УДК 073.585:004.89
https://doi.org/10.47093/2713-069X.2025.6.4.55-63

CC BY 4.0

Оригинальная статья

Использование методов машинного обучения для диагностики заболеваний на основе неструктурированных медицинских текстов

А.Д. Ермак[1,*], Е.А. Макарова[1], А.Н. Кафтанов[1], Д.В. Гаврилов[1], Р.Э. Новицкий[1], А.В. Гусев[2]

[1]*ООО «К-Скай», наб. Варкауса, д. 17, г. Петрозаводск, 185910, Россия*

[2]*ФГБУ «Центральный научно-исследовательский институт организации и информатизации здравоохранения» Министерства здравоохранения Российской Федерации, ул. Добролюбова, д. 11, г. Москва, 127254, Россия*

Аннотация

Современные методы машинного обучения открывают новые возможности для анализа медицинских текстов. Использование неструктурированных данных позволяет улучшить качество поддержки принятия врачебных решений и развивать персонализированные подходы к лечению пациентов. ***Цель исследования:*** разработка оптимального алгоритма прогнозирования заболеваний с помощью мультиметочной классификации на основании медицинских текстов из отобранных случаев лечения пациентов. ***Материалы и методы.*** В исследовании использовались анонимизированные электронные медицинские карты 387 590 пациентов. Для анализа текстовой информации применялись методы лемматизации и векторизации на основе предобученной модели FastText. Разработана мультиметочная модель классификации, предсказывающая 156 диагностических категорий, сгруппированных по основным группам заболеваний. Для построения моделей применялись нейросетевые архитектуры и ансамбли деревьев решений. ***Результаты.*** Предложенные модели показали высокую эффективность. Использование различных методов агрегации текстовых векторов позволило повысить качество прогнозирования. Модель продемонстрировала стабильность и клиническую интерпретируемость результатов, обеспечивая возможность применения в реальной медицинской практике. ***Заключение.*** Разработанный подход к анализу неструктурированных медицинских текстов с помощью методов машинного обучения является перспективным инструментом для поддержки диагностики заболеваний. Дальнейшие исследования направлены на улучшение интерпретируемости моделей и их адаптацию к различным клиническим источникам данных.

Ключевые слова: машинное обучение; неструктурированные данные; мультиметочная классификация; нейронные сети; персонализированная медицина

Для цитирования: Ермак А.Д., Макарова Е.А., Кафтанов А.Н., Гаврилов Д.В., Новицкий Р.Э., Гусев А.В. Использование методов машинного обучения для диагностики заболеваний на основе неструктурированных медицинских текстов. Национальное здравоохранение. 2025; 6 (4): 55–63. https://doi.org/10.47093/2713-069X.2025.6.4.55-63

Контактная информация:

* Автор, ответственный за переписку: Ермак Андрей Дмитриевич. E-mail: aermak@webiomed.ru

Статья поступила в редакцию: 16.06.25 **Статья принята к печати:** 13.08.25 **Дата публикации:** 30.12.25

Disease diagnosis from unstructured medical texts using machine learning techniques

Andrey D. Ermak[1,*], Elena A. Makarova[1], Alexey N. Kaftanov[1], Denis V. Gavrilov[1], Roman E. Novitskiy[1], Alexandr V. Gusev[2]

[1]*K-SkAI LLC, Varkaus Embankment, 17, Petrozavodsk, 185910, Russia*

[2]*Russian Research Institute of Health, Dobrolyubova str., 11, Moscow, 127254, Russia*

Abstract

Modern machine learning methods open new opportunities for analyzing medical texts. The use of unstructured data enables improved clinical decision support and the development of personalized patient treatment approaches. ***The aim***

© А.Д. Ермак, Е.А. Макарова, А.Н. Кафтанов, Д.В. Гаврилов, Р.Э. Новицкий, А.В. Гусев, 2025

of the study: to develop an optimal algorithm for disease prediction using multi-label classification based on medical texts from selected patient treatment cases. ***Materials and methods.*** The study utilized anonymized electronic medical records of 387 590 patients. Textual data were processed using lemmatization and vectorization based on a pretrained FastText model. A multi-label classification model was developed to predict 156 diagnostic categories grouped by major disease classes. Neural network architectures and decision tree ensembles were applied for model building. ***Results.*** The proposed models demonstrated high effectiveness. The use of various text vector aggregation methods improved prediction quality. The model showed stability and clinical interpretability, supporting its applicability in real-world medical practice. ***Conclusion.*** The developed approach to analyzing unstructured medical texts using machine learning methods is a promising tool for disease diagnosis support. Further research will focus on improving model interpretability and adapting models to diverse clinical data sources.

Keywords: machine learning; unstructured data; multi-label classification; neural networks; personalized medicine
For citation: Ermak A.D., Makarova E.A., Kaftanov A.N., Gavrilov D.V., Novitskiy R.E., Gusev A.V. Disease diagnosis from unstructured medical texts using machine learning techniques. National Health Care (Russia). 2025; 6 (4): 55–63. https://doi.org/10.47093/2713-069X.2025.6.4.55-63

Contacts:
* Corresponding author: Andrey D. Ermak. E-mail: aermak@webiomed.ru

The article received: 16.06.25 **The article approved for publication:** 13.08.25 **Date of publication:** 30.12.25

ВВЕДЕНИЕ

Развитие технологий машинного обучения и обработки естественного языка открыло новые возможности для анализа медицинских текстов, автоматического извлечения данных и поддержки принятия решений [1]. Включение текстовых записей в алгоритмы анализа позволяет выявлять новые клинические закономерности, осуществлять прогнозы и оценивать эффективность стратегий лечения. Это расширяет возможности систем поддержки принятия врачебных решений, способствуя появлению персонализированных подходов к ведению пациентов и повышению качества медицинской помощи [2].

