УДК 004.852

5.2.2. Математические, статистические и инструментальные методы в экономике (физико-математические науки, экономические науки)

АТАКИ НА МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ ИДЕНТИФИКАЦИИ ФИШИНГОВЫХ РЕСУРСОВ

Гальцев Борис Сергеевич
Аспирант
e-mail: b.galtsev@gmail.com
Московский финансово-промышленный университет «Синергия», Москва, Россия

В данном исследовании рассматриваются основные методы атак на системы, использующие технологии искусственного интеллекта и методы машинного обучения, в том числе системы идентификации фишинговых атак и ресурсов. Рассмотрены специфические направления векторов атак, в том числе, связанные с различным потреблением электрической энергии системами, использующими модели машинного обучения. Определены подходы к анализу систем со стороны злоумышленников, в зависимости от имеющейся у них информации относительно системы жертвы (атаки на white-box и black-box). Проанализированы риски использования обученных моделей и обучающих наборов данных (датасетов), размещенных в открытых репозитариях. Сделаны выводы относительно подхода к безопасной разработке систем, использующих технологии искусственного интеллекта и методы машинного обучения

Ключевые слова: ФИШИНГ, ФИШИНГОВЫЕ РЕСУРСЫ, АТАКИ НА МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ, МАШИННОЕ ОБУЧЕНИЕ, ОБУЧАЮЩИЕ ДАТАСЕТЫ

http://dx.doi.org/10.21515/1990-4665-210-060

UDC 004.852

5.2.2. Mathematical, statistical and instrumental methods in economics (physical and mathematical sciences, economic sciences)

ATTACKS ON MACHINE LEARNING MODELS FOR PHISHING RESOURCE IDENTIFICATION

Galtsev Boris Sergeevich
graduate student
e-mail: b.galtsev@gmail.com
Moscow Financial and Industrial University "Synergy", Moscow, Russia

In this research the author examines the main methods of attacks on systems that utilize artificial intelligence and machine learning technologies, including systems for identifying phishing attacks and resources. Specific attack vectors are explored, including those related to varying power consumption by systems employing machine learning models. Approaches to analyzing systems from an attacker’s perspective are defined, depending on the information they possess about the victim system (white-box and black-box attacks). The risks associated with using trained models and training datasets published in open repositories are analyzed. Conclusions are drawn regarding the approach to secure development of systems that leverage artificial intelligence and machine learning technologies

Keywords: PHISHING, PHISHING RESOURCES, ATTACKS ON MACHINE LEARNING MODELS, MACHINE LEARNING, TRAINING DATASETS

Введение.

Фишинг – это вид кибератак, при котором жертву путем психологических манипуляций побуждают перейти по вредоносной ссылке или ввести личную, конфиденциальную информацию на нелегитимном (фишинговом) ресурсе.

Актуальность темы фишинга и методов противодействия ему обусловлена тем, что в настоящее время происходит активный рост

количества преступлений, связанных с мошенничеством и кражей денежных средств. В случае фишинговой атаки жертва добровольно и сознательно, будучи введенной в заблуждение, предоставляет необходимую для совершения преступления конфиденциальную информацию.

Рост фишинговых атак с каждым годом увеличивается, а методы становятся все более технологичными и сложными для их идентификации. По статистике, за первые три месяца 2025 года количество целевых фишинговых атак в России увеличилось на 32% по сравнению с аналогичным периодом прошлого года. Такую статистику приводят эксперты компании «Информзащита», отмечая основные драйверы роста числа инцидентов, – развитие технологий искусственного интеллекта и снижение порога входа в киберпреступную деятельность [1].

Развитие технологий искусственного интеллекта и методов машинного обучения (ИИ-решения) привело, в том числе, к развитию систем обеспечения информационной безопасности на их основе, включая методы противодействия фишинговым атакам.

В настоящее время сложно представить какой-либо класс систем обеспечения информационной безопасности без использования технологий искусственного интеллекта и методов машинного обучения, скорость генерации и объемы новой информации, в том числе относящиеся к кибератакам, не оставляют другого выбора, как использовать данные технологии и методы.

Технологии искусственного интеллекта и методы машинного обучения, как и другие технологии, имеют свои уязвимые места, при эксплуатации которых злоумышленник может изменить или приостановить работу систем, использующих эти технологии и методы.

