data

MDPI

Data Descriptor

Stimulated Microcontroller Dataset for New IoT Device Identification Schemes through On-Chip Sensor Monitoring

Alberto Ramos [1,*], Honorio Martín [1], Carmen Cámara [2] and Pedro Peris-Lopez [2]

[1] Electronic Technology Department, University Carlos III of Madrid, 28911 Leganés, Spain; hmartin@ing.uc3m.es

[2] Computer Science and Engineering Department, University Carlos III of Madrid, 28911 Leganés, Spain; macamara@inf.uc3m.es (C.C.); pperis@inf.uc3m.es (P.P.-L.)

* Correspondence: alramosg@ing.uc3m.es

Abstract: Legitimate identification of devices is crucial to ensure the security of present and future IoT ecosystems. In this regard, AI-based systems that exploit intrinsic hardware variations have gained notable relevance. Within this context, on-chip sensors included for monitoring purposes in a wide range of SoCs remain almost unexplored, despite their potential as a valuable source of both information and variability. In this work, we introduce and release a dataset comprising data collected from the on-chip temperature and voltage sensors of 20 microcontroller-based boards from the STM32L family. These boards were stimulated with five different algorithms, as workloads to elicit diverse responses. The dataset consists of five acquisitions (1.3 billion readouts) that are spaced over time and were obtained under different configurations using an automated platform. The raw dataset is publicly available, along with metadata and scripts developed to generate pre-processed T–V sequence sets. Finally, a proof of concept consisting of training a simple model is presented to demonstrate the feasibility of the identification system based on these data.

Dataset: https://doi.org/10.5281/zenodo.10042177.

Dataset License: The dataset is available under CC-BY 4.0 licence

Keywords: on-chip; sensors; identification; microcontrollers; machine learning; deep learning; hardware security; IoT; fingerprinting; PUF

check for updates

Citation: Ramos, A.; Martín, H.; Cámara, C.; Peris-Lopez, P. Stimulated Microcontroller Dataset for New IoT Device Identification Schemes through On-Chip Sensor Monitoring. *Data* **2024**, *9*, 62. https://doi.org/10.3390/data9050062

Academic Editor: Rüdiger Pryss

Received: 15 March 2024
Revised: 13 April 2024
Accepted: 27 April 2024
Published: 28 April 2024

1. Background and Summary

The Internet of Things (IoT) comprises all those devices that, whether for the purpose of actuation, acquisition, processing, or data exchange, are connected to the Internet or other communication networks [1]. This ubiquitous paradigm encompasses a wide range of applications today, ranging from wearable devices to Industry 4.0, including home automation and many others [2]. This proliferation is expected to increase significantly in the coming years, with the number of active devices worldwide estimated to reach 21.5 billion by 2025 [3,4]. This scenario highlights the need to strengthen the security frameworks on which IoT environments are built. In this regard, device ID authentication within an ecosystem plays a crucial role in ensuring trust in the associated services. Commonly adopted approaches include solutions such as whitelisting based on device MAC addresses, the use of IoT communication protocols themselves, as well as leveraging statistical features of network traffic supported by machine learning (ML) [5,6]. Despite these alternatives, a number of suggested solutions have become compromised when IoT devices were exposed. Recent research has also revealed vulnerabilities by employing malicious inputs to ML-based approaches considered robust [7].

Identification techniques based on differences in the physical properties of devices are gaining prominence, commonly referred to as physically unclonable functions (PUF).

These methods leverage the inherent physical variations in hardware, stemming from manufacturing processes, to achieve unique device identification. A challenge-response pair scheme (CRPs) is typically employed to exploit these physical variations for device identification [8–10].

A current trend that is gaining momentum involves the development of PUF-based identification systems utilizing artificial intelligence (AI) algorithms, such as machine learning (ML) or deep learning (DL). The literature presents various approaches, including the enrollment process of CRPs from existing PUFs [11], and authentication of wireless nodes through DL modeling of transmission parameters [12]. Despite the promising results from such proposals, the majority of studies have primarily been conducted using Monte Carlo-style simulations [13], and a lack of datasets for replicating these methodologies is prevalent.

Overall, the utilization of on-chip sensors, typically embedded in devices such as microcontrollers, SoCs, or FPGAs, remains largely unexplored as a means of device identification [14]. Internal sensors are commonly integrated into these devices by manufacturers to monitor the chip's status under different operational conditions. Their placement varies according to the specifics of the chip's architecture [15]. These embedded sensors are subject to intrinsic physical variations introduced during manufacturing processes, similarly to the rest of the components of the computing device. By employing hardware-specific strategies to stimulate their electronic activity (challenge), we can exploit the dual variability of the devices and collect this information through the sampling of the on-chip sensors (responses). Ultimately, our objective is to use the collected responses to uniquely identify devices by modeling their behavior with the assistance of ML/DL-based algorithms. This approach represents an initial stride towards the development of AI-based identification systems for embedded devices, leveraging the responses from on-chip sensors to infer specific device footprints. These responses embody an innovative source of variability, as the exploitation of electronic activity from the microcontroller architecture alongside on-chip sensors remains uncharted territory, to the best of our understanding.

In order to facilitate future investigations exploring these mechanisms, this work presents a novel dataset [16] consisting of readings from the on-chip temperature and voltage sensors embedded into STM32L152RCT6 microcontrollers. These microcontrollers were incorporated in 20 STM32L-DISCOVERY development boards used to generate a real-world dataset, thereby offering a compelling alternative to the simulation-driven datasets that are predominantly found in the field. The data were collected during the execution of five different algorithms (including matrix multiplication with different types and sorting and cryptographic algorithms) chosen to trigger different architectural blocks. The dataset comprises five acquisitions conducted under conditions characterized by low variability and involving diverse supply equipment, employing distinct acquisition strategies, at different periods. The data are provided in raw form, ready for manipulation, and can be pre-processed for experimentation with data-dependent identification algorithms, with the tools provided for this purpose [17]. Additionally, the dataset includes the calibration values of the internal sensors involved, as well as the unique identifiers (UID) of the chips, which contain relevant information for future meta-analysis, such as the manufacturing batch or the X–Y position of the silicon wafer [18]. The data collection was performed using a platform designed to automate and enhance the flexibility of the acquisition processes (refer to Figure 1).

