A novel holistic approach for hardware trojan detection powered by deep learning (HERO)

Serafeim. P. Moustakidis[1*], Konstantinos G. Liakos[2], Georgios K. Georgakilas[2], Nikolaos Sketopoulos[2], Stavros Seimoglou[2], Patrik Karlsson[1], Fotios Plessas[2]

[1]AIDEAS OÜ, Narva mnt 5, Tallinn, Harju maakond, 10117, Estonia; [2]Department of Electrical and Computer Engineering School of Engineering, University of Thessaly Volos, 38221 Greece
*Corresponding author: s.moustakidis@aideas.eu

ABSTRACT

A novel artificial intelligence (AI) empowered methodology is presented in this paper providing a holistic approach to the problem of hardware trojan (HT) detection in integrated circuits (IC). The proposed HERO pipeline demonstrates a unique potential in identifying HTs by combing multi-modal data coming from different stages of the IC manufacturing process. Apart from HERO's innovative character and the project results, this manuscript also focuses on our future project vision, presenting the main steps required for scaling up the HERO technology readiness level as well as our plans for technology commercialisation.

Keywords: Hardware trojan detection; integrated circuits; deep learning; multi-modal data

1. INTRODUCTION

Considering the propagation of integrated circuits in today's world – including domestic appliances like washing machines, means of transportation (e. g. cars or airplanes), clinical devices (e. g. enabling precise diagnosis) and military appliances (e. g. enhancing the effectiveness of weapons) – hardware trojans impact our everyday life and may even cause life threatening situations. Unlike other errors and malfunctions, trojans are inserted deliberately. Apart from insider attacks, the economically driven outsourcing of production steps to third party contractors enlarges the attack surface dramatically. Contractors, their employees, and intruders potentially modify the design without the designer's or customer's knowledge.

Ideally, any undesired modification made to an IC should be detectable by presilicon verification/simulation and post-silicon testing. However, existing HT detection techniques require either the use of golden models or time-consuming exhaustive verification procedures [1-4]. Today, there is no single tool that can provide a holistic approach for hardware trojan detection taking advantage of multi-modal data coming from different stages of the IC manufacturing process. HERO tries for the first time to develop an AI empowered multiscale approach for HT detection overcoming all the aforementioned challenges.

The proposed methodology goes a step forward from the detection approaches that are available today providing a holistic global solution to the HT detection problem. The combined effect of known HT detection approaches with the latest advances in data mining algorithms and deep learning networks has led to the development of a novel gold standard HT detection tool that is expected to revolutionise the IC sector.

As reported in our results, HERO demonstrated unique potential in detecting trojan-infected IC constituting a first proof of concept of the proposed AI pipeline. The proposed holistic methodology may facilitate and accelerate HT detection bringing a breakthrough in numerous IC-related sectors. Further experimentation on bigger datasets is essential to validate the detection capacity of HERO.

The rest of the manuscript is organised as follows. State of the art (SoA) is given in Section 2 whereas the innovative character of HERO with respect to the current SoA is presented in Section 3. Section 4 presents the core results achieved and Section 5 focuses on the future project vision. Acknowledgement and references are given in Sections 6 and 7.

2. STATE OF THE ART

The HT detection consists a modern and challenging problem that lacks a meaningful solution to deal with it. Advanced data-driven studies are limited and especially are focused on providing a partial solution to the problem either on the design verification phase or after fabrication. Support vector machines [5] and other SoA Machine Learning (ML) techniques [6-7] have been investigated with respect to their predictive capacity to recognise HT-infected designs using pre-determined and manually selected features. Multi-neural networks, decision trees, Bayesian and k-nearest neighbours' classifiers have been also applied for detecting HT during and after fabrication using side-channel parameters such as delay or power values [8].

As can be seen from the overall picture given in [9], the use of methods based on ML techniques for the detection of HTs is at a primary stage. We have also observed a great variability on the features employed as well as in the type of benchmarks that were used for the development of the training models. Although it is noticeable that none of the studies has ever attempted to work on a holistic approach that could detect HTs combining information from heterogeneous data sources.