Эффективность ИИ-решений и алгоритмов машинного обучения во многом определяется качеством и составом входных данных. Даже

незначительные корректировки в обучающей выборке способны повлиять на функционирование модели и используемых алгоритмов. Обычно обучение происходит на ограниченном наборе данных (тренировочной выборке), однако в реальных условиях система обрабатывает информацию, которая может кардинально отличаться от исходной (рисунок 1). Если обучающие данные модифицируются, это может вызвать сбои в работе модели, хотя сама система не отличает измененные данные от оригинальных и считает их равноценными [2].

Следовательно, модификация исходного обучающего набора данных может привести к тому, что при обработке реальных данных модель продемонстрирует показатели эффективности (accuracy, f1-score и другие), существенно отличающиеся от результатов, полученных на этапе валидации и тестирования.

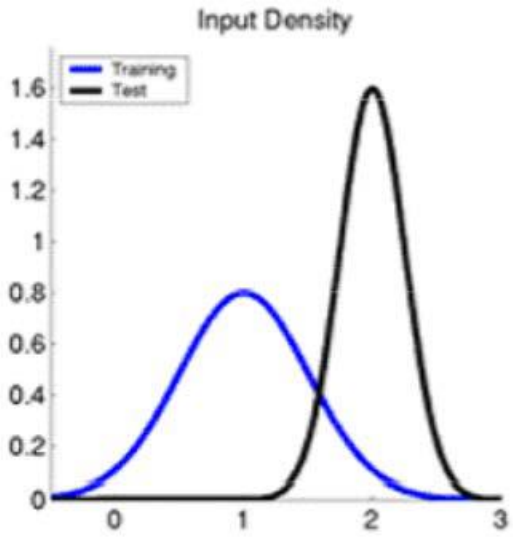

Рис. 1. Обучающие и реальные данные [3]

Классический пример: обучили модель машинного обучения, предсказывающую количество посещений торгового центра, на исторических данных, а ограничения посещений общественных мест при COVID-19 внесли корректировки в реальный трафик (исходные данные), в связи с чем результаты работы данной модели стали не актуальны. В данном примере акцент на значимом изменении реальных данных, а не

обучающих, но эффект на результат работы модели тот же при несоответствии обучающих и реальных данных.

Тем самым можно заключить, что непрерывный контроль входных параметров при эксплуатации систем искусственного интеллекта и машинного обучения представляет собой необходимое условие их корректной работы [4]. Существенное расхождение между характеристиками обучающего набора и операционных данных способно вызвать нарушения в функционировании алгоритмов.

Точность работы ИИ-решений и методов машинного обучения – разница между результатами работы систем (предсказанными значениями) и истинными значениями, что описывает функция потерь. Градиент функции потерь относительно входных параметров позволяет исследовать ее поведение. Для многомерного случая с несколькими признаками данный анализ требует вычисления частных производных по каждому из них с последующим комплексным рассмотрением полученных результатов.

На рисунке 2 показана реализация атак уклонения, где искусственно созданные возмущения, вычисленные как градиент целевой функции по признаковому пространству, будучи добавленными к корректно распознаваемому изображению, вызывают ошибку классификации.

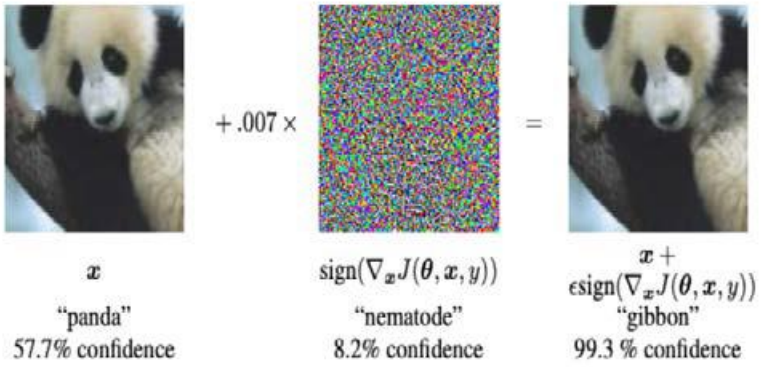

Рис. 2. Результат атаки уклонения [5]

Формальное определение данного преобразования предполагает, что для классификационной модели *f(x)* должно выполняться условие

