Revisiting Deep Learning Models:
Effective Baselines for Medicine Tabular Tasks

С. П. Высоцкий[1,*], М.Н. Устинова[1], Т.А. Ремизова[1]

[1] Центральный университет
[*] *email: s.vysotskiy@edu.centraluniversity.ru*

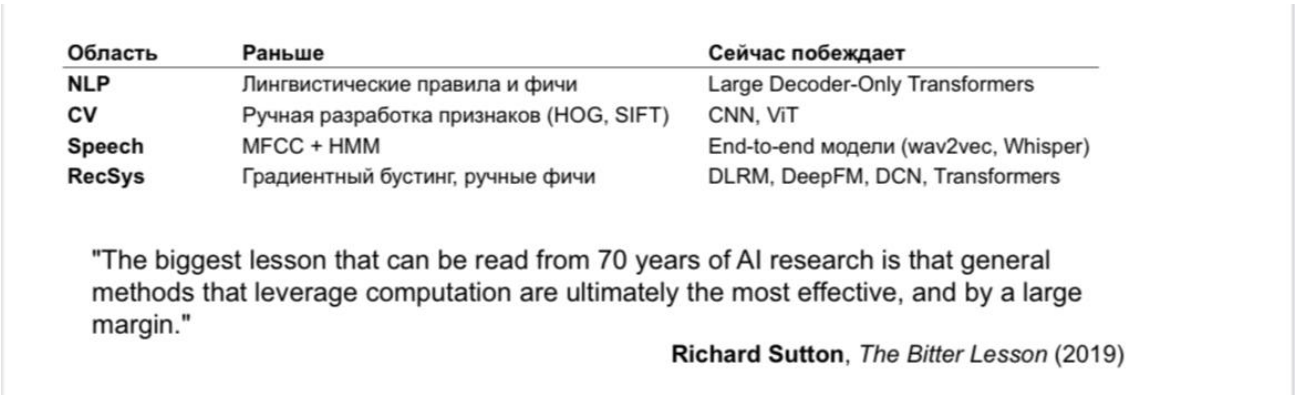
Несмотря на широкое распространение табличных данных в клинической практике, методы глубинного обучения по-прежнему редко применяются в реальных медицинских решениях. Это связано как с доминированием бустинга, так и с недостаточной интерпретируемостью и универсальностью нейросетевых подходов. В данной работе мы переосмысливаем потенциал глубинных моделей для табличных данных, начиная с простых MLP-решений и анализируя их эффективность на фоне более сложных архитектур. Обзор охватывает четыре ключевых направления текущих инноваций: сложные модели, новые архитектуры, foundation-подходы и бенчмарки. Мы показываем, что компактные и параметр-эффективные модели могут успешно конкурировать с бустингом как по качеству, так и по удобству применения. Далее мы исследуем интерпретируемую архитектуру GateNet и применяем её к задаче раннего выявления онкологических заболеваний на табличных медицинских данных. Модель демонстрирует высокую точность и прозрачность выбора признаков, что особенно важно для доверия и применения в клинической практике.

1 Введение

До недавнего времени ведущим инструментом для работы с табличными данными оставались методы градиентного бустинга над решающими деревьями: XGBoost, LightGBM, CatBoost. Они обеспечивают высокую точность, интерпретируемость и относительно низкие вычислительные затраты, что сделало их стандартом индустрии и базовой линией для сравнения. В Яндексе, например, CatBoost — метод, который хорошо работает с гетерогенными признаками и особенностями табличных данных, такими как пропуски и шум.

С 2020 года всё активнее исследуются нейросетевые подходы к табличным данным. Хотя бустинг пока чаще остаётся лидером по точности, современные нейросети быстро догоняют и в ряде задач уже конкурируют с ними. К 2023–2024 годам наблюдается заметное улучшение результатов нейросетевых моделей и рост их популярности в табличной модальности.