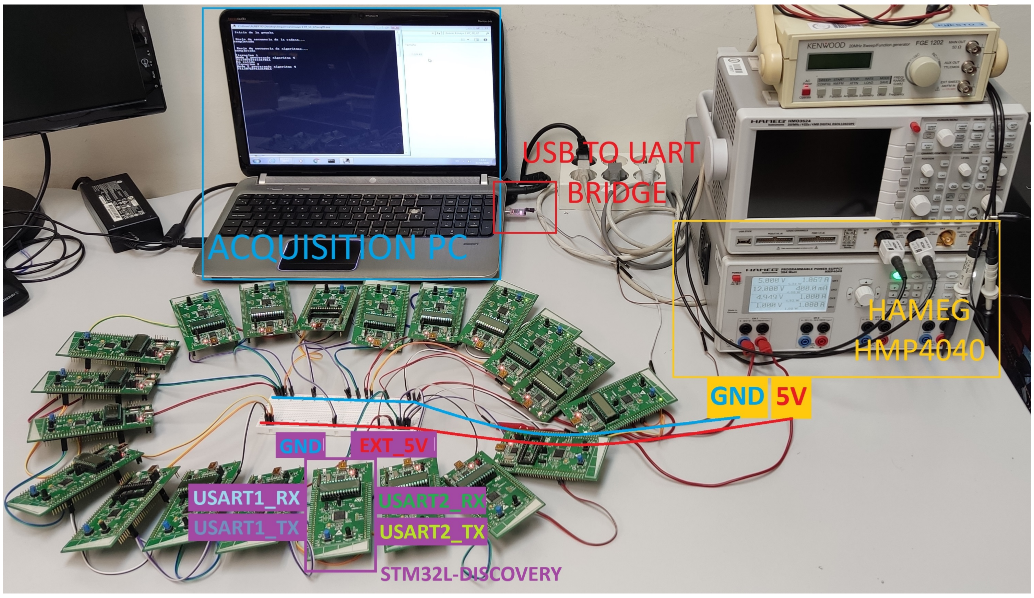

Figure 1. Picture of the experimental acquisition setup during the execution of a trial with a HAMEG HMP4040 as power source.

The purpose of this dataset is to verify the capabilities of on-chip voltage and temperature sensors incorporated in battery-free computing devices, along with the elicitation of their electronic activity through varying workloads. The selection of these workloads was made seeking a compromise between representing operations commonly found in real-world IoT applications (i.e., matrix operations are common during image processing, as well as AES in communication encryption), while also stimulating different parts of the typical hardware architecture of a microcontroller. The main goal was to collect data demonstrating the feasibility of utilizing on-chip sensors in novel AI-based identification schemes. Consequently, the dataset serves as a platform to address fundamental inquiries within this domain. These inquiries include the importance of the algorithm in obtaining responses from internal sensors, the impact of the data used during algorithm execution on their readings, the relevance of workload duration, and the effect of different scenarios on the performance of the identification system, among other relevant aspects. This dataset seeks to grant in-depth knowledge of the scope and scalability of this technology from a practical perspective in the field of hardware security, which is of great importance for both existing and future devices.

The remaining sections of this article are organized as follows: Section 2 presents a detailed description of the dataset. The experimentation methodology and data collection process are described in Section 3, along with a technical overview of the measures taken to ensure the quality of the collected data. Section 4 provides guidance on the potential uses of the dataset, along with the tools provided to facilitate its analysis. A proof of concept as an identification system is presented in Section 5, to demonstrate the usability of the dataset in the related field of application. In Section 6, the limitations of the dataset are discussed. Finally, the conclusions are presented in Section 7.

2. Data Description

The MOSID [16] (microcontroller on-chip sensor identification) dataset consists of five acquired data subsets (1.3 billion readouts totaling 6.72 GB in size, compressed into various .rar files occupying a total of 560 MB), collected during different experiments and periods using various equipment and acquisition strategies. Each of them gathers readings from the internal voltage and temperature monitoring sensors of 20 STM32L-DISCOVERY devices, which were captured at the moment of their elicitation under workloads (refer to Section 3.2 for further details). The structure of the dataset is the following one:

```
Zenodo
└── MOSID Dataset
    ├── ACQ1.rar
    │   └── ACQ1
    │       ├── 1_1
    │       │   ├── data_1.txt
    │       │   ├── ...
    │       │   └── data_20.txt
    │       ├── 1_2
    │       ├── ...
    │       └── 20_5
    ├── ACQ2.rar
    │   └── ACQ2
    ├── ACQ3.rar
    │   └── ACQ3
    ├── ACQ4.rar
    │   └── ACQ4
    ├── ACQ5.rar
    │   └── ACQ5
    └── Table_UIDS.csv
```

- **ACQ1**: The first subset, resulting from the initial randomized experiment powered by a HAMEG HMP4040 power supply, utilizing a daisy-chain topology. Twenty files were obtained for each of the proposed algorithms per board (20 out of 20 boards, resulting in 2000 files), yielding 6.824.000 pairs of raw temperature and voltage sensor readings per board, summing up to a total of 136.480.000 pairs of readings.
- **ACQ2**: The second subset, arising from the second randomized experiment powered by a HAMEG HMP4040 power supply, also employing a daisy-chain topology with a different order from ACQ1. Similarly, twenty files were obtained for each of the proposed algorithms per board (20 out of 20 boards, resulting in 2000 files total), providing 6.824.000 pairs of raw temperature and voltage sensor readings per board, resulting in 136.480.000 pairs of readings.
- **ACQ3**: The third subset, generated from individual acquisitions initiated from a rest state, powered by a HAMEG HMP4040 power supply. Once again, twenty files for each proposed algorithm per board were obtained (20 out of 20 boards, resulting in 2000 files total), producing 6.824.000 pairs of raw temperature and voltage sensor readings per board, summing to a total of 136.480.000 pairs of readings.
- **ACQ4**: The fourth subset, originated from an acquisition experiment powered by a GOLD SOURCE DF1731SB power supply, utilizing a partial daisy-chain setup with two devices at a time. As before, twenty files for each proposed algorithm per board were obtained (18 out of 20 boards, excluding candidates 14 and 15, resulting in 1800 files total), yielding 6.824.000 pairs of raw temperature and voltage sensor readings per board, totaling 122.832.000 pairs of readings.
- **ACQ5**: The last of the subsets, derived from an acquisition experiment powered by a HANMATEK HM305 power supply, also using a partial daisy-chain setup with two devices at a time. Similarly to ACQ4, twenty files for each proposed algorithm per board were obtained (18 out of 20 boards, excluding candidates 14 and 15, resulting in 1800 files total), providing 6.824.000 pairs of raw temperature and voltage sensor readings per board, summing up to a total of 122.832.000 pairs of readings.

All subsets are structured according to the folder format "X_Y", where X is the number manually assigned to the board, and Y is the corresponding number for the executed algorithm. Within each of these folders, TXT files are presented in the format "data_Z.txt", where Z represents the iteration number to which the file belongs. In each file, starting from the fifth line, temperature and voltage values are provided. Samples are captured during the execution of the stimulus in successive lines until an EOT (end of transmission) frame is received (considered as the end of the file). The final element of the data record, is the CSV-formatted table "Table_UIDS.csv" that contains all meaningful metadata from the boards. The headers of this table include the calibration values and unique identifiers, as described in Table 1, along with an additional header named BOARD_NUM, which provides the manually assigned number X for each board.

Table 1. Description of the data exchanged during computation by each node.

