СОВРЕМЕННЫЕ ТЕНДЕНЦИИ ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОБНАРУЖЕНИЯ АППАРАТНЫХ ЗАКЛАДОК В ЦИФРОВЫХ УСТРОЙСТВАХ

Воронов А.Ю.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Стемпицкий В.Р. – канд. тех. наук.

В работе рассматриваются основные направления в области обнаружения скрытого функционала в цифровых устройствах с использованием приемов применения методов машинного обучения для анализа результатов, полученных с помощью классических подходов. Дано краткое представление результатов последних работ. Статья подчеркивает перспективность применения нейросетей для анализа результатов классических подходов.

В условиях роста информатизации общества повышается спрос на электронные компоненты и устройства, обладающие высокой производительностью и улучшенными эксплуатационными характеристиками. Увеличивающаяся цифровизация сфер государственного управления, оборонной отрасли, а также товарно-сервисных отношений диктует необходимость постоянного присутствия на рынке всё более мощных и при этом универсальных цифровых или гибридных интегральных микросхем (ИМС). В свою очередь, развитие концепций цифровых экосистем требует не только масштабного применения существующей электроники, но и создания новых решений с расширенным функционалом, большими объёмами хранения и передачи данных. Это обусловливает высокую скорость разработки и значительные объёмы выпуска продукции, которые зачастую невозможно обеспечить в рамках одной компании или даже одного государства.

Стремясь снизить экономические издержки при создании каждой последующей итерации ИМС, разработчики интегрируют в проекты IP-блоки (Intellectual Property) сторонних фирм, специализирующихся на конкретных дизайнерских решениях, а также обращаются к полупроводниковым фабрикам, расположенным за рубежом. Однако перечисленные реалии современного цикла разработки цифровых устройств существенно увеличивают риски внесения несанкционированных изменений — аппаратных троянов. Такие закладки могут возникать как при использовании внешних IP-блоков, так и на этапе производства, и способны приводить к нарушению функциональности устройства, утечке данных или выходу прибора из строя.

Аппаратные трояны принято классифицировать по пяти основным признакам: на каком этапе разработке была внедрена закладка; какой уровень архитектуры подвергся атаке; механизму активации закладки; результату активации трояна; геометрическому расположения закладки [1, с. 497]. Результатом активации аппаратных закладок в часто является изменение функциональных параметров, отличных от изначальной спецификации цифрового устройства. Аппаратные закладки, рассматриваемые в данных работах внедрены на этапе проектирования на уровне регистровых передач (Register Transfer Level, RTL). Аппаратные трояны имеют внутренний, который срабатывает по истечению определенного времени (счетчики), и внешний, который срабатывает с помощью определенной комбинации байт на входе устройства, механизмы активации.

Верификацию внесения сторонних изменений в оригинальный дизайн цифровой схемы проводят при помощи методов, которые можно разделить на две группы: с последующим разрушением микросхемы и без ее разрушения [1, с. 718]. Первый способ существует на протяжении долгого времени и основывается на том, что поочередно изучается топология слоев готовой микросхемы с помощью оптической или электронно-лучевой микроскопии. Данный метод обладает высокой точностью, однако является долгим и дорогостоящим, а также требует специализированную лабораторию с обученным персоналом. Неинвазивные способы обнаружения аппаратных закладок более предпочтительны в связи с низкой стоимостью и возможностью обнаружения трояна на этапе разработки. Одним из неинвазивных методов является анализ по стороннему каналу (Side-channel analysis), суть которого в отслеживании изменений физических характеристик микросхемы во время проектирования или работы: потребляемая мощность, температура и временные задержки.

Перечисленные классические подходы для обнаружения внедренного скрытого функционала в чистом виде являются малоэффективными, из-за большого количества данных в современных микросхемах, количество элементов в которых не меньше нескольких миллионов. По этой причине, для анализа большого количества данных с малыми изменениями применяются методы машинного обучения. Основные методы обнаружения по этапу разработки цифрового устройства можно разделить на те, которые применяются на этапе проектирования, и на те, которые используются при проверке готовой продукции или инженерного образца.