При разработке подобных решений необходимо учитывать ряд ограничений. Важными требованиями являются высокая скорость обработки запросов, возможность локального развертывания решений в медицинских организациях, а также соответствие требованиям информационной безопасности. Само применение такого рода технологий сопряжено с рядом технических вызовов, включая сложность терминологии, неоднородность и различия в форматах документации. Значительная часть медицинских данных представлена в неструктурированном виде, например в виде текстовых протоколов осмотров, выписок, заключений, что усложняет их обработку и интерпретацию. Все эти факторы требуют комплексного подхода к разработке решений, объединяющего методы машинного обучения, лингвистического анализа и медицинской экспертизы.

Современные исследования охватывают широкий спектр подходов по классификации медицинских текстов, различаясь по источникам данных, методам представления/обработки текстов (TF-IDF, Word2Vec, ELMo, fastText) и по алгоритмам классификации (логистическая регрессия, случайный лес, сверточные и рекуррентные нейронные сети) [3, 4]. Разброс метрик качества подтверждает это разнообразие: F1-score в опубликованных работах варьируется от 0,67 для традиционных моделей машинного обучения до 0,98 для нейросетевых архитектур, чувствительность достигает 0,99, а AUROC колеблется от 0,79 до 0,86 [5–9].

Таким образом, современные исследования показывают, что для успешной классификации медицинских текстов критически важен выбор подхода к их обработке, алгоритма классификации и метода векторизации. Однако остаются нерешенные вопросы, связанные с интерпретируемостью моделей, адаптацией алгоритмов к различным языкам и источникам данных, а также необходимостью обеспечения их устойчивости к ошибкам и вариативности медицинской терминологии. В данном исследовании мы стремимся внести вклад в решение этих проблем, предлагая подход, адаптированный к специфике реальных клинических русскоязычных данных.

Цель исследования: разработка оптимального алгоритма прогнозирования заболеваний с помощью мультиметочной классификации на основании медицинских текстов из отобранных случаев лечения пациентов.

МАТЕРИАЛЫ И МЕТОДЫ

Источник данных. Проведено многоцентровое когортное ретроспективное исследование с использованием базы данных платформы Webiomed (https://webiomed.ru/), содержащей обезличенные электронные медицинские карты 50 миллионов пациентов

© Andrey D. Ermak, Elena A. Makarova, Alexey N. Kaftanov, Denis V. Gavrilov, Roman E. Novitskiy, Alexandr V. Gusev, 2025

из Российской Федерации. Информация о каждом пациенте в базе данных представлена в виде множества записей с учетом временных изменений и включает клинические данные о состоянии здоровья, лабораторные и инструментальные исследования, зарегистрированные заболевания.

Участники. Каждая запись в исследованном наборе представляла собой первичный осмотр пациента, относящийся к отдельному случаю оказания медицинской помощи. Для включения осмотра в набор данных использовались следующие критерии.

1. Случай оказания медицинской помощи, к которому относится осмотр, представляет собой амбулаторное посещение.
2. Указаны пол и дата рождения пациента, его возраст 18 лет и старше на момент обращения.
3. В медицинской карте есть данные о завершении случая лечения, включая основной диагноз, а также даты открытия и закрытия случая.
4. Дата осмотра находится в пределах от дня открытия случая до следующего календарного дня включительно.
5. В осмотре заполнены разделы жалоб и объективных данных.

Полученные данные были разделены на обучающую (72 %) и тестовую выборки (20 %), 8 % были использованы для контроля переобучения алгоритмов.

Исходы. В исследовании решалась задача прогнозирования заболеваний с помощью мультиметочной классификации на основании медицинских текстов из отобранных случаев лечения пациентов. Поскольку один пациент мог иметь несколько диагностированных состояний в рамках одного случая оказания медицинской помощи, предполагалось, что ему может быть присвоено несколько меток одновременно.

Процесс присвоения меток основывался на зарегистрированных диагнозах в виде кодов МКБ-10, включая основной диагноз, сопутствующие заболевания и осложнения. Для уменьшения размерности пространства прогнозируемых меток и повышения устойчивости модели коды были сгруппированы в более обобщенные диагностические категории. Группировка проводилась на основе медицинских знаний, что позволило объединять схожие состояния в более крупные классы. Такой подход не только упрощал сложность задачи, но и повышал клиническую интерпретируемость результатов, поскольку предсказания модели соответствовали логически связанным группам заболеваний.

В результате отбора были сформированы 156 диагностических меток, каждая из которых относилась к одной из выделенных групп заболеваний: сердечно-сосудистые, эндокринные, онкологические, респираторные, желудочно-кишечные, инфекционные, аутоиммунные и неврологические (приложение 1 https://doi.org/10.47093/2713-069X.2025.6.4.55-64-annex). При этом из анализа и разметки были исключены случаи, в которых диагностировались травмы, наркологические и алкогольные заболевания, состояния, связанные с беременностью.

Предикторы. Для прогнозирования исходов в данном исследовании использовались ограниченный набор факторов, включающий только возраст пациента на момент визита и его пол, а также тексты из жалоб, объективного осмотра и локального статуса (если он был зафиксирован). Возраст пациента учитывался как важный фактор, поскольку он может влиять на вероятность возникновения различных заболеваний и состояние здоровья в целом. Пол также был принят во внимание, так как многие заболевания имеют половые различия проявлений и частоты.