Applying ML for the detection of HTs on ICs is expected to be a very effective solution of the near future. A possible way to achieve this is by using and combining many different types of features from a variety of benchmarks towards the creation of models that can detect the HTs both in pre-silicon and post-silicon phases. This approach will create practical tools that will be able to detect and prevent accurately the infected HT circuits, with the ultimate objective to optimize and achieve a higher level of security in the circuit industry.

3. BREAKTHROUGH CHARACTER OF THE PROJECT

The breakthrough character of HERO is highlighted below. HERO is highly innovative due to its:

- *ability to handle and combine multi-modal data*

An innovative AI-driven pipeline has been designed in HERO allowing the incorporation and fusion of heterogeneous data as they have been generated in different manufacturing stages (design, verification, post-silicon etc). Every data source is treated separately via the proposed HERO analytics pipeline that brings all available data into a uniform format (single-channel images) and deep learning networks are finally applied to identify HT circuits.

- *ability to overcome class imbalance problems*

The trojan infected and trojan-free IC classes are not equally represented given that for every trojan-free IC there are hundreds HT-infected variations of the circuit. This poses a severe class imbalance problem that makes the training of any ML or DL technique difficult. HERO overcomes this challenge by integrating in its pipeline data augmentation algorithms such as Generative Adversarial Networks (GANs) [10], Synthetic Minority Over sampling TEchnique (SMOTE) [11] and ADAptive SYNthetic (ADASYN) [12].

- *ability to defeat high dimensionality*

HERO relies on a highly effective and computationally efficient feature selection (FS) algorithm to reduce the increased feature dimensionality and finally select / rank features with respect to their relevance with the existence of HT in ICs. The employed FS algorithm [13] pays due attention on the complementary properties between features and thus selects highly relevant features while on the other hand overcomes the feature redundancy problems by promoting co-operation between features.

- *ability of building powerful models without the need of huge training datasets*

HERO implements a modality transformation where non-interpretable feature vectors are being transformed to the image domain. Looking at these visual representations of data, a human can extract relevant information and a sense of the meaning it conveys. DL has been recently proven to be extremely successful in various domains especially the ones involving images [14]. Thus, transforming features to the visual domain enables the application of a large number of powerful CNN architectures including pre-trained ones bypassing the requirement of large amounts of labelled data to facilitate the training of these very deep networks.

Tab. 1. Progress beyond the state-of-the-art by HERO

Challenge to overcome	SoA	HERO
Handling of multi-modal data	-	✓
Severe class imbalance	Partially	✓
Curse of dimensionality	Partially	✓
Need of massive datasets for training	-	✓

4. PROJECT RESULTS

The main project result is the design of the innovative holistic HERO AI-empowered detection pipeline. After extensive experimentation with numerous tools and algorithms, we finally concluded with the methodology presented in **Fig.1**. The HERO HT detection methodology includes an advanced analytics pipeline applied at each data source along with a deep learning-based fusion and decision-making mechanism. More information about the HERO processing components is provided below, whereas some of our core results are finally given in the second part of Section 4.

HERO analytics pipeline

The HERO analytics pipeline comprises of all the necessary processing steps to: (i) extract informative features, (ii) overcome the class imbalance problem and (iii) bring all the available data into a common representation form (images in our case).

Dataset creation and feature extraction

Approximately 1000 circuit benchmarks from the Trust-HUB library [15] were utilised in our experimentation to generate an informative training dataset for our algorithms. Moreover, we have also designed and manufactured our own HT-infected and HT-free circuits using Design Compiler NXT software from Synopsys Design Compiler and Cadence Innovus Implementation System software. Various HT detection techniques were considered in our analysis:

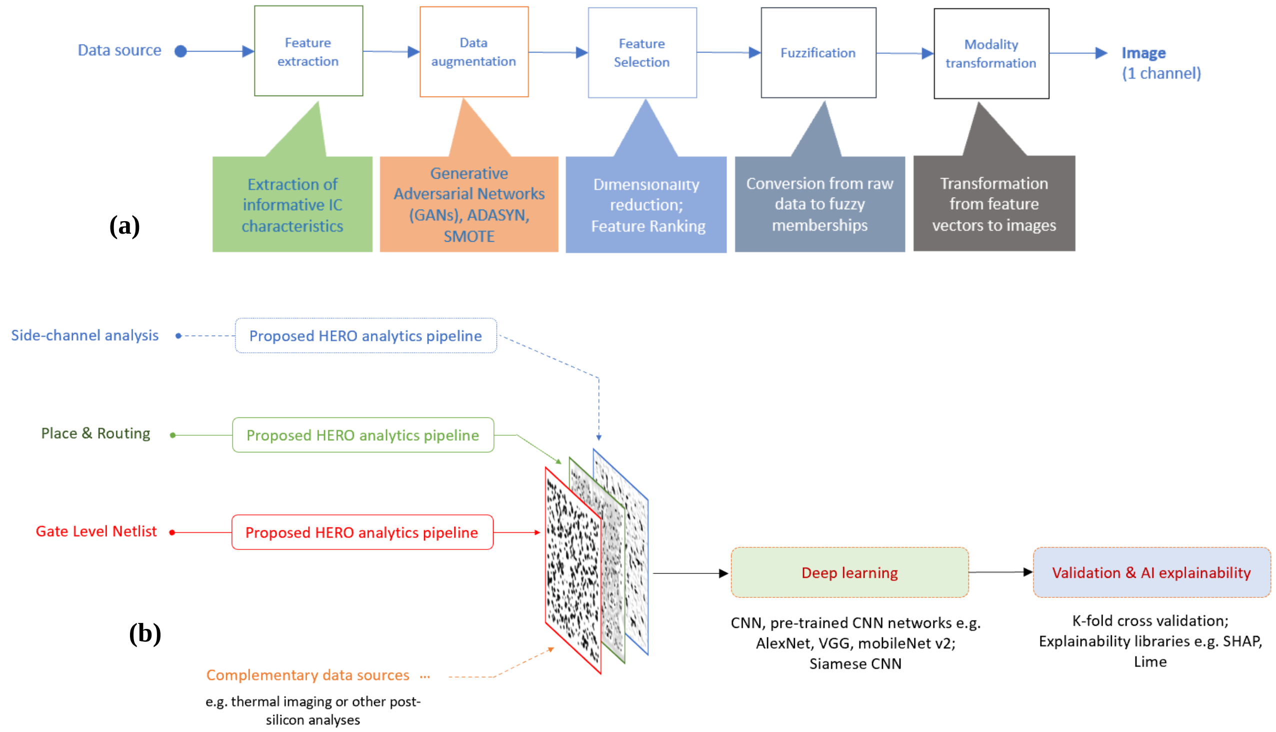
Data source
Feature extraction
Data augmentation
Feature Selection
Fuzzification
Modality transformation
Image
(1 channel)
Extraction of informative IC characteristics
Generative Adversarial Networks (GANs), ADASYN, SMOTE
Dimensionality reduction; Feature Ranking
Conversion from raw data to fuzzy memberships
Transformation from feature vectors to images
(a)
Side-channel analysis
Proposed HERO analytics pipeline
Place & Routing
Proposed HERO analytics pipeline
Gate Level Netlist
Proposed HERO analytics pipeline
(b)
Complementary data sources ...
e.g. thermal imaging or other post-silicon analyses
Deep learning
CNN, pre-trained CNN networks e.g. AlexNet, VGG, mobileNet v2; Siamese CNN
Validation & AI explainability
K-fold cross validation; Explainability libraries e.g. SHAP, Lime

Fig. 1. a) The proposed HERO analytics pipeline applied at each data source, b) The proposed holistic HERO approach

(i) Gate level netlist analysis (GLN) that describes the connections between circuit components; (ii) Placement and Routing (PnR) that decides the exact placement of circuitry and the design of all the wires needed to connect the placed components; (iii) Side channel analysis including feature related to power consumption or the delay of certain paths of the circuit.

