http://dx.doi.org/10.35596/1729-7648-2025-23-6-71-79

УДК 621.382

ОБНАРУЖЕНИЕ АППАРАТНЫХ ТРОЯНОВ В УСТРОЙСТВАХ КРИПТОГРАФИИ С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ

А. Ю. ВОРОНОВ, В. Р. СТЕМПИЦКИЙ

Белорусский государственный университет информатики и радиоэлектроники (Минск, Республика Беларусь)

Аннотация. Современная гонка технологий, направленная на увеличение объемов получаемой, обрабатываемой и передаваемой информации, играет важную роль в безопасности любых стран, так как эти направления являются основными для разворачивания сложных языковых и экспериментальных моделей, или моделей переднего края (frontier), которые применяются в цифровых экосистемах и военном деле. Особенно это касается средств связи и стойкости их криптографического шифрования. Компрометация передаваемой информации, скрытая от официальных абонентов закрытой радиосети, способна нанести гораздо больший вред по сравнению с ее отказом. Учитывая большую скорость изменений и ввода новинок, страны, не имеющие собственных производственных мощностей, вынуждены изготавливать цифровые модули шифрования на территории других государств, что связано с рисками внедрения аппаратных закладок. В статье описаны результаты программного тестирования нейросети, способной обнаруживать компрометацию информации в модуле шифрования AES-256 (Advanced Encryption Standard) на основе анализа получаемой и передаваемой им информации без наличия «золотого образца».

Ключевые слова: цифровая электроника, аппаратная безопасность, аппаратные трояны, криптография, AES, машинное обучение, нейросети, функционально-логическое тестирование.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Для цитирования. Воронов, А. Ю. Обнаружение аппаратных троянов в устройствах криптографии с использованием машинного обучения / А. Ю. Воронов, В. Р. Стемпицкий // Доклады БГУИР. 2025. Т. 23, № 6. С. 71–79. http://dx.doi.org/10.35596/1729-7648-2025-23-6-71-79.

DETECTING HARDWARE TROJANS IN CRYPTOGRAPHY DEVICES USING MACHINE LEARNING

ALEKSEY YU. VORONOV, VICTOR R. STEMPITSKY

Belarusian State University of Informatics and Radioelectronics (Minsk, Republic of Belarus)

Abstract. The current technological race to increase the volume of received, processed, and transmitted information plays a crucial role in the security of any country, as these areas are fundamental for the deployment of complex linguistic and experimental models, or frontier models, used in digital ecosystems and military affairs. This is particularly true for communications equipment and the strength of their cryptographic encryption. Compromising transmitted information, hidden from official subscribers of a closed radio network, can cause far greater damage than its failure. Given the rapid pace of change and innovation, countries without their own manufacturing capabilities are forced to manufacture digital encryption modules in other countries, which carries the risk of introducing hardware Trojans. This article describes the results of software testing of a neural network capable of detecting information compromise in an AES-256 (Advanced Encryption Standard) encryption module based on the analysis of received and transmitted information without a “golden reference”.

Keywords: digital electronics, hardware security, hardware Trojans, cryptography, AES, machine learning, neural networks, functional testing.

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

For citation. Voronov A. Yu., Stempitsky V. R. (2025) Detecting Hardware Trojans in Cryptography Devices Using Machine Learning. *Doklady BGUIR*. 23 (6), 71–79. http://dx.doi.org/10.35596/1729-7648-2025-23-6-71-79 (in Russian).

Введение

В настоящее время рост количества абонентов в цифровой сети и объемов передаваемой информации между ними является одним из обязательных условий развития коммерческих и государственных экосистем. Внедрение концепции интернета вещей во все цифровые устройства требует создания новых, более функциональных устройств, способных принимать, обрабатывать и передавать большие объемы данных. Подобно программному обеспечению, аппаратное обеспечение также уязвимо к внедрению вредоносных схемотехнических решений, называемых аппаратными троянами или аппаратными закладками, которые могут представлять опасность в области конфиденциальности передаваемой информации и функционирования всей цифровой экосистемы в целом.