Для анализа строковых данных применялись методы обработки естественного языка. Из текстов были удалены пунктуация и служебная информация, после чего, с использованием библиотеки pymorphy2 [10], был проведен процесс лемматизации, что помогло уменьшить вариативность слов в тексте и улучшить качество анализа. Для дальнейшего разделения текста на отдельные элементы был использован метод токенизации из библиотеки razdel[1].

Потом каждый полученный токен был переведен в числовой вектор длиной 50 элементов с помощью предобученной на основе технологии FastText[2] модели. Данная модель была заранее обучена на корпусе обезличенных медицинских записей на русском языке, что позволило создать численное представление слов, характерных для медицинской тематики. Итоговое представление текста представляло собой матрицу размерностью 50*количество слов в тексте, где каждая строка соответствовала вектору для отдельного токена, что позволяло учесть контекстуальные зависимости и семантику в полном объеме.

После этапа получения матрицы векторов для преобразования всей информации о тексте в один вектор фиксированной длины были использованы и оценены несколько подходов (рис. 1):

1. Подход с динамическим пулингом (Dynamic Pooling)[3]

В этом методе текст делится на сегменты, каждый из которых обрабатывается с использованием операции «Max pooling», которая находит максимальное значение по каждому признаку в сегменте. Итоговое представление текста получается путем усреднения результатов всех сегментов. В результате получается вектор длиной 50 элементов.

[1] Razdel – библиотека для токенизации русскоязычных текстов. URL: https://github.com/natasha/razdel (дата обращения: 17.11.2025).
[2] Bojanowski P., Grave E., Joulin A., Mikolov T. Enriching Word Vectors with Subword Information. 2016. URL: http://arxiv.org/abs/1607.04606 (дата обращения: 01.11.2025).
[3] Nawrot P., Chorowski J., Łańcucki A., Ponti E.M. Efficient Transformers with Dynamic Token Pooling. Труды Ежегодной конференции Ассоциации вычислительной лингвистики (ACL 2023). URL: https://doi.org/10.18653/v1/2023.acl-long.353 (дата обращения: 01.11.2025).

2. Подход с комбинацией пулингов (Combined Pooling)[4]

В данном методе, помимо стандартного динамического пулинга, используется метод вычисления энергетических векторов. Векторы взвешиваются в зависимости от их «энергии», которая определяется как квадрат нормы каждого вектора. Итоговое представление текста формируется путем объединения двух результатов: энергетического и динамического пулинга. После объединения итоговый вектор имеет длину 100 элементов.

3. Встроенный в нейросеть метод внимания (Attention Pooling)[5]

Каждый токен в тексте получает коэффициент внимания, который определяет, насколько важен данный токен для общего смысла текста. Итоговое представление текста формируется путем взвешивания векторов слов в зависимости от их значимости для задачи. В результате данного подхода итоговый вектор также имеет длину 50 элементов.

После получения вектора для текста к нему присоединялись данные о возрасте и поле пациента, что позволяло улучшить представление о пациенте с учетом демографических факторов. В зависимости от выбранного подхода итоговый вектор имел размерность 52 либо 102 элемента.

Моделирование. Статистический анализ и построение моделей машинного обучения выполняли на языке программирования Python, версия 3.10. В качестве архитектур моделей использовали нейросетевые и ансамблевые методы. Используемая нейронная сеть включала последовательность полносвязных слоев с нелинейными активациями LeakyReLU, слоями Batch-Normalization и механизмом регуляризации