Data augmentation and FS

To cope with the class imbalance problem, a number of data augmentation algorithms was investigated. GANs, ADASYN and SMOTE were employed to enhance the minority class and overcome potential bias on the generated detection models. Subsequently, FS was applied to: (a) rank features with respect to their relevance HT and (b) remove features with no or redundant information.

Modality transformation on the fuzzy domain

The selected features were then gone through a modality transformation process. Specifically, raw feature values were converted to fuzzy memberships and then they were transformed into the image domain where the distinguishable visual patterns could be captured by the CNN-based networks that follows.

Holistic deep learning-based HT detection

Applying the proposed analytics pipeline on different data sources leads to the creation of multi-channel images which were then supplied to a CNN-based deep learning network. The pipeline has been carefully designed to also incorporate other complementary data e.g. thermal imaging acquired in the post-silicon phase. It should be noted that a variety of networks could be used here (either pre-trained or not) to implement the detection task (e.g. CNN, autoencoders or Siamese-based networks) taking into account all the available information from the various phases of the manufacturing process (expressed in the form of images).

Results

Indicative results from our experimentation are presented below. Specifically, **Tab. 1** cites the obtained classification outcomes of the proposed AI pipeline using data from GLN and PnR analyses. Common performance metrics were employed such as confusion matrix, rates such as true-negatives (TN), false-negatives (FN), false-positives (PP), true-positives (TPs), precision and overall classification accuracy. **Fig.2** shows the detection accuracies obtained from: (i) HERO, (ii) HERO without data augmentation (*w/o DA*), (iii) HERO without FS (*w/o FS*) and (iv) HERO without modality transformation (*w/o MT*).

The following remarks could be drawn from **Tab. 1** and **Fig.2**.

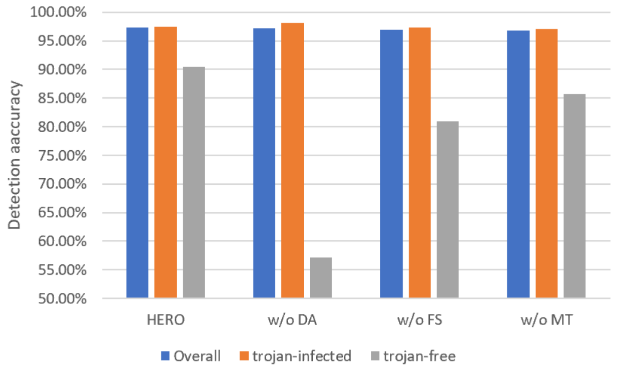

Fig. 2. Comparative analysis

Tab. 1. Detection performance achieved by HERO

	Trojan-free	*Trojan-infected*
Trojan free	90.48% (TN)	2.49% (FN)
Trojan-infected	9.52% (FP)	97.51% (TP)
precision	99.77%	
Overall accuracy	**97.35%**	
Selected num. of features	28	

- A 97.35% detection performance was achieved by HERO in the specific dataset. Inclusion of more data samples (thousands) should be considered to validate and verify the effectiveness of the proposed methodology.

- Data augmentation plays a crucial role since the minority class (trojan-free) cannot be properly recognised without DA.

- Reducing the dimensionality of the initial feature space has also a positive effect on the classification accuracies (especially for the minority class)

- Modality transformation has a smaller contribution to the final detection capacity of HERO, however it is also a necessary processing step because it enables the incorporation of imaging data on the same AI pipeline.

5. FUTURE PROJECT VISION

5.1. Technology Scaling

During ATTRACT Phase 1 the holistic HERO AI-empowered detection pipeline has been validated on approximately 1000 circuits. In order to reach TRL 5-6 the HERO pipeline will be used on additional circuit designs from the different stages of the IC manufacturing process. Additional data such as thermal imaging and X-ray inspection data will be incorporated in the training data set to improve the ML and DL techniques. We also intend to extend the scope from IC HT detection to SoC (system on a chip) HT detection.

5.2. Project Synergies and Outreach

The consortium will be extended to include the following additional partners: a second ML/AI company, a thermal imaging organisation, an X-ray electronics inspection organisation, IC/SoC design company, and a semiconductor manufacturer.