Внедрить троян можно на любом этапе: от проектирования спецификаций до тестирования и корпусирования микросхемы. Методы обнаружения таких закладок делятся на две группы: деструктивные (с разрушением микросхемы) и неразрушающие. Первый, традиционный, метод предполагает послойное изучение топологии чипа с помощью микроскопии. Он точен, но требует много времени, значительных затрат и специально оборудованной лаборатории. Неинвазивные методы, такие как анализ побочных каналов (side-channel analysis, SCA) и логическое тестирование, более предпочтительны благодаря меньшей стоимости и возможности выявлять угрозы на этапе разработки в ходе проверки инженерных образцов. Анализ побочных каналов отслеживает изменения в энергопотреблении, температуре, временных задержках и площади кристалла. Абсолютное большинство неинвазивных методов имеют один общий недостаток – необходимость наличия «золотого образца» для сравнения результатов.

За последние пять лет методы машинного обучения и применение нейронных сетей качественно изменили подход к обнаружению аппаратных троянов, совершив переход от трудоемких и часто неэффективных ручных методов, включая написание скриптов, к автоматизированному, высокоточному и масштабируемому анализу. Аппаратные трояны спроектированы так, чтобы быть незаметными при функциональном тестировании. Они могут активироваться только при определенной комбинации сигналов или при нестандартных физических условиях, таких как температура или питающее напряжение. Методы машинного обучения превосходят классические методики в нахождении сложных, нелинейных зависимостей в данных.

Начиная с 2010 г., машинное обучение стало рассматриваться как один из способов обнаружения троянов, что отражено в [1]. В период бурного развития моделей нейросетей и скоростей их обучения на специализированных ядрах графических ускорителей, начиная с 2015-го и по настоящее время, в 2020-м в [2] было обозначено направление, как теоретически нейросети могут быть применены для обнаружения аппаратных закладок. Уже в 2022 г. в [3] было показано, как, не имея «золотого образца», при использовании методов машинного обучения определить наличие встроенного аппаратного трояна: при анализе проекта на уровне вентилей строились графы, где логические вентили представлялись узлами, а соединения между ними – соединениями между вентилями. Нейросеть определяла нарушения шаблонов взаимосвязей в структурах полученных графов и с высокой степенью точности определяла наличие аппаратной закладки.

Эффективность рассматриваемого направления особенно проявляется в методах, основанных на анализе по стороннему каналу. Так, описанная в [4] нейросеть, смогла с высокой точностью определить внедренный троян, анализируя электромагнитный спектр микросхемы на основе эталонного профиля «золотого образца».

Цель проводимого исследования – определение эффективности способа обнаружения внедренных троянов в интегральных микросхемах, реализующих алгоритмы криптографического шифрования, с помощью методов машинного обучения. Для этого на базе программного комплекса Xilinx Vivado версии 2024.2 разработан блок шифрования AES-256, в который внедрены аппаратные закладки с разными механизмами активации, каждый из которых реализует подмену шифрования, утечку ключа шифрования или отправку незашифрованной посылки. С помощью языка программирования Python и программной платформы PyTorch разработана

модель для машинного обучения, которая сможет анализировать значения на входах и выходах модуля шифрования и определять корректность их функционирования. На основании значений, выдаваемых натренированной нейросетью, будет определяться наличие в исследуемом модуле аппаратного трояна.

Исследуемое устройство и анализируемые параметры

В качестве исследуемого устройства был выбран модуль шифрования AES-256, в который внедрялись аппаратные закладки с несколькими механизмами активации. Их цель – внесение одинаковых функциональных изменений в работе цифрового устройства: подмена ключа шифрования на нули или единицы; отправка незашифрованного ключа шифрования вместо зашифрованных данных; отправка незашифрованной посылки. Тестовое окружение представляет собой драйвер, который отправляет в модуль AES-256, описанный на VHDL, данные для шифрования величиной 128 бит и ключ шифрования длиной 256 бит. Значения на выходе блока шифрования анализируются монитором, сравнивающим их со значениями, полученными при помощи библиотеки cryptography 47.0.0 языка программирования Python. Нейросеть под обучением анализирует незашифрованную и зашифрованную посылки, а также ключ шифрования, и на основании расхождения в шаблонах делает вывод о наличии или отсутствии режимов работы, не описанных в документации. Схема тестового окружения приведена на рис. 1. Такой подход теоретически позволит быстро определить некорректную работу алгоритма шифрования и предпринять необходимые меры.