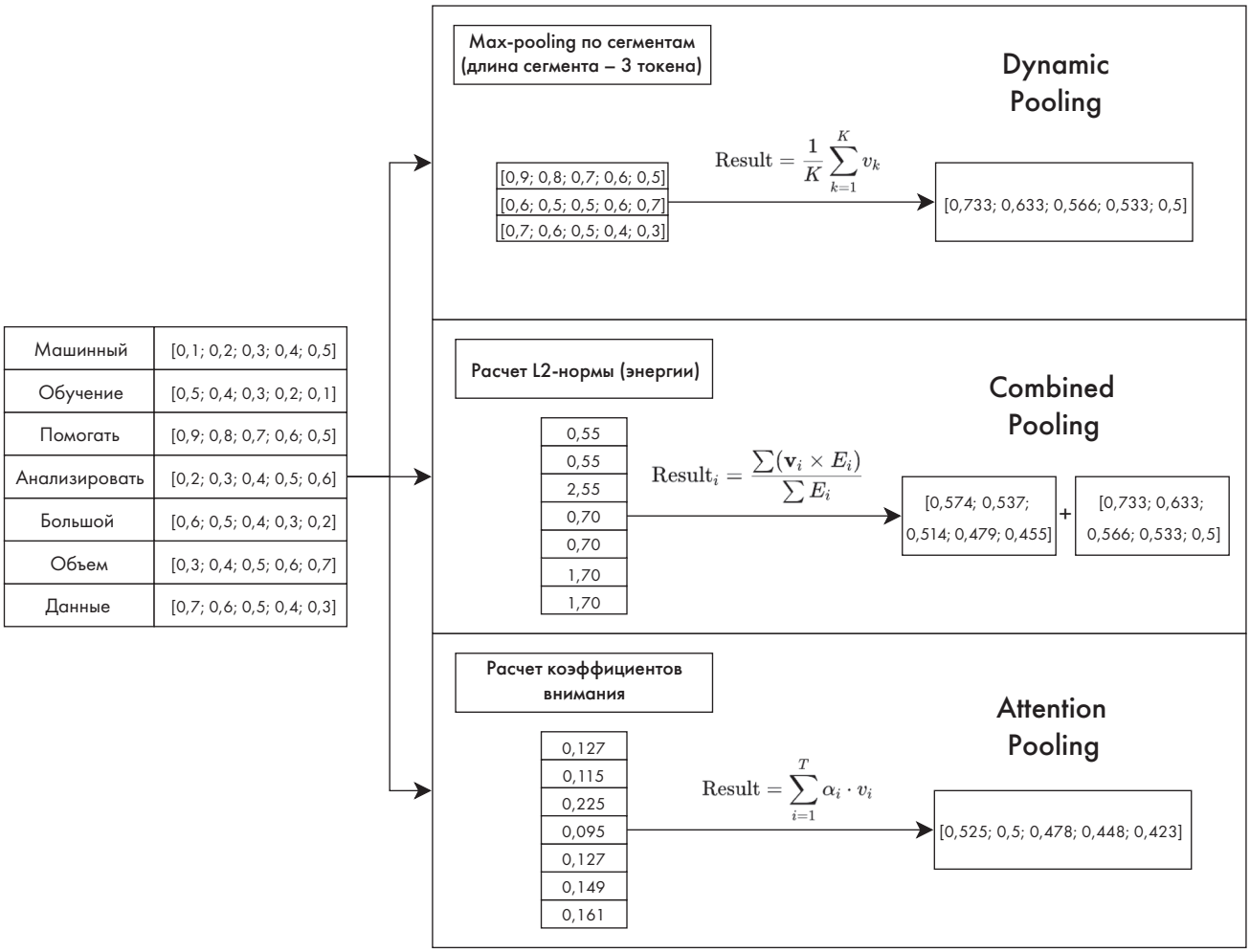

Рис. 1. Примеры преобразования матрицы векторов в единый агрегированный вектор текста с использованием различных исследованных подходов

Fig. 1. Examples of transforming a matrix of vectors into a single aggregated text vector using various explored approaches

[4] Xing J., Luo D., Xue C., Xing R. Comparative Analysis of Pooling Mechanisms in LLMs: A Sentiment Analysis Perspective. 2024. URL: http://arxiv.org/abs/2411.14654 (дата обращения: 01.11.2025).

[5] Sun X., Lu W. Understanding Attention for Text Classification. Труды Ежегодной конференции Ассоциации вычислительной лингвистики (ACL 2020). URL: https://doi.org/10.18653/v1/2020.acl-main.312 (дата обращения: 01.11.2025).

Dropout (p = 0,15). Обучение модели проводилось с минимизацией функции потерь на основе кросс-энтропии, используя Adam-оптимизатор. Для сравнения также применялась модель на основе ансамбля деревьев решений – RandomForestClassifier, объединенная в MultiOutputClassifier из библиотеки scikit-learn[6]. Поиск наилучших гиперпараметров (глубины деревьев, числа деревьев в ансамбле, минимального количества объектов для разбиения узла и минимального количества объектов в листе) выполнялся с помощью RandomizedSearchCV [11], используя KFold кросс-валидацию с тремя фолдами.

Для исследования эффективности моделей использовали набор классических метрик для мультиметочной классификации [12, 13]. Оценка производилась по следующим показателям: средняя площадь под характеристической кривой (macro AUROC), средняя площадь под кривой «точность–полнота» (macro AUPRC), точность (macro accuracy), средняя чувствительность (macro sensitivity), средняя специфичность (macro specificity) и средняя F1-мера (macro F1-score).

В качестве порога классификации результата работы моделей использовали максимумы индекса Юдена, рассчитанные для каждой из 156 меток. Однако фиксированные пороги классификации не позволяли оценить степень уверенности модели в каждом предсказании, что особенно важно в задаче мультиметочной классификации. Для решения этой проблемы была применена дополнительная интерпретация предсказанных вероятностей с использованием Bayesian Confidence Score [14]. Этот метод позволяет скорректировать предсказанные вероятности с учетом индивидуальных порогов, рассчитанных по индексу Юдена, и привести их к единой шкале уверенности в диапазоне от 0 до 1. Такой подход делает возможным корректное сравнение уверенности предсказаний между различными метками и позволяет более точно идентифицировать случаи, в которых модель демонстрирует высокую определенность.

РЕЗУЛЬТАТЫ

Для анализа был собран набор данных, включающий 1 000 000 первичных осмотров 387 590 пациентов за период с 2002 по 2025 г. Распределение возраста пациентов и пола в зависимости от размеченных на основании кодов МКБ-10 меток отражено в приложении 1. Количественные характеристики наборов данных после разделения на обучающий, тестовый и валидационный представлены в таблице 1.

Результаты всех проведенных экспериментов с использованием различных архитектур и подходов к векторному представлению текстов представлены в таблице 2.

Первые попытки с использованием ансамблевого метода на основе MultiOutputClassifier и Dynamic Pooling показали средний уровень предсказательной способности на тестовом наборе данных. Данная модель имела значительный размер – 1,5 ГБ, что ограничивало ее практическое применение, особенно в условиях недостатка вычислительных мощностей в практическом здравоохранении. В отличие от этого базовая нейросетевая модель, использующая Dynamic Pooling, показала значительно более высокую дискриминативную способность, с AUROC 0,881 и точностью 0,797, при этом ее размер составил всего 5 МБ. Этот фактор делает ее более подходящей для применения в реальных условиях, где важна не только точность, но и эффективность развертывания, что стало основанием для проведения всех дальнейших экспериментов именно на ее основе. После коррекции

Таблица 1. Описание наборов данных, использовавшихся в исследовании

Table 1. Description of the datasets used in the study

Показатель	Обучающий набор	Тестовый набор	Валидационный набор
Количество пациентов	284 672	61 205	41 713
Количество первичных осмотров	719 813	200 093	80 094

