Архитектура "трансформер" для анализа рисков группового влияния нутриентов пищи

Баландина[1], Груздев[1], Савельев[2], Будакян[1], Кисиль[3], Грачев[1]

1) Физический факультет МГУ имени М. В. Ломоносова
2) ПАО “Вымпелком”
3) Институт русского языка и культуры МГУ имени М. В. Ломоносова

Задача

Актуальность

- Метаболический синдром (МС) - медицинское состояние, которое характеризуется висцеральным ожирением. Основной причиной МС является неправильный образ жизни. МС приводит к сердечно-сосудистым и прочим заболеваниям.

- Раннее выявление и лечение метаболического синдрома может снизить риск развития серьезных заболеваний в будущем.

- Ранняя диагностика нарушений становится одной из ключевых задач в обследовании пациентов с подозрением на МС.

Задача

Данные

- Центром Нутрициологии и Адаптивного Питания были предоставлены анонимизированные данные анамнезов наблюдаемых пациентов.
- Данные были формализованы и приведены в табличном виде, где каждому анамнезу пациента соответствует строка таблицы.
- Данные были предобработаны и очищены от дубликатов. Реальные анамнезы были разбиты на классы: имеет диагноз, связанный с метаболическим синдромом / здоровый.
    - 2885 строк
    - Данные разбиты 70% / 30% (обучающая / тестовая выборки)
    - 85 признаков, 67 – состав диеты, 18 – антропометрия и проч.
    - Баланс классов 1872 / 1013 (~65% / 35%, здоровые / больные)

Возраст	Рост	Профессия	Спорт	ОТ	ОБ
39.0	180.0	Работники преимущественно умственного труда	Не занимаюсь	130.0	120.0
45.0	177.0	Работники преимущественно умственного труда	Регулярный спорт	0.0	0.0
56.0	181.0	Работники среднего по тяжести труда	Не занимаюсь	50.0	150.0
33.0	184.0	Работники среднего по тяжести труда	Регулярный спорт	106.0	101.0
35.0	182.0	Работники преимущественно умственного труда	Легкий спорт	0.0	0.0
31.0	0.0	Работники среднего по тяжести труда	Не занимаюсь	0.0	0.0

Таблица с реальными данными пациентов.

Задача

Цель работы и этапы выполнения

- Задача **бинарной** классификации

- Обучить нейронную сеть оценивать вероятность наличия метаболического синдрома на основе признаков;

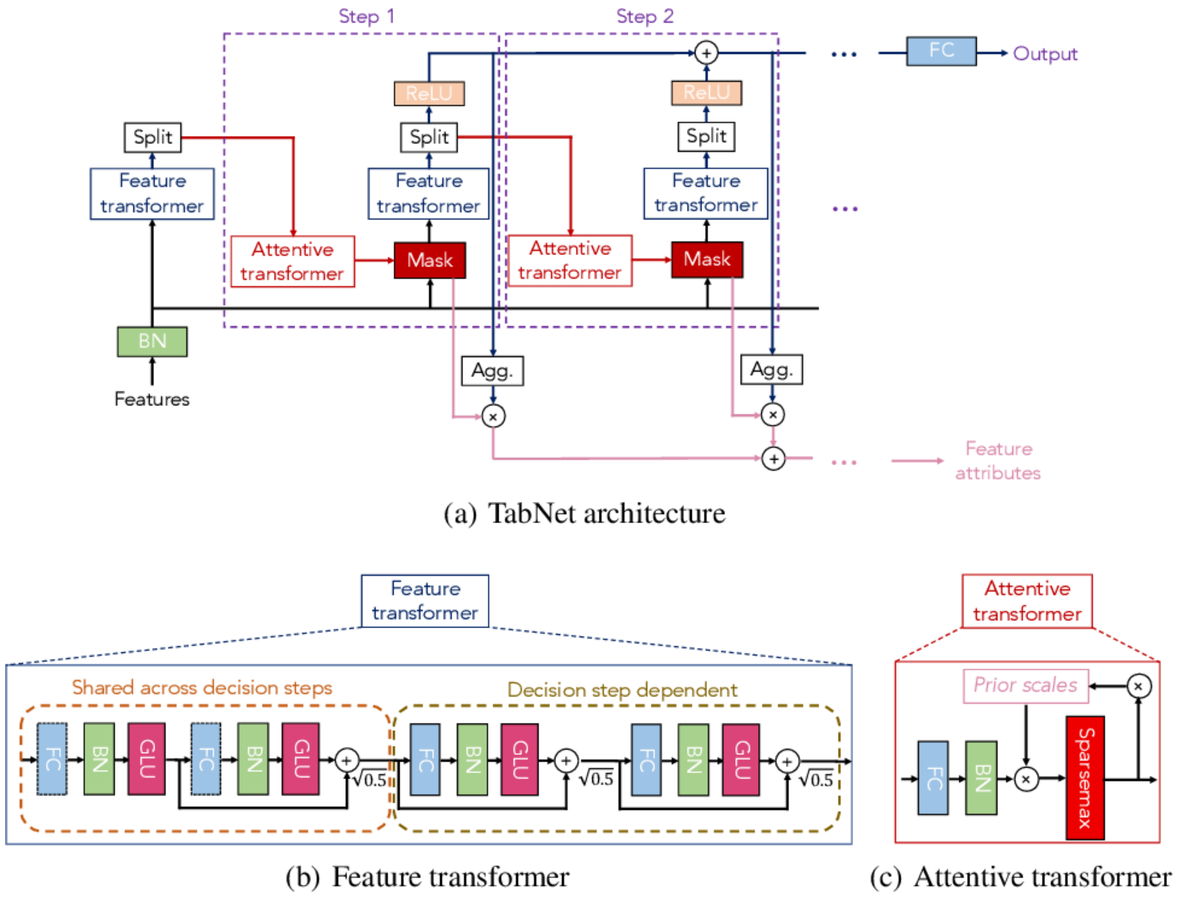
- Оценить точность предсказаний;