Data Type	Description	Size	Format
UID	Unique 96-bit identifier of the microcontroller embedded by the manufacturer.	12 B	"0x" + 24 hexadecimal characters.
$\text{VREFINT}_{\text{CAL}}$	Calibration value of the internal core reference voltage sensor.	2 B	Unsigned integer number in the range [0, 4095].
TS_{CAL_1}	Calibration value of the core temperature sensor acquired at 30 °C.	2 B	Unsigned integer number in the range [0, 4095].
TS_{CAL_2}	Calibration value of the core temperature sensor acquired at 110 °C.	2 B	Unsigned integer number in the range [0, 4095].
$\text{VREFINT}_{\text{DATA}}$	Value of the voltage sensor output converted by the ADC.	2 B	Unsigned integer number in the range [0, 4095].
TS_{DATA}	Value of the temperature sensor output converted by the ADC.	2 B	Unsigned integer number in the range [0, 4095].
EOT	end-of-transmission (EOT) frame from the node upon completion of algorithm computation.	4 B	Fixed value of 32 bits.

Given that the intention of this work is to provide the data in the rawest form, while aiming to facilitate their usage, aspects such as normalization and the generation of new sets for testing will be addressed in Section 4.

3. Methods

In order to construct this dataset, a new platform was developed as an integral strategy to encompass both the automatic data acquisition and experimentation of the target devices. To achieve this, a serial topology setup was designed, forming an acquisition daisy chain that meets the requirements considered essential for studying all the aforementioned ideas.

First and foremost, the chain enables the automation of algorithm computation and the acquisition of internal sensor readings during these stimuli, eliminating the tedious task of manual execution. This results in a significant saving in the time and personnel resources that would otherwise be required to carry out these tasks individually. Furthermore, the chain offers flexible scalability in terms of the number of devices to be experimented with, both in quantity and compatibility with development boards from various manufacturers. As long as the devices have two universal synchronous and asynchronous serial receiver and transmitter (USART) interfaces, which are supported by most varieties available on the market, they can be integrated into the chain topology. Furthermore, this topology is also compatible with any kind of device equipped with USART, such as FPGAs and SOCs. The customization of the algorithms used in successive experiments was achieved through firmware flashing of the devices, allowing them to be adapted to the specific needs of each experiment. Additionally, the chain provides control over the execution order of the algorithms and the order of the boards during the experiments. If necessary, randomization of these orders is also allowed. Moreover, control over the number of iterations that the devices perform during a trial is provided, allowing extensive computation of a workload over time.

To ensure data quality, efforts were made to minimize any bias that could arise from the equipment feeding the boards in the chain. To facilitate future reference, orderly storage of the acquired data was ensured.

3.1. Experimental Platform Design

The developed platform (shown in Figure 1) is primarily aimed at conducting experiments on 20 STM32L-DISCOVERY development boards manufactured and sold by STMicroelectronics®, which incorporate the ultra-low-power STM32L152RCT6 microcontrollers. Among the numerous and diverse peripherals they possess, efforts were made to minimize the use of hardware to only what is strictly necessary, as the extent of its influence on the internal sensors is unknown:

1. 32-bit Central Processing Unit (CPU) ARM® Cortex® M3 (operating at 32 MHz): this supports the functionality of the node in the chain, managing other peripherals as needed, as well as performing the computation of the workloads.
2. Analog to digital converter (ADC) with 25 channels, 12-bit resolution, and a sampling rate of 1 Msp/s. This peripheral is responsible for acquiring samples from the internal sensors of core temperature and internal voltage reference, which are connected to channels 16 and 17 of the ADC, respectively.
3. 2 × USART communication interfaces. In the case of the proposed system, the use of USART1 and USART2 peripherals is necessary to enable bidirectional communication of each node backward and forward in the chain, respectively.
4. Direct memory address (DMA) controller with 12 channels. The DMA controller is used to offload the burden of memory access from the CPU during the acquisition and storage of sensor readings.

Moreover, additional elements are incorporated for the implementation of the acquisition chain topology. A key component is the acquisition PC, responsible for the initialization and randomization (if necessary) of the experiment's execution order, as well as for managing its execution. Additionally, it is responsible for properly collecting and indexing the data transmitted through the serial port. To physically enable communication between the nodes and the host PC, a USB to UART Bridge Virtual COM Port (VCP) bongle based on the CP2102 circuit is used. Furthermore, a ROHDE & SCHWARZ® HAMEG HMP4040 programmable high-precision power supply with four channels of 0/32V–0/10A is employed for acquisitions 1, 2, and 3 (named ACQ1, ACQ2, and ACQ3, respectively). This power supply is utilized to ensure proper power supply to the devices under study and to minimize the effects of the disturbances typically present. For acquisitions ACQ4 and ACQ5, the Gold Source® DF1731SB and HANMATEK® HM305 sources were employed, respectively. The specific objectives include, in the case of ACQ4, obtaining samples from the boards with a different power supply, thereby enabling the study of their impact on the robustness of identification schemes. Lastly, the purpose of the ACQ5 acquisition is to enhance the variability of the dataset obtained from devices powered by a low-budget source, extending the exploration of its implications. Utilizing these three different power supply scenarios established a solid foundation to serve as a reference point for the study of the technique. Throughout the utilization of the platform in these experiments, an average power demand of approximately 5.067 W was observed. This translates to roughly 253 mW per node, a value also noted during the individual experiments.

Finally, two guiding principles were followed regarding communication and power supply wiring: first, minimizing the length of connections to protect the setup against electromagnetic interference (EMI); second, designing a symmetrical arrangement for both the communication and power supply of all devices, aiming to minimize the influence of wiring on the experiments.

3.2. Operation Description

The platform startup process begins with the firmware loading of the STM32L-DISCOVERY boards, which will be referred to as nodes from now on. This firmware allows the configuration of the previously described hardware of the participating nodes, as well as the management of different processes and functionalities of the node itself during the experiment through a state machine. It also embeds the algorithms and their test values, which serve as stimuli to obtain data from the internal temperature and voltage sensors, enabling the evaluation of the impact of data type and stimulus nature on the unique identification of the devices. The candidate stimuli include:

1. **20 × 20 Long-type matrix product**: Multiplication of two square matrices 20 × 20 of type Long, calculating each element as the sum of the products of corresponding elements from the original matrices.
2. **20 × 20 Float-type matrix product**: Similar to the previous, multiplication of two square matrices 20 × 20 of type Float, calculating each element as the sum of products of the corresponding elements from the original matrices.
3. **Algorithm for ascending sorting, Bubble Sort**: Sorting an array of integers in ascending order by iterating over the array and comparing adjacent elements, swapping them if they are in the wrong order, until no swaps are made in a complete iteration.
4. **Algorithm for 2D-point clustering, Convex Hull**: Finds the smallest convex contour around a set of points by iteratively selecting points with the smallest angle relative to the current point. It first forms an upper chain and then joins it to a lower chain that connects the contour.
5. **Encryption algorithm AES 128-bit**: Symmetric encryption algorithm that uses 128-bit keys to encrypt and decrypt data in 128-bit blocks, employing a combination of substitution, permutation, and data mixing operations across multiple rounds.