Таблица 2. Результаты экспериментов

Table 2. Experimental results

Алгоритм	AUROC	Точность	Чувствительность	Специфичность
MultiOutputClassifier + Dynamic Pooling	0,832	0,788	0,803	0,795
Нейронная сеть + Dynamic Pooling	0,881	0,797	0,813	0,797
Нейронная сеть + Dynamic Pooling + коррекция дисбаланса	0,850	0,778	0,776	0,778
Нейронная сеть + Combined Pooling	0,908	0,837	0,843	0,837
Нейронная сеть + Attention Pooling	0,917	0,852	0,856	0,852

[6] Buitinck L., Louppe G., Blondel M., et al. API design for machine learning software: experiences from the scikit-learn project. 2013. URL: http://arxiv.org/abs/1309.0238 (дата обращения: 01.11.2025).

дисбаланса классов при обучении (методами SMOTE [15], ADASYN [16]) наблюдалось небольшое снижение метрик качества, что тесно связано с изменением структуры выборки. Применение методов агрегации нескольких векторных представлений текста (Combined Pooling) дало более впечатляющие результаты. При использовании этого подхода на тестовой выборке AUROC вырос до 0,908, точность – до 0,837.

Лучший результат был получен с помощью нейросетевой архитектуры, оснащенной механизмом внимания (Attention Pooling). Ее архитектура представлена на рисунке 2.

Метрики качества модели, полученные по отдельным меткам, представлены на рисунке 3 для двадцати пяти наиболее часто встречавшихся заболеваний. Значительное преимущество данной архитектуры достигается за счет возможности исследования значимости исходных токенов для принятия моделью окончательных решений. Для этого используется механизм внимания, который учитывает внутреннюю фокусировку модели на ключевых элементах текста и отражает чувствительность предсказания модели к изменению конкретных входных признаков. В результате удается получить взвешенную оценку важности токенов, где учитываются прямой вклад каждого слова в предсказание с точки зрения обученной архитектуры. Такой подход позволяет выявлять наиболее информативные слова и фразы, определяющие

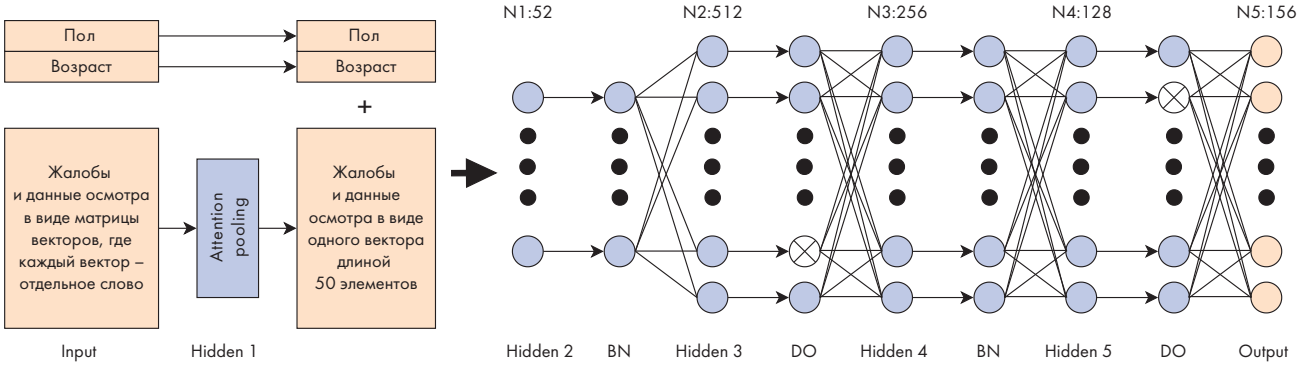

Рис. 2. Архитектуры нейросети с встроенным модулем Attention Pooling, показавшей лучшие результаты

Примечание: Input – input data, входные данные; Hidden – hidden layer, скрытый слой; BN – Batch Normalization (normalizes data within the layer, нормализует данные внутри слоя); DO – DropOut (randomly deactivates neurons during training, случайно отключает нейроны слоя во время обучения), Output – output predictions, выходные предсказания.

Fig. 2. Neural network architectures with an integrated Attention Pooling module that demonstrated the best performance

