД также могут расширяться во времени, что порождает новые вызовы при работе не только с МНД, но и с поступлением новых данных, которые могут не принадлежать ранее известному распределению данных в обучающей выборке. Вообще говоря, было бы полезно детально проанализировать и подходы к определению МНД [Abualigah, 2025]. К сожалению, на данный момент единой методологии отнесения данных к МНД не существует. Каждое направление деятельности характеризуется своим потоком данных и их размерностью.

Авторы чаще рассматривают МНД с точки зрения проблемы, которую они могут вызвать при использовании различных методов машинного обучения. Так, в статье [Wang, 2023] предлагается считать, что малая выборка это – такой набор данных, в котором мало аннотированных данных, или же аннотация является сложной или дорогостоящей. Описание малой выборки носит скорее качественный характер, нежели количественный. В статье [Hollmann et al., 2025], к МНД относят табличные данные, содержащие менее 10 000 строк. В работе [Сафонова, 2023] предлагается конкретное определение МНД в контексте глубокого обучения: МНД – это наборы данных, которые содержат менее 1,000 аннотированных примеров или те, которые плохо покрывают распределение признаков. Часто это приводит к недостаточности информации для эффективного извлечения значимых признаков методами глубокого обучения. Существуют также случаи, называемые авторами «extra-small» (сверхмалые данные), когда набор данных включает от 1 до 10 аннотированных примеров (например, при изучении редких природных катастроф [13]). В работе по изучению Байесовской модели для локализации модификаций белков (PTMProphet), модели глубокого обучения для контроля инсулина, иммунные анализы на МНД, сегментация сердечных МРТ с использованием 3D моделей на маленьком датасете (150 пациентов) столкнулись с ограничениями применения ИЭМ в медицине для создания клинических и программных решений. Основные проблемы здесь – низкая статистическая значимость, ограниченность признаков, высокая вероятность переобучения, ограниченная обобщаемость, и наличие смещений [Mikołajewski et al, 2023]. Медицинские приложения являются ярким примером использования МНД. Медицинские МНД часто состоят из нескольких десятков или сотен прецедентов, что препятствует обучению больших моделей без переобучения. Так в одной из работ, посвященных построению модели для диагностики опухолей мозга [Piffer, 2024], описывают МНД как ограниченное количество аннотированных образцов, обусловленное сложностью, дороговизной и трудностью сбора данных, часто ограниченное несколькими десятками или сотнями пациентов. Малый размер определяется в относительных терминах по отношению к сложности модели и числу параметров, подлежащих обучению (когда данных недостаточно для надежного обучения глубоких нейросетей). Авторами проведен систематический обзор 77 исследований с МНД (в среднем ~16 600 образцов, минимально 16). Представленные исследования относительно достаточности наборов данных для обучения искусственных нейронных сетей (ANN) показали, что правило «в 10 раз больше параметров» [Pasini, 2015] недостаточно консервативно для дискретного выбора. Обширные эксперименты Монте-Карло на синтетических данных с разной сложностью и уровнем шума доказали, что для стабильного обучения рекомендуют правило «в 50 раз

больше параметров», особенно при оценке качества модели по логарифму правдоподобия [Alwosheel et al, 2018]. В свою очередь в контексте работы с пищевыми продуктами, МНД возникают по причине «усталости вкуса», когда эксперты не могут обеспечить значительное количество экспериментов. Кроме того, уточняется, что такие данные, которые потребители собирают, пробуя разные продукты, обычно представлены таблицами, вытянутыми вправо. Количество признаков значительно, а количество наблюдений не велико [Corneu, 2002]. Гетерогенность данных приводит к проблеме, когда внутри набора данных, как это было в пищевых исследованиях, находятся разнородные признаки. Предложенный метод работы с разбивкой данных на кластеры в соответствии с их природой косвенно можно считать с переходом на работу с МНД [Noroosi, 2023]. В обзорной статье [Nivedhaa, 2024] дефицит данных (small data) понимается как недостаточное количество, низкое качество или предвзятость (bias) тренировочного набора, что приводит к плохой обобщающей способности моделей, а отсутствие репрезентативности и дисбаланс классов усиливают риски переобучения и некорректных прогнозов. МНД также можно понимать как недостаточно полные, нерепрезентативные или с сильными дисбалансами, что ведет к системным ошибкам и смещениям в моделях [de Miguel Beriain, 2022]. Это ограниченный объем данных с очень низкой долей неудачных примеров [Pajić, 2023], где количество доступных меток и образцов недостаточно для применения традиционных ML-моделей большого объема. В данном случае речь о нескольких сотнях размеченных примеров при 1.4% неудач. Ряд авторов подчеркивают важность «data centric AI» – подхода, где фокус смещается с улучшения алгоритмов на улучшение качества данных. [Nisheva-Pavlova, 2022]. Данные должны быть связаны с людьми через своевременные, значимые инсайты, часто визуализированные и структурированные, чтобы быть понятными и полезными для повседневных задач. В ряде научных областей МНД являются нормой [Mendes, 2020].