The major focus of the Dissemination & Communication plan will be to ensure that the project activities and outcomes are widely spread among the appropriate target communities, at appropriate times, via appropriate methods, as well as to identify potential contributors to the development, evaluation, uptake and exploitation of HERO Phase 2 outcomes, encouraging their participation on a systematic and regular basis.

The HERO Phase 2 communication strategy will be structured in three main phases:

- “Phase 1 – Preliminary Project Promotion phase” aims at: (i) Agreeing upon the communication strategy and future activities; (ii) Creating initial awareness in the markets related with the Project’s objectives and scope.
- “Phase 2 – Project Commercialisation phase” aims at: (i) Creating more “targeted awareness” regarding HERO technologies with key players and potential users; (ii) Informing the target market about the technological benefits.
- “Phase 3 – Business Strategy phase” aims at: (i) Maximizing target market and industry awareness regarding the HERO holistic solution; (ii) Thus contributing to ensure the project sustainability and full exploitation.

5.3. Technology application and demonstration cases

The HERO AI-empowered HT detection pipeline will be validated and demonstrated pre-silicon on a multitude of IC designs and ten SoC designs provided by the IC/SoC design company. The designs will cover ICs used for different applications, e.g. medical, control systems, and IoT devices.

In the second phase three IC designs for different applications will be selected and fabricated. For each design, a trojan-free IC and an IC with HT will be produced. The ICs will be inspected by thermal imaging and X-ray to provide additional data to the detection algorithms.

For the fabrication we will use EUROPRACTICE IC Services (https://europractice-ic.com) which is a consortium of five renowned European research organizations, who support academic institutions and medium-sized companies with IC prototyping service. At least 250,000 Euro of the budget will be dedicated to IC prototyping.

By improving HT detection, HERO Phase 2 will enable provision of security assurances for electronics that are used everywhere in our society. From critical infrastructure, Industry 4.0, global supply chains, to home electronics, smart phones, wearable medical devices and IoT devices. Therefore, HERO Phase 2 will bring benefit to all the areas of Scientific Research, Industry and Societal Challenges.

5.4. Technology commercialization

The commercialisation plan is divided in three main phases:

- "*Phase 1* – exploitation activities during the project duration to liaison with potential end-users in order to get information for the optimal commercialisation route. These activities are complimentary to Phase 3 of the communication strategy.
- "*Phase 2* – based on phase 1, focusing on the development of HERO business plan. It will detail the marketing channels and strategy, exploitation role of each partner, economic calculations and forecasts for different business scenarios.
- "*Phase 3* – following the end of the project, will bring the prototype from TRL7 to TRL9 by seeking additional investment, both internal and external. Possible option might also be to enter a joint partnership agreement with large IC manufactures.

5.5. Envisioned risks

The core risks foreseen are the following:

Risk1: not having enough data to train the models

Mitigation: Inclusion of an IC/SoC design company and a semiconductor manufacturer will minimize this risk. Additional data will also be provided by the inspection techniques. Furthermore, the data augmentation techniques developed during HERO Phase 1 have proven to be able to cope with the problem of limited data.

Risk 2: many possible variables and very complex interactions exist which makes the detection difficult

Mitigation: The applied big data and FS technique have shown proven capabilities in coping with problems of high dimensionality. The use of GPU-based deep learning techniques is expected to cope the computational issue.

5.6. Liaison with Student Teams and Socio-Economic Study

A team of MSc students from University of Thessaly was involved in the project providing ideas and technological expertise in the field of IC design. The role of MSc will be enhanced in Phase 2. The consortium is willing to contribute to the socio-economic study.

6. ACKNOWLEDGEMENT

This project has received funding from the ATTRACT project funded by the EC under Grant Agreement 777222.