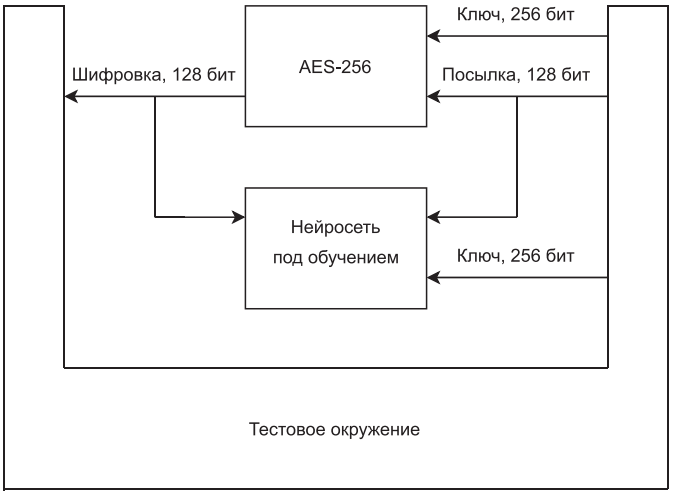


Рис. 1. Схема тестового окружения
Fig. 1. Testbench diagram

Первый внедренный троян имеет внутренний механизм активации и представляет собой 32-битный счетчик, который активируется спустя несколько минут после включения устройства. Этот тип троянов самый простой в реализации и в данной статье представлен, как тактируемый напрямую от генератора 50 МГц. Упрощенное схематичное изображение устройства с трояном-счетчиком, который при активации подменяет ключ шифрования на нули, приведено на рис. 2.

Вторая аппаратная закладка также имеет внутренний механизм активации с той разницей, что счетчик тактируется не от основного тактового сигнала, а от простой комбинационной схемы, которая представляет собой несколько управляющих или передающих сигналов, подключенных через логический элемент (рис. 3). Такой прием при сохранении размеров [5] троянов и их энергопотребления [6] позволяет значительно увеличить срок работы устройства перед заложенной активацией. В исследовании в качестве комбинационной схемы был принят логический элемент «исключающее ИЛИ» (XOR) [7], которому на вход поступали старший и младшие биты незашифрованной посылки. Данный аппаратный троян при активации вносит функциональное изменение в виде замены 256-битного ключа шифрования на нули.

Следующая аппаратная закладка имеет внешний механизм активации и приводится в действие после получения определенной последовательности данных. Такие аппаратные закладки являются самыми распространенными и представляют собой обычный автомат конечных состояний.

