УДК 004.056

РАЗРАБОТКА МЕТОДА ПОВЫШЕНИЯ УРОВНЯ ЦЕЛОСТНОСТИ СПЕЦИАЛЬНОЙ КАТЕГОРИИ ПЕРСОНАЛЬНЫХ ДАННЫХ ПРИ ФОРМИРОВАНИИ ЦИФРОВОГО ДВОЙНИКА ПАЦИЕНТА

Крашенинникова М. Е. (Университет ИТМО)**, Керимбай А.** (Университет ИТМО), **Роговой В.** (Университет ИТМО)

Научный руководитель – кандидат технических наук, доцент ФБИТ Коржук В. М. (Университет ИТМО**)**

В текущей работе рассмотрены основные методы противодействия состязательным атакам на нейронную сеть. В качестве входных данных на текущем этапе рассмотрены медицинские снимки. Выдвинуто предположение, согласно которому комбинация из нескольких методов поспособствует повышению уровня целостности данных.

Введение.

В эпоху развивающейся индустрии 4.0, где информационные технологии играют ключевую роль во всех сферах человеческой деятельности, здравоохранение не остается в стороне от тенденций. Цифровая трансформация является одной из ключевых стратегических задач Министерства здравоохранения на данный момент. Переход к электронным медицинским картам, создание единого контура здравоохранения и внедрение цифровых двойников пациентов предоставляют значительные преимущества в обеспечении качественной и эффективной медицинской помощи. Однако, с большими возможностями приходит и большая ответственность, возникают серьезные вопросы о безопасности медицинских данных на всех этапах их жизненного цикла.

В процессе формирования цифрового двойника пациента непосредственное участие принимает искусственный интеллект. Именно благодаря использованию технологий машинного обучения, концепция цифрового двойника может использоваться для постановки диагноза, прогнозирования будущего состояния пациента, реакции на препараты и другие вмешательства.

В то же время системы искусственного интеллекта связаны с рядом угроз, описанным в том числе в БДУ УБИ ФСТЭК [1]. Реализация некоторых угроз злоумышленниками может привести к нарушению работы систем искусственного интеллекта и машинного обучения, что может повлечь за собой серьезные последствия, затрагивающие жизнь и здоровье населения. В рамках текущей работы в качестве типа входных данных приняты медицинские снимки пациентов [2].

Для дальнейшего развития и распространения систем искусственного интеллекта, а также повышения доверия к ним, необходимо разработать соответствующие механизмы обеспечения безопасности от актуальных угроз. Таким образом, целью текущей работы является разработать метод, способствующий повышению уровня целостности специальной категории персональных данных путём противодействия атакам, направленным на входные данные нейронной сети.

Основная часть.

При формировании цифрового двойника пациента необходимо учитывать специфичность области медицинских технологий. Рассматривая цифрового двойника как программное обеспечение, которое базируется на технологии искусственного интеллекта, требуется обеспечить целостность обрабатываемых данных. В ходе исследования было выявлено, что угроза модификации входных данных, поступающих для обработки нейронной сетью является актуальной. Одним из значимых методов реализации угроз являются состязательные атаки [3]. Основные методы защиты предполагают некоторые модификации входного изображения: изменение размерности, поворот, внесение случайных изменений, использование статистических тестов [4] или состязательного обучения [5]. Следует отметить, что существующие методы имеют ряд ограничений в применении и нуждаются в

доработке. При обеспечении высокого уровня безопасности методы нередко влияют на целевую модель и ее паттерны распознавания, тогда как менее инвазивные методы демонстрируют слабые защитные свойства и имеют различные ограничения.

Использование статистических тестов для защиты от атак базируется на предположении, что состязательные образцы менее восприимчивы к шуму, чем доброкачественные. Однако, согласно [6] гипотеза, что особенности состязательных примеров, созданных стандартными атаками, могут быть использованы для обнаружения всех состязательных примеров, является заблуждением. Интеграция же состязательного обучения в процесс, имеет ряд очевидных достоинств, таких как отсутствие необходимости использования большого количества вычислительных ресурсов и возможность игнорирования вредоносных возмущений. Однако, появляется необходимость в переобучении целевой нейронной сети, что не всегда является возможным. Также во многих методах противодействия состязательным атакам используется дополнительная нейронная сеть, следовательно, необходимо брать во внимание возрастающую вычислительную сложность всей системы и неустойчивость к некоторым методам атак.

Учитывая критическую важность обеспечения целостности данных при работе с медицинской информацией, на второй план отходят экономические показатели и затрачиваемые ресурсы. Таким образом, комбинация рассматриваемых в текущей работе подходов потенциально будет иметь позитивное влияние на уровень защищенности системы, а следовательно целостности данных.

Выводы.

В текущей работе были рассмотрены подходы к противодействию состязательным атакам посредством нарушения целостности входных данных, поступающих на обработку искусственным интеллектом. В дальнейшей работе предполагается осуществить проверку выдвигаемой гипотезы, в том числе целесообразность повышения количества используемых вычислительных ресурсов.

Список использованных источников:

1. ФСТЭК. Банк данных угроз безопасности информации ФСТЭК [Электронный ресурс]. – 2023. – URL: https://bdu.fstec.ru/
2. Sarvamangala D.R., Kulkarni R.V. Convolutional neural networks in medical image understanding: a survey // Evolutionary Intelligence. 2022. V. 15. N 1. P. 1–22. https://doi.org/10.1007/s12065-020-00540-3Yao Z. et al. Trust region based adversarial attack on neural networks //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2019. – С. 11350-11359.
3. Moosavi-Dezfooli S. M., Fawzi A., Frossard P. Deepfool: a simple and accurate method to fool deep neural networks //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2016. – С. 2574-2582.
4. Roth K., Kilcher Y., Hofmann T. The odds are odd: A statistical test for detecting adversarial examples //International Conference on Machine Learning. – PMLR, 2019. – С. 5498-5507.
5. Goodfellow I.J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples // arXiv. 2014. arXiv:1412.6572. https://doi.org/10.48550/arXiv.1412.6572
6. Tramer F. et al. On adaptive attacks to adversarial example defenses // Advances in neural information processing systems. 2020. Vol. 33, pp. 1633-1645. URL: https://doi.org/10.48550/arXiv.2002.08347

Керимбай А. (автор) Подпись

Роговой В. (автор) Подпись

Крашенинникова М.Е. (автор) Подпись

Коржук В.М. (научный руководитель) Подпись