- Использовать обученную модель для того, чтобы найти группы признаков, влияющих на результат.

Предыдущие работы

Предыдущие работы по нейросетевым методам обработки табличных данных

TabNet[1] — рекурентная нейронная сеть, разработанная для анализа табличных данных.



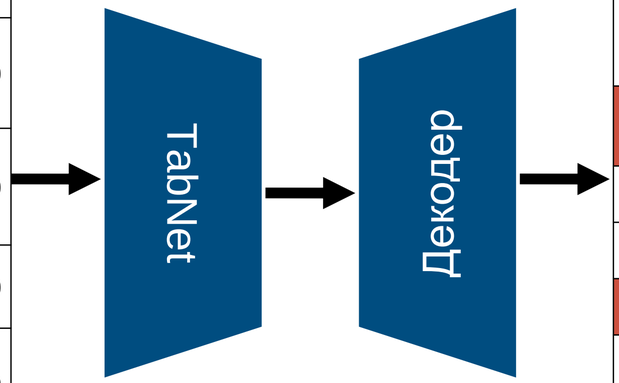
Общая схема работы TabNet

[1]Arik, S. O., & Pfister, T. (2019). TabNet: Attentive Interpretable Tabular Learning. arXiv preprint arXiv:1908.07442.

Предыдущие работы

Предыдущие работы по нейросетевым методам обработки табличных данных. TabNet. Предобучение

Возраст	Рост	Профессия	Спорт	ОТ	ОБ
39.0	180.0	?	Не занимаюсь	130.0	120.0
?	177.0	Работники преимущественно умственного труда	?	0.0	0.0
56.0	181.0	Работники среднего по тяжести труда	Не занимаюсь	50.0	150.0
33.0	?	Работники среднего по тяжести труда	Регулярный спорт	?	101.0
?	?	Работники преимущественно умственного труда	?	0.0	0.0
31.0	0.0	?	Не занимаюсь	0.0	0.0



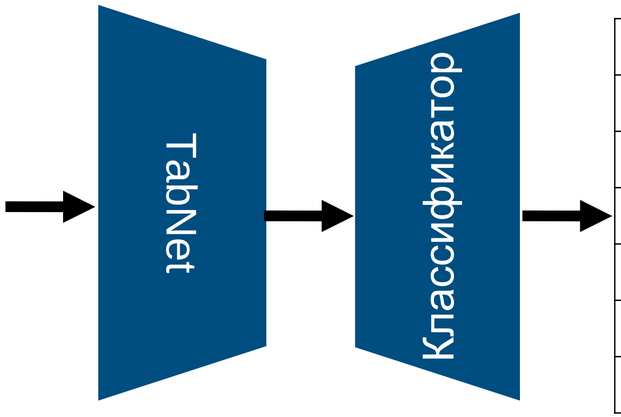
Возраст	Рост	Профессия	Спорт	ОТ	ОБ
		Работники преимущественно умственного труда			
45.0			Регулярный спорт		
	184.0			106.0	
35.0	182.0		Легкий спорт		
		Работники среднего по тяжести труда			

