СРАВНЕНИЕ АНСАМБЛЯ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ И BERT ДЛЯ АНАЛИЗА ТЕКСТОВЫХ ОПИСАНИЙ КТ ГОЛОВНОГО МОЗГА НА ПРЕДМЕТ НАЛИЧИЯ ВНУТРИЧЕРЕПНЫХ КРОВОИЗЛИЯНИЙ

DOI: 10.17691/stm2024.16.1.03
УДК 616.831–005.1–073756.8:004.8
Поступила 13.10.2023 г.

© **А.Н. Хоружая,** младший научный сотрудник отдела инновационных технологий[1];
Д.В. Козлов, младший научный сотрудник отдела медицинской информатики, радиомики и радиогеномики[1];
К.М. Арзамасов, к.м.н., руководитель отдела медицинской информатики, радиомики и радиогеномики[1];
Е.И. Кремнева, к.м.н., ведущий научный сотрудник отдела инновационных технологий[1];
старший научный сотрудник[2]

[1]Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы, ул. Петровка, 24, стр. 1, Москва, 127051;
[2]Научный центр неврологии, Волоколамское шоссе, 80, Москва, 125367

Цель исследования — обучить и протестировать ансамбль моделей машинного обучения, а также сравнить характеристики его работы с предобученной на медицинских данных языковой моделью BERT в задаче простой бинарной классификации наличия/отсутствия признаков внутричерепного кровоизлияния (ВЧК) в протоколах описаний КТ головного мозга.

Материалы и методы. В качестве моделей, с помощью которых решалась задача бинарной классификации, было выбрано 7 алгоритмов машинного обучения и 3 метода векторизации текста. Обучение моделей проводили на текстовых данных, которые представляли собой протоколы описаний 3980 КТ-исследований головного мозга из 56 медицинских организаций стационарной медицинской помощи Москвы. Эти тексты были векторизованы тремя способами: «мешок слов», TF-IDF и Word2Vec. Далее к ним применяли следующие алгоритмы машинного обучения: дерево решений, случайный лес, логистическая регрессия, метод ближайших соседей, метод опорных векторов, CatBoost и XGBoost. Анализ данных, а также их предварительную обработку осуществляли с использованием библиотеки NLTK (Natural Language Toolkit, версия 3.6.5) и библиотеки Scikit-learn (версия 0.24.2). В качестве предообученной на медицинских данных модели-трансформера BERT была взята версия MedRuBertTiny2.

Результаты. По результатам обучения и тестирования семи алгоритмов машинного обучения выбраны три алгоритма с наиболее высокими метриками (чувствительность, специфичность): CatBoost, логистическая регрессия и метод ближайших соседей. Самые высокие метрики получены при использовании метода векторизации текста «мешок слов». Эти алгоритмы были собраны в ансамбль методом стекинга (stacking). Показатели чувствительности и специфичности на тестовом наборе данных из исходной выборки составили 0,93 и 0,90 соответственно. Далее ансамбль и модель BERT были обучены на независимом наборе данных, содержащем 9393 текстовых протокола диагностических описаний, разделенных также на обучающую и тестовую выборки. При тестировании на этом наборе данных ансамбля алгоритмов машинного обучения чувствительность и специфичность составили 0,92 и 0,90 соответственно. Тестирование на этих данных модели BERT продемонстрировало чувствительность 0,97 и специфичность 0,90.

Заключение. Обученный ансамбль показал высокие результаты точности работы при анализе текстовых протоколов описаний КТ головного мозга с признаками внутричерепных кровоизлияний, но все равно при его использовании необходимо обеспечить ручной пересмотр результатов для контроля качества. Предобученная модель-трансформер BERT, дополнительно обученная на диагностических текстах, продемонстрировала более высокие метрики точности ($p<0,05$). Это говорит о перспективности модели в задачах бинарной классификации и для поиска информации по неструктурированным медицинским записям.

Ключевые слова: компьютерная томография; диагностические описания; внутричерепное кровоизлияние; обработка естественного языка; машинное обучение; BERT.

Как цитировать: Khoruzhaya A.N., Kozlov D.V., Arzamasov K.M., Kremneva E.I. Comparison of an ensemble of machine learning models and the BERT language model for analysis of text descriptions of brain CT reports to determine the presence of intracranial hemorrhage. Sovremennye tehnologii v medicine 2024; 16(1): 27, https://doi.org/10.17691/stm2024.16.1.03

Для контактов: Хоружая Анна Николаевна, e-mail: KhoruzhayaAN@zdrav.mos.ru

English

Comparison of an Ensemble of Machine Learning Models and the BERT Language Model for Analysis of Text Descriptions of Brain CT Reports to Determine the Presence of Intracranial Hemorrhage

A.N. Khoruzhaya, Junior Researcher, Department of Innovative Technologies[1];
D.V. Kozlov, Junior Researcher, Department of Medical Informatics, Radiomics and Radiogenomics[1];
K.M. Arzamasov, MD, PhD, Head of the Department of Medical Informatics, Radiomics and Radiogenomics[1];
E.I. Kremneva, MD, PhD, Leading Researcher, Department of Innovative Technologies[1]; Senior Researcher[2]

[1]Scientific and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Department of Health, Bldg 1, 24 Petrovka St., Moscow, 127051, Russia;
[2]Research Center for Neurology, 80 Volokolamskoye Shosse, Moscow, 125367, Russia

The aim of this study is to train and test an ensemble of machine learning models, as well as to compare its performance with the BERT language model pre-trained on medical data to perform simple binary classification, i.e., determine the presence/absence of the signs of intracranial hemorrhage (ICH) in brain CT reports.