Итак, мы рассмотрели некоторые подходы к определению понятия «малые данные» с количественной и качественной точек зрения. Объем таких data set'ов недостаточен для обучения искусственных нейронных сетей, а также порождает проблему переобучения в классических моделях ML. При этом авторы сходятся во мнении, что МНД неизбежны, так как всегда будут существовать редкие явления, узкоспециализированные направления деятельности, требующие глубокой экспертизы в сборе и разметке данных. Представленные методы интеллектуального анализа данных и получение надежных, устойчивых и объяснимых результатов работы над данными не обеспечивают качества, достаточного для принятия надежных решений.

2. Методы обучения интерполяционно-экстраполяционных моделей на малых выборках данных

Так как избежать работы с МНД не удастся, то исследователи ИИ создают методологии и методы для эффективного решения задач ИАД и прогнозирования искомых свойств на данных. Любое использование программного обеспечения, которое обладает продвинутыми интеллектуальными особенностями не обходится без набора обучающих данных (примеров).

За годы исследований методов работы МНД, накоплено значительное количество подходов [Wang, 2023]: Data augmentation (1990-е); Transfer Learning (1990-е); Self-Supervised Learning (2010-е); Semi-Supervised Learning (1970-е); Few-Shot / Zero-Shot Learning (2000-е); Active Learning (1990-е); Weakly Supervised Learning (2000-е); Multi-Task Learning (1997-е); Ensemble Learning (1990-е); Process-Aware / Physics-Informed Learning (2010-е); Spatial Cross-Validation (2010-е); Causal Discovery (2020-е); Data-Centric AI (2020-е).

Не смотря на разнообразие методов улучшения качества работы моделей на МНД существует ряд проблем, которые на данный момент не решены надежным способом. Так проблемой является выявление, аудит и смягчение различных видов предвзятости (bias) и искажений в наборах данных, используемых для обучения алгоритмов принятия решений. Малое количество примеров для каждого целевого свойства влияют на появление некорректности, дискриминации и потери доверия к ИИ-системам. Примечательно, что помимо классических методов преодоления проблемы МНД, некоторые авторы предлагают интерактивный подход с «человеком в петле» (human-in-the-loop) и формирование разнообразных, мультидисциплинарных команд. Привлечение экспертов для повышения надежности работы моделей на МНД набирает популярность в ряде направлений применения ИЭМ [de Miguel Beriain, 2022]. Классические нейросетевые модели подвержены переобучению и нестабильности из-за малого числа примеров относительно сложности модели. Подчеркивается, что МНД имеют низкую вариативность, невысокую скорость обновления и ограниченный объём, но содержат высоко структурированные и информативные сведения, критичные для понимания анализируемых явлений. Улучшение процессов международного обмена данными позволит исследователям накапливать достаточного размера выборки для использования современных методов анализа [Mendes, 2020]. Глубокие нейросети требуют десятков тысяч размеченных образцов, которые часто недоступны [Piffer, 2024]. Важная проблема, по мнению авторов, заключается в несбалансированности классов. МНД часто характеризуются значительной несбалансированностью классов, что усложняет обучение модели. Несба-