Схема самообучения TabNet

Предыдущие работы

Предыдущие работы по нейросетевым методам обработки табличных данных. TabNet. Дообучение

Возраст	Рост	Профессия	Спорт	ОТ	ОБ
39.0	180.0	Работники преимущественно умственного труда	Не занимаюсь	130.0	120.0
45.0	177.0	Работники преимущественно умственного труда	Регулярный спорт	0.0	0.0
56.0	181.0	Работники среднего по тяжести труда	Не занимаюсь	50.0	150.0
33.0	184.0	Работники среднего по тяжести труда	Регулярный спорт	106.0	101.0
35.0	182.0	Работники преимущественно умственного труда	Легкий спорт	0.0	0.0
31.0	0.0	Работники среднего по тяжести труда	Не занимаюсь	0.0	0.0

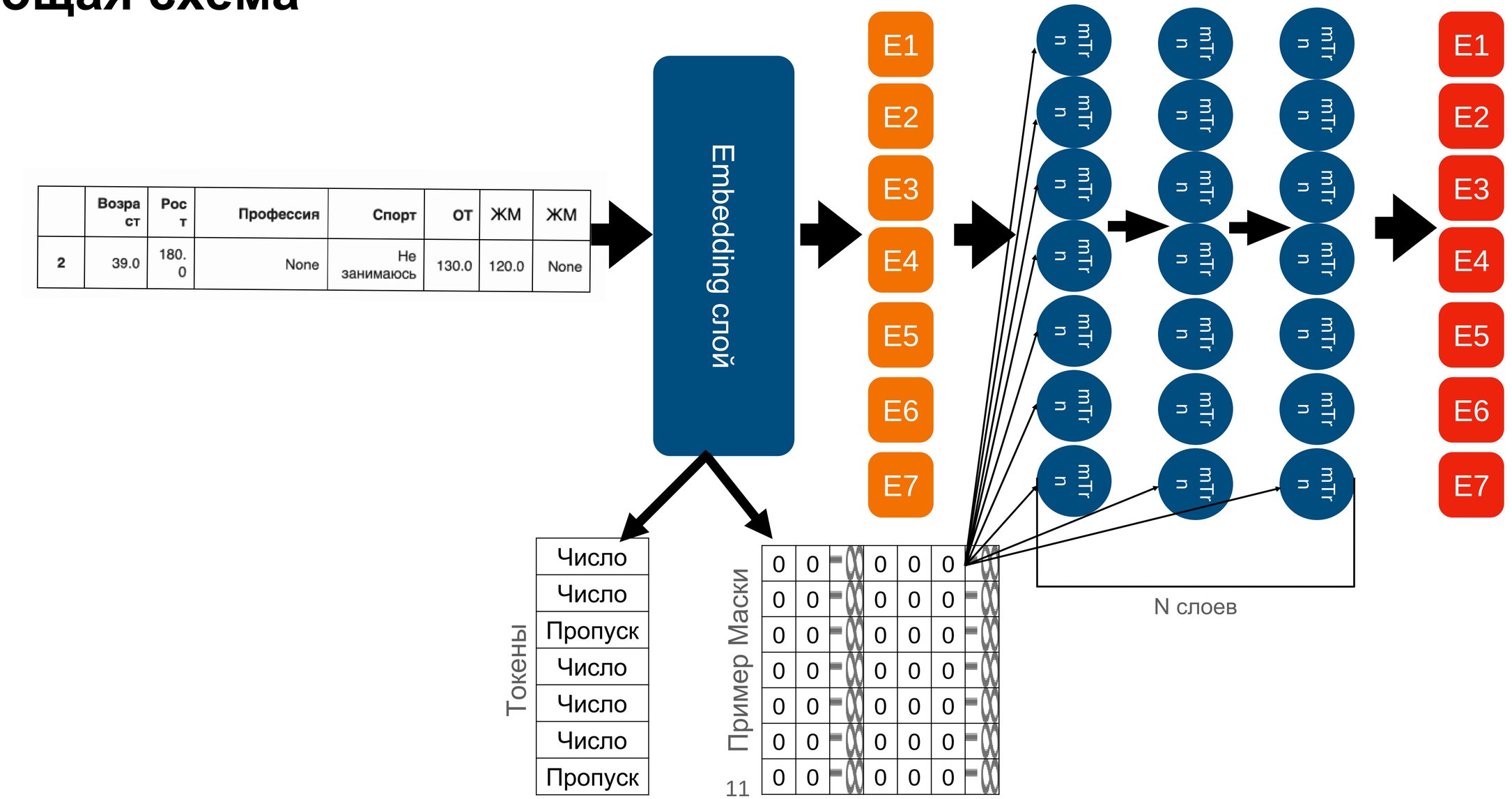


	Диагноз
2	0
3	1
4	1
5	1
6	0
7	0

Схема дообучения TabNet

Разработанная архитектура

Общая схема



Разработанная архитектура

Embedding слой

A. В качестве Embedding слоя для чисел взят слой PAF[4] (Periodic Activation Functions): $f(x) = Periodic(x) = concat[sin(v), cos(v)]$, где $v = [2\pi c_1 x, \ldots, 2\pi c_k x]$