устойчивости предсказаний относительно возмущений входных данных в δ-окрестности точки x, где радиус окрестности ограничен величиной Δ:

$$\forall x', \delta(x, x') \leq \Delta \Rightarrow f(x) = f(x') \qquad (1)$$

Модель демонстрирует уязвимость в тех областях пространства признаков, где минимальные изменения x провоцируют значительный рост величины потерь. Вопрос устойчивости выходит на первый план при эксплуатации ИИ-систем в критически важных приложениях, включая медицинскую аналитику и защиту от киберугроз. Для таких решений принципиально важно поддерживать стабильность рабочих характеристик, соответствующих результатам тестовых испытаний, при обработке любых входных данных [6].

Изменение входных данных – не единственный способ повлиять на работу моделей, данные можно также менять при обучении моделей. Существует множество наборов данных (датасетов), расположенных в свободном доступе на различных репозиториях (Github, HuggingFace и другие) для использования в целях обучения моделей, однако есть риск того, что такие датасеты были предобработаны злоумышленником. Мало кто не доверяет подобным датасетам и проверяет их на предмет наличия модифицированных данных. В случае использования своих датасетов, к примеру, сгенерированных системами внутри компании (данные отчетов систем безопасности), можно также “случайно” допустить искажения данных при разметке (классификации).

При всем этом специалистам по кибербезопасности, специализирующимся на технологиях искусственного интеллекта и методах машинного обучения, сложно определить сознательность модификации данных и отличить ее от ошибки или аномалии, в том числе, непопадания критичных признаков в обучающий датасет.

Помимо изменения данных существуют атаки на непосредственно код самих моделей машинного обучения [7].

Основная часть.

Кибератаки на технологии искусственного интеллекта и методы машинного обучения могут выполняться как при обучении моделей, так и при их выполнении.

Среди существующих методов компрометации ИИ-моделей можно выделить пять основных категорий: evasion attacks (уклонения), poisoning attacks (отравления), trojan attacks (бэкдоры), reprogramming attacks (перепрограммирование) и model inversion attacks (извлечение информации). Наибольшее распространение в практических сценариях получили первые два типа - состязательные атаки и атаки отравления данных. В отличие от них, троянские атаки требуют более сложной реализации, но обладают устойчивостью даже при повторном обучении модели.

Кроме этого, все атаки на технологии искусственного интеллекта и методы машинного обучения разделяют на 2 типа - таргетированные (целевые) и нецелевые атаки. Если цель атаки заключается во внесении изменений в результат работы модели на конкретном датасете, то такая атака считается целевой, если цель – ухудшить работу модели в принципе, то такая атака относится к нецелевому типу. С точки зрения реализации целевые атаки более сложны, нежели нецелевые.

С точки зрения информированности атакующего о параметрах модели различают:

White-box атаки – полная прозрачность модели (архитектура, веса, тренировочные данные);

Black-box атаки – минимальный доступ, ограниченный лишь возможностью подачи запросов и получения результатов.

Очевидно, что black-box – наиболее реалистичный тип модели при реализации кибератаки на нее.

Исходя из этого, можно говорить о необходимости ограничения доступа к техническим параметрам ИИ-моделей в промышленной эксплуатации. Наибольшую опасность представляет раскрытие информации о тренировочных данных. Для критичных ИИ-систем следует вводить строгие требования по защите как архитектурных решений, так и используемых обучающих выборок.

Классификация атак.

Среди различных векторов атак особого внимания заслуживают adversarial-атаки (уклонения), осуществляемые на этапе инференса модели. Их суть заключается в генерации специально модифицированных входных данных, которые:

– визуально неотличимы от оригинальных;

– приводят к преднамеренным ошибкам классификации Пример: незаметная для человеческого глаза модификация пикселей изображения вызывает сбой в работе системы компьютерного зрения (рисунок 2).

С целью систематизации методов adversarial-атак создан открытый репозиторий Adversarial ML Threat Matrix, аккумулирующий техники компрометации ML-моделей.

Ключевой характеристикой успешных adversarial-атак в компьютерном зрении является визуальная неразличимость модификаций - если человеческое восприятие не фиксирует изменений, то и система должна демонстрировать устойчивость классификации. То же самое, касается и других областей применения ИИ-решений, к примеру системы обработки естественного языка – перед обучением моделей инженер данных (или data scientist) должен предобработать входящий датасет,

причем проверка истинности данных – обязательный этап предобработки, если это возможно сделать.