Materials and Methods. Seven machine learning algorithms and three text vectorization techniques were selected as models to solve the binary classification problem. These models were trained on textual data represented by 3980 brain CT reports from 56 inpatient medical facilities in Moscow. The study utilized three text vectorization techniques: bag of words, TF-IDF, and Word2Vec. The resulting data were then processed by the following machine learning algorithms: decision tree, random forest, logistic regression, nearest neighbors, support vector machines, Catboost, and XGboost. Data analysis and pre-processing were performed using NLTK (Natural Language Toolkit, version 3.6.5), libraries for character-based and statistical processing of natural language, and Scikit-learn (version 0.24.2), a library for machine learning containing tools to tackle classification challenges. MedRuBertTiny2 was taken as a BERT transformer model pre-trained on medical data.

Results. Based on the training and testing outcomes from seven machine learning algorithms, the authors selected three algorithms that yielded the highest metrics (i.e. sensitivity and specificity): CatBoost, logistic regression, and nearest neighbors. The highest metrics were achieved by the bag of words technique. These algorithms were assembled into an ensemble using the stacking technique. The sensitivity and specificity for the validation dataset separated from the original sample were 0.93 and 0.90, respectively. Next, the ensemble and the BERT model were trained on an independent dataset containing 9393 textual radiology reports also divided into training and test sets. Once the ensemble was tested on this dataset, the resulting sensitivity and specificity were 0.92 and 0.90, respectively. The BERT model tested on these data demonstrated a sensitivity of 0.97 and a specificity of 0.90.

Conclusion. When analyzing textual reports of brain CT scans with signs of intracranial hemorrhage, the trained ensemble demonstrated high accuracy metrics. Still, manual quality control of the results is required during its application. The pre-trained BERT transformer model, additionally trained on diagnostic textual reports, demonstrated higher accuracy metrics ($p<0.05$). The results show promise in terms of finding specific values for both binary classification task and in-depth analysis of unstructured medical information.

Key words: computed tomography; diagnostic reports; intracranial hemorrhage; natural language processing; machine learning; BERT.

Введение

Качественный анализ клинических данных с использованием различных алгоритмов машинного обучения приобретает все большее значение в научных исследованиях и практическом здравоохранении. Увеличивающиеся с каждым годом объемы текстовой информации создают трудности для врачей, осуществляющих сбор и статистическую обработку медицинских данных; исследователей, которые анализируют эти данные с целью получения новых научных знаний; а также для разработчиков программного обеспечения [1, 2].

Неструктурированные тексты, включая медицинские карты, диагностические протоколы описаний, отзывы пациентов, комментарии в социальных сетях, являются богатейшим источником данных для научных исследований. Однако анализ таких текстов вручную занимает много времени и чреват ошибками. Особенно важно быстрое и качественное извлечение необходимой информации из рентгенологических заключений. Эта информация и ее последующая автоматическая обработка способны помогать в минимальные сроки принимать эффективные решения при диагностике той или иной патологии, что особенно актуально в сфере экстренной и неотложной медицинской помощи — например, для диагностики внутричерепных кровоизлияний (ВЧК) [3, 4].

Для преобразования письменного текста в интерпретируемые компьютером наборы данных применя-

ют различные методы обработки естественного языка (natural language processing, NLP) [5]. Полученные таким образом данные могут быть проанализированы с помощью моделей машинного обучения (machine learning, ML) [6], в том числе с помощью более прогрессивных подходов, включающих глубокое обучение (deep learning, DL). Глубокое обучение — это подмножество методов машинного обучения, в котором для анализа данных используются искусственные нейронные сети. Алгоритмы глубокого обучения могут применяться для анализа медицинских текстов, а также для выявления закономерностей и взаимосвязей в данных. Тем не менее каждый единичный алгоритм обладает своими минусами и не слишком точно работает, например, с низкоструктурированными текстами [7]. По этой причине рекомендуется создавать ансамбли алгоритмов, которые бы сочетали в себе лучшие качества всех отдельных моделей.

Эффективность ансамблей алгоритмов машинного обучения для задач NLP в медицине была продемонстрирована в ограниченном количестве работ, посвященных анализу медицинской информации и извлечению из текста определенных признаков [8, 9]. Однако те данные, которые имеются, свидетельствуют о том, что этот подход позволяет справляться с бинарной или мультиклассовой классификацией с достаточно высокой точностью, более существенной, чем у единичного алгоритма. Например, показатель AUROC для признака «интракраниальный масс-эффект» в текстовых протоколах КТ головного мозга составил 0,96 для коллективной модели XGBoost при векторизации текста методом TF-IDF (term frequency-invers document frequency) [8].

В последнее время для задач обработки естественного языка, в том числе низкоструктурированных медицинских текстов, имеющих высокую вариативность в описаниях, применяется модель двунаправленных презентаций кодировщика для трансформеров (Bidirectional Encoder Representations from Transformers, BERT) [10]. Языковая модель BERT способна решать широкий спектр задач обработки естественного языка за счет того, что считывает текстовые данные как справа налево, так и слева направо (двунаправленно). Поэтому она демонстрирует лучшие качества по сравнению со своими предшественниками, которые были однонаправленными. BERT состоит из нескольких слоев, образующих «трансформер», который изучает контекстуальные связи и близость между различными словами из текстовых данных. Трансформеры нацелены на анализ слов: они связывают слова, чтобы распознать семантику предложения для лучшего понимания его общего смысла [11]. Даже без дополнительного обучения на специфических данных медицинских текстов модель BERT способна предъявлять довольно высокие показатели точности за счет предварительного обучения на большом объеме данных с другими задачами (например, анализа изображений), а с дообучением и вовсе может превзойти имеющиеся способы автоматической текстовой обработки [12, 13].