лансированность классов приводит к тому, что модель чаще предсказывает преобладающий класс, что ведёт к ухудшению качества предсказаний редких классов [Safonova, 2023]. Существуют подходы, которые характеризуются меньшей зависимостью от объёма данных, что важно в прикладных областях с ограниченными данными. В описаниях этих подходов, авторы [Jiao, 2024] приводят в своем обзоре преимущества методов в различных задачах: визуальное представление, обработка текста, мультимодальные данные, медицинские приложения. Некоторые авторы [Kimpimäki, 2023] используют внедрение и адаптацию методов вычислительной абдукции (комбинация вычислительных методов и абдуктивного рассуждения) в области менеджмента и организационных исследований, в частности в поисках устойчивой стратегии управления.

Многие идеи, разработанные еще в прошлом веке, только сегодня стали активно применяться для работы с накопленными данными, тем не менее, в 2020-е годы все чаще появляются работы, связанные с исследованием причинно-следственных связей в данных. Дело в том, что «объяснимость» как одно из свойств надежного ИИ должна отвечать пользователю на вопрос «*Почему?*», что связано с поиском причин получения системой ИИ конкретного прогноза или ответа. Большинство методов здесь посвящены корреляционным моделям, опирающимся на различные манипуляции со статистическим подходами. Авторы пытаются увеличить надежность, сделать модели устойчивыми к аномалиям, а также увеличить набор данных для построения уверенных корреляций. Вместе с тем, существуют и другие подходы, которые опираются на логику рассуждений и позволяют преодолевать барьеры МНД не только в статике, но и в динамике их изменений. К сожалению, как показало углубленное изучение (в том числе – экспериментальный анализ на представленных ниже данных НМИЦ НХ им. Н.Н. Бурденко), рассмотренные в Разделах 1-2 известные подходы к обработке МНД не позволяют получать неоспариваемые прикладные результаты.

3. Метод интеллектуального анализа данных с учетом причинно-следственных связей

ДСМ-подход и базирующаяся на нем методология ИАД (ДСМ-ИАД) используют эвристику причинного сходства для идентификации факторов влияния, «вынуждающих» наличие целевых свойств у тех прецедентов в анализируемом *data set*'е, которые таковыми свойствами обладают.

Логико-математическими средствами ДСМ-рассуждения обеспечивается восстановление скрытых, т.е. представленных в неявном виде в анализируемых исходных эмпирических данных, причинно-следственных зависимостей вида *набор значений параметров => целевые свойства*.

ДСМ-метод может рассматриваться как методология организации интеллектуального анализа данных, ориентированного на выделение в анализируемых данных эмпирических зависимостей каузального типа, и позволяет формировать результативные решения для преодоления всех пяти типов проблем («барьеров») – классов ограничений математических ИЭ техник компьютерного анализа данных, о которых шла речь в Разделе I. При этом ДСМ-подход обеспечивает использующему его исследователю гибкость в выборе инструментария формализованного описания и анализа данных [Финн, 2010].

В части предлагаемых в ДСМ-подходе инструментальных средств представления данных и знаний это, в первую очередь это – возможности единообразной «логики» (и алгоритмикой рассуждения) обрабатывать разнотипные данные, в том числе: булевские данные; значения в шкалах наименований; значения в порядковых шкалах; значения в метрических шкалах; числовые значения параметров; текстовые описания; семантически связанные между собою значения в описании каждого конкретного анализируемого прецедента.

Не менее существенна также и гибкость при выборе конкретных ДСМ-инструментов интеллектуального анализа данных [Финн, 2021].

Завершая рассмотрение возможностей и ограничений использования в компьютерном анализе данных наиболее распространенных математических моделей, выделим ДСМ-подход, еще раз фокусируясь на его характеристиках и преимуществах для эксперта в области ИАД. Суммируя рассмотренные выше доводы и аргументы, отметим, что ДСМ-подход предоставляет возможности для:

- интеграции в процесс экспертизы всех тех данных, которые эксперт считает релевантными наличием или, наоборот, отсутствием целевых эффектов, а также проактивной идентификации таких эффектов;
- обеспечения неформальной интерпретации и объяснения результатов ДСМ-ИАД-экспертизы с помощью эмпирических зависимостей причинно-следственного типа, выделяемых из анализируемых данных;
- оперирования актуальными, постоянно пополняемыми новыми элементами, обучающими коллекциями прецедентов, в том числе – ограниченными по своим текущим размерам [Забежайло, 2023].