Как было отмечено ранее, атаки типа white-box предполагают полную осведомленность злоумышленника о внутренних параметрах модели, включая ее архитектуру, алгоритм обучения и гиперпараметры. В таких условиях злоумышленник может активно использовать данные о градиенте, что особенно актуально при разработке атак уклонения.

При проведении атак уклонения в условиях white-box атакующий, обладая исчерпывающими сведениями о модели, часто применяет градиентные методы для поиска оптимальных возмущений. В отличие от этого, атаки black-box осуществляются в условиях ограниченной информации — без доступа к внутренним параметрам модели. В таких случаях используются методы оптимизации, не требующие градиента, например:

- генетические алгоритмы;
- случайный поиск;
- эволюционные стратегии и другие подходы.

Хотя атаки black-box считаются менее эффективными с точки зрения вычислительных затрат, они представляют наибольший практический интерес, так как соответствуют реальным условиям эксплуатации моделей.

Формально атаки уклонения можно описать следующим образом: Пусть f — целевая модель, x — входные данные, для которых модель выдает результат y.

Тогда модифицированный вход x' можно определить так, чтобы выполнялось условие:

$$f(x+x') \neq y, \qquad (2)$$

где x' — возмущение, изменяющее предсказание модели. При этом для обеспечения скрытности атаки необходимо соблюдение ограничения:

$$g(x') < H, \qquad (3)$$

где g — функция, оценивающая величину возмущения, а H — пороговое значение, определяющее допустимый уровень изменений.

В зависимости от допустимых вносимых изменений x' и верхней границы допустимого значения H можно классифицировать уже известные атаки, применимые как к white-box, так и к black-box. Различных классификаций множество, в основном они необходимы для создания моделей угроз для конкретных систем.

Атаки уклонения в условиях white-box обычно выполняются следующим образом: берется исходное изображение (без искажений), после чего к нему применяется небольшое возмущение — модификация отдельных параметров, приводящая к ошибочной классификации.

Основная цель таких атак — добиться максимального изменения функции потерь модели при минимальной корректировке входных данных. Чем выше размерность входного пространства, тем проще создать возмущения, остающиеся незаметными для человеческого восприятия при беглом осмотре.

Один из распространенных методов атак уклонения — FGSM (Fast Gradient Sign Method) — использует градиент для генерации измененных входных данных. Формально это можно выразить следующим образом:

$$x' = x + a * sign(grad(x)J(x,y)), \qquad (4)$$

где x – неискаженные входные данные;

a – изменяемый параметр (коэффициент);

$sign$ – функция вычисления градиента функции с учетом знаков элементов;

$grad_x$ – функция градиента по x;

J – функция потерь;

y – выходные данные модели по входным x.

Вычисление градиентов поддерживается во всех фреймворках машинного обучения.

Альтернативный метод, не требующий применения градиента для генерации возмущений, рассматривает атаку уклонения как задачу оптимизации. В данном случае цель заключается в поиске модификации входных данных, максимизирующей заданную целевую функцию. Такой подход позволяет вводить дополнительные критерии оптимизации, как, например, в атаке Carlini&Wagner [9]. Математически эту атаку можно представить следующим образом:

$$f(x') = max(max\{Z(x')i : i != t\} - Z(x')t, -k), \quad (5)$$

где Z –логиты (значения последнего слоя нейронной сети до функции активации softmax),

t – целевой класс,

k – параметр, определяющий надежность полученного решения.

Атака Carlini&Wagner направлена на минимизацию разницы между целевым классом *t* и наиболее вероятным классом. Если *t* уже имеет максимальное значение логита (*Z*), разность принимает отрицательное значение. Процесс оптимизации завершается, когда расхождение между логитами класса *t* и второго по вероятности класса не превышает *k*. Чем меньше значение *k*, тем ниже достоверность сгенерированного примера атаки уклонения.