В статье «The Bitter Lesson» (2019) [1] Ричард Саттон подчеркнул, что ключ к успешному машинному обучению — это использование вычислительной мощности и масштабируемых методов, которые позволяют моделям автоматически извлекать полезные представления из данных. Он показал, что подходы, основанные на жёстко запрограммированных знаниях и ручной инженерии признаков, со временем уступают универсальным и гибким моделям, которые учатся напрямую из сырого входа. Эта идея особенно актуальна для нейросетей, которые демонстрируют способность к эффективному обучению на больших объёмах данных и самообучению, что делает их перспективными для решения сложных задач, включая табличные данные.



Область	Раньше	Сейчас побеждает
NLP	Лингвистические правила и фичи	Large Decoder-Only Transformers
CV	Ручная разработка признаков (HOG, SIFT)	CNN, ViT
Speech	MFCC + HMM	End-to-end модели (wav2vec, Whisper)
RecSys	Градиентный бустинг, ручные фичи	DLRM, DeepFM, DCN, Transformers

"The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin."

Richard Sutton, *The Bitter Lesson* (2019)

Рис. 1: State-of-the-art -подходы в различных областях машинного обучения. Видно, что везде побеждают нейронные сети.

Несмотря на заметный прогресс в табличном дип-обучении, до сих пор отсутствует универсальный подход, способный устойчиво превосходить бустинг во всех сценариях. Нейросети уже научились конкурировать с деревьями по качеству предсказаний, особенно на задачах с высоким уровнем шума, сложной зависимостью между признаками или в условиях необходимости end-to-end обучения. Однако, как показывают независимые сравнения, бустинг по-прежнему остаётся мощной базовой линией, особенно при ограниченных

данных и строгих требованиях к интерпретируемости. Развитие в этой области — скорее постепенная эволюция, чем революционный переход, однако направление движения очевидно: нейросети уверенно сокращают отставание.

На этом фоне особое внимание привлекли попытки перенести успехи больших языковых моделей (LLM) на табличную модальность. Под влиянием успеха GPT-подобных архитектур сформировалась гипотеза, что в будущем любые данные — от текста и изображений до биомедицинских сигналов и таблиц — смогут обрабатываться универсальной трансформерной моделью. В 2023 году было предложено несколько подходов, предполагающих токенизацию таблиц с последующей генерацией табличных выходов в парадигме autoregressive modeling. Однако практика показала: подобные generative-подходы не демонстрируют конкурентного качества на предсказательных задачах. В частности, в отчётах по AutoGluon (Amazon) [5] прямо указано: "Generative models are not ready for predictive ML on tabular data." Даже при достижении технической готовности таких моделей остаётся открытым вопрос о целесообразности их применения, учитывая вычислительные издержки и неэффективность на малых датасетах.

В свете этих вызовов и ограничений мы выделяем три ключевых направления для развития методов табличного машинного обучения: 1. Разработка специализированных архитектур, способных извлекать сложные зависимости в условиях табличной модальности. Сюда входят такие подходы, как:

• KNN-подобные модели с латентным сравнением и поисковыми модулями (NCA, ModernNCA, 2025) [2]

• Архитектуры с внутренним ансамблированием, такие как TabM [3] , реализующие батч-ансамбли в MLP;

• Многостадийные сети с адаптивным отбором признаков, такие как GateNet [4], где применяется идея soft-выбора и селекции фичей.

2. Интеграция foundation-моделей с возможностью дообучения под конкретные downstream-задачи. Хотя универсальные модели для табличной модальности пока не достигли зрелости, идея создания "материнских моделей" MotherNet [6], с последующей адаптацией к конкретной задаче остаётся крайне перспективной.

3. Создание качественных и репрезентативных бенчмарков, отражающих реальные производственные сценарии: наличие шума, категориальных

фичей, пропусков, смещений и доменных сдвигов. Недостаточная стандартизация в оценке мешает честному сравнению новых архитектур и затрудняет репликацию результатов.