Цель исследования — создание, обучение и тестирование ансамбля моделей машинного обучения, который бы достигал максимальных значений характеристик точности, а также сравнение работы этого ансамбля с предобученной на медицинских данных языковой моделью BERT в задаче простой бинарной классификации наличия/отсутствия признаков внутричерепного кровоизлияния в протоколах описаний компьютерной томографии головного мозга.

Материалы и методы

Исходные данные представляют собой выгрузку из Единого радиологического информационного сервиса Единой медико-информационной аналитической системы (ЕРИС ЕМИАС) [14], содержащую 34 188 исследований, полученных в результате проведения бесконтрастной КТ головного мозга в 56 медицинских организациях стационарной медицинской помощи. Анализ данных, а также их предварительную обработку осуществляли с использованием NLTK (Natural Language Toolkit, версия 3.6.5) — библиотеки для символьной и статистической обработки естественного языка и Scikit-learn (версия 0.24.2) — библиотеки для машинного обучения, содержащей инструменты для задач классификации. Автоматический отбор протоколов описаний и их последующую экспертную верификацию проводили по 14 ключевым словам, специфичным для ВЧК, а также 64 стоп-фразам, наличие которых в тексте подразумевало отсутствие этого состояния. Отбор текстов с искомой патологией осуществляли при наличии следующих ключевых слов (в том числе содержащих указание на тип кровоизлияния): кровоизлиян-, гематом-, геморрагическ-, внутримозгов-, субарахноидальн-, эпидуральн-, субдуральн-, внутрижелудочков-, САК (субарахноидальное кровоизлияние), ЭДК (эпидуральное кровоизлияние), СДК (субдуральное кровоизлияние), ВМК (внутримозговое кровоизлияние), ВЖК (внутрижелудочковое кровоизлияние), паренхиматозн-. При этом в текстах должны были отсутствовать стоп-фразы: например, «КТ-данных за внутричерепную гематому и ушиб головного мозга не получено», «признаков внутричерепного кровоизлияния не выявлено» и др. Описание визуальной картины любой крови, в том числе постоперационной или посттравматической, также приравнивалось к наличию искомой патологии. К описанию кровоизлияния относилось в том числе указание плотности содержимого от 40 до 90 единиц Хаунсфилда (HU). Например, следующее описание было приравнено к содержащему искомую патологию: «На серии компьютерных томограмм в левой височной области отмечаются геморрагические очаги размерами до 20, 11, 8, 6, 4 мм плотностью до 65 HU. В левой теменной

области отмечается геморрагическое содержимое, повторяющее контуры борозд».

В результате был получен набор данных (датасет 1) с двумя классами текстовых протоколов: с описанием ВЧК и без такового. Использовались полные тексты рентгенологических протоколов, содержащие как описание, так и заключение, их длина колебалась от 310 до 3554 символов с пробелами. Подробнее с алгоритмом отбора можно ознакомиться в нашем предыдущем исследовании [15].

Для оценки производительности модели протоколы из датасета 1 были разделены случайным образом на выборки в соотношении 7:3, поскольку именно такое соотношение обучающего/тестового набора данных позволяет получить наиболее оптимальные метрики качества работы алгоритма [16]. Из 3980 протоколов 2786 были отнесены к обучающему набору данных, 1194 — к тестовому. Из 1194 тестовых наборов 927 не содержали признаков ВЧК, 267 имели такие признаки. Все протоколы имели уникальный идентификатор, позволяющий исключить утечку данных из обучающего в тестовый набор.

В качестве моделей, с помощью которых решалась задача бинарной классификации, выбрано семь алгоритмов машинного обучения и три метода векторизации текста. Использованы следующие алгоритмы: логистическая регрессия (logistic regression), случайный лес (random forest), библиотека градиентного бустинга (CatBoost, версия 1.1.1), метод опорных векторов (support vector machine, SVM), метод k-ближайших соседей (k-nearest neighbors, KNN), библиотека градиентного бустинга (XGBoost, версия 1.7.1) из библиотеки Scikit-learn на языке программирования Python (версия 3.9.7). Для каждого алгоритма проводили поиск оптимальных гиперпараметров методом перебора.

Помимо алгоритмов машинного обучения из библиотеки Scikit-learn использовали разные методы векторного представления текстовых протоколов на естественном языке: «мешок слов» (bag of words), TF-IDF и Word2Vec.

Метод векторизации «мешок слов» создает таблицу (словарь), где каждое уникальное слово в тексте представлено отдельным столбцом, а строки соответствуют предложениям. Если слово встречается в предложении, в ячейке таблицы ставится 1; если его там нет — 0. TF-IDF рассчитывает ценность слова для строки и текста в целом по встречаемости уникального слова в строке. С математической точки зрения формула для определения TF-IDF имеет следующий вид:

$$TF\text{-}IDF=TF\cdot IDF,$$

где TF — встречаемость уникального слова в строке, IDF — обратная частота документа (показывает, какое количество раз слово встречается во всем наборе данных).

Word2Vec представляет собой нейронные сети, которые способны оценивать косинусную близость векторов слов.

В качестве предообученной модели BERT была взята версия MedRuBertTiny2 [17]. Она отлажена на основе специально собранного набора данных из более чем 30 000 медицинских анамнезов на русском языке. Эта модель создана в рамках проекта по разработке метода исправления опечаток в историях болезни с использованием моделей BERT в качестве ранжирования кандидатов (т.е. им присваивались оценка или вес для определения того, какие из них наиболее релевантны для конкретной задачи и имеют большее значение). MedRuBertTiny2 была дообучена со следующими техническими параметрами: скорость обучения — lr=1e-5, n_splits=4, epoch=10.