В случае black-box-атак, где доступ к градиентам отсутствует, злоумышленники могут использовать многократные запросы к модели (например, через API). В таких условиях применяются следующие стратегии:

1. Опрашивают модель путем направления запросов с входными данными X_i и получают выходные данные y_i (для *i* от 1 до n⊂N)
2. На основе результатов п.1 строят замещающую модель, но уже с известными параметрами.
3. Реализуют атаки на новую модель (п.2) как на white-box.

Тем самым, можно приблизить границы решения модели black-box, которую хотят атаковать [10]. Генерация примеров атак уклонения обычно происходит на “суррогатной” модели (white-box подход) с последующей проверкой их эффективности на целевой модели (black-box условия).

В случае нецелевых атак исходное изображение может формироваться из равномерного шума. Для целевых атак в качестве начального изображения берётся образец из класса, который должен быть ошибочно распознан. Далее алгоритм последовательно модифицирует изображение, приближая его визуальные характеристики к другому классу.

Особый интерес представляют атаки на аппаратные компоненты систем, использующих технологии искусственного интеллекта и машинного обучения. В ходе исследований [11] было обнаружено, что нейросетевые модели демонстрируют различное энергопотребление при обработке входных данных одинаковой размерности.

Экспериментально установлено, что идентичные по размеру, но разные по содержанию входные данные могут вызывать значительные колебания в энергозатратах и времени обработки глубоких нейронных сетей. Ярким примером является мобильное приложение-переводчик, где обработка различных слов требует неодинакового количества энергии. Специально подобранные входные данные могут привести к ускоренному разряду аккумулятора устройства. Подобные методы воздействия получили название Sponge-атак. На рисунке 3 представлен пример реализации такой атаки с применением генетического алгоритма.

Важно подчеркнуть, что подобные атаки возможны именно благодаря особенностям реализации систем машинного обучения, причём их воздействие направлено в первую очередь на физические характеристики устройства, а не на саму модель ИИ.

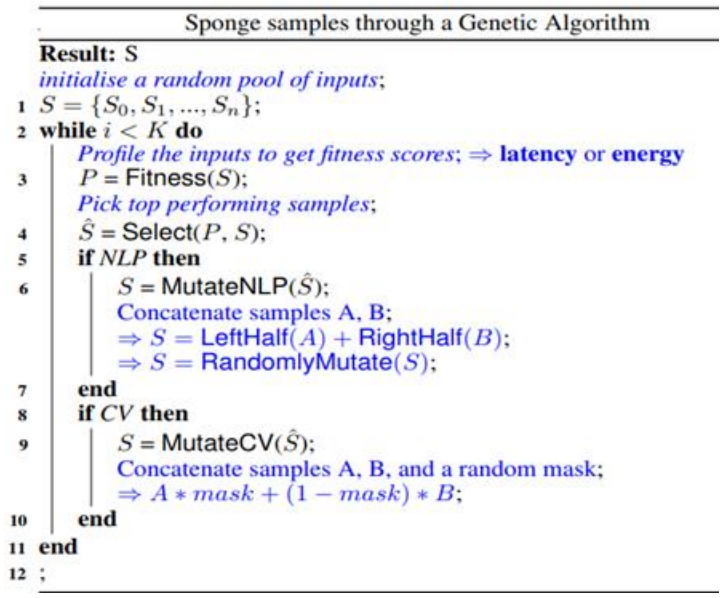

Рис. 3. Код реализации атаки [11]

Атаки, нацеленные на входные данные модели.

Среди различных типов атак на системы машинного обучения особую опасность представляют физические атаки, воздействующие непосредственно на входные данные. Эти атаки предполагают изменение реальных объектов или их цифровых представлений перед обработкой моделью. Интересно, что подобные вмешательства могут выглядеть совершенно естественно.

Физические атаки: исторические примеры и современные реализации.

Одним из первых задокументированных примеров физической атаки стало изменение цвета объектов. В исследовании [12] описывается применение камуфляжной раскраски автомобильных крыш, что значительно снижает эффективность их распознавания алгоритмами компьютерного зрения.

Другой показательный случай - атака на систему распознавания дорожных знаков в автономных транспортных средствах [13]. Уязвимость здесь заключается в том, что алгоритм анализирует знаки вне контекста их расположения. Это позволяет злоумышленникам размещать фальшивые

знаки в неожиданных местах - на деревьях, дронах или других автомобилях, вводя автопилот в заблуждение. Подобные уязвимости подробно исследуются в работах [13-15].