7. REFERENCES

[1] Abramovici, M. & Bradley, P, 2009, Integrated circuit security: new threats and solutions. In Proceedings of the 5th Annual Workshop on Cyber Security and Information Intelligence Research: Cyber Security and Information Intelligence Challenges and Strategies (CSIIRW '09). Association for Computing Machinery, New York, NY, USA, Article 55, pp. 1–3. doi: https://doi.org/10.1145/1558607.1558671

[2] Jin, Y. & Makris, Y., 2008, Hardware Trojan detection using path delay fingerprint, IEEE International Workshop on Hardware-Oriented Security and Trust, Anaheim, CA, 2008, pp. 51-57. doi: 10.1109/HST.2008.4559049

[3] Potkonjak, M., Nahapetian, A., Nelson, M. & Massey, T., 2009, Hardware Trojan horse detection using gate-level characterization, 2009 46th ACM/IEEE Design Automation Conference, San Francisco, CA, 2009, pp. 688-693. doi: 10.1145/1629911.1630091.

[4] Banga, M. & Hsiao, M.S., 2009, A Novel Sustained Vector Technique for the Detection of Hardware Trojans, 2009 22nd International Conference on VLSI Design, New Delhi, 2009, pp. 327-332. doi: 10.1109/VLSI.Design.2009.22.

[5] Inoue, T., Hasegawa, K, Yanagisawa, M. & Togawa, N., 2017, Designing hardware trojans and their detection based on a SVM-based approach, 2017 IEEE 12th International Conference on ASIC (ASICON), Guiyang, 2017, pp. 811-814. doi: 10.1109/ASICON.2017.8252600.

[6] Lodhi, F.K. et al., 2016, A self-learning framework to detect the intruded integrated circuits, 2016 IEEE International Symposium on Circuits and Systems (ISCAS), Montreal, QC, 2016, pp. 1702-1705. doi: 10.1109/ISCAS.2016.7538895.

[7] Hasegawa, K., Shi, Y. & Togawa, N., 2018, Hardware Trojan Detection Utilizing Machine Learning Approaches, 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/ BigDataSE), New York, NY, 2018, pp. 1891-1896. doi: 10.1109/TrustCom/BigDataSE.2018.00287.

[8] Iwase, T., Nozaki, Y., Yoshikawa, M. & Kumaki, T., 2015, Detection technique for hardware Trojans using machine learning in frequency domain, 2015 IEEE 4th Global Conference on Consumer Electronics (GCCE), Osaka, 2015, pp. 185-186. doi: 10.1109/GCCE.2015.7398569.

[9] Liakos, K.G., Georgakilas, G.K., Moustakidis, S., Karlsson, P. & Plessas, F.C., 2019, Machine Learning for Hardware Trojan Detection: A Review, 2019 Panhellenic Conference on Electronics & Telecommunications (PACET), Volos, Greece, 2019, pp. 1-6. doi: 10.1109/PACET48583.2019.8956251.

[10] Zhang, H., Yu, X., Ren, P., Luo, C. & Min, G., 2019, Deep Adversarial Learning in Intrusion Detection: A Data Augmentation Enhanced Framework, Pre-print 2019, arXiv:1901.07949

[11] He, H., Bai, Y., Garcia, E. & Li., S., 2008, ADASYN: Adaptive synthetic sampling approach for imbalanced learning," 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, 2008, pp. 1322-1328. doi: 10.1109/IJCNN.2008.4633969.

[12] Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W., 2002, SMOTE: Synthetic Minority Over-sampling Technique, Journal Of Artificial Intelligence Research, 16, pp. 321-357. doi: 10.1613/jair.953

[13] Moustakidis, S., & Theocharis, J., 2010, SVM-FuzCoC: A novel SVM-based feature selection method using a fuzzy complementary criterion. Pattern Recognition, 43(11), pp. 3712-3729. doi: 10.1016/j.patcog.2010.05.007

[14] Bengio, Y., LeCun, Y. & Hinton, G., 2015, Deep Learning, Nature, 521, pp. 436–444. doi: https://doi.org/10.1038/nature14539

[15] Shakya, B. et al., 2017, Benchmarking of Hardware Trojans and Maliciously Affected Circuits. Journal of Hardware and Systems Security 1, pp. 85–102. doi: https://doi.org/10.1007/s41635-017-0001-6