B. Позиционное кодирование, необходимое, чтобы модель понимала к какому признаку относится embedding.

C. Производится по следующей формуле:
$PE_{pos,2i} = sin(pos/1000^{2i/d_{model}})$,
$PE_{pos,2i+1} = cos(pos/1000^{2i/d_{model}})$ [5]

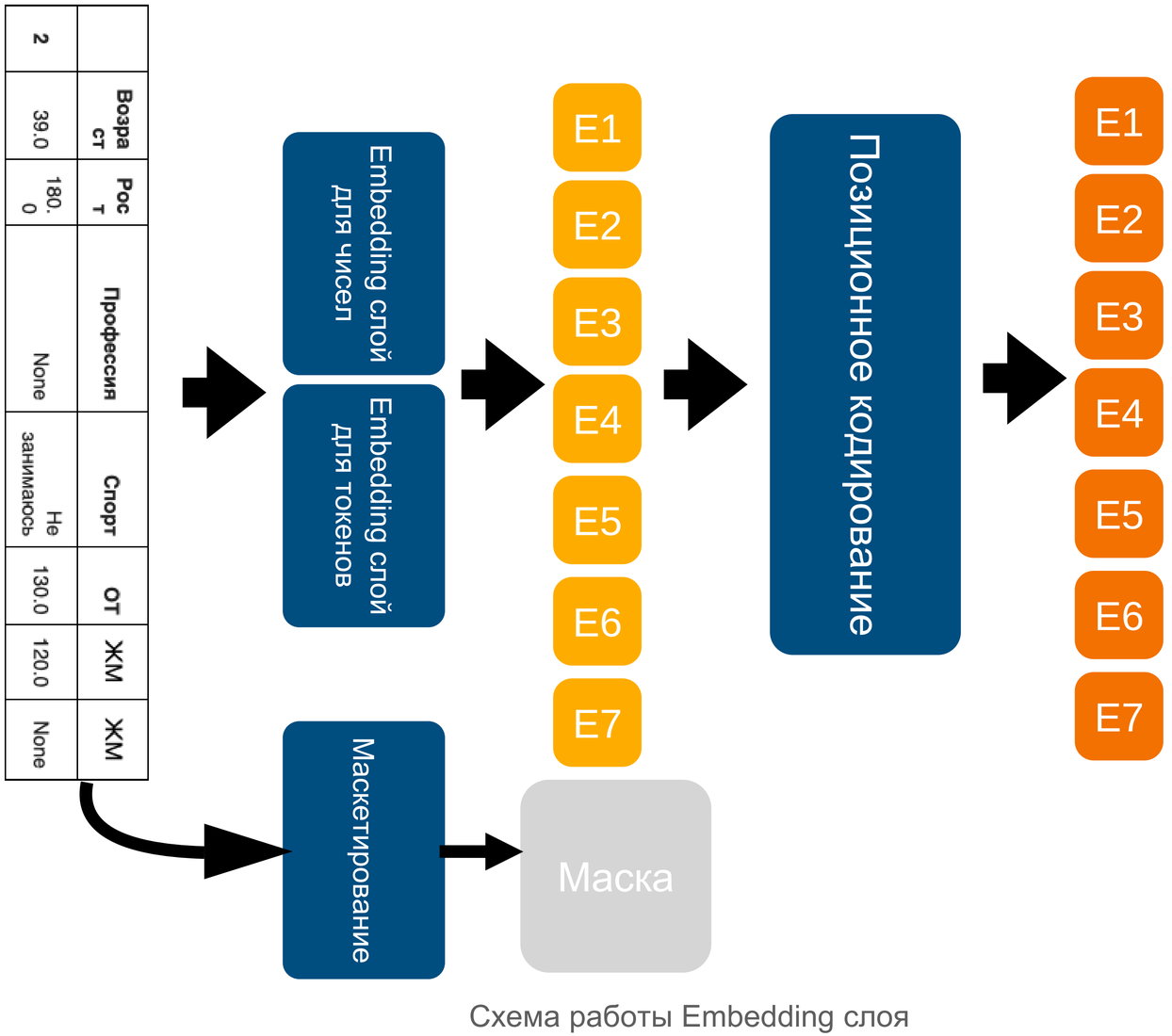


Схема работы Embedding слоя

[4] Yury Gorishniy and Ivan Rubachev and Artem Babenko(2022). On Embeddings for Numerical Features in Tabular Deep Learning. arXiv 2203.05556

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010.