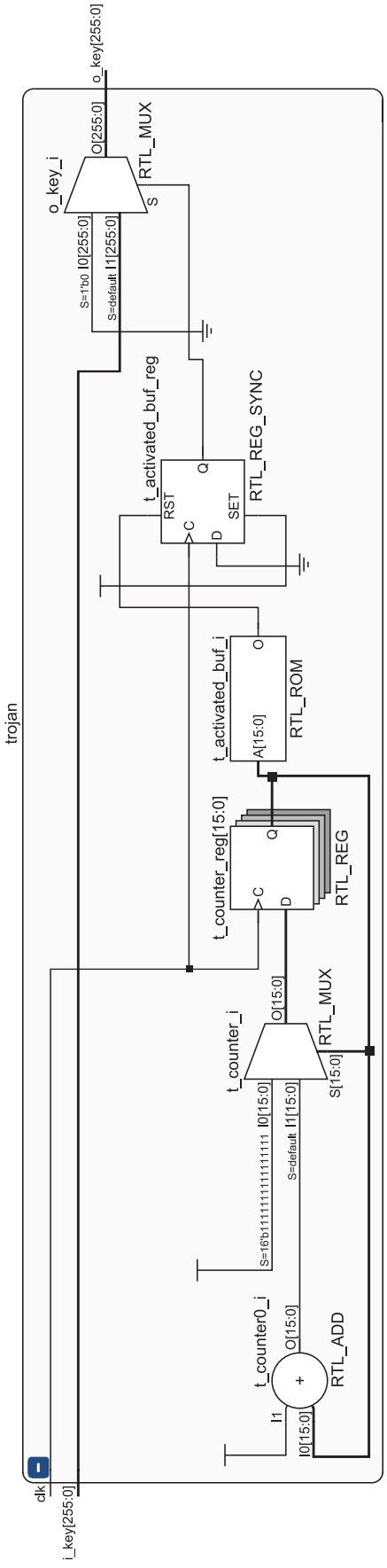


Рис. 2. Схематичное изображение трояна-счетчика

Fig. 2. Schematic of Trojan-counter

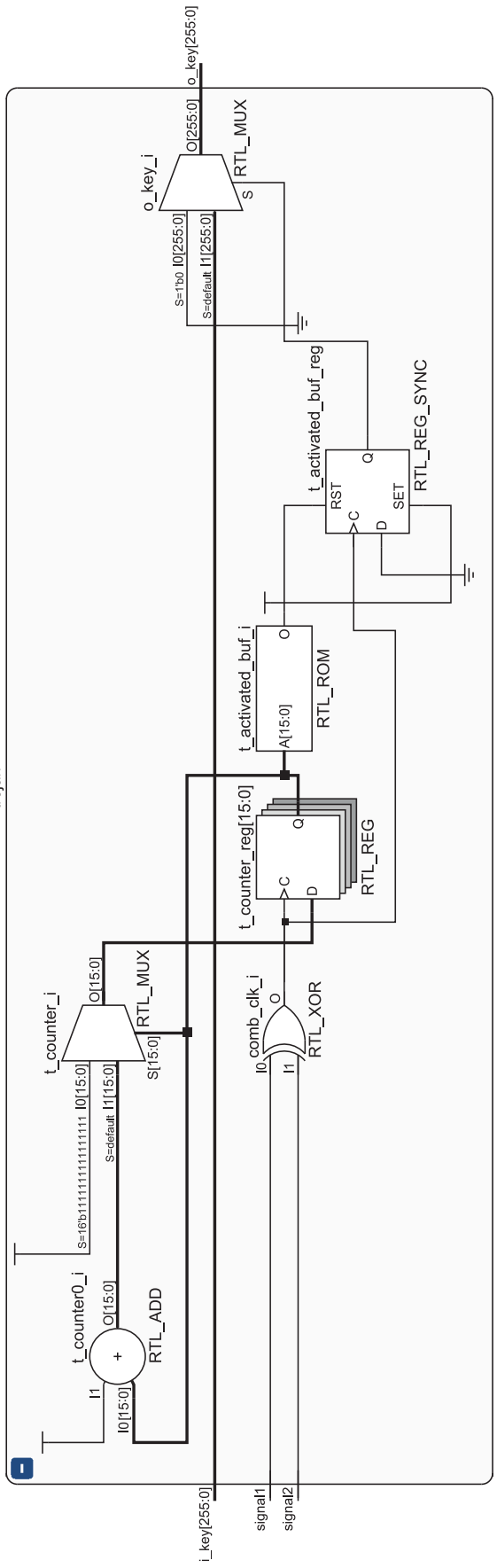


Рис. 3. Схематичное изображение комбинационного трояна-счетчика

Fig. 3. Schematic of combinational Trojan-counter

После получения всей последовательности данных троян переходит в замкнутое состояние конечного автомата (deadlock), при котором активируется скрытый функционал. В данной реализации у трояна отсутствует энергонезависимая память для хранения последнего состояния конечного автомата, и отключение устройства от питающего напряжения сбросит автомат в начальное состояние [8]. Для активации трояна ему необходимо получить значения «15», «1F» и «AF» в шестнадцатеричном формате в старших байтах незашифрованных посылок. При активации аппаратная закладка подменит ключ шифрования на нули. Схематичное изображение трояна показано на рис. 4.

Каждый из механизмов активации троянов также будет выполнять операции подмены ключа шифрования на единицы, отправки незашифрованного ключа вместо зашифрованной посылки или отправки незашифрованной посылки. Их схематичные изображения не приведены ввиду малой целесообразности, поскольку в них будут изменены лишь входные линии, за исключением тех троянов, где осуществляется отправка ключа шифрования. В этих троянах отправка ключа происходит за два цикла шифрования, потому что ключ составляет 256 бит, а посылка – 128 бит. Также сама отправка ключа повторяется 10 раз при каждой смене ключа шифрования после активации трояна.

При помощи библиотеки cryptography 47.0.0 Python сформирован набор тестовых векторов из 12 миллионов значений, которые включали в себя значения как нормальной работы устройства, так и значения, соответствующие выходным параметрам рассматриваемого блока шифрования при активированной аппаратной закладке.