В рамках настоящей работы мы сосредотачиваемся на первом направлении — исследовании архитектур — с применением моделей в реальной прикладной задаче из области медицины: детекция прогрессии онкологического заболевания по клиническим и радиологических данным пациентов. Мы рассмотрим как базовые архитектурные решения, так и сравним их с современные нейросетевыми подходами. В частности, особое внимание будет уделено применению многостадийной архитектуры GateNet, сочетающей интерпретируемый отбор признаков и глубокую обработку информации.

2 Постановка задачи

В данной работе мы рассматриваем задачу обучения с учителем на табличных биомедицинских данных, характеризующихся ограниченным объёмом выборки (до 1000 наблюдений) и высокой размерностью признакового пространства. Наша задача предсказать таргет прогрессии (есть/нет), исходя из признакового описания пациента.

Работа с табличными данными сопряжена с рядом ключевых сложностей:

• Гетерогенность признаков. В одной таблице часто сочетаются числовые, категориальные, бинарные и ординальные признаки. Это усложняет создание универсальных архитектур, так как разные типы признаков требуют разного подхода к представлению и обработке. Например, категориальные признаки нуждаются в эмбеддингах, числовые — в нормализации, а бинарные — в специальных индукциях.

• Отсутствие локальности и структуры. В изображениях и тексте присутствуют пространственные или временные зависимости — например, соседние пиксели или слова формируют смысловые блоки. Благодаря этому появились архитектуры с индуктивными смещениями — свёрточные нейросети для CV и позиционные эмбеддинги для NLP. В табличных данных никакой явной «локальной» или «последовательной» структуры нет — признаки могут быть связаны самым разным образом, часто непредсказуемо.

• Повышенный уровень шума и сложность интерпретации. Метки в табличных данных часто зависят от сложных, трудно формализуемых правил. Человеческий эксперт даже с опытом иногда не может однозначно интерпретировать, почему объект получил ту или иную метку. Это значительно усложняет оценку качества модели и понимание её решений.

• Небольшой объём обучающих данных. Особенно актуально для чувствительных сфер — медицина, фармацевтика, биотехнологии, где сбор и маркировка данных обходятся дорого и требуют времени. Обучающие выборки в несколько сотен примеров — не редкость. Решая задачу на медицинских данных, мы приходим к выводу, что мы обладаем лишь небольшими сырыми обучающами выборками, собранными вручную.

Целью исследования является разработка метода, способного обеспечивать высокое качество предсказаний в условиях малых данных, что типично для многих задач в медицине и фармацевтике.

Мы стремимся построить модель , удовлетворяющую следующим ключевым требованиям:

• P1: Способность эффективно обобщать при ограниченном количестве обучающих примеров (размер обучающей выборки ≤ 1000);

• P2: Устойчивость к шуму и нерелевантным признакам, с возможностью автоматически исключать мешающие переменные;

• P3: Достаточная выразительная мощность для захвата сложных нелинейных зависимостей, характерных для биомедицинских данных.

Мы рассматриваем подходы, основанные на современных архитектурах глубокого обучения для табличной модальности, с акцентом на интерпретируемость, устойчивость и адаптивность к малым выборкам.

3 Модели и методы

В исследовании мы сосредоточились на подборе архитектур моделей. Выборка была предобработана методами очистки, стандартизации, а также дополнительно обогащена новыми признаками. Также мы провели дополнительный up-sampling технологией SMOTE, для того, чтобы взвесить семплы каждого класса (их было неодинаковое количество). Так, мы получили обучающий сетии из почти 800 объектов, описанных около 20 признаками.

Мы протестировали основные модели машинного обучения, начиная с простых линейных моделей, заканчивая CatBoost и простыми MLP-решениями. Подробный обзор качества будут показаны далее. Задача интерпретируемых предсказаний остро встает в задаче медицинского предсказания. Так, модели, которые могут отбирать фичи для предсказания могут быть полезны. В алгормитмах бустинга и случайного леса отбор осуществляется на этапе жадного разбиения и метода случайных подпространств. В случае нейронных сетей отбор происходит неявно, зависимости получаются не локально разреженными как в случае алгоритмах на решающих деревьях.