Once the chain is properly programmed, connected, and powered, the next step is to start the script, which is executed on the acquisition computer (developed using Python 3.9 and compatible with Windows 7 and later versions) when conducting an experiment. Through this script, the user specifies the number of nodes to evaluate, which algorithms from the available set will be tested, and the number of repetitions for each algorithm. If configured, the script performs sequence randomization for both the execution order of nodes in the chain and the order in which algorithms are executed. These randomized sequences are then transmitted through the serial port.

As shown in Figure 2, during the initial configuration phase, each node receives two arrays with the sequences that will define the current experiment (i.e., a first sequence for the order of the nodes and second one for the order of the algorithms) and passes them on to the next node, subtracting the corresponding position in the chain and the order in which it will be called. Once all nodes have been configured, they enter into a waiting state, marking the beginning of the command phase, where they wait for the message from the acquisition PC with the position of the node that should execute the first algorithm in the sequence. This instruction travels from node to node until it reaches the desired position, promoting the action phase along the way. If the receiving node matches the specified node in the instruction, it will compute the specified algorithm, sample the sensors during the process, and send these readings along with other previous metadata to the host PC (details in Table 1). Otherwise, its role is to transmit all the received information to the next node towards the acquisition computer, acting as a bridge. Once all the samples for an algorithm execution have been collected and a end-of-transmission frame has been sent, the nodes that were in the action phase recognize it and return to the command phase, awaiting for new commands. This process is iterated for the configured number of repetitions, continuing with the remaining nodes in the sequence until all the algorithms have been computed according to the established order.

During each iteration of the nodes, .txt files are created to store all the information related to the experiment. This allows for indexing of the collected data from each board based on their unique ID, the algorithm, and the corresponding repetition. This indexing facilitates control and pre-processing of the data after the experiment has been completed.

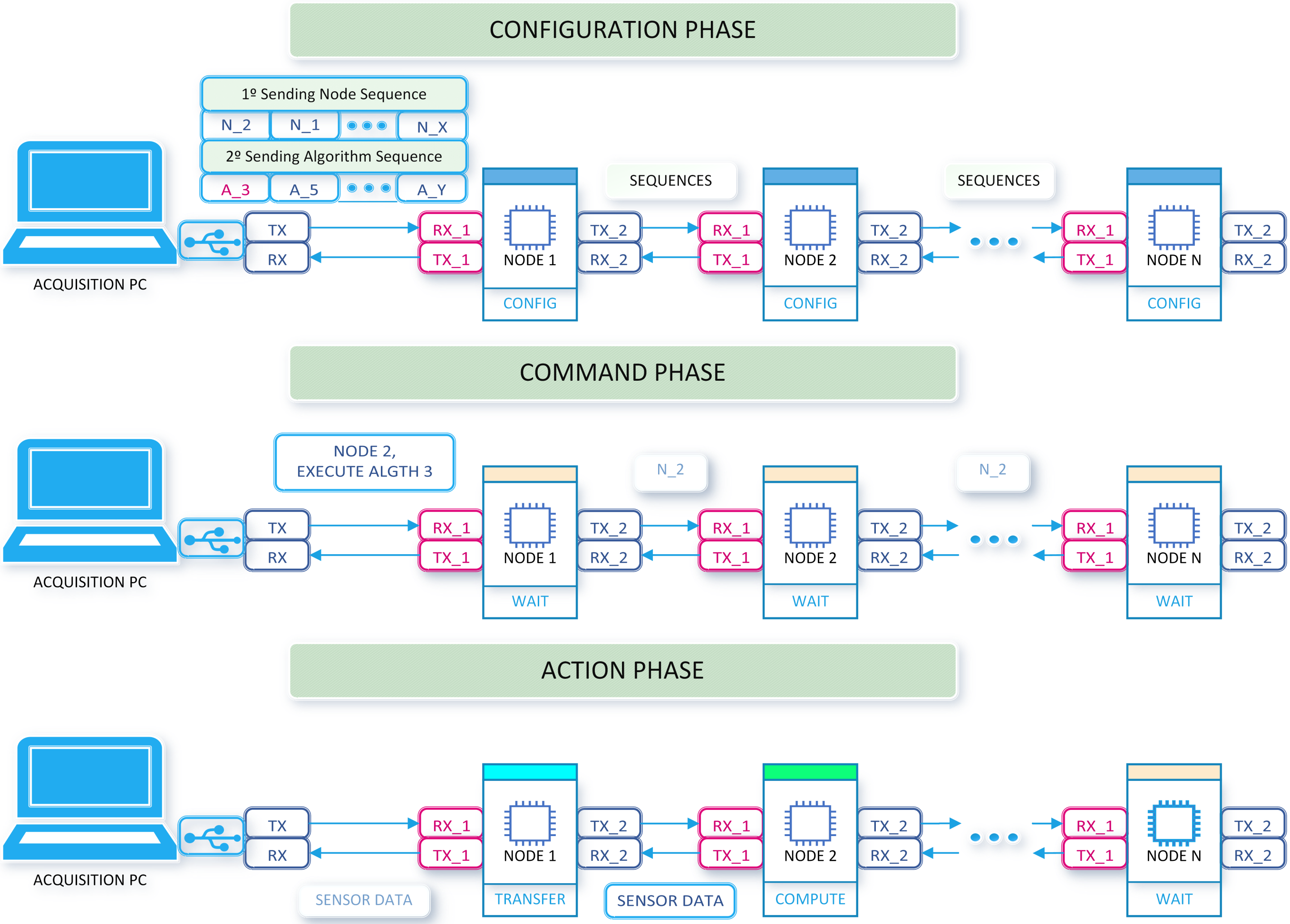

Figure 2. Visualization of the different stages during the operation of the platform.

3.3. Data Acquisition Trials

The data collection was performed on different days in a climate-controlled laboratory room under normal temperature conditions (25 °C). The same equipment described earlier was strategically used to minimize any potential effects of external factors. As anticipated in Section 2, a total of 5 acquisitions were carried out for the construction of the dataset, involving 20 boards, with the 5 previously mentioned algorithms, and 20 repetitions per node and algorithm.

3.3.1. Randomized Experiments

Two out of the five data collection sessions (corresponding to the subsets ACQ1 and ACQ2) were conducted using the setup depicted in Figure 1 (utilizing the power supply ROHDE & SCHWARZ® HAMEG HMP4040), with variations in the physical arrangement of the boards. Initially, an ID number (can be found at *Table_UIDS.csv*) was assigned to each board according to the physical order of ACQ1. Then, in ACQ2, the devices were physically mixed and randomly chosen to form the chain in a new order of selection. Software randomization was carried out by generating two random vectors; one contained the randomized sequence of the 20 values corresponding to the acquisition order in the chain, while the other contained the randomized sequence of the order in which the algorithms would be executed. This additional layer of software randomization of the sequences of nodes and algorithms at the beginning of each experiment enabled the gathering of sufficient data to observe any potential biases related to the chain's topology. Following each approximately 50-hour experiment (totaling 100 h), a total of 2.8 GB of data from the internal sensors of the microcontrollers were collected during the randomized experiments.