В медицине онкологических заболеваний головного мозга, количество прецедентов, как правило, незначительное и позволяет накапливать новые данные крайне медленно. Такая ситуация требует использования адекватных средств анализа данных. Авторы работы [Аментес, 2024] представляют результаты ИАД накопленного в НМИЦ НХ им. Н.Н.Бурденко (г. Москва) примерно за 15 лет клинической практики data set'a размером в 250 прецедентов, каждый из которых характеризуется более чем 225 признаками. Такой набор данных (как таблица ОБЪЕКТЫ x ПРИЗНАКИ)

сильно «вытянут вправо». Кроме того, набор данных пополняется новыми примерами крайне редко (всего 10-15 случаев в год). Малая по числу прецедентов выборка не позволяет использовать классические статистические методы анализа данных. Представленная в докладе модель ИАД на базе ДСМ-метода, предлагает результативное решение этой проблемы. Интерполяционно-экстраполяционная модель на базе эвристики причинного сходства позволяет интерполировать обучающий `data_set` каузальными эмпирическими зависимостями и «диагностировать» новые объекты про вер ко йкстраполируемости на ка ж д ы й из них тех причинно-следственных зависимостей (биомаркеров исследуемого эффекта), которые получены при интерполяции анализируемого `data_set`'а.

Методология ДСМ-метода реализована через стандартные методы программного языка `python`. Так, для построения сходства используются базовые модули фильтров данных, а для получения замыкания Галуа – алгоритмы перебора и сортировок столбцов и строк датафреймов. Универсальность методологии ДСМ позволяет решать задачи интеллектуального анализа данных существующими библиотеками, вместе с тем получая достоверные (проверяемые) и надежные (устойчивые) результаты сравнения внутри группы прецедентов методом рассуждений. Алгоритмика проверки «запрета на контрпримеры (ЗКП)» реализуется через порождения подмножеств на данных, обладающих противоположным свойством. Использование комбинации датафреймов, хранящихся в переменной среды обработки данных позволяет быстро проводить необходимые манипуляции с табличными данными. Механизмы сохранения найденных неподвижных точек, формируют локальную базу знаний (атлас зависимостей – маркеров целевого эффекта ¹). Такими инструментальными средствами выполняется весь пайплайн работы методологии рассуждения ДСМ-метода: поиск сходств, построение неподвижных точек, проверка на контрпримеры, построение интерпретируемого, объяснимого в терминах постановки задачи анализа, доступного для понимания эксперту, работающему с когнитивным интерфейсом программы.

Совместно с экспертами НМИЦ НХ Н.Н. Бурденко разработано программное обеспечение, реализующее методологию ДСМ-метода работы с малыми выборками [Аментес, 2024]. В ходе проведенных испытаний удалось получить прикладные результаты, которые не удавалось сформировать другими (см. Разделы 1-2 выше) средствами:

а) сформированы причинно-следственные зависимости – маркеры исследуемых эффектов (логические условия, выполненные на примерах и невыполненные на контрпримерах отдельно для двух эффектов – позитивного и негативного исходов операций);

¹ В обсуждаемых экспериментах – позитивного или, наоборот, негативного исхода нейрохирургической операции.

б) неоспариваемым образом (т.е. с выполнением условия ЗКП) найденными в процессе ДСМ-анализа факторами причинности (маркерами позитивного и негативного исходов операций) разделены примеры и контрпримеры из анализируемого *data set*'а.

Результаты, выдаваемые работой используемой компьютерно-ориентированной модели ИАД, позволяет не только отнести исследуемый объект к одному из целевых классов, но и выдать заключение о причинах наличия целевого свойства, объясняя эту причинность в содержательном контексте терминов и понятий языка экспертов-медиков, предоставивших исходные данные для проводимого анализа.