Так, мы приходим к мотивации использования алгоритма GateNet: сети с отбором признаков. GateNet — это двухстадийная нейросетевая архитектура для работы с табличными данными, которая реализует поэтапный отбор и связывание признаков. На первой стадии используются стохастические гейты (шлюзы), которые динамически выбирают релевантные признаки, фильтруя шум и нерелевантные данные. Вторая стадия отвечает за связывание выбранных признаков и построение предсказательной модели с высокой выразительной способностью.

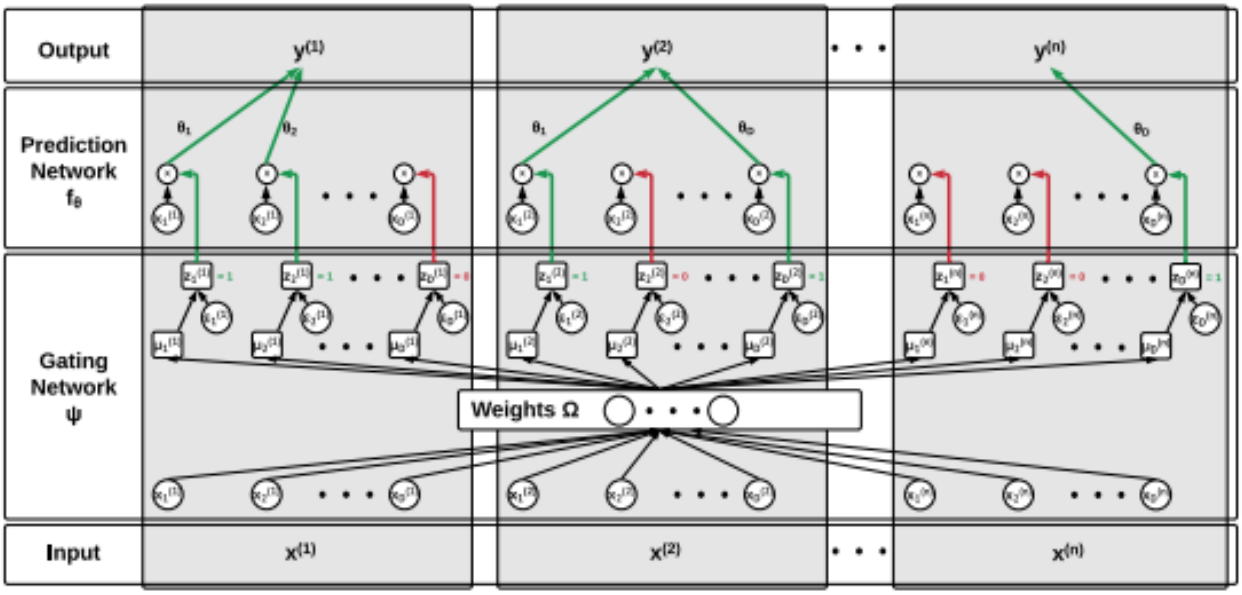


Рис. 2: GateNet: двухстадийная нейронная сеть с отбором признаков

Обучение моделей в многостадийном ансамбле просходит совместным образом: обучаем модель на кросс-энтропийный критерий классификации (детекции прогрессии) с добавочным членом l_0-регуляризации. Подробный алго-

ритм обучения можно прочитать в статье Locally Sparse Neural Networks for Tabular Biomedical Data [4].

Такой алгоритм помогает добавить интерпретируемость, а также упростить сложность модели. Так, часть модели отвечает за связывание части фичей, что в свою очередь дает разреженность информации даже в очень больших MLP-модулях GateNet. Так, далее мы покажем, что сложность архитектуры модели GateNet не влияет на переобучение.

Подбор оптимальных гиперпараметров (внутренних слоев сети отбор и сети-предсказателя, параметры обучения) мы проводили при помощи библиотеки Optuna.

4 Результаты

Лучшие результаты показала модель глубокого обучения GateNet. Подробнее по гиперпараметром можно посмотреть в техническом отчете приложенным к статье. Там же вы найдете код для обучения модели и работу с данными.