Для дообучения и повторного тестирования ансамбля алгоритмов и модели BERT использовался новый, независимый размеченный набор данных (датасет 2), собранный по подобию датасета 1, но с бóльшим количеством текстов. Этот набор содержал 9393 протокола описаний (5443 без описания патологии и 3950 с описанием ВЧК), которые были разделены на обучающий (6790) и тестовый (2603) наборы. Тексты в датасетах 1 и 2 не повторяются.

Оценка качества работы алгоритмов проводилась при помощи функции classification_report. Для статистического анализа использовался Mc Nemar's test. Проверялась нулевая гипотеза об отсутствии статистически значимых различий между показателями чувствительности и специфичности алгоритмов машинного обучения и их ансамблей против альтернативной — об их наличии.

Для улучшения качества работы моделей была проведена предварительная обработка текста, которая заключалась в приведении всех букв в словах к нижнему регистру (А→а), удалении лишних символов и слов (предлоги, союзы, частицы), лемматизации и разбиении текста на токены (разделение предложений на слова-компоненты). Затем предобработанный текст был векторизован тремя способами: «мешок слов», TF-IDF и Word2Vec.

Результаты

К предобработанному и векторизованному тексту были применены все семь исследуемых алгоритмов машинного обучения из библиотеки Scikit-learn. Каждый из алгоритмов машинного обучения тестировался с поочередным применением всех трех методов векторизации текстов. Результаты тестирования приведены в табл. 1–3.

В результате анализа полученных метрик было принято решение о применении метода сбора ансамбля (stacking) из алгоритмов с наиболее высокими метриками, при котором обучение происходило на двух моделях и передача результата осуществлялась на вход третьей. Обучение и тестирование происходило на датасете 1 с использованием поочеред-

Таблица 1

Результаты тестирования алгоритмов машинного обучения с применением метода векторизации текста «мешок слов»

Алгоритм	Точность	Полнота	F1-score	Чувствительность	Специфичность
Дерево решений					
Кровоизлияние	0,78	0,74	0,76	0,93	0,77
Норма	0,93	0,95	0,94		
Логистическая регрессия					
Кровоизлияние	0,80	0,85	0,82	0,95	0,85
Норма	0,96	0,95	0,95		
Случайный лес					
Кровоизлияние	0,86	0,13	0,22	0,99	0,13
Норма	0,82	0,99	0,90		
Метод ближайших соседей					
Кровоизлияние	0,63	0,86	0,73	0,87	0,86
Норма	0,96	0,87	0,92		
CatBoost					
Кровоизлияние	0,76	0,78	0,77	0,94	0,78
Норма	0,94	0,94	0,94		
XGBoost					
Кровоизлияние	0,86	0,79	0,83	0,79	0,97
Норма	0,95	0,97	0,96		
Метод опорных векторов					
Кровоизлияние	0,80	0,86	0,83	0,94	0,86
Норма	0,96	0,94	0,95		

Таблица 2

Результаты тестирования алгоритмов машинного обучения с применением метода векторизации текста TF-IDF

Алгоритм	Точность	Полнота	F1-score	Чувствительность	Специфичность
Дерево решений					
Кровоизлияние	0,67	0,69	0,68	0,81	0,65
Норма	0,91	0,90	0,9		
Логистическая регрессия					
Кровоизлияние	0,87	0,78	0,82	0,96	0,78
Норма	0,94	0,96	0,95		
Случайный лес					
Кровоизлияние	0,88	0,50	0,64	0,98	0,50
Норма	0,87	0,98	0,92		
Метод ближайших соседей					
Кровоизлияние	0,77	0,76	0,77	0,93	0,76
Норма	0,93	0,93	0,93		
CatBoost					
Кровоизлияние	0,82	0,79	0,81	0,94	0,78
Норма	0,94	0,94	0,94		

Окончание табл. 2

Алгоритм	Точность	Полнота	F1-score	Чувствительность	Специфичность
Метод опорных векторов					
Кровоизлияние	0,84	0,82	0,83	0,50	0,82
Норма	0,95	0,95	0,95		
XGBoost					
Кровоизлияние	0,94	0,95	0,94	0,79	0,95
Норма	0,82	0,79	0,80		

Таблица 3

Результаты тестирования алгоритмов машинного обучения с применением метода векторизации текста Word2Vec

Алгоритм	Точность	Полнота	F1-score	Чувствительность	Специфичность
Дерево решений					
Кровоизлияние	0,80	0,59	0,68	0,95	0,59
Норма	0,88	0,95	0,91		
Логистическая регрессия					
Кровоизлияние	0,81	0,69	0,75	0,95	0,69
Норма	0,91	0,95	0,93		
Случайный лес					
Кровоизлияние	0,86	0,76	0,81	0,96	0,76
Норма	0,93	0,96	0,94		
Метод ближайших соседей					
Кровоизлияние	0,86	0,77	0,81	0,96	0,77
Норма	0,93	0,96	0,94		
CatBoost					
Кровоизлияние	0,79	0,69	0,73	0,94	0,78
Норма	0,90	0,94	0,92		
Метод опорных векторов					
Кровоизлияние	0,81	0,73	0,77	0,95	0,73
Норма	0,92	0,95	0,93		
XGBoost					
Кровоизлияние	0,92	0,92	0,92	0,73	0,92
Норма	0,71	0,73	0,72		

но трех методов векторизации текста. Результаты приведены в табл. 4.

Основываясь на данных, представленных в табл. 4, можно отметить, что ансамбль алгоритмов машинного обучения, состоящий из CatBoost, логистической регрессии и k-ближайших соседей с методом векторизации текста «мешок слов» (Stacking CatBoost, Random LogReg & KNN, bag of words), предъявил наилучшие результаты по специфичности ($p<0,05$), тогда как показатели чувствительности во всех трех способах векторизации текста статистически значимо не различались ($p>0,05$).