Заключение

Существуют предметные области, где накопление данных происходит медленно, а количество параметров, описывающих свойства даже одного явления-прецедента достаточно велико. Так, в медицине онкологических заболеваний головного мозга, количество прецедентов, как правило, незначительное и позволяет накапливать новые данные крайне медленно. Малая по числу прецедентов выборка не позволяет использовать классические статистические методы, а также глубокие нейронные сети. Даже деревья решений не дают здесь устойчивого, интерпретируемого результата требуемой точности.

Представленная в настоящей статье компьютерно-ориентированная модель интеллектуального анализа данных на базе ДСМ-метода предлагает результативное решение этой проблемы. Ее возможности в реальных приложениях продемонстрированы на примере интеллектуального компьютерного анализа реальных данных НМИЦ НХ им. Н.Н. Бурденко (г. Москва). Интерполяционно-экстраполяционная модель на базе эвристики причинного сходства позволяет интерполировать обучающий *data_set* каузальными эмпирическими зависимостями и «диагностировать» новые объекты проверкой экстраполируемости на каждый из них тех причинно-следственных зависимостей, которые получены при интерполяции анализируемого *data_set*'а. Результат, выдаваемый работой такой модели интеллектуального анализа данных, позволяет не только отнести исследуемый объект к одному из целевых классов, но и сформировать заключение о наличии целевого свойства, объясняя причины его наличия у конкретного объекта в содержательном контексте терминов и понятий языка экспертов (медиков), предоставивших реальные исходные данные для выполняемого анализа.

Список литературы

- [Аментес, 2024] Аментес А.В., Забейайло М.И. Об опыте разработки атласа биомаркеров исхода нейрохирургических операций // НТИ. Сер. 2. Инф. процессы и системы/ ВИНТИ РАН. – 2024. – № 8. – ISSN 0548-0027.
- [Забейайло, 2023] Забейайло М.И., Аментес А.В. О некоторых особенностях интеллектуального анализа коллекций эмпирических данных, пополняемых новыми сведениями, но ограниченных по своим размерам // Научно-техническая информация. Сер. 2. – 2023. – № 6. – С. 19-24.
- [Финн, 2010] Финн В.К. Индуктивные методы Д.С. Милля в системах искусственного интеллекта // Искусственный интеллект и принятие решений. Ч. I. – 2010. – № 3. – С. 3-21; Ч. II // Там же. – № 4. – С. 14-40.
- [Финн, 2021] Финн В.К. Искусственный интеллект (методология, применения, философия). – 2-е изд., испр. – М.: ЛЕНАНД, 2021. – 468 с.
- [Abualigah, 2025] Abualigah L. Enhancing Real-Time Data Analysis through Advanced Machine Learning and Data Analytics Algorithms // International Journal of Online and Biomedical Engineering (iJOE). – 2025. – Vol. 21, No. 1. – P. 4-25.
- [Alwosheel, 2018] Alwosheel A., van Cranenburgh S., Chorus C.G. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis // Journal of Choice Modelling. – 2018. – Vol. 28. – P. 167-182.
- [Ballantyne, 2023] Ballantyne P., Berragan C. Overture poi data for the united kingdom: a comprehensive, queryable open data product // A PREPRINT. – Liverpool: Geographic Data Science Lab, University of Liverpool, 2023.
- [Conant, 2015] Conant D.D. Developing a Scalable Model to Analyze Expanding Data Sets // INFORMS Transactions on Education. – 2015. – Vol. 15, No. 3. – P. 215-223.
- [Cordelli, 2023] Cordelli E., Guarrasi V., Iannello G., Ruffini F., Sicilia R., Soda P., Tronchin L. Making AI trustworthy in multimodal and healthcare scenarios // CEUR Workshop Proceedings. – 2023. – Vol. 3265. – P. 1-18.
- [Corney, 2002] Corney D.P. A. Intelligent analysis of small data sets for food design: PhD: Computer Science / David Peter Alfred Corney; University College London. – London, 2002. – 168 p. – URL: ProQuest Dissertations.
- [Faraway, 2017] Faraway J., Augustin N. When small data beats big data // Department of Mathematical Sciences, University of Bath. – Bath, 2017.
- [Frontoni, 2022] Frontoni E., Paolanti M., Lauriault T.P., Stiber M., Duranti L., Abdul-Mageed M. Trusted Data Forever: Is AI the Answer? // Proceedings of the 25th EDBT. – 2022. – P. 1-12.
- [Hollmann, 2025] Hollmann N., Müller S., Purucker L., Krishnakumar A., Körfer M., Hoo S.B., Schirmmeister R.T., Hutter F. Accurate predictions on small data with a tabular foundation model // Nature. – 2025. – Vol. 637, No. 7979. – P. 319-326.
- [Jiao, 2024] Jiao L., Wang Y., Liu X., Li L., Liu F., Ma W., Guo Y., Chen P., Yang S., Hou B. Causal Inference Meets Deep Learning: A Comprehensive Survey // Research. – 2024. – Vol. 7. – Article 0467.
- [Kimpimäki, 2023] Kimpimäki J.-P. From observation to insight: Computational abduction and its applications in sustainable strategy research: dissertation. – Lappeenranta-Lahti University of Technology LUT, 2023. – 116 p. – ISBN 978-952-412-004-3.