Основная проблема рассматриваемого подхода, а именно – обычный анализ данных на входах и выходах цифрового блока, заключается в том, что считается невозможным сколько-нибудь точное предсказание нейросетью результатов шифрования, особенно побитового. Начальный подход в исследовании базировался на обучении нейронной сети без учителя, как это было проделано в [3], используя технику автоэнкодера. Автоэнкодер является базовой техникой машинного обучения и позволяет модели самостоятельно понять, как лучше представить оригинальные значения в более сжатой форме. Ожидалось, что данный формат позволит нейросети находить аномалии при анализе входов и выхода блока AES-256 при активированном трояне. Для проверки этой теории с использованием набора тестовых векторов модель обучалась десять тысяч эпох, где в каждой эпохе вероятность подстановки неверного значения зашифрованного текста изменялась от 5 до 15 %. По итогу обучения логарифм потерь (log loss) составил 0,693, что соответствовало модели, которая с вероятностью 50 % отличает два класса (в рассматриваемом случае – нормальная работа и аномалия). Применение данного подхода не позволило обучить модель: даже после длительной тренировки она продолжала выдавать значения случайным образом.

На основании результатов применения техники автоэнкодера было решено изменить подход к рассматриваемым наборам данных. Учитывая, что шифрование AES-256 подразумевает побитовые операции при расширении ключа шифровании и шифровании самой посылки, это не позволяет применить классический подход для обнаружения некорректной работы цифрового блока. Было принято решение применить подход, используемый для обучения классификации изображений. В этом методе нейросеть при своей работе получает 128-битное значение незашифрованной посылки, 128-битное – зашифрованной посылки и 256-битное значение ключа шифрования, а на выходе выдает значение, к какому классу принадлежит данный набор бит: нормальная работа, подмена ключа шифрования, утечка ключа шифрования, отправка незашифрованной посылки. Ожидалось, что новый подход полностью копирует задачу побитовой классификации изображений, для чего нейросети сейчас активно и успешно применяются. Машинное обучение проводилось на тех же векторах значений на протяжении 50 эпох. По итогу обучения нормальная работа предсказывается правильно в 81 % случаев (F1-score: 0,88), подмена ключа шифрования – в 99 % (F1-score: 0,99), утечка ключа шифрования – в 99 % (F1-score: 0,86), отправка незашифрованной посылки – в 100 % случаев (F1-score: 1,00). Такой результат значительно превосходит первоначальный подход с автоэнкодером для задачи обнаружения аномалий в работе блока шифрования.

Для полученной модели был сформирован новый набор тестовых векторов с другим ядром генерации (seed) случайных чисел. Этот набор включал один миллион значений незашифрованных посылок с одним ключом шифрования на каждую одну тысячу посылок. Данные на выходе (зашифрованные посылки) получались при симуляции работы AES-256 в Xilinx Vivado 2024.2