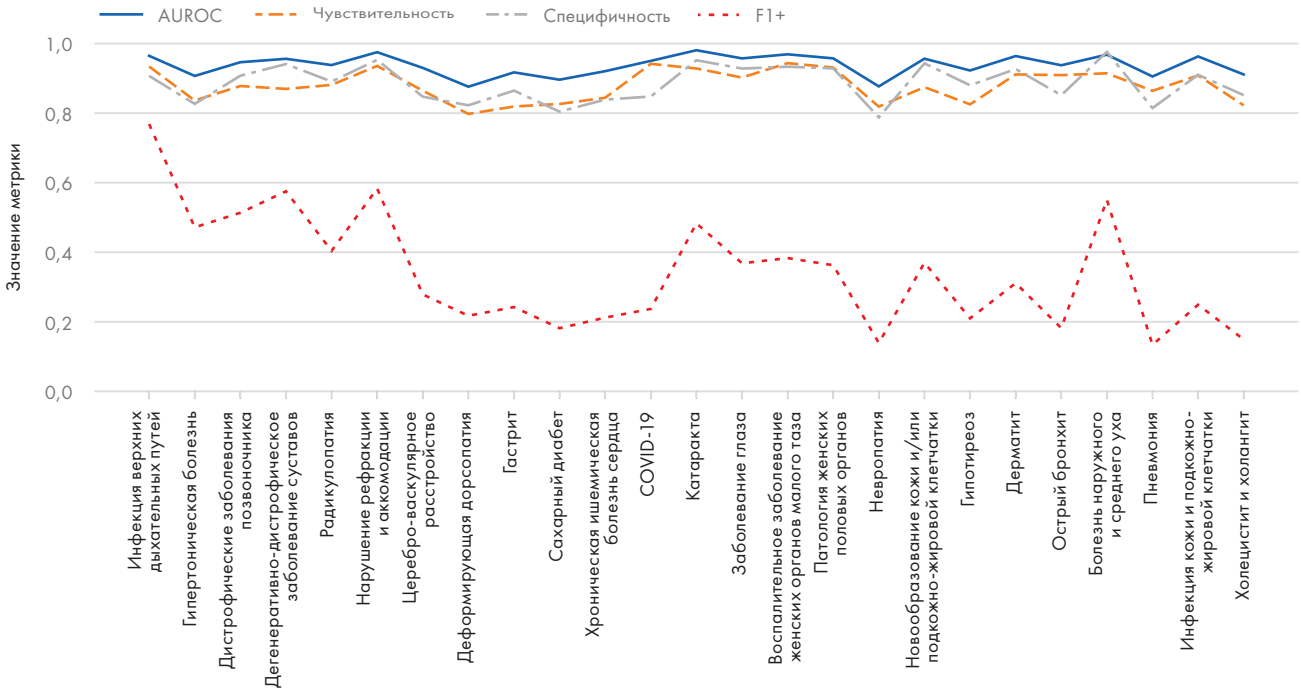

Рис. 3. Микро-AUROC, специфичность, чувствительность и F1-score для двадцати пяти наиболее часто встречавшихся нозологий

Fig. 3. Micro-AUROC, specificity, sensitivity, and F1-score for the twenty-five most frequent nosologies

классификацию, что способствует улучшению доверия к решениям в критически значимых приложениях.

ОБСУЖДЕНИЕ

Анализ медицинских текстов для прогнозирования заболеваний остается актуальной задачей, поскольку автоматизированные методы позволяют значительно улучшить процесс диагностики и связанного с этим принятия клинических решений. Одним из ключевых аспектов подобных работ является источник данных. Например, J.H.B. Masud и соавт. [5] использовали амбулаторные записи университетской больницы Тайбэя, J. Huang и M. Zeng работали с данными MIMIC-III [6, 7], включающими выписные эпикризы пациентов реанимационных отделений. В исследовании A. Blanco и соавт. [8] были задействованы записи на испанском языке, что накладывало дополнительные особенности на обработку текста, а K. Zhang и соавт. [9] сосредоточились на акушерских картах китайских пациентов. Источники данных важны для определения применимости разработки и сценария ее использования.

Сама обработка текста является этапом, на котором исследования разнились как по применяемым методам, так и по их глубине. В большинстве работ базовая предобработка включала удаление пунктуации и стоп-слов, однако ключевые различия связаны с выбором алгоритмов представления текста. Например, J.H.B. Masud и соавт. [5] использовали Word2Vec для векторизации, а команда J. Huang – TF-IDF и Word2Vec [6]. A. Blanco с соавт. расширили этот подход, применив контекстные векторные представления ELMo [8]. K. Zhang и соавт. сравнили несколько методов векторизации и показали, что использование контекстных эмбеддингов, таких как ELMo и fastText, позволяет лучше учитывать семантические особенности медицинского текста, что особенно важно при многометочной классификации [9].

В плане алгоритмов классификации исследования также демонстрируют разнообразие подходов. В ранних работах использовались традиционные модели, такие как случайный лес, наивный байесовский классификатор и логистическая регрессия [6, 9], но современные исследования делают акцент на глубоком обучении. Например, J.H.B. Masud и соавт. применили сверточные нейронные сети [5], а M. Zeng и соавт. использовали глубокое трансферное обучение, что позволило достичь лучших результатов [7].

Результаты нашего исследования согласуются с выводами предыдущих работ, посвященных анализу медицинских текстов, показавших, что использование нейросетевых моделей позволяет достичь высокой точности классификации медицинских записей. В нашей работе также применялись методы обработки естественного языка и машинного обучения для классификации медицинских текстов, однако с учетом специфики русскоязычной медицинской документации. В отличие от исследований, основанных на англоязычных базах данных, наш подход учитывает особенности русскоязычных записей, что делает его более адаптированным к локальной клинической практике. Применение обобщенных диагностических категорий позволило уменьшить сложность классификации без потери информативности, что также подтверждается результатами аналогичных работ, где использовалась группировка заболеваний для повышения устойчивости моделей [8].

Применение предложенной модели классификации медицинских текстов может способствовать своевременному выявлению заболеваний, повышению точности медицинских прогнозов и автоматизации обработки медицинской документации. Автоматизированная система может использоваться для поддержки медицинских работников при анализе текстовых записей и вынесении предварительных решений о вероятных диагнозах (рис. 4). Кроме того, диагностика заболеваний на основе текстовых данных может улучшить маршрутизацию пациентов и помочь в мониторинге групп риска. Это особенно актуально в условиях ограниченных ресурсов здравоохранения, когда автоматизированные инструменты могут снизить нагрузку на медицинский персонал и повысить эффективность принятия решений.