3.3.2. Individual Data Collection

For the third acquisition experiment (subset ACQ3), each node was individually assembled and tested using the same supply equipment and number of repetitions. This was conducted for every board, starting from a rested state, one at a time, algorithm by algorithm, under identical ambient conditions. This allowed contrasting, in case it existed, the magnitude of any bias that the chain's topology might introduce, as the nodes closest to the acquisition PC were theoretically subject to a higher load during the experiments, more frequently serving as data transfer bridges. This experiment involved the collection of an additional 1.4 GB of readings, which also served for identification purposes.

3.3.3. Alternative Powered Collection

The fourth acquisition experiment (subset ACQ4) involved simultaneously assembling a chain of two boards. The same 5 algorithms were tested with 20 repetitions per algorithm, using the Gold Source® DF1731SB power supply in this case. In this case, a different supply scenario of the boards was presented, both in terms of equipment and the utilization of topology. The provided data offer valuable information for assessing variances in the responses of on-chip monitoring sensors, as well as their influence on the performance of proposed identification solutions. Eighteen out of the twenty available boards were evaluated in this experiment, allowing the collection of 1.25 GB of readings.

3.3.4. Low-range Powered Collection

Lastly, a final acquisition experiment (corresponding to subset ACQ5) was conducted following the same strategy described for subset ACQ4, powered this time by the low-range HANMATEK® HM305 power supply. This last subset, together with the preceding one, allow the evaluation of the robustness of the identification schemes developed against the impact of variability in the behaviors of the boards when powered by sources with different stability. In this experiment, eighteen out of the twenty available boards were also evaluated, providing an additional 1.25 GB of readings.

3.4. Technical Validation

Various measures were taken to ensure the highest quality in the obtained dataset. Regarding the acquisition platform and the devices and conditions used during data acquisition, it was mentioned earlier that the experiments were conducted in the same spatial location under nearly identical temperature and power conditions, without additional elements or equipment that could have had an additional impact (EMC, temperature) on the collected data. It is worth noting that the influence of the acquisition equipment is considered negligible as it does not share power with the chain. The acquisition equipment had a resolution of 1 mV and 1 mA, and a precision error of 0.01% and 0.1%, respectively.

In terms of communication, the USART communication interface was used with standards such as RS-232 or RS-485 (up to 1000 m), providing robust communication. The data transmission from end to end was performed at a baud rate of 115,200 baud, which is equivalent to 115,200 bit/s. This baud rate significantly exceeded the transmission and reception needs and ensured a practically negligible probability of communication error. Additionally, using a parity bit during communication further enhanced the error probability reduction.

Regarding the assembly of the chain, as mentioned earlier, a symmetrical and equidistant power distribution setup was arranged to minimize the effect of connectors on electronic activity. These measures ensured that any inductive-capacitive bias introduced by the wiring was thus leveled, akin to the power connection of a conventional device. The interconnection was performed using Dupont cables of the same length and from the same manufacturer. Regardless of the acquisition trials, identical lengths of interconnections between devices and power connections were used, to preserve the fidelity between test conditions. In the case of separate acquisitions, these were also performed in a rest state, meaning that a reasonable time interval was respected between algorithm and board acquisitions.

In relation to the exploitation of the embedded sensors, it is important to note that their measurements were acquired using both channels of the ADC, as explained in previous sections. The sampling process was conducted with a sampling period far below the manufacturer's limit (1 µs), as compared to the sampling windows detailed in Table 2. It is important to emphasize that this sampling window resulted from sampling at specific key points of each algorithm. Coupled with the deterministic behavior of the devices under test, this implicitly guaranteed a fixed sampling period between samples throughout the workload, as represented in the following table.

Table 2. Data gathered and sampling windows per algorithm and iteration.

Algorithm	Nº T–V Pairs	Sampling Window	File Size per Iteration
Long-type matrix product	79,600	1.68 ms	856 kB
Float-type matrix product	79,600	1.28 ms	856 kB
Bubble Sort	100,000	1.20 ms	1075 kB
Convex Hull	70,000	1.50 ms	752 kB
AES 128-bit	12,000	8.42 ms	129 kB

From the perspective of data analysis, we observed a qualitative relationship between the normalized data obtained and the environmental conditions under which the experiments were conducted. Upon inspecting the acquired data, it is evident that the temperature sensor measurements consistently exceeded the ambient temperature of the laboratory, attributed to the workload effect. Furthermore, no outliers or inconsistent values were identified that deviated significantly from the expected magnitudes being measured. Additionally, it is worth noting that the acquired files exhibited identical sizes for each algorithm, indicating a consistent data collection process.

4. User Notes

There are numerous practices regarding the workflow, depending on the object of study. Concerning the provided data collection, one can speculate about the numerous and varied analyses that could be performed (such as analyzing the effect of the algorithm used on sensor responses or studying the variability of responses over time, to name a few examples). These analyses may require certain peculiarities in terms of data processing. In a general sense and in line with the overarching goal of creating the dataset (which was the study of the usability of device data for identification purposes), two Python scripts [17] were developed to handle the data and generate outputs for easy exploration.

The first script, DataBuilder, is a command line interface that allows the adaptation of all data within an acquisition directory to the selected output format, generating one or multiple CSV files (whose data are organized in columns according to the format specified in the Table 3). It enables the exclusion of algorithms and boards as needed from the dataset, as well as downsampling by the required factor for the output(s). Finally, it facilitates the conversion of raw values (ADC conversions without contextualization) into temperature (°C) and voltage (V) values at the output by normalizing them using calibration values (described in Table 1) extracted from each microcontroller and equations provided by the manufacturer [18]:

$$Temperature\ (^{\circ}\mathrm{C}) = \frac{110\ ^{\circ}\mathrm{C} - 30\ ^{\circ}\mathrm{C}}{TS_{CAL_2} - TS_{CAL_1}} \times (TS_{DATA} - TS_{CAL_1}) + 30\ ^{\circ}\mathrm{C} \tag{1}$$

$$Voltage\ (\mathrm{V}) = 3\mathrm{V} \times \frac{VREFINT_{CAL}}{VREFINT_{DATA}} \tag{2}$$

To facilitate the study of using fixed sequences for the identification of devices based on their electronic activity and through the use of artificial intelligence, the sequencer script allows the construction of sequences of pairs of temperature–voltage values of a desired length, along with their corresponding board label.

Table 3. Description of the headers in the generated CSV files.

Field	Description
Voltage Value	Internal voltage sensor sample acquired by the board.
Temperature Value	Internal temperature sensor sample acquired by the board.
Board Number	Manually assigned board number from the set.
Algorithm	Number assigned to the algorithm during which the sample was acquired.
Iteration	Iteration of the algorithm to which the sample belongs.