Разработанная архитектура

Слой трансформера

Главное в трансформере - это механизм внимания (Attention), с помощью которого входные вектора представляются через все прочие вектора (получают контекст).

$$Attention(Q, K, V, MASK) = softmax(\frac{QK^T}{\sqrt{d_{model}}} + MASK)V$$

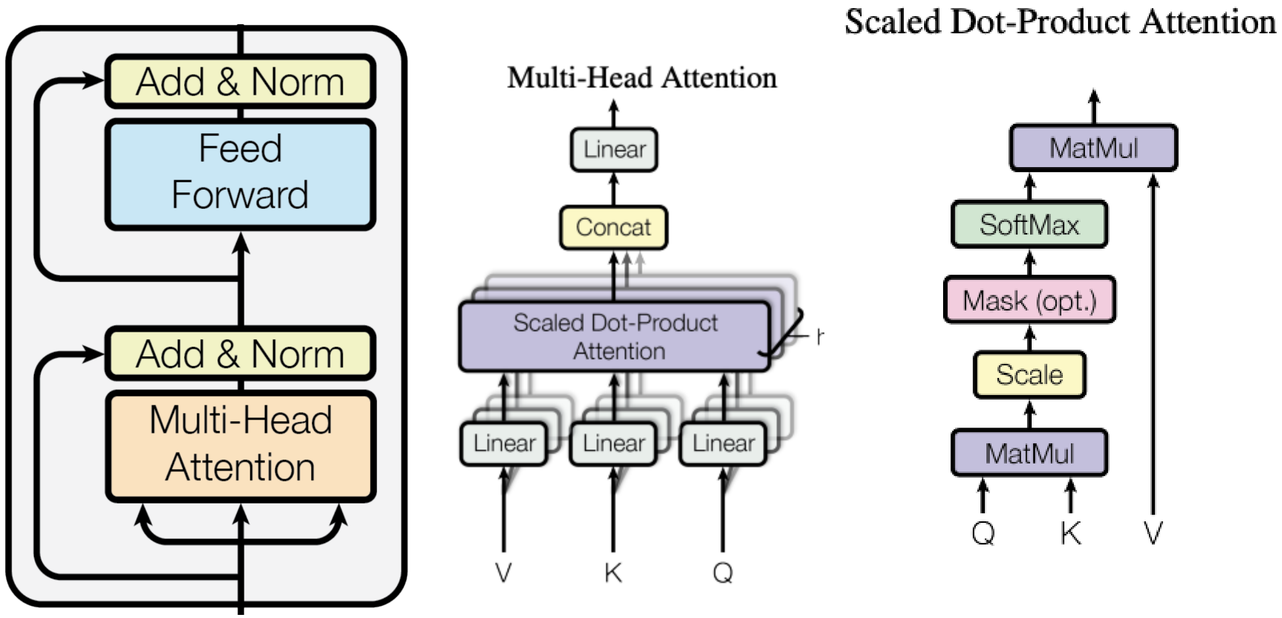
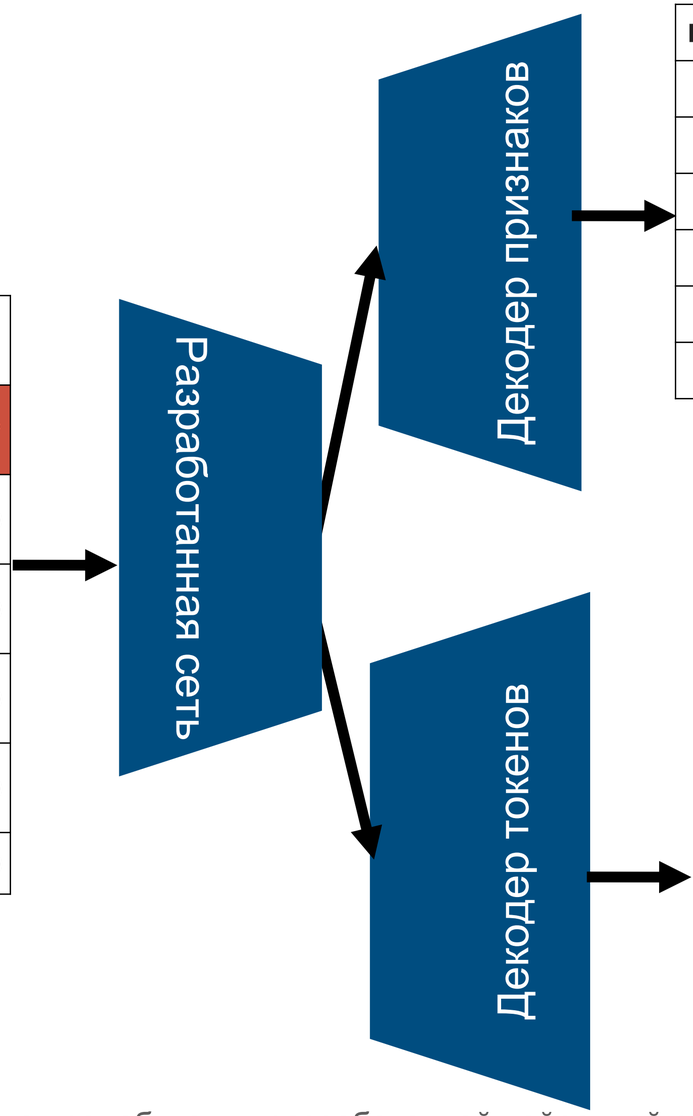