Несмотря на достигнутые результаты, исследование имеет ряд ограничений. Во-первых, модель обучалась на информации из одной базы данных (Webiomed), куда собирались и структурировались данные из различных медицинских учреждений различных регионов. Методы сбора данных имеют свои ограничения – правильность сбора и точность извлечения признаков не являются идеальными. Таким образом, эти особенности могут ограничивать обобщаемость модели на другие источники медицинской информации. Во-вторых, в качестве предикторов использовался ограниченный набор входных параметров (возраст, пол, текстовые записи), что не учитывает лабораторные и инструментальные исследования, потенциально влияющие на точность прогнозов. В-третьих, для анализа использовались тексты амбулаторных осмотров, которые имеют различия с документами из круглосуточных стационаров. В-четвертых, целевое событие определялось только кодом МКБ-10 и дополнительно не валидировалось. Также, несмотря на предпринятые меры по обработке текстов, возможны ошибки, связанные с неоднозначностью медицинских терминов и различиями в стилях ведения медицинской документации.

В дальнейшем целесообразно расширить набор используемых данных, включив структурированную информацию о пациентах. Также перспективным направлением является тестирование модели на независимых выборках, что позволит оценить ее устойчивость и адаптируемость к разным клиническим

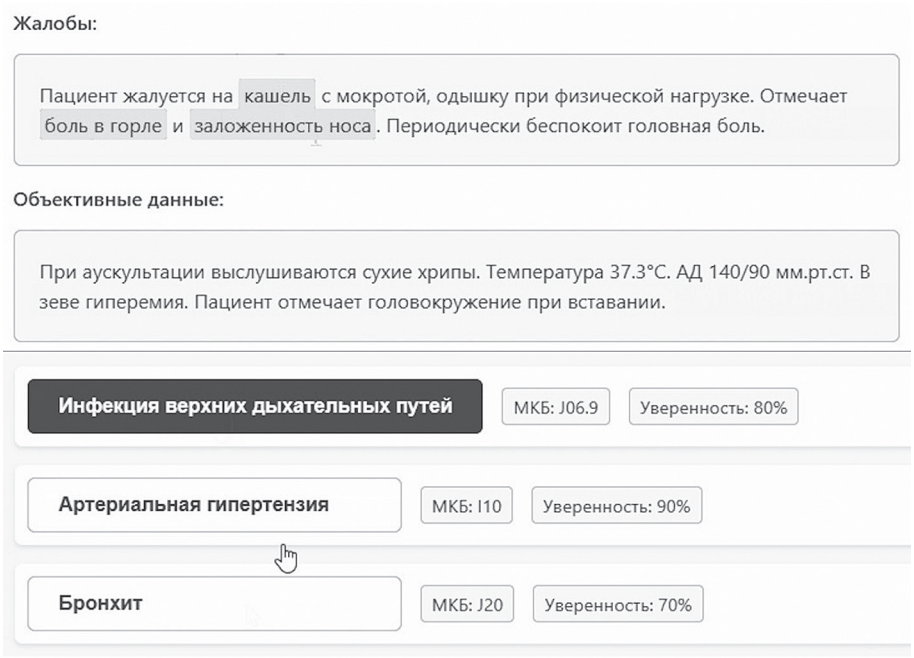

Рис. 4. Пример сервиса на основе разработанной модели машинного обучения с возможностью прогнозирования у пациента заболеваний на основе данных жалоб и объективного осмотра с функцией отображения значимых слов в исходном тексте
Fig. 4. Example of a service based on the developed machine learning model with the capability to predict patient diseases using complaint and physical examination data, including a feature for highlighting important words in the original text

условиям. Дополнительным направлением может стать разработка специализированных языковых моделей для русскоязычных медицинских текстов, аналогичных BERT, адаптированных под специфику медицинской терминологии.

ЗАКЛЮЧЕНИЕ

Анализ медицинских текстов с применением методов обработки естественного языка и алгоритмов машинного обучения представляет собой перспективное направление для повышения качества клинических решений. Результаты настоящего исследования подтверждают, что использование нейросетевых моделей, адаптированных под русскоязычную медицинскую документацию, обеспечивает высокую точность классификации и может служить эффективным инструментом поддержки принятия решений в здравоохранении. Предложенный подход, основанный на обобщенных диагностических категориях и учете локальных языковых особенностей, обладает высоким потенциалом для интеграции в существующие клинические процессы.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Conflict of interests. The authors declare that there is no conflict of interests.

Финансирование. Исследование не имело спонсорской поддержки (собственные ресурсы).

Financial support. The study was not sponsored (own resources).

ВКЛАД АВТОРОВ

А.Д. Ермак – обработка материала, анализ и интерпретация данных, написание текста рукописи.
Е.А. Макарова – анализ и интерпретация данных, редактирование статьи.
А.Н. Кафтанов – обработка материала, анализ и интерпретация данных.
Д.В. Гаврилов – обзор публикаций по теме статьи, анализ и интерпретация данных.
Р.Э. Новицкий – разработка дизайна исследования, сбор материала.
А.В. Гусев – разработка дизайна исследования, редактирование статьи.
Все авторы утвердили окончательную версию статьи.

AUTHOR CONTRIBUTIONS

Andrey D. Ermak – material processing, data analysis and interpretation, manuscript writing.
Elena A. Makarova – data analysis and interpretation, article editing.
Alexey N. Kaftanov – material processing, data analysis and interpretation.
Denis V. Gavrilov – publication review on the topic of the study, data analysis and interpretation.
Roman E. Novitskiy – study design development, data collection.
Alexandr V. Gusev – study design development, article editing.
All the authors approved the final version of the article.

