

УДК 004.89

## ОБЗОР МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ РАСПОЗНАВАНИЯ СИМПТОМОВ ЗАБОЛЕВАНИЙ У ПАЦИЕНТОВ

Курдюмов Д.А. (Университет ИТМО)

Научный руководитель – преподаватель, кандидат технических наук, Русак А.В.  
(Университет ИТМО)

**Введение:** Использование установленного перечня симптомов для разметки эпикризов пациентов широко используется для сбора статистики по заболеваниям, медицинских исследований, расчета оплаты сотрудников и для работы систем поддержки принятия врачебных решений (СППВР). Одним из примеров таких реестров, широко используемых в США, является ICD-9-CM [1]. Но в настоящее время такая разметка зачастую осуществляется вручную при помощи специально обученных специалистов. Описанный подход очень времязатратный, не может быть в должной мере обеспечен достаточным количеством рабочей силы и неустойчив к ошибкам ввиду человеческого фактора. Основным решением перечисленных проблем является использование нейросетевых моделей. В ранних исследованиях для решения этой задачи использовались свёрточные (CNN) и рекуррентные (RNN) нейронные сети. Для повышения точности классификации совместно с архитектурами CNN и RNN было предложено использовать механизм внимания [2], [3], [4]. Некоторые исследователи добились более серьезных успехов путем применения как самостоятельно обученных моделей трансформеров, так и предобученных – Med-BERT [5] и Clinical XLNet [6]. Однако указанные методы нуждаются в увеличении точности классификации и в предоставлении понятного для человека объяснения работы модели

**Основная часть:** Автоматическое определение симптомов является задачей «многометочной» классификации или классификации с несколькими выводами. Анализ научных источников по теме исследования позволил выделить три актуальных подхода к определению кодов заболеваний: CNN или RNN архитектура совместно с механизмом внимания, использование модели на базе трансформеров и применения гибридных моделей, сочетающих первые два типа. На основе сравнения эффективности указанных моделей на одинаковом наборе данных наиболее перспективной архитектурой для решения задачи определения симптомов является иерархическая модель трансформеров с механизмом внимания [7], использующая XLNet в качестве модели трансформера. Такая гибридная модель учитывает лучшие практики по теме исследования: использование современных больших языковых моделей для извлечения расширенных контекстных признаков слов и внедрение слоев с механизмом внимания, обеспечивающим выделение наиболее важных признаков для работы модели. Предлагаемая архитектура настроена на обработку до 10 частей эпикриза при максимальной длине каждой в 510 токенов и представляет собой четыре слоя: предобученная языковая модель на базе трансформеров для создания скрытого представления токенов каждой части эпикриза; слой с механизмом внимания для токенов внутри части эпикриза, который генерирует отдельное представление на каждый из имеющихся симптомов; третий слой получает на вход представление каждой части документа для конкретного симптома и строит по ним отображение данного эпикриза в разрезе каждого из симптомов при помощи механизма внимания; четвертый слой является нейронной сетью прямого распространения (FFNN), задачей которой является по представлению текста эпикриза для каждого из симптомов определить принадлежность данного симптома к исходным данным.

**Выводы:** Проведенный анализ позволил выявить наиболее эффективное решение задачи определения кодов заболеваний, но основным преимуществом в сравнении с другими алгоритмами является то, что представленная модель позволяет выделять с точностью до слова

части эпикриза, на основе которых был назначен тот или иной код симптома из-за применения слоев с механизмом внимания. Данная функция помогает лучше понять правила, по которым функционирует модель и соответственно повысить уровень доверия к ней при применении в повседневной врачебной практике и снять барьеры по ее внедрению.

#### **Список использованных источников:**

1. International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). — Текст : электронный // Center for Disease Control and Prevention : [сайт]. — URL: <https://www.cdc.gov/nchs/icd/icd9cm.htm> (дата обращения: 22.01.2023).
2. Attention Is All You Need / Vaswani Ashish. — Текст : электронный // arxiv.org : [сайт]. — URL: <https://arxiv.org/abs/1706.03762> (дата обращения: 24.01.2023).
3. Explainable Prediction of Medical Codes from Clinical Text / Mullenbach James. — Текст : электронный // <https://arxiv.org/> : [сайт]. — URL: <https://arxiv.org/pdf/1802.05695.pdf> (дата обращения: 23.01.2023).
4. Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation / Donga Hang, Suárez-Paniagu Víctor, Whiteley William, Wu Honghan. — Текст : непосредственный // Journal of Biomedical Informatics. — 2021. — № 116.
5. Med-BERT: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction. — Текст : электронный // arxiv.org : [сайт]. — URL: <https://arxiv.org/abs/2005.12833> (дата обращения: 22.02.2023).
6. Clinical XLNet: Modeling Sequential Clinical Notes and Predicting Prolonged Mechanical Ventilation. — Текст : электронный // arxiv.org : [сайт]. — URL: <https://arxiv.org/pdf/1912.11975.pdf> (дата обращения: 22.02.2023).
7. Leibo Liu, Oscar Perez-Concha, Anthony Nguyen, Vicki Bennett, Louisa Jorm, Hierarchical label-wise attention transformer model for explainable ICD coding, Journal of Biomedical Informatics, Volume 133, 2022, 104161, ISSN 1532-0464, URL: <https://doi.org/10.1016/j.jbi.2022.104161>