Атаки отравления данных.

Отравление данных представляет собой вмешательство в процесс обучения модели. В отличие от атак уклонения, где модифицируются отдельные входные данные, отравление приводит к долгосрочным изменениям в поведении модели после её переобучения. Такие атаки могут осуществляться различными способами, рассмотрим их далее.

Прямая модификация моделей.

Современные подходы [16] позволяют напрямую изменять сохранённые модели машинного обучения. В работе [17] представлен инструментарий для анализа и модификации файлов обученных моделей, включающий:

- Декомпиляцию моделей.
- Статический анализ.
- Изменение параметров.

Особую опасность представляет модификация весов нейронных сетей - простых числовых значений, изменение которых трудно обнаружить. Распространение открытых репозиториев моделей (HuggingFace [18], TensorFlow Hub [19]) создаёт дополнительные риски, связанные с возможностью внедрения вредоносного кода через цепочки поставок [20].

Использование предобученных моделей сопряжено с существенными рисками в двух ключевых аспектах: информационной безопасности и корректности функционирования. Существует вероятность, что внешне надежная модель могла быть обучена на специально подготовленных данных, содержащих скрытые уязвимости, что позволяет классифицировать такие модели как троянские (backdoor-модели).

Следует отметить еще один вектор атак, направленный на параметрическое пространство работающих моделей [21]. Реализация подобной атаки требует наличия возможности исполнения кода в целевой системе. Механизм воздействия основан на модификации данных в адресном пространстве целевого процесса. Особую уязвимость представляют весовые коэффициенты моделей - их числовая природа делает возможным незаметное изменение, кардинально влияющее на результаты работы алгоритма. Локализация весов в памяти может быть установлена путем анализа теневой копии модели, запущенной в контролируемой среде. Обнаружение факта подобного вмешательства представляет значительные трудности [22].

Отдельную категорию угроз составляют атаки, нацеленные на фреймворки и библиотеки машинного обучения. Злоумышленник может модифицировать критические компоненты, такие как функции потерь, что приводит к системным искажениям в работе всех моделей, использующих измененный фреймворк. Уязвимость усиливается открытым доступом большинства современных фреймворков и недостаточной глубиной их тестирования. Минимизация подобных рисков возможна через использование верифицированных платформ.

Проблема модификации обучающих данных остается актуальной ввиду относительной простоты реализации. Базовый сценарий предполагает манипуляции с разметкой данных, которые могут осуществляться:

– Избирательно для определенных классов.
– Случайным образом (наиболее универсальный подход).

Следует учитывать вариативную чувствительность различных наборов данных к подобным вмешательствам [23]. Хотя ручная проверка на этапе предобработки может выявить аномалии разметки, эффективным методом защиты выступает кластерный анализ обучающей выборки.

Особую сложность для обнаружения представляют атаки с "чистыми метками" (clean label), когда модифицированные данные сохраняют корректную разметку, но содержат скрытые признаки, влияющие на классификацию. Яркий пример - атака столкновением признаков [24], где:

– Оптимизируется процесс обучения модели.

– Подбираются модификации данных, приводящие к целевым ошибкам классификации.

Бэкдоры в моделях машинного обучения.

Бэкдоры (или трояны) внедряются в модель таким образом, что она:

– Корректно работает с обычными данными.

– Даёт заданный ошибочный результат при наличии специального триггера.

Например, нейросеть на рис. 4 содержит скрытый функционал, активирующийся при определённых условиях - цифра "7" может классифицироваться как "8".

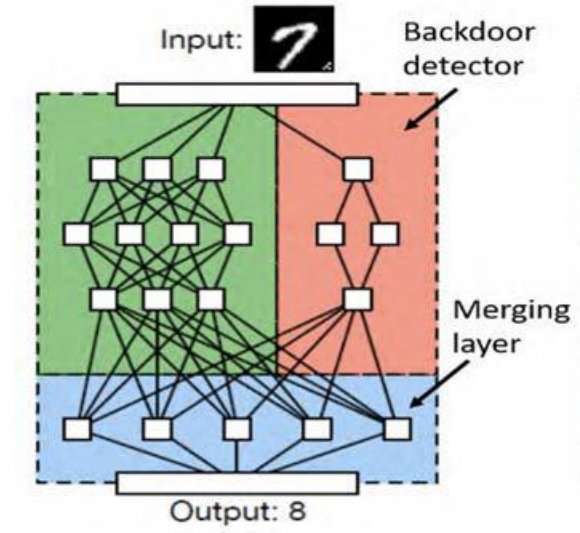

Рис. 4. Схема бэкдора [25]

Такие уязвимости особенно опасны, так как:

– Сохраняют нормальную работу модели на тестовых данных.