Данный ансамбль был дообучен и протестирован на датасете 2. Чувствительность при этом составила 0,92, специфичность — 0,90. Обращает на себя внимание тот факт, что метрики значимо не изменились ($p>0,05$). При этом в кривой обучения отмечается замедление прогресса и выход на плато, что говорит о достижении некоторого предела при использовании подобного подхода. На рис. 1 представлена его матри-

Таблица 4

Результаты тестирования ансамблей алгоритмов машинного обучения с применением трех методов векторизации текстов

Алгоритм	Точность	Полнота	F1-score	Чувствительность	Специфичность
Stacking CatBoost, Random LogReg & KNN, TF-IDF					
Кровоизлияние	0,82	0,84	0,83	0,94	0,84
Норма	0,95	0,94	0,95		
Stacking CatBoost, Random LogReg & KNN, Word2Vec					
Кровоизлияние	0,42	0,14	0,21	0,94	0,14
Норма	0,78	0,94	0,84		
Stacking CatBoost, Random LogReg & KNN, bag of words					
Кровоизлияние	0,78	0,90	0,84	0,93	0,90
Норма	0,97	0,93	0,95		

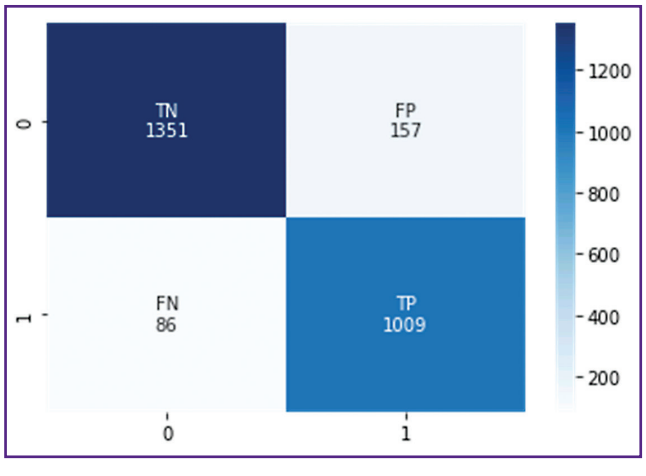

Рис. 1. Матрица ошибок для ансамбля алгоритмов CatBoost, Random LogReg & KNN, bag of words
По вертикали истинная оценка исследования: *0* — исследование без признаков внутримозговых кровоизлияний (истинно-отрицательный результат); *1* — исследование с признаками кровоизлияний (истинно-положительный результат). По горизонтали оценки ансамбля: *0* — наличие патологии выявлено ошибочно (ложноположительный результат), *1* — отсутствие патологии указано ошибочно (ложноотрицательный результат)

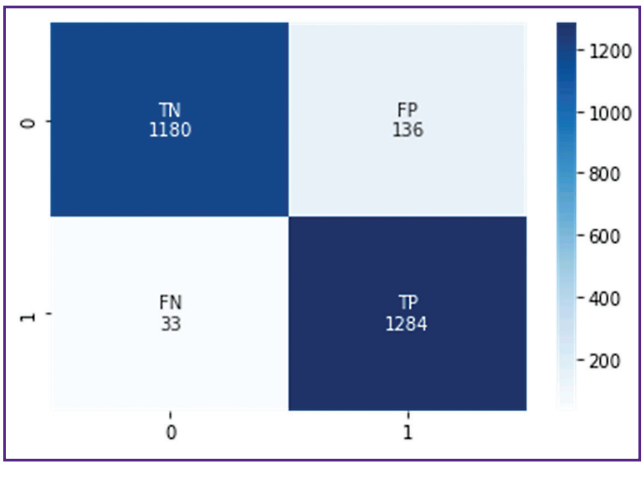

Рис. 2. Матрица ошибок для модели BERT
По вертикали истинная оценка исследования: *0* — исследование без признаков внутримозговых кровоизлияний (истинно-отрицательный результат); *1* — исследование с признаками кровоизлияний (истинно-положительный результат). По горизонтали оценки BERT: *0* — наличие патологии выявлено ошибочно (ложноположительный результат), *1* — отсутствие патологии указано ошибочно (ложноотрицательный результат)

ца ошибок, демонстрирующая количество истинно- и ложноположительных и отрицательных результатов.

На том же независимом наборе данных (датасет 2) прошла дообучение и тестирование предобученная медицинская модель BERT. Чувствительность составила 0,97, специфичность — 0,90. Данные метрики статистически значимо отличаются в лучшую сторону ($p<0,05$) от метрик, полученных в результате дообучения и тестирования ансамбля алгоритмов машинного обучения на том же наборе, несмотря на то, что модель BERT суммарно была обучена на меньшем корпусе диагностических рентгенологических текстов. На рис. 2 представлена ее матрица ошибок, демонстрирующая количество истинно- и ложноположительных, а также отрицательных результатов.

Обсуждение

В предыдущем исследовании [15] мы обсуждали возможность использования алгоритма дерева решений для задачи бинарной классификации протоколов заключений КТ головного мозга на предмет наличия ВЧК. У данного алгоритма самая высокая интерпретируемость среди других методов машинного обучения

в совокупности с простотой и возможностью автоматического обучения [18]. Это послужило причиной выбора данного алгоритма на первом, пилотном этапе для создания программы автоматического анализа диагностических текстов. Однако в процессе исследования выяснилось, что у него имеются существенные ограничения в виде ложных срабатываний; сложностей с классификацией текстов, в которых описание наличия и отсутствия искомой патологии сильно варьирует; а также необходимости ручного пересмотра исследований для обеспечения контроля качества [15].