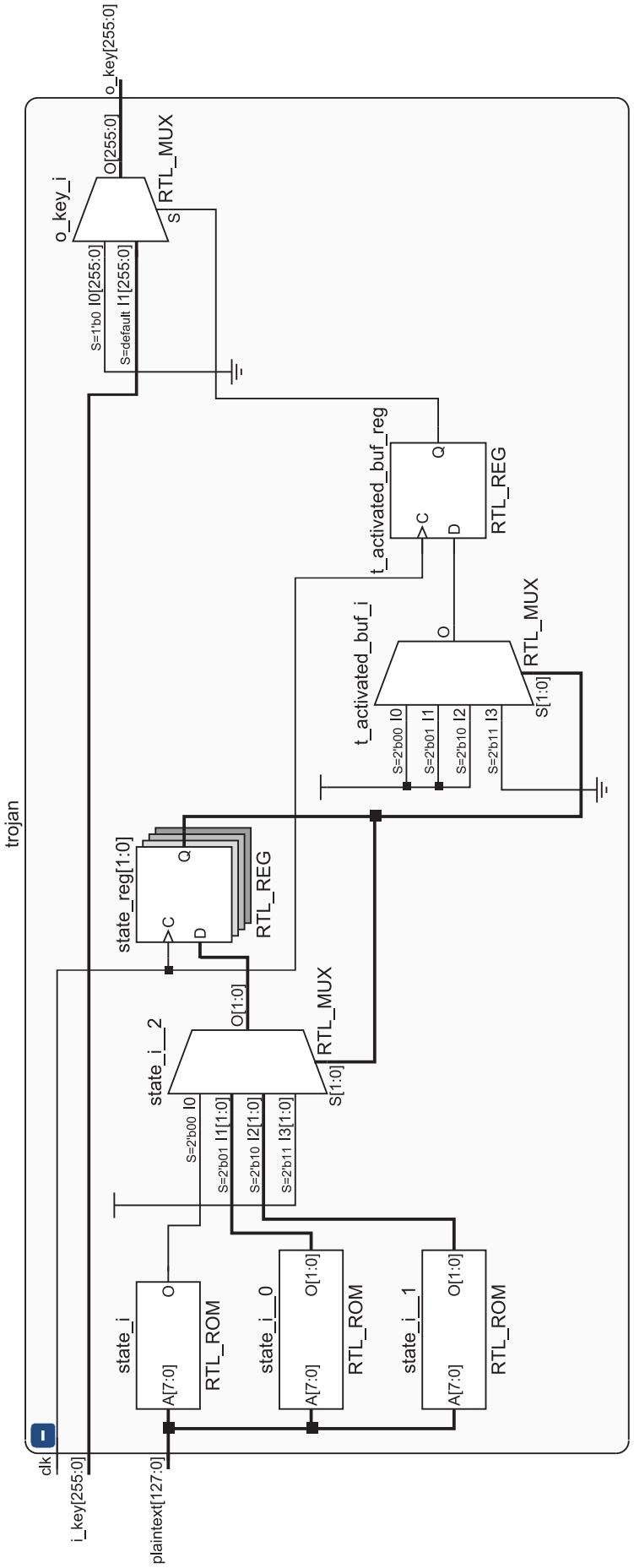


Рис. 4. Схематичное изображение трояна с автоматом конечных состояний
Fig. 4. Schematic representation of a Trojan with a finite state machine

как для нормальной работы, так и для реализаций со встроенным аппаратным трояном для каждого способа его активации и вида функционального изменения.

Результаты исследований и их обсуждение

Полученные значения анализировались обученной нейросетью для каждой реализации модуля шифрования AES-256. Результаты анализа обученной нейросети приведены в табл. 1.

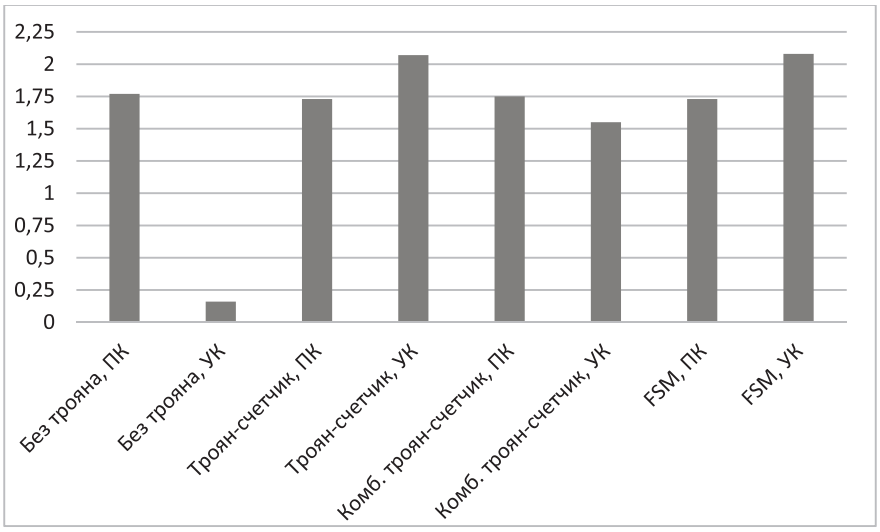
Таблица 1. Результаты анализа обученной нейросети
Table 1. Results of the trained neural network