The script explores CSV files, which should have been generated in a multiple format located in the directory specified by a variable called *folder* and will generate sequences with the length specified in the *sequence_length* variable, as it iterates through all the files it finds. As a result of this processing, an HDF5 file is generated for each board, containing the generated sequences (Equation (3)) and their respective labels for later use. It is worth noting that the main objective of the script is to generate training/validation/test sets for various models. As an intermediate step before creating .hdf5 files, Z-score data normalization is performed on the temperature and voltage values. If this is not desired, it can be fixed by following the instructions available in the repository.

$$\left[(T_{norm_1}, V_{norm_1}) \quad (T_{norm_2}, V_{norm_2}) \quad \cdots \quad (T_{norm_{n-1}}, V_{norm_{n-1}}) \quad (T_{norm_n}, V_{norm_n}) \right] \tag{3}$$

5. Proof of Concept

The paramount consideration in validating the constructed dataset lies in its usability. To address this, a proof-of-concept experiment was undertaken, wherein a 1D convolutional neural network (CNN) model was trained on the ACQ1 subset. The objective was to demonstrate the efficacy of the dataset in identifying 20 distinct devices.

Before commencing training, the data underwent preparation utilizing commented and openly available scripts [17]. The DataBuilder configuration employed for the output data included multiple files, T–V normalization, without any exclusions or downsampling. Subsequently, the resulting files were utilized to generate sequences of T–V pairs, as depicted in Equation (3), with a fixed length of 100. A total of 1,364,800 sequences were utilized, distributed according to the ratios of 70% for training, 15% for validation, and 15% for testing.

The model's construction and training were conducted in Python, employing the Keras API. It contains 102,005 parameters and occupies a size of 398.46 KB. Training utilized a learning rate of 0.0001 for 40 epochs, with an 'early stop' callback implemented after 4 epochs. Additional details are available in the repository [19].

Figure 3 illustrates the evolution of the training and validation accuracy and loss error throughout the epochs. Subsequently, the model's performance was assessed against the test subset (Figures 4 and 5), achieving an average accuracy of 93.71%. These results highlight the robustness of the preliminary model and validate the data collection process.

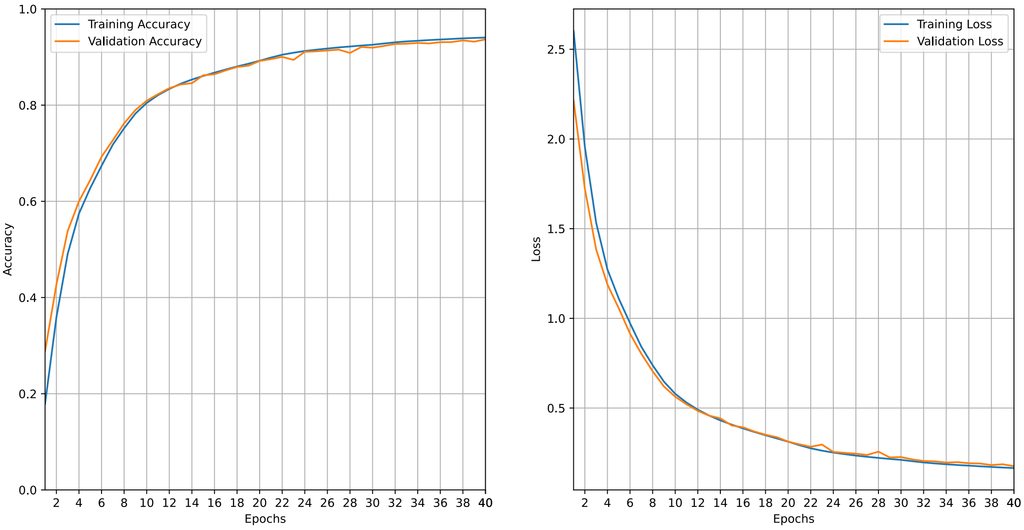

Figure 3. Accuracy and loss per epoch, training versus validation.

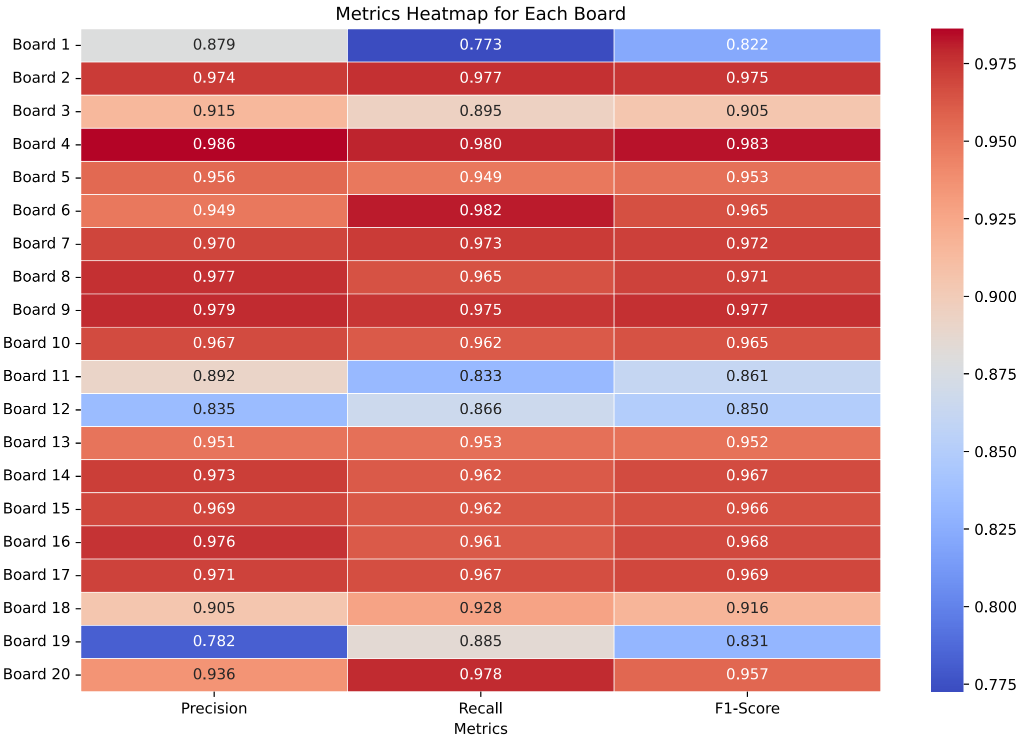

Figure 4. Heatmap of precision, recall, and F1-score from test predictions.

As is evident, this proof of concept was far from optimized. Tuning hyperparameters and exploring more suitable architectures are essential to unlock its full potential. This can be appreciated in the slight confusion that the model incurred at times, as is the case with boards 1 and 19. The sub-optimal performance and inadequacy of the model, combined with the use of a small portion of the dataset, limited its capacity to extract generalizations about the differences in hardware responses of these ultra-low power devices. However, since delving into these ideas is beyond the scope of this paper, the extent of its capabilities will be explored in future research endeavors.