Поэтому было решено перейти к усложнению классификатора. Для этого мы избрали подход с созданием ансамблей из нескольких алгоритмов машинного обучения и применили несколько вариантов векторизации текста, которые переводят письменную речь в формат, доступный для автоматической обработки. Обученный ансамбль показал довольно высокие результаты относительно точности работы при анализе текстовых протоколов описаний КТ головного мозга с признаками внутричерепных кровоизлияний. Однако даже в этом случае оказался необходим ручной пересмотр для контроля качества.

По нашему мнению, основанному на ручном пересмотре массива неправильно интерпретируемых автоматически протоколов описаний, основная причина ошибок кроется в том, что ансамбль алгоритмов машинного обучения все равно не учитывает семантические особенности конструкции рентгенологических протоколов и контекстуальную близость терминов, в них встречающихся. Например, следующий протокол описания был ошибочно размечен как содержащий картину искомых патологических изменений в головном мозге:

«При КТ-исследовании в веществе головного мозга патологических очагов ушибов не определяется. В левой теменной и затылочной области субдуральной гематомы и пневмоцефалии не определяется. В базальных отделах лобных долей САК четко не определяются. В остальном без динамики. Срединные структуры не смещены. Боковые желудочки симметричны, содержимое однородное. Цистерны основания мозга прослеживаются, не деформированы. Щели субарахноидальных пространств и конвекситальные борозды не расширены. <…> При сравнении с КТ ГМ вот 27.12.2022 г положительная динамика: отмечается регресс очага ушиба и САК в в базальных лобных отделах слева, пневмоцефалии и пластинчатой субдуральной гематомы. Перелом височной кости слева. Перелом затылочной кости. Патологическое содержимое в ячейках сосцевидного отростка слева. Полисинусит.».

При чтении протокола специалисту становится ясно, что патологические изменения в виде очага ушиба, субдуральной гематомы и САК регрессировали и более не определяются на КТ головного мозга, соответственно данное исследование можно интерпретировать как «норму» с точки зрения ВЧК. Однако анализ его описания по ключевым словам и стоп-конструкциям, которые принимают во внимание алгоритмы машинного обучения, затруднен.

Кроме того, случаются и неверные интерпретации в виде ложноотрицательных ответов. Например, таким образом ансамблем был размечен следующий протокол:

«В базальных ядрах и в островковой доле слева с переходом на базальные отделы височной области определяется гиподенсный участок плотностью +16…+19 ед.Н., размерами 50х28х35мм, несколько краниальнее данного участка в лобной области точечный гиперденсный участок до 5мм в диаметре. Снижение дифференцировки серого и белого вещества, сглаженность борозд в лобно-теменно-височной области слева. ASPECTS в бассейне левой СМА суммарно 5 баллов. <…> Свежих костно-травматических изменений достоверно не выявлено. Заключение: ранние КТ-признаки ишемии в лобно-теменно-височной области слева. Подострая ишемия в базальных ядрах, в островковой доли и в височной области слева. Гиперденсный очаг в лобной области слева- точечное кровоизлияние? гиперденсный сосуд? Рекомендуется КТ в динамике.».

По данному заключению можно сделать вывод, что врач описал некую гиперденсную область, но не уверен в ее субстрате. Тем не менее она может являться кровоизлиянием, и ошибочное исключение данного протокола, в зависимости от цели, будет нежелательным.

Следует отметить, что на появление такого рода неточностей могла повлиять предобработка набора данных. Этот факт является одним из ограничений данной работы и требует отдельного исследования.

Дообученная на корпусе диагностических текстов модель-трансформер BERT продемонстрировала более высокие метрики точности, получив специфические семантические и контекстуальные связи, характерные для рентгенологических протоколов описаний. Донастройка гиперпараметров модели и ее прицельное дообучение на наборах данных с большим количеством протоколов описаний могут позволить еще больше повысить ее производительность, тогда как дообучение ансамбля алгоритмов машинного обучения в рамках данной задачи, по-видимому, к значимому улучшению результата не приведет [19].

Отраженные в работе инструменты могут хуже работать на заключениях врачей, которые описывают рентгеновские снимки отлично от стандарта, поддерживаемого в медицинских организациях Департамента здравоохранения Москвы, или на заключениях с грамматическими ошибками. Данный аспект также является ограничением и требует дополнительного исследования (возможно, с корпусом текстов из иных медицинских учреждений).

На данный момент имеется множество сообщений об адаптации модели BERT для анализа медицинских текстов, представленных на разных языках: на араб-

ском [20], немецком [21], турецком [22], корейском [23], китайском [24] и других. Сообщается также, что для достижения максимальной точности в задачах NLP в области медицины можно объединять традиционные подходы машинного обучения в качестве первичной классификации текстов и BERT для более точного анализа на выявление в текстах искомых признаков или смыслов [25].

Стремление достичь самых высоких показателей точности алгоритмов для анализа неструктурированных медицинских текстов продиктовано текущими проблемами и ограничениями, в целом характерными для использования искусственного интеллекта в медицине. И в первую очередь к ним относится качество данных, используемых для обучения, например, алгоритмов компьютерного зрения. Создание качественных наборов данных — процесс трудоемкий и небыстрый. Неструктурированные медицинские тексты, по которым происходит отбор самих диагностических изображений, могут содержать ошибки, несоответствия и пропущенные данные, и это в конечном счете повлияет на точность результатов [26]. Чем больше в руках врачей и экспертов будет высокопроизводительных инструментов автоматического отбора для создания подобных наборов данных, тем лучше.