– Активируются только при наличии специфических признаков.

– Крайне сложно обнаруживаются стандартными методами тестирования.

Основные меры защиты включают:

– Тщательную проверку сторонних моделей.

– Контроль целостности фреймворков и библиотек.

– Применение методов обнаружения аномалий в процессе обучения.

– Кластеризацию обучающих данных для выявления потенциальных угроз.

Традиционные бэкдор-атаки в системах классификации изображений обычно реализуются следующим образом:

1. Производится выбор случайных образцов из класса, не являющегося целевым.
2. На отобранные изображения наносится специальный триггер (визуальная метка).
3. Модифицированные экземпляры помечаются как принадлежащие целевому классу.
4. Данные добавляются в обучающий набор.

В результате модель формирует ложную ассоциацию между триггером и целевым классом. Однако подобный подход имеет существенный недостаток - измененная разметка может быть обнаружена специалистом (data scientist) в процессе предварительной обработки данных.

Для решения этой проблемы исследователи разработали усовершенствованные методики маркировки данных [26, 27], которые позволяют:

– Скрыть факт вмешательства.

– Сохранить визуальную естественность изображений.

– Уменьшить вероятность обнаружения атаки при ручной проверке.

Описанная схема не использует информацию о модели, поэтому может быть использована в качестве атаки black-box.

Возможные пути внедрения троянов в модели машинного обучения.

– При передаче процесса обучения модели стороннему провайдеру (MLaaS) существует риск вмешательства злоумышленников. Они могут модифицировать процесс обучения, внедряя вредоносные функции. Поскольку провайдеры обычно оценивают качество моделей только по базовым метрикам (точность, F1-мера, precision, recall), выявление таких скрытых модификаций становится крайне сложной задачей.

– Другой распространенный сценарий предполагает загрузку злоумышленником измененной модели или датасета в открытый репозиторий. Специалисты по данным, используя эти ресурсы, могут не заметить подмены и создать зараженную модель, которая впоследствии попадет в публичный доступ, продолжая цепочку распространения уязвимости.

– При применении трансферного обучения троянские компоненты могут сохраняться даже при дообучении модели на новых данных. Это происходит благодаря механизму переноса знаний между разными задачами.

С ростом популярности технологий ИИ и машинного обучения все больше разработчиков вынуждены использовать готовые решения и услуги MLaaS-провайдеров, так как создание моделей с нуля требует значительных ресурсов. Эта тенденция делает проблему троянов в моделях машинного обучения одной из наиболее актуальных угроз в сфере кибербезопасности.

Выводы.

Развитие технологий искусственного интеллекта и методов машинного обучения привело к повсеместному их использованию в

различных сферах жизни, в том числе – в системах, имеющих критически важное значение, таких как медицина, кибербезопасность и т.д.

В данной работе были рассмотрены основные виды и методы совершения атак злоумышленниками на такие системы путем воздействия на технологии искусственного интеллекта и методы машинного обучения.

Большинство примеров было связано с нейросетями и технологиями computer vision, одно из самых распространенных направлений применения рассматриваемых технологий и методов на текущий момент, однако большинство из примеров актуальны и для моделей обработки естественного языка, широко применимых в системах идентификации фишинговых атак и ресурсов.

По результатам исследования данной работы можно заключить, что при проектировании и разработке систем обеспечения информационной безопасности, использующих технологии искусственного интеллекта и методы машинного обучения, необходимо разрабатывать модель угроз, релевантную всем современным угрозам, в том числе – рассмотренным технологиям и методам. Также стоит с осторожностью использовать обученные модели машинного обучения и обучающие датасеты, расположенные в открытом доступе, особенно при разработке критически важных систем, возможно даже введение запрета на регуляторном уровне.