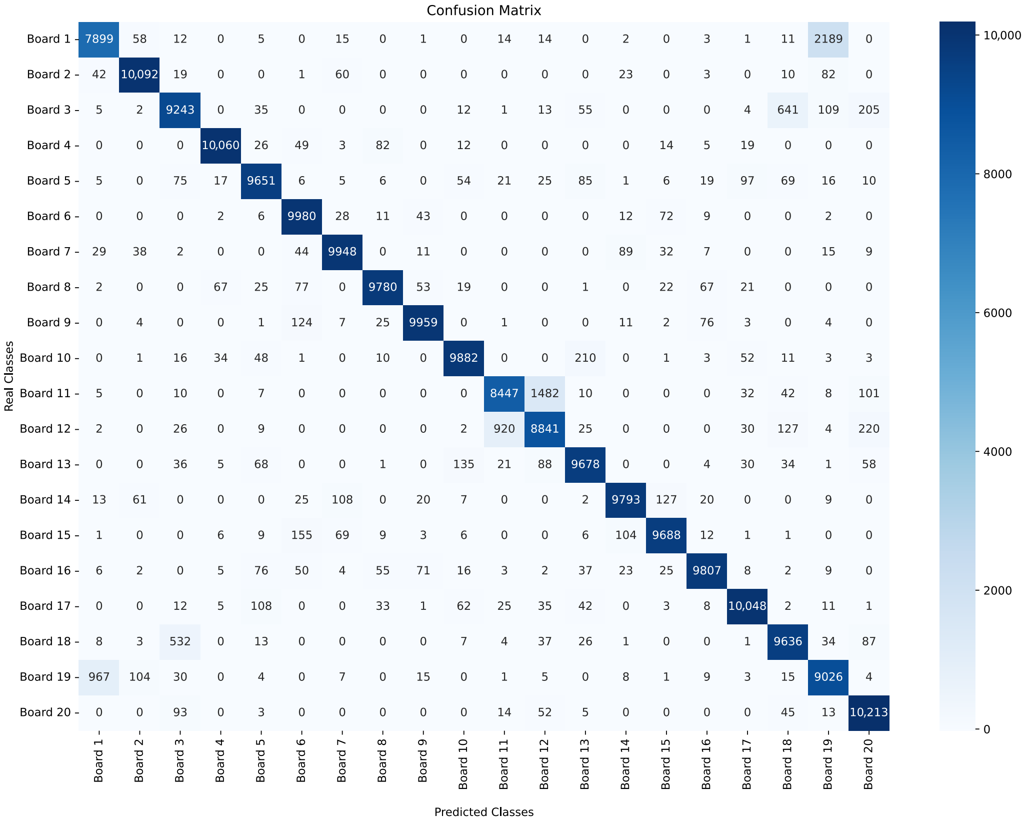

Figure 5. Confusion matrix of the predictions made by the trained model on the test subset.

6. Limitations of the Dataset

The presented work centers on the creation of a dataset acquired via ultra-low-power microcontrollers, resulting in bounded variations in the obtained readings during workload execution due to their low consumption feature. The inherent simplicity of the microcontrollers contributed to a restricted range of variations observed in the dataset, contrasting with the potential information that sensors integrated into more powerful and sophisticated multi-core architectures could offer. Additionally, this simplicity poses a barrier to implementing more effective elicitation strategies that could enhance the dataset further.

A core constraint encountered during the construction of this dataset involved limiting its focus exclusively to devices not powered by batteries. In ecosystems like IoT or Industry 4.0, battery-operated devices, along with those integrating energy harvesting measures, represent a substantial portion of this category. However, the intricate nature of these technologies introduces a multitude of factors that contribute to variability, thereby significantly complicating the quest for the ground truth. Among these factors, we can find the effect on discharge discrepancies, irregular charging, as well as the aging process they undergo, not to mention the variations introduced by the PMIC associated with these types of power sources. Consequently, it was determined to exclude them from the scope of this study.

Furthermore, it would have been highly meaningful to conduct experiments under different temperature and/or power supply conditions, to expand the capabilities for studying the presented data. Thus, incorporating the capability to study the aging effects of boards could provide valuable insights into long-term performance characteristics. Although this extension is planned for future work, the current results from the experiments already provide a solid foundation for analyzing their applicability, among other aspects.

Moreover, conducting experiments under diverse temperature and/or power supply conditions would have provided significant value in broadening the scope for studying the presented data. Hence, integrating the capability to examine the aging effects of the boards could offer valuable insights into their long-term performance characteristics. While this extension is earmarked for future investigation, the current findings from the experiments establish a robust foundation for analyzing their applicability, among other considerations.

7. Conclusions

This paper introduced a novel dataset that, to the best of our knowledge, offers unprecedented opportunities for studying microcontroller-based device identification through data obtained from on-chip monitoring sensors. For data collection, we designed an automated acquisition platform capable of scaling and configuring certain aspects of the experiments as needed. The methodology, materials, and tools used and/or developed for constructing the dataset are extensively detailed for replication purposes. The data were collected from 20 devices based on the STM32L152RTXX microcontroller, during the arousal of their electronic activity with five different workloads. Furthermore, various power supplies and topologies were employed to collect the five subsets comprising the dataset.

This dataset facilitates investigation into questions such as the robustness of device identification, the impact of workload on electronic activity and sensor readings, the influence of the power source, potential biases resulting from the acquisition setup, and the effects of prolonged stimulus on the obtained responses. Concurrently, the development of machine learning or deep learning techniques to effectively utilize this information for security purposes remains an open area for future research. In contrast, this raises questions concerning the potential profiling devices, applications, or users for malicious intents. Such actions could unveil unforeseen vulnerabilities in security and privacy, thus emphasizing the necessity for further exploration in this domain.