Кроме того, подобные инструменты могут иметь исключительную важность для организации здравоохранения. Например, они способны упрощать подготовку различных статистических отчетов и помогать отслеживать работу медицинских информационных систем, предназначенных для автоматизации лечебно-диагностических, административных, вспомогательных и иных процессов [27].

Заключение

Обученный ансамбль алгоритмов машинного обучения показал высокие результаты работы при анализе текстовых протоколов описаний КТ-исследований головного мозга с признаками внутричерепных кровоизлияний и в целом может использоваться в задачах бинарной классификации. Однако мы не можем отказаться от ручного пересмотра для контроля качества. Предобученная медицинская модель-трансформер BERT продемонстрировала при дообучении на том же наборе данных статистически значимо более высокие метрики точности, которые при дальнейшем подборе гиперпараметров и дообучении модели на большем количестве диагностических текстов могут оказаться еще выше. Это говорит о ее высоком потенциале и возможности дальнейшего совершенствования для анализа неструктурированной медицинской информации с целью выявления определенных значений: например, факта оперативного вмешательства или кровоизлияний в разных стадиях развития.

Тем не менее наиболее эффективный инструмент для анализа текстовых протоколов диагностических описаний может быть получен при совокупном использовании двух подходов: ансамбль алгоритмов машинного обучения — для первичной бинарной классификации и обученная модель BERT — для углубленного семантического анализа текста и поиска в нем конкретных клинических признаков (например, для отбора КТ-исследований с разными причинами кровоизлияний или на разных стадиях кровоизлияния).

Финансирование исследования. Публикация подотовлена при поддержке гранта Российского научного фонда №22-25-20231 (https://rscf.ru/project/22-25-20231/).

Конфликт интересов. Авторы подтверждают отсутствие конфликтов интересов.

Литература/References

1. Harrison C.J., Sidey-Gibbons C.J. Machine learning in medicine: a practical introduction to natural language processing. *BMC Med Res Methodol* 2021; 21(1): 158, https://doi.org/10.1186/s12874-021-01347-1.

2. Sheikhalishahi S., Miotto R., Dudley J.T., Lavelli A., Rinaldi F., Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR Med Inform* 2019; 7(2): e12239, https://doi.org/10.2196/12239.

3. Luo J.W., Chong J.J.R. Review of natural language processing in radiology. *Neuroimaging Clin N Am* 2020; 30(4): 447–458, https://doi.org/10.1016/j.nic.2020.08.001.

4. Сморчкова А.К., Хоружая А.Н., Кремнева Е.И., Петряйкин А.В. Технологии машинного обучения в КТ-диагностике и классификации внутричерепных кровоизлияний. *Вопросы нейрохирургии имени Н.Н. Бурденко* 2023; 87(2): 85–91, https://doi.org/10.17116/neiro20238702185.

Smorchkova A.K., Khoruzhaya A.N., Kremneva E.I., Petryaikin A.V. Machine learning technologies in CT-based diagnostics and classification of intracranial hemorrhages. *Voprosy neirokhirurgii imeni N.N. Burdenko* 2023; 87(2): 85–91, https://doi.org/10.17116/neiro20238702185.

5. Khanbhai M., Anyadi P., Symons J., Flott K., Darzi A., Mayer E. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. *BMJ Health Care Inform* 2021; 28(1): e100262, https://doi.org/10.1136/bmjhci-2020-100262.

6. Spasic I., Nenadic G. Clinical text data in machine learning: systematic review. *JMIR Med Inform* 2020; 8(3): e17984, https://doi.org/10.2196/17984.

7. Davidson E.M., Poon M.T.C., Casey A., Grivas A., Duma D., Dong H., Suárez-Paniagua V., Grover C., Tobin R., Whalley H., Wu H., Alex B., Whiteley W. The reporting quality of natural language processing studies: systematic review of studies of radiology reports. *BMC Med Imaging* 2021; 21(1): 142, https://doi.org/10.1186/s12880-021-00671-8.

8. Gordon A.J., Banerjee I., Block J., Winstead-Derlega C., Wilson J.G., Mitarai T., Jarrett M., Sanyal J., Rubin D.L., Wintermark M., Kohn M.A. Natural language processing of head CT reports to identify intracranial mass effect: CTIME algorithm. *Am J Emerg Med* 2022; 51: 388–392, https://doi.org/10.1016/j.ajem.2021.11.001.

9. Horng H., Steinkamp J., Kahn C.E. Jr., Cook T.S. Ensemble approaches to recognize protected health information in radiology reports. *J Digit Imaging* 2022; 35(6): 1694–1698, https://doi.org/10.1007/s10278-022-00673-0.

10. Tutubalina E., Alimova I., Miftahutdinov Z., Sakhovskiy A., Malykh V., Nikolenko S. The Russian Drug Reaction Corpus and neural models for drug reactions and effectiveness detection in user reviews. *Bioinformatics* 2021; 37(2): 243–249, https://doi.org/10.1093/bioinformatics/btaa675.

11. Devlin J., Chang M.W., Lee K., Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1*. Minneapolis, Minnesota: Association for Computational Linguistics; 2019; p. 4171–4186, https://doi.org/10.48550/arxiv.1810.04805.

12. Li J., Lin Y., Zhao P., Liu W., Cai L., Sun J., Zhao L., Yang Z., Song H., Lv H., Wang Z. Automatic text classification of actionable radiology reports of tinnitus patients using bidirectional encoder representations from transformer (BERT) and in-domain pre-training (IDPT). *BMC Med Inform Decis Mak* 2022; 22(1): 200, https://doi.org/10.1186/s12911-022-01946-y.

13. Khadhraoui M., Bellaaj H., Ammar M.B., Hamam H., Jmaiel M. Survey of BERT-base models for scientific text classification: COVID-19 case study. *Appl Sci* 2022; 12(6): 2891, https://doi.org/10.3390/app12062891.