Схема работы трансформер-слоя

Разработанная архитектура

Предобучение

Возраст	Рост	Профессия	Спорт	ОТ	ОБ
39.0	180.0	?	Не занимаюсь	130.0	?
?	177.0	Работники преимущественно умственного труда	?	0.0	0.0
56.0	181.0	Работники среднего по тяжести труда	Не занимаюсь	50.0	150.0
33.0	?	None	Регулярный спорт	?	101.0
?	?	Работники преимущественно умственного труда	?	0.0	0.0
31.0	0.0	?	None	0.0	0.0



Возраст	Рост	Профессия	Спорт	ОТ	ОБ
39.0	180.0			130.0	120.0
	177.0			0	0
56.0	181.0			50.0	150.0
33.0	184.0			106.0	101.0
35.0	182.0			0.0	0.0
31.0	0.0			0.0	0.0

0	0	0	0	0	0
1	0	0	0	0	0
0	0	0	0	0	0
0	0	1	0	0	0
0	0	0	0	0	0
0	0	1	1	0	0

Схема предобучения разработанной нейронной сети.

Разработанная архитектура

Обучение

Возраст	Рост	Профессия	Спорт	ОТ	ОБ
39.0	180.0	Работники преимущественно умственного труда	Не занимаюсь	130.0	120.0
45.0	177.0	Работники преимущественно умственного труда	Регулярный спорт	0.0	0.0
56.0	181.0	Работники среднего по тяжести труда	Не занимаюсь	50.0	150.0
33.0	184.0	Работники среднего по тяжести труда	Регулярный спорт	106.0	101.0
35.0	182.0	Работники преимущественно умственного труда	Легкий спорт	0.0	0.0
31.0	0.0	Работники среднего по тяжести труда	Не занимаюсь	0.0	0.0