- [**de Miguel Beriain, 2022**] de Miguel Beriain I., Nicolás Jiménez P., Rementería M.J., Cirillo D., Cortés A., Saby D., Lazcoz Moratinos G. Auditing the quality of datasets used in algorithmic decision-making systems // Panel for the STOA, European Parliamentary Research Service. – Brussels, 2022. – 41 p.
- [**Mendes, 2020**] Mendes P.S.F., Siradze S., Pirro L., Thybaut J.W. Open data in catalysis: from today's big picture to the future of small data // *ChemCatChem*. – 2020.
- [**Mikolajewski, 2023**] Mikolajewski D., Mikołajewska E. Artificial intelligence-based analysis of small data sets in medicine // *Studia i Materiały Informatyki Stosowanej*. – 2023. – Vol. 15, No. 2. – P. 18-23.
- [**Nisheva-Pavlova, 2022**] Nisheva-Pavlova M., Dobрева B. Small Data and Data Centric AI: Case Study from the Master's Program in Artificial Intelligence at Sofia University // *Proceedings of the Fifteenth International Conference on Information Systems and Grid Technologies (ISGT'2022)*. Sofia, Bulgaria, May 27–28, 2022. CEUR Workshop Proceedings. Vol. 3154. – P. 171-179.
- [**Nivedhaa, 2024**] Nivedhaa N. A comprehensive review of AI's dependence on data // *International Journal of Artificial Intelligence and Data Science (IJADS)*. – 2024. – Vol. 1, No. 1. – P. 1-11.
- [**Noroozi, 2023**] Noroozi G. Data Heterogeneity and Its Implications for Fairness: MSc: Computer Science. Western University. – Ontario, 2023.
- [**Pajić, 2023**] Pajić N., Djapan M., Bulushek E., Fahrenbruch W., Đorđević A., Stefanović M. Machine Learning Prediction Model for Small Data Sets Instead of Destructive Tests for a Case of Resistance Brazing Process Verification // *IJIE*. – 2023. – Vol. 30, No. 3. – P. 797-814.
- [**Piffer, 2024**] Piffer S., Ubaldi L., Tangaro S., Retiko A., Talamonti H. Tackling the small data problem in medical image classification with artificial intelligence: a systematic review // *Progress in Biomedical Engineering*. – 2024. – Vol. 6, No. 032001.
- [**Pasini, 2015**] Pasini A. Artificial neural networks for small dataset analysis // *Journal of Thoracic Disease*. – 2015. – Vol. 7, No. 5. – P. 953-960.
- [**Safonova, 2023**] Safonova A., Ghazaryan G., Stiller S., Main-Knorn M., Nendel C., Ryo M. Ten deep learning techniques to address small data problems with remote sensing // *International Journal of Applied Earth Observation and Geoinformation*. – 2023. – Vol. 125. – Art. 103569.
- [**Wang, 2023**] Wang H., Duentsch I., Guo G., Khan S.A. Special issue on small data analytics // *International Journal of Machine Learning and Cybernetics*. – 2023. – Vol. 14, No. 1. – P. 1-2.