Таблица 1: Величина метрики Balanced Accuracy предсказания моделей

Model	**Balanced Accuracy**
Bagging KNN	0.6312
RF	0.77272
MLP	0.63636
LSPIN(GateNet)	0.81818

Результаты обнадеживают и дают понять, что в глубокого обучения большой потенциал в решении медицинских задач табличного формата. При небольшом тюнинге этого подхода, мы можем добиваться выразительного качества. Так, добавляя степени свободы в модель, мы добавляем интерпретируемость, а также способность выучивать более сложные зависимости. Нейронные сети способны как запоминать, так и аппроксимировать реальные зависимости «см. табл. 1».

5 Обсуждение

Предложенная нами модель демонстрирует устойчиво высокое качество на различных выборках. Так, на отложенной выборке соревнования Kaggle она показывает хорошие результаты вне зависимости от глубины архитектуры. Даже глубокую модель удаётся эффективно обучить на небольшом количестве данных без признаков переобучения. Аналогичную устойчивость демонстрируют и компактные модели при длительном обучении: метрика качества продолжает расти, что позволяет гибко дообучать модель под конкретные особенности тренировочной выборки.

Достичь этого удаётся благодаря архитектуре отбора признаков, которая позволяет сети автоматически адаптировать внутренние представления и выбирать наиболее релевантные признаки, избегая переобучения.

В качестве дальнейших шагов мы видим расширение предложенного подхода на другие задачи табличного глубокого обучения, а также апробацию в реальных задачах прогнозирования в медицине. Кроме того, мы планируем исследовать возможности интеграции поисковых модулей (например, NCA) и внутреннего ансамблирования в рамках архитектуры GateNet. Эти идеи вдохновляют нас на разработку нового класса моделей — нейронных лесов, в которых пространство признаков выбирается аналогично случайным лесам, а предсказания объединяются через методы батч-ансамблирования, как в TabM. Такой подход открывает перспективы для построения новых универсальных и интерпретируемых моделей в табличной модальности.

6 Заключение

В рамках данной работы мы продемонстрировали эффективность архитектуры GateNet — нейросетевого подхода к отбору признаков и построению предсказательной модели для табличных данных. Разработанный нами пайплайн показывает высокую стабильность и качество как на валидационных, так и на отложенных выборках. Мы показали, что принципы, извлечённые из «горького урока» (The Bitter Lesson), применимы и в табличном машинном обучении: универсальные модели, построенные на обучении и масштабируемых архитектурах, постепенно превосходят специализированные ручные решения. При этом возможность интеграции интерпретируемых механизмов

делает такие модели не только точными, но и понятными.

В дальнейшем мы планируем развивать направление табличного глубокого обучения, применяя его как к медицинским задачам, так и к более общим случаям анализа табличных данных.

Благодарности

Авторы хотят поблагодарить преподавателей за интересный материал, который позволил по-новому взглянуть на науку и окружающий мир. Отдельно хотим сказать спасибо команде ЦУ за чуткость в течение курса.

Список литературы

[1] Bitter Lesson, Rich Sutton, — URL: — https://www.incompleteideas.net/IncIdeas/BitterLesson.html

[2] REVISITING NEAREST NEIGHBOR FOR TABULAR DATA: A DEEP TABULAR BASELINE TWO DECADES LATERб 2025 — URL: — https://arxiv.org/pdf/2407.03257v2

[3] TABM: ADVANCING TABULAR DEEP LEARNING WITH PARAMETER-EFFICIENT ENSEMBLING, 2024, Yandex Reseacrh — URL: — https://arxiv.org/pdf/2410.24210v2

[4] Locally Sparse Neural Networks for Tabular Biomedical Data, 2022 — URL: — https://arxiv.org/pdf/2106.06468

[5] autogluon, Amazon — URL: — https://github.com/autogluon

[6] MotherNet: A Foundational Hypernetwork for Tabular Classification, 2023 — URL: — https://arxiv.org/pdf/2312.08598v1