ЛИТЕРАТУРА/REFERENCES

1 Spasic I., Nenadic G. Clinical text data in machine learning: Systematic review. JMIR Medical Informatics. 2020; 8(3): e17984. https://doi.org/10.2196/17984. PMID: 32229465
2 Hossain E., Rana R., Higgins N., et al. Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: A systematic review. Computers in Biology and Medicine. 2023; 155: 106649. https://doi.org/10.1016/j.compbiomed.2023.106649. PMID: 36805219
3 Wu S., Roberts K., Datta S., et al. Deep learning in clinical natural language processing: A methodical review. Journal of the American Medical Informatics Association. 2020; 27(3): 457–470. https://doi.org/10.1093/jamia/ocz200. PMID: 31794016
4 Kesiku C.Y.Y., Chaves-Villota A., Garcia-Zapirain B. Natural Language Processing Techniques for Text Classification of Biomedical Documents: A Systematic Review. Information (Switzerland). 2022; 13(10): 499. https://doi.org/10.3390/info13100499.
5 Masud J.H.B., Kuo C.C., Yeh C.Y., et al. Applying Deep Learning Model to Predict Diagnosis Code of Medical Records. Diagnostics. 2023; 13(13): 2297. https://doi.org/10.3390/diagnostics13132297
6 Huang J., Osorio C., Sy L.W. An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes. Computers Methods and Programs in Biomedicine. 2019; 177: 141–153. https://doi.org/10.1016/j.cmpb.2019.05.024. PMID: 31319942
7 Zeng M., Li M., Fei Z., et al. Automatic ICD-9 coding via deep transfer learning. Neurocomputing. 2019; 324: 43–50. https://doi.org/10.1016/j.neucom.2018.04.081
8 Blanco A., Perez-de-Viñaspre O., Pérez A., et al. Boosting ICD multi-label classification of health records with contextual embeddings and label-granularity. Computers Methods and Programs in Biomedicine. 2020; 188: 105264. https://doi.org/10.1016/j.cmpb.2019.105264. PMID: 31851906
9 Zhang K., Ma H., Zhao Y., et al. The Comparative Experimental Study of Multilabel Classification for Diagnosis Assistant Based on Chinese Obstetric EMRs. Journal of Healthcare Engineering. 2018; 2018: 7273451. https://doi.org/10.1155/2018/7273451. PMID: 29666671
10 Korobov M. Morphological analyzer and generator for Russian and Ukrainian languages. Communications in Computer and Information Science. Springer Verlag; 2015; 542: 320–332. https://doi.org/10.1007/978-3-319-26123-2_31
11 Bergstra J., Bengio Y. Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research. 2012; 13: 281–305.
12 Sokolova M., Lapalme G. A systematic analysis of performance measures for classification tasks. Information Processing and Management. 2009; 45(4): 427–437. https://doi.org/10.1016/j.ipm.2009.03.002
13 Hinojosa Lee M.C., Braet J., Springael J. Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores. Applied Sciences. 2024; 14(21): 9863. https://doi.org/10.3390/app14219863
14 Maltoudoglou L., Paisios A., Lenc L., et al. Well-calibrated confidence measures for multi-label text classification with a large number of labels. Pattern Recognition. 2022; 122: 108271. https://doi.org/10.1016/j.patcog.2021.108271
15 Chawla N.V., Bowyer K.W., Hall L.O., et al. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 2002; 16(1): 321–357. https://doi.org/10.1613/jair.953
16 He H., Bai Y., Garcia E.A., et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. Proceedings of the International Joint Conference on Neural Networks. 2008: 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

Информация об авторах

Ермак Андрей Дмитриевич – аналитик данных направления искусственного интеллекта ООО «К-Скай».
ORCID: https://orcid.org/0000-0002-0513-8557
Макарова Елена Андреевна – канд. техн. наук, руководитель направления искусственного интеллекта ООО «К-Скай».
ORCID: https://orcid.org/0000-0002-5410-5890
Кафтанов Алексей Николаевич – канд. мед. наук, аналитик данных направления искусственного интеллекта ООО «К-Скай».
ORCID: https://orcid.org/0000-0001-6898-8009
Гаврилов Денис Владимирович – руководитель медицинского направления ООО «К-Скай».
ORCID: https://orcid.org/0000-0002-8745-857X
Новицкий Роман Эдвардович – генеральный директор ООО «К-Скай».
ORCID: https://orcid.org/0000-0002-2350-977X
Гусев Александр Владимирович – канд. техн. наук, старший научный сотрудник отдела научных основ организации здравоохранения ФГБУ «Центральный научно-исследовательский институт организации и информатизации здравоохранения» Минздрава России.
ORCID: https://orcid.org/0000-0002-0513-8557

Information about the authors

Andrey D. Ermak – Data Analyst, Artificial Intelligence Division, K-SkAI LLC.
ORCID: https://orcid.org/0000-0002-0513-8557
Elena A. Makarova – Cand. of Sci. (Technical), Head of the Artificial Intelligence Division, K-SkAI LLC.
ORCID: https://orcid.org/0000-0002-5410-5890
Aleksey N. Kaftanov – Cand. of Sci. (Medicine), Data Analyst, Artificial Intelligence Division, K-SkAI LLC.
ORCID: https://orcid.org/0000-0001-6898-8009
Denis V. Gavrilov – Head of the Medical Division, K-SkAI LLC.
ORCID: https://orcid.org/0000-0002-8745-857X
Roman E. Novitsky – general manager, K-SkAI LLC.
ORCID: https://orcid.org/0000-0002-2350-977X
Alexandr V. Gusev – Cand. of Sci. (Technical), Senior Research Fellow, Department of Scientific Foundations of Health Care Organization, Russian Research Institute of Health.
ORCID: https://orcid.org/0000-0002-0513-8557