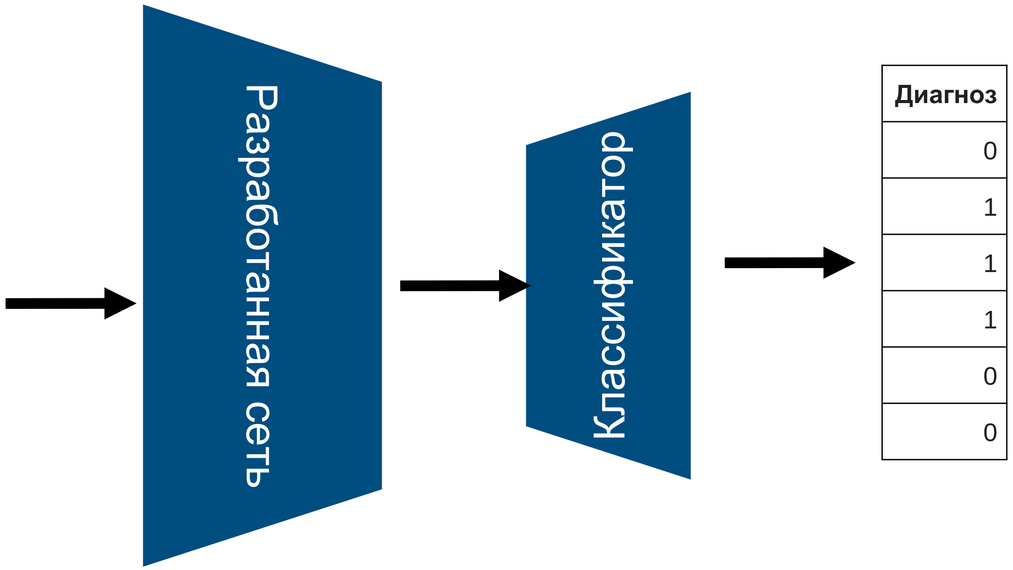


Схема дообучения разработанной нейронной сети.

Результаты

1. При обучении моделей различных архитектур были получены следующие показатели accuracy на тестовой выборке

- Accuracy=$\frac{TP+TN}{TP+TN+FP+FN}$
- TP –количество верно найденых диагнозов с метаболическим нарушением;
- FP –количество раз, когда модель считала, что нарушение есть, однако его не было;
- TN –количество верно найденых диагнозов с отсутствием метаболических нарушений;
- FN –количество раз, когда модель считала, что нарушения нет, однако оно было.

Разработанная сеть	TabNet	CatBoost
72%	68%	**74%**

Значения Accuracy на тестовой выборке исследуемых моделей

Марковская кластеризация

1) Граф преобразуется в смежную матрицу;
2) Смежная матрица преобразуется в стохастическую матрицу;
3) Пока матрица изменяется повторять:
   a) Расширение матрицы;
   b) Инфляция матрицы.

- Хорошо масштабируется с увеличением размера графа.
- Работает как с взвешенными, так и невзвешенными графами.
- Выдает хорошие результаты кластеризации. Устойчив к шуму в данных графа.
- Количество кластеров не задается заранее, но можно настроить гранулярность кластера с помощью параметров.
- Не подходит для кластеров с большим диаметром.

Результаты

В ходе проведения исследования были кластеры из слоя Attention.

Нездоровые	Здоровые
Масса тела, холестерин, насыщенный жир, магний	Возраст, рост, общая жидкость
Рост, объем талии, галактоза	Возраст, рост, жировая масса
Объем талии, жировая масса, общая жидкость, изолейцин, гамма-линеленовая кислота	Пол, рост, объем талии, холестерин
Объем талии, жировая масса, кремний, изолейцин	Масса тела, насыщенные жиры, сахара, магний
Обхем талии, мононенасыщенные жиры, бетта-каротин	Насыщенные жиры, сахара, углеводы, фолаты
Обхем бедер, жировая масса, мышечная масса, спорт	Насыщенные жиры, сахара олеиновая кислота
Объем талии, жировая масса, йод	Насыщенные кислоты, сахара, лейцин
Жировая масса, мышечная масса, общая жидкость, изолейцин, никель	Натрий, жиры, β-каротин
Жировая масса, мышечная масса, общая жидкость, изойлецин, фтор	Фолаты, омега-3, биотин
Жировая масса, мышечная масса, медь	Фолаты, омега-3, фенилаланин+тирозин
Жировая масса, мышечная масса, молибден	β-каротин, фосфор, метилонин, отсутствие спорта, ИМТ

Полученные кластеры для людей с/без метаболического синдрома

Результаты

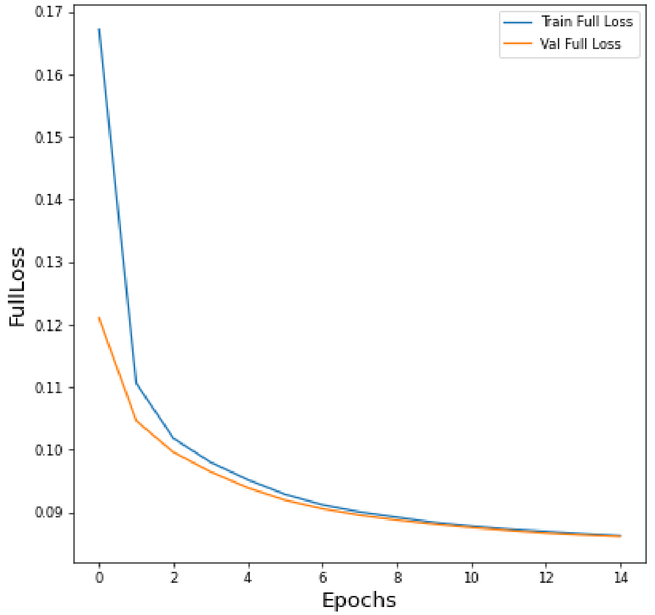
1. Обучена нейронная сеть архитектуры транформер на задаче бинарной классификации;