К первой категории относятся применение графовых нейронных сетей (Graph Neural Network, GNN) для топографического анализа RTL, который представлен в виде абстрактного синтаксического

дерева, и использование больших языковых моделей (Large Language Model, LLM) при семантической поверке исходного кода на языке описания аппаратуры.

В работе [2] авторы используют исходный RTL-код, полученный из бенчмарков TrustHub и открытых репозиториев, который при помощи инструментария Pyverilog языка программирования Python весь проект преобразуется в граф потоков данных (Data Flow Graph, DFG), каждый узел которого наделяется поведенческими признаками. В дальнейшем сравнивалась эффективность трех архитектур GNN: графовой сверточной сети (Graph Convolutional Network, GCN), графовой сети внимания (Graph Attention Network, GAT) и графовой изоморфной сети (Graph Isomorphism Network, GIN). По показателям самую высокую точность определения обеспечивает архитектура GCN с двумя слоями с показателем 98,66 %. Авторы отдельно отмечают достаточно высокий показатель точности 98,20 % у GIN с пятью слоями. GAT же показала максимальную точность 96,80 % в реализации с двумя слоями.

В работе [3] продемонстрирована возможность обнаружения трояна при помощи анализа входного и выходного траффика в устройстве криптографии AES 256. Для обучения использовалась полносвязанная нейронная сеть, которая обучена на пользовательском датасете из свыше 4 миллиона наборов данных. Обученная модель проверялась на новых данных и показала точность не менее 95 % во всех случая, кроме одного случая подмены ключа на одинаковые биты. Данная методика еще находится в активной доработке для возможности работы с более сложными случаями подмены ключа.

Ко второй категории часто относится проверка физических характеристик изделия по сторонним каналам. К примеру, в работе [4], используя данные временных задержек и преобразований Вейвлета и Фурье для изменений напряжения, показана эффективность применения Глубинной нейронной сети (Deep Neural Network, DNN) и модель случайного леса (Random Forest Classifier, RFC). Модели обучались на основе пользовательского датасета и показали свою эффективность при тестах AES-1000, AES-1100, AES-1300 и AES-1400 показав полноту обнаружения в 90 % и 97 % соответственно.

В работе [5] показано, как при помощи электромагнитной микроскопии можно обнаружить встроенный аппаратный троян в микросхему, используя модель автоэнкодера для сверточной нейронной сети. Для обучения использован пользовательской датасет. Методика показала 87 % точность при ее проверке на FPGA со встроенными троянами из тестов AES TrustHub. Автор предполагает использование этого метода в связке с физически неклонируемыми функциями (Physical Unclonable Function, PUF).

Таким образом можно сделать вывод, что применение методов машинного обучения для обнаружения аппаратных закладок в заказных микросхемах является крайне перспективным дополнением для анализа данных, получаемых классическими методами проверок, таких как функционально-логическое тестирование, проверка по стороннему каналу и анализ топологии проекта.

***Список использованных источников**:*

1. Белоус А. И., Солодуха В. А., Шведов С. В. Программные и аппаратные трояны – способы внедрения и методы противодействия. Первая техническая энциклопедия. Москва, Техносфера, 2019. 630 с

2. V. T. Hayashi and W. V. Ruggiero, "Hardware Trojan Detection in Open-source Hardware Designs Using Machine Learning," IEEE Trans. VLSI Syst., 2025.

3. 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.

4. Bhatta, N.P., Amsaad, F. MI assisted techniques in power side channel analysis for trojan classification. Cluster Comput 28, 157 (2025). https://doi.org/10.1007/s10586-024-04715-w

5. Maitreyi Ashok, Matthew J. Turner, Ronald L. Walsworth, Edlyn V. Levine, and Anantha P. Chandrakasan. 2022. Hardware Trojan Detection Using Unsupervised Deep Learning on Quantum Diamond Microscope Magnetic Field Images. J. Emerg. Technol. Comput. Syst. 18, 4, Article 67 (October 2022), 25 pages. https://doi.org/10.1145/3531010.