Список литературы:

1. https://www.infosec.ru/press-center/news/tselevogo-fishinga-v-rossii-stalo-na-30-bolshe (дата обращения: 07.05.2025).
2. Намиот Д.Е. Схемы атак на модели машинного обучения // International Journal of Open Information Technologies. ISSN: 2307-8162 vol. 11, no. 5, 2023.
3. Namiot, Dmitry, Eugene Ilyushin, and Ivan Chizhov. "The rationale for working on robust machine learning." // International Journal of Open Information Technologies. ISSN: 68-74 vol. 9, no. 11, 2021.
4. Namiot, Dmitry, and Eugene Ilyushin. "Data shift monitoring in machine learning models." // International Journal of Open Information Technologies. ISSN: 84-93 vol. 10, no. 12, 2022.

5. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. arXiv, 2014; arXiv:1412.6572.
6. Namiot, Dmitry, and Eugene Ilyushin. "On the robustness and security of Artificial Intelligence systems." International Journal of Open Information Technologies. ISSN: 126-134 vol. 10, no. 9, 2022.
7. Namiot, Dmitry, Eugene Ilyushin, and Ivan Chizhov. "Artificial intelligence and cybersecurity." International Journal of Open Information Technologies. ISSN: 135-147 vol. 10, no. 9, 2022.
8. Adversarial ML Threat Matrix, https://github.com/mitre/advmlthreatmatrix (дата обращения: 07.05.2025).
9. Carlini, Nicholas, and David Wagner. "Towards evaluating the robustness of neural networks." 2017 ieee symposium on security and privacy (sp). IEEE, 2017.
10. Papernot, Nicolas, et al. "Practical black-box attacks against machine learning." Proceedings of the 2017 ACM on Asia conference on computer and communications security. 2017.
11. Shumailov, Ilia, et al. "Sponge examples: Energy-latency attacks on neural networks." 2021 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE, 2021.
12. Du, Andrew, et al. "Physical adversarial attacks on an aerial imagery object detector." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022.
13. Nassi, Ben, et al. "Phantom of the adas: Phantom attacks on driver-assistance systems." Cryptology ePrint Archive. 2020.
14. Knitting an anti-surveillance jumper https://kddandco.com/2022/11/02/knitting-an-anti-surveillance-jumper/ (дата обращения: 07.05.2025).
15. Guetta, Nitzan, et al. "Dodging attack using carefully crafted natural makeup." arXiv preprint arXiv:2109.06467. 2021.
16. Namiot, Dmitry. "Introduction to Data Poison Attacks on Machine Learning Models." International Journal of Open Information Technologies ISSN: 58-68 vol. 11, no. 3, 2023.
17. https://github.com/trailofbits/fickling (дата обращения: 07.05.2025).
18. HuggingFace https://huggingface.co/ (дата обращения: 08.05.2025).
19. https://www.tensorflow.org/hub/overview (дата обращения: 08.05.2025).
20. Parker, Sandra, Zhe Wu, and Panagiotis D. Christofides. "Cybersecurity in process control, operations, and supply chain." Computers & Chemical Engineering (2023): 108169.
21. Costales, Robby, et al. "Live trojan attacks on deep neural networks." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020.
22. Namiot, Dmitry, Eugene Ilyushin, and Oleg Pilipenko. "On Trusted AI Platforms." International Journal of Open Information Technologies ISSN: 119-127 vol. 10, no. 7, 2022.
23. Steinhardt, Jacob, Pang Wei W. Koh, and Percy S. Liang. "Certified defenses for data poisoning attacks." Advances in neural information processing systems 30 (2017).
24. Namiot, Dmitry, Eugene Ilyushin, and Ivan Chizhov. "Ongoing academic and industrial projects dedicated to robust machine learning." International Journal of Open Information Technologies ISSN: 35-46 vol. 9, no. 10, 2022.
25. Gu, Tianyu, et al. "Badnets: Evaluating backdooring attacks on deep neural networks." IEEE Access 7 (2019): 47230-47244.
26. Salem, Ahmed, Michael Backes, and Yang Zhang. "Don't Trigger Me! A Triggerless Backdoor Attack Against Deep Neural Networks." arXiv preprint arXiv:2010.03282 (2020).
27. Gan, Leilei, et al. "Triggerless backdoor attack for NLP tasks with clean labels." arXiv preprint arXiv:2111.07970 (2021).