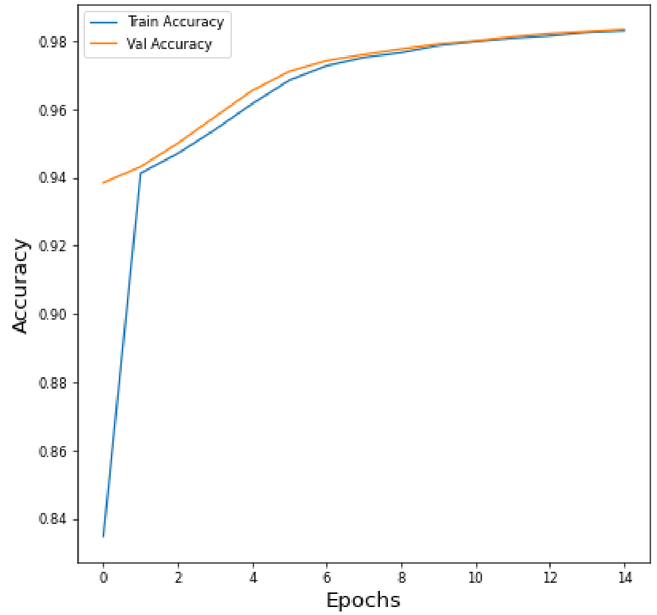
2. Обученная сеть сравнима по качеству с другими, используемыми при задаче бинарной классификации;

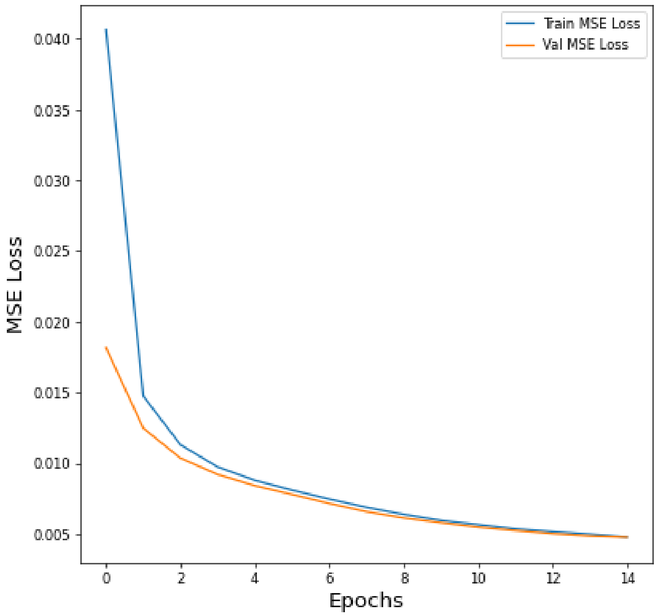
3. При помощи марковской кластеризации из слоя внимания сети получены кластеры признаков для людей, классифицированных как имеющих метаболический синдром, так и для людей, классифицированных как не имеющих метаболический синдром.

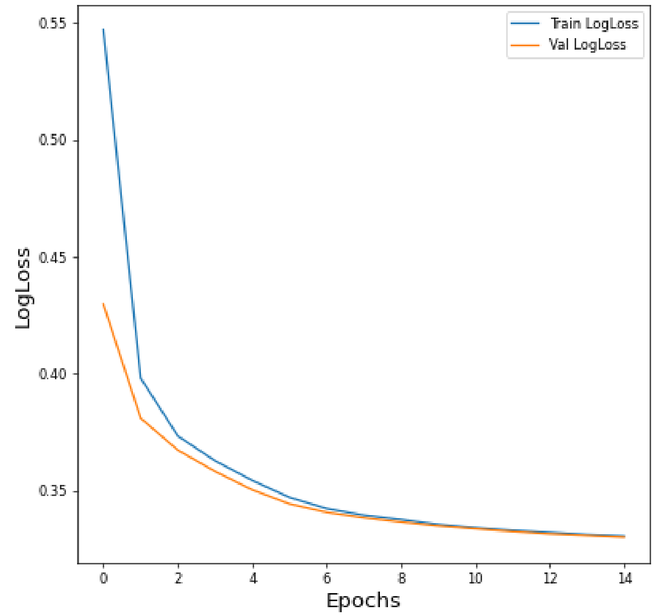
Графики обучения

Графики самообучения









Графики обучения

Графики дообучения