Анализируемый проект	Нормальная работа, %	Утечка ключа, %	Подмена ключа, %	Незашифрованная посылка, %
Без трояна	98,06	0,16	1,77	0,01
Троян-счетчик, подмена ключа на 0	0,17	0	99,83	0
Троян-счетчик, подмена ключа на 1	0,16	0	99,84	0
Троян-счетчик, незашифрованная посылка	0,16	0	0	99,83
Троян-счетчик, отправка ключа	96,19	2,07	1,73	0,01
Комбинационный троян-счетчик, подмена ключа на 0	0,16	0	99,84	0
Комбинационный троян-счетчик, подмена ключа на 1	26,87	0,05	73,08	0
Комбинационный троян-счетчик, незашифрованная посылка	26,86	0,05	0,55	72,54
Комбинационный троян-счетчик, отправка ключа	96,69	1,55	1,75	0,01
FSM, подмена ключа на 0	0,02	0	99,98	0
FSM, подмена ключа на 1	0,02	0	99,98	0
FSM, незашифрованная посылка	0,02	0	0	99,98
FSM, отправка ключа	96,19	2,08	1,73	0,01

В связи с невысокой плотностью ошибок в самом наборе данных для операции типа «отправка ключа», а также с сопоставимой погрешностью обнаружения операций типа «подмена ключа» классификатором на рис. 5 приведена диаграмма, отображающая долю ошибок для троянов с операциями «подмена ключа» и «утечка ключа» с каждым вариантом активации, а также для AES-256 без аппаратной закладки. Обозначения на рис. 5: УК – утечка ключа; ПК – подмена ключа; Комб. – комбинационный.

На рис. 5 отчетливо видно, что во всех случаях отправки ключа подмены доля обнаружения классификатором наличия операций типа «подмена ключа» – в среднем 1,73, что соответствует такому же значению при работе устройства шифрования без аппаратной закладки. Следует также отметить, что плотность событий отправки ключа составляет в среднем 2 % среди общего числа тестовых векторов для каждого проекта, а именно – 20 тысяч операций отправки ключа шифрования на один миллион операций нормального шифрования.

Полученные результаты позволяют предположить, что при помощи машинного обучения возможно определить подмену ключа шифрования в модулях AES-256 и его разновидностях при использовании более разнообразных комбинаций подмененных ключей, а применение более сложных архитектур нейронных сетей увеличит эффективность предлагаемой методики обнаружения аппаратной закладки. И хотя данный метод не позволяет обнаружить троян на этапе проектирования, с его помощью на этапе эксплуатации можно обнаружить утечку секретной информации в системах передачи данных и сразу перейти на резервные каналы связи. Дешевизна и простота внедрения данной методики проверки обеспечивают возможность ее применения при реализации пользовательских алгоритмов шифрования на интегральных микросхемах специального назначения (ASIC), произведенных на мощностях третьих стран.



УК – утечка ключа, ПК – подмена ключа, Комб. – комбинационный

Рис. 5. Доля ошибок для AES-256 без аппаратной закладки
и для троянов при операциях «утечка ключа» и «подмена ключа»
Fig. 5. Error rates for AES-256 without hardware backdoors
and for Trojans during “key leakage” and “key substitution” operations

Заключение

1. По итогам исследования можно отметить, что методика, основанная на анализе входных и выходных данных с помощью машинного обучения, позволяет обнаруживать функционирующую аппаратную закладку в криптографических устройствах шифрования в 90 % случаев, в том числе и для событий малой плотности. Нейросеть-классификатор способна распознавать как простые случаи (например, утечку ключа или незашифрованную посылку), так и нормальную работу устройства шифрования и подмену ключа на одинаковые биты с достаточной точностью.

2. Полученные результаты позволяют сделать вывод о перспективности предлагаемого метода, а также доработать архитектуру нейронной сети с целью добавления возможности обнаружения иных комбинаций подмененного ключа и увеличения ее эффективности для обнаружения и классификации скрытого функционала цифровых блоков со встроенным аппаратным трояном.