Author Contributions: A.R. conceived and designed the experiments, developed the software, conducted the experimentation, curated data, validated concepts, and drafted the original manuscript. H.M. contributed to the experiment design, conducted experimentation, provided supervision, and participated in manuscript editing. C.C. contributed to the experiment design, curated data, validated concepts, provided supervision, and participated in manuscript editing. P.P.-L. contributed to the experiment design, curated data, validated concepts, provided supervision, and participated in manuscript editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received support from the Madrid Government under the project PUCFA-CM-UC3M, within the framework of the V Regional Plan of Scientific Research and Technological Innovation (V-PRICIT), as well as from the CYCAD grant (PID2022-140126OB-I00) provided by the Spanish Ministry of Science, Innovation, and Universities. Additionally, this endeavor received funding from the NextGeneration-EU research and innovation program through the DIPSY-TECH (TED2021-131401A-C22) and CIOMET (TED2021-131681B-I00) projects.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The **dataset** can be accessed via https://doi.org/10.5281/zenodo.10042177, along with the **tools** provided for processing the data, which are available at https://doi.org/10.5281/zenodo.10042160 (both accessed on 27 April 2024). Additionally, the CNN-based **model** used in Section 5 and the script used to generate it can be found at https://doi.org/10.5281/zenodo.10042163 (accessed on 27 April 2024), for transparency purposes. The content of these repositories is licensed under the GNU GPL v3.0.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

IoT	Internet of Things
AI	Artificial Intelligence
SOC	System On Chip
PUF	Physically Unclonable Function
MAC	Media Access Control
ML	Machine Learning
DL	Deep Learning
CRP	Challenge-Response Pair
FPGA	Field Programmable Gate Array
AES	Advanced Encryption Standard
UID	Unique Identifier
MOSID	Microcontroller On-chip Sensor IDentification
CSV	Comma-Separated Values
ADC	Analog-Digital Converter
CNN	Convolutional Neural Network
API	Application Programming Interface
PMIC	Power Management Integrated Circuit

References

1. Dicholkar, S.V.; Sekhar, D. Review-IoT Security Research Opportunities. In Proceedings of the 2020 International Conference on Convergence to Digital World-Quo Vadis (ICCDW), Mumbai, India, 18–20 February 2020; pp. 1–4. [CrossRef]
2. Ul Rehman, S.; Singh, P.; Manickam, S.; Praptodiyono, S. Towards Sustainable IoT Ecosystem. In Proceedings of the 2020 2nd International Conference on Industrial Electrical and Electronics (ICIEE), Lombok, Indonesia, 20–21 October 2020; pp. 135–138. [CrossRef]
3. Tournier, J.; Lesueur, F.; Mouël, F.L.; Guyon, L.; Ben-Hassine, H. A survey of IoT protocols and their security issues through the lens of a generic IoT stack. *Internet Things* **2021**, *16*, 100264. [CrossRef]
4. Yaqoob, I.; Ahmed, E.; Hashem, I.A.T.; Ahmed, A.I.A.; Gani, A.; Imran, M.; Guizani, M. Internet of Things Architecture: Recent Advances, Taxonomy, Requirements, and Open Challenges. *IEEE Wirel. Commun.* **2017**, *24*, 10–16. [CrossRef]
5. Miettinen, M.; Marchal, S.; Hafeez, I.; Asokan, N.; Sadeghi, A.R.; Tarkoma, S. IoT SENTINEL: Automated Device-Type Identification for Security Enforcement in IoT. In Proceedings of the 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS), Atlanta, GA, USA, 5–8 June 2017; pp. 2177–2184. [CrossRef]
6. Sivanathan, A.; Gharakheili, H.H.; Loi, F.; Radford, A.; Wijenayake, C.; Vishwanath, A.; Sivaraman, V. Classifying IoT Devices in Smart Environments Using Network Traffic Characteristics. *IEEE Trans. Mob. Comput.* **2019**, *18*, 1745–1759. [CrossRef]
7. Kotak, J.; Elovici, Y. Adversarial Attacks Against IoT Identification Systems. *IEEE Internet Things J.* **2023**, *10*, 7868–7883. [CrossRef]
8. Ahmed, M.K.; Yanambaka, V.P.; Abdelgawad, A.; Yelamarthi, K. Physical Unclonable Function Based Hardware Security for Resource Constraint IoT Devices. In Proceedings of the 2020 IEEE 6th World Forum on Internet of Things (WF-IoT), New Orleans, LA, USA, 2–16 June 2020; pp. 1–2. [CrossRef]
9. Abdallah, A.; de Abreu, M.F.B.; Vieira, F.H.T.; Cardoso, K.V. Toward Secured Internet of Things (IoT) Networks: A New Machine Learning based Technique for Fingerprinting of Radio Devices. In Proceedings of the 2022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC), Las Vegas, NV, USA, 8–11 January 2022; pp. 955–956. [CrossRef]
10. Zhang, J.; Shen, C.; Guo, Z.; Wu, Q.; Chang, W. CT PUF: Configurable Tristate PUF Against Machine Learning Attacks for IoT Security. *IEEE Internet Things J.* **2022**, *9*, 14452–14462. [CrossRef]
11. Khalfaoui, S.; Leneutre, J.; Villard, A.; Gazeau, I.; Ma, J.; Danger, J.L.; Urien, P. Water- PUF: An Insider Threat Resistant PUF Enrollment Protocol Based on Machine Learning Watermarking. In Proceedings of the 2021 IEEE 20th International Symposium on Network Computing and Applications (NCA), Boston, MA, USA, 23–26 November 2021; pp. 1–10. [CrossRef]
12. Chatterjee, B.; Das, D.; Maity, S.; Sen, S. RF-PUF: Enhancing IoT Security Through Authentication of Wireless Nodes Using In-Situ Machine Learning. *IEEE Internet Things J.* **2019**, *6*, 388–398. [CrossRef]
13. Cui, Y.; Gu, C.; Wang, C.; O'Neill, M.; Liu, W. Ultra-Lightweight and Reconfigurable Tristate Inverter Based Physical Unclonable Function Design. *IEEE Access* **2018**, *6*, 28478–28487. [CrossRef]
14. Martin, H.; Vatajelu, E.I.; Natale, G.D. Identification of Hardware Devices based on Sensors and Switching Activity: A Preliminary Study. In Proceedings of the 2021 Design, Automation & Test in Europe Conference & Exhibition (DATE), Grenoble, France, 1–5 February 2021; pp. 1496–1499. [CrossRef]
15. Kornaros, G.; Pnevmatikatos, D. A Survey and Taxonomy of On-Chip Monitoring of Multicore Systems-on-Chip. *ACM Trans. Des. Autom. Electron. Syst.* **2013**, *18*, 1–38 [CrossRef]

16. Ramos García, A.; Martin Gonzalez, H.; Cámara Núñez, M.C.; Peris-López, P. MOSID: A dataset of readings from the internal monitoring sensors of STM32L152RTXX microcontrollers during the stimulation of their electronic activity. Zenodo. 2023. Available online: https://zenodo.org/records/10042177 (accessed on 5 April 2024).
17. Ramos, A. DeviceFingerprinting-Tools: Tools v1.1.0. Zenodo. 2023. Available online: https://zenodo.org/records/10042160 (accessed on 5 April 2024).
18. STMicroelectronics©. RM0038 STM32L1XXXX Reference Manual. Rev 18. 2023. Available online: https://www.st.com/resource/en/reference_manual/rm0038-stm32l100xx-stm32l151xx-stm32l152xx-and-stm32l162xx-advanced-armbased-32bit-mcus-stmicroelectronics.pdf (accessed on 5 April 2024).
19. Ramos, A.; Cámara, M. DeviceFingerprinting-ID-Models: Models v1.0.0. Zenodo. 2023. Available online: https://zenodo.org/records/10042163 (accessed on 5 April 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.