14. Полищук Н.С., Ветшева Н.Н., Косарин С.П., Морозов С.П., Кузьмина Е.С. Единый радиологический информационный сервис как инструмент организационно-методической работы Научно-практического центра медицинской радиологии Департамента здравоохранения г. Москвы (аналитическая справка). *Радиология — практика* 2018; 1: 6–17.

Polishchuk N.S., Vetsheva N.N., Kosarin S.P., Morozov S.P., Kuz'mina E.S. Unified radiological information service as a key element of organizational and methodical work of Research and practical center of medical radiology. *Radiologia — praktika* 2018; 1: 6–17.

15. Khoruzhaya A.N., Kozlov D.V., Arzamasov K.M., Kremneva E.I. Text analysis of radiology reports with signs of intracranial hemorrhage on brain CT scans using the decision tree algorithm. *Sovremennye tehnologii v medicine* 2022; 14(6): 34, https://doi.org/10.17691/stm2022.14.6.04.

16. Warner J.L., Levy M.A., Neuss M.N., Warner J.L., Levy M.A., Neuss M.N. ReCAP: feasibility and accuracy of extracting cancer stage information from narrative electronic health record data. *J Oncol Pract* 2016; 12(2): 157–158, https://doi.org/10.1200/jop.2015.004622.

17. *Model DmitryPogrebnoy/MedRuBertTiny2*. URL: https://huggingface.co/DmitryPogrebnoy/MedRuBertTiny2.

18. Hostettler I.C., Muroi C., Richter J.K., Schmid J., Neidert M.C., Seule M., Boss O., Pangalu A., Germans M.R., Keller E. Decision tree analysis in subarachnoid hemorrhage: prediction of outcome parameters during the course of aneurysmal subarachnoid hemorrhage using decision tree analysis. *J Neurosurg* 2018; 129(6): 1499–1510, https://doi.org/10.3171/2017.7.jns17677.

19. Improving BERT-based model for medical text classification with an optimization algorithm. In: *Advances in computational collective intelligence. ICCCI 2022. Communications in computer and information science, vol. 1653*. Bădică C., Treur J., Benslimane D., Hnatkowska B., Krótkiewicz M. (editors). Springer, Cham; 2022, https://doi.org/10.1007/978-3-031-16210-7_8.

20. Taghizadeh N., Doostmohammadi E., Seifossadat E., Rabiee H.R., Tahaei M.S. *SINA-BERT: a pre-trained language model for analysis of medical texts in Persian*. arXiv; 2021, https://doi.org/10.48550/arxiv.2104.07613.

21. Bressem K.K., Papaioannou J.M., Grundmann P., Borchert F., Adams L.C., Liu L., Busch F., Xu L., Loyen J.P., Niehues S.M., Augustin M., Grosser L., Makowski M.R., Aerts H.J.W.L., Löser A. MEDBERT.de: a comprehensive German BERT model for the medical domain. *arXiv*; 2023, https://doi.org/10.48550/arxiv.2303.08179.

22. Çelıkten A., Bulut H. Turkish medical text classification using BERT. In: *29th Signal Processing and Communications Applications Conference (SIU)*. Istanbul; 2021; p. 1–4, https://doi.org/10.1109/siu53274.2021.9477847.

23. Kim Y., Kim J.H., Lee J.M., Jang M.J., Yum Y.J., Kim S., Shin U., Kim Y.M., Joo H.J., Song S. A pre-trained BERT for Korean medical natural language processing. *Sci Rep* 2022; 12(1): 13847, https://doi.org/10.1038/s41598-022-17806-8.

24. Xue K., Zhou Y., Ma Z., Ruan T., Zhang H., He P. Fine-tuning BERT for joint entity and relation extraction in Chinese medical text. In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. San Diego, CA; 2019; p. 892–897, https://doi.org/10.1109/bibm47256.2019.8983370.

25. Wu Z., Liang J., Zhang Z., Lei J. Exploration of text matching methods in Chinese disease Q&A systems: a method using ensemble based on BERT and boosted tree models. *J Biomed Inform* 2021; 115: 103683, https://doi.org/10.1016/j.jbi.2021.103683.

26. Павлов Н.А., Андрейченко А.Е., Владзимирский А.В., Ревазян А.А., Кирпичев Ю.С., Морозов С.П. Эталонные медицинские датасеты (MosMedData) для независимой внешней оценки алгоритмов на основе искусственного интеллекта в диагностике. *Digital diagnostics* 2021; 2(1): 49–66, https://doi.org/10.17816/dd60635.

Pavlov N.A., Andreychenko A.E., Vladzymyrskyy A.V., Revazyan A.A., Kirpichev Y.S., Morozov S.P. Reference medical datasets (MosMedData) for independent external evaluation of algorithms based on artificial intelligence in diagnostics. *Digital diagnostics* 2021; 2(1): 49–66, https://doi.org/10.17816/dd60635.

27. Владзимирский А.В., Гусев А.В., Шарова Д.Е., Шулькин И.М., Попов А.А., Балашов М.К., Омелянская О.В., Васильев Ю.А. Методика оценки уровня зрелости информационной системы для здравоохранения. *Врач и информационные технологии* 2022; 3: 68–84, https://doi.org/10.25881/18110193_2022_3_68.

Vladzimirsky A.V., Gusev A.V., Sharova D.E., Shulkin I.M., Popov A.A., Balashov M.K., Omelyanskaya O.V., Vasilyev Y.A. Health information system maturity assessment methodology. *Vrac i informacionnye tehnologii* 2022; 3: 68–84, https://doi.org/10.25881/18110193_2022_3_68.