Список литературы

1. Tehranipoor, M. A Survey of Hardware Trojan Taxonomy and Detection / M. Tehranipoor, F. Koushanfar // IEEE Design & Test of Computers. 2010. Vol. 27, Iss. 1. P. 10–25. DOI: 10.1109/MDT.2010.
2. Chen, Z. Deep Learning for Cybersecurity: A Review / Z. Chen // 2020 International Conference on Computing and Data Science (CDS), Stanford, CA, USA, 2020. P. 7–18. DOI: 10.1109/CDS49703.2020.00009.
3. Hardware Trojan Detection Using Graph Neural Networks / R. Yasaei [et al.] // IEEE Trans. Comput.-Aided Design Integr. Circuits Syst. 2025. Vol. 44, No 1. P. 25–38.
4. Hardware Trojan Detection Using Unsupervised Deep Learning on Quantum Diamond Microscope Magnetic Field Images / M. Ashok [et al.] // ACM Journal on Emerging Technologies in Computing Systems. 2022. Vol. 18, No 4. P. 1–25.
5. Xilinx Power Estimator User Guide (UG440). Santa Clara, California: AMD/ Xilinx, 2023.
6. Vivado Design Suite User Guide (UG901). Santa Clara, California: AMD/ Xilinx, 2020.
7. Тарасов, И. Е. ПЛИС Xilinx. Языки описания аппаратуры VHDL и Verilog, САПР, приемы проектирования / И. Е. Тарасов. М.: Горячая линия – Телеком, 2021.
8. Белоус, А. И. Программные и аппаратные трояны – способы внедрения и методы противодействия. Первая техническая энциклопедия в 2-х книгах / А. И. Белоус, В. А. Солодуха, С. В. Шведов. М.: Техносфера, 2019.

Поступила 14.10.2025 Принята в печать 25.11.2025

References

1. Tehranipoor M., Koushanfar F. (2010) A Survey of Hardware Trojan Taxonomy and Detection. *IEEE Design & Test of Computers*. 27 (1), 10–25. DOI: 10.1109/MDT.2010.
2. Chen Z. (2020) Deep Learning for Cybersecurity: A Review. *2020 International Conference on Computing and Data Science (CDS), Stanford, CA, USA*. 7–18. DOI: 10.1109/CDS49703.2020.00009.
3. Yasaei R., Chen L., Yu S.-Y., Al Faruque M. A. (2025) Hardware Trojan Detection Using Graph Neural Networks. *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst*. 44 (1), 25–38.
4. Ashok M., Turner M. J., Walsworth R. L., Levine E. V., Chandrakasan A. P. (2022). Hardware Trojan Detection Using Unsupervised Deep Learning on Quantum Diamond Microscope Magnetic Field Images. *ACM Journal on Emerging Technologies in Computing Systems*. 18 (4), 1–25.
5. *Xilinx Power Estimator User Guide (UG440).* Santa Clara, California: AMD/ Xilinx, 2023.
6. *Vivado Design Suite User Guide (UG901).* Santa Clara, California: AMD/ Xilinx, 2020.
7. Tarasov I. E. (2021) *Xilinx FPGAs. VHDL and Verilog Hardware Description Languages, CAD, and Design Techniques*. Moscow, Goryachaya Liniya – Telekom Publ. (in Russian).
8. Belous A. I., Soloduha V. A., Shvedov S. V. (2019) *Software and Hardware Trojans – Methods of Deployment and Countermeasures. The First Technical Encyclopedia in Two Books*. Moscow, Technosphera Publ. (in Russian).

Received: 14 October 2025 Accepted: 25 November 2025

Вклад авторов / Authors' contribution

Авторы внесли равный вклад в написание статьи / The authors contributed equally to the writing of the article.

Сведения об авторах

Воронов А. Ю., асп. каф. микро- и наноэлектроники, Белорусский государственный университет информатики и радиоэлектроники

Стемпицкий В. Р., канд. тех. наук, доц., проректор по научной работе, научный руководитель НИЛ «Компьютерное проектирование микро- и наноэлектронных систем», Белорусский государственный университет информатики и радиоэлектроники

Адрес для корреспонденции

220013, Республика Беларусь,
Минск, ул. П. Бровки, 6
Белорусский государственный университет
информатики и радиоэлектроники
Тел.: +375 29 353-52-74
E-mail: voronov.drawtoon@gmail.com
Воронов Алексей Юрьевич

Information about the authors

Voronov A. Yu., Postgraduade of the Department of Micro- and Nanoelectronics, Belarusian State University of Informatics and Radioelectronics

Stempitsky V. R., Cand. Sci. (Tech.), Associate Professor, Vice-Rector for Academic Affairs, Advicer of the R&D Laboratory "Computer-Aided Design of Micro- and Nanoelectronic Systems", Belarusian State University of Informatics and Radioelectronics

Address for correspondence

220013, Republic of Belarus,
Minsk, P. Brovki St., 6
Belarusian State University
of Informatics and Radioelectronics
Tel.: +375 29 353-52-74
E-mail: voronov.drawtoon@gmail.com
Voronov